UNIVERSIDAD DE ALMERÍA

ESCUELA POLITÉCNICA SUPERIOR Y FACULTAD DE CIENCIAS EXPERIMENTALES

DEPARTAMENTO DE MATEMÁTICAS

Trabajo fin de máster. Máster en matemáticas.

EL MÉTODO DE LA ENTROPÍA CRUZADA. ALGUNAS APLICACIONES.

Autor: Sophie Helene Bischel

Dirigido por: Antonio Salmerón Cerdán

Septiembre 2013

Índice

1. Introducción **3**

2. Qué es el método de la entropía cruzada. **4**
2.1. Muestreo por importancia 4
2.2. La distancia Kullback-Leibler 6
2.3. Estimación de probabilidades *rare-event* 7
2.4. CE para optimización 9
2.5. Ejemplo de CE aplicado, clasificación con K-means. 12

3. Regresión mediante entropía cruzada. **15**
3.1. El método CE para regresión lineal. 15
3.2. Ejemplos de CE aplicado a regresión lineal. 18

4. El método CE para clasificación Naïve Bayes. **25**
4.1. Clasificación Naïve Bayes. 25
4.2. Elección de la distribución a partir de la que generar los parámetros. 28
4.3. Algunos ejemplos. 33

5. Conclusiones **39**

6. Anexos **40**

7. Bibliografía **43**

PARTE 1

Introducción

En esta memoria de trabajo veremos cómo funciona y cómo se puede aplicar el método de la *entropía cruzada (CE)*, a diversos problemas. Este método se utilizó por primera vez (*Rubinstein* en [12]) para la estimación de probabilidades muy bajas, llamadas rare-events, en 1997. Pero pronto se descubrió, (*Rubinstein,* $1999 - 2001$), como una eficaz herramienta para resolver no sólo problemas de simulación de *rare events*, ver [1], sino también para problemas complejos de optimización. El nombre de cross-entropy, o entropía cruzada, se debe a que utiliza la distancia Kullback-Leibler, véase [16]. Esta distancia es una medida entre dos funciones de densidad g y h, es también conocida como entropía cruzada entre g y h.

$$\begin{aligned}\mathcal{D}(g,h) &= \int g(x) \ln \frac{g(x)}{h(x)} \mu(dx) \\ &= \int g(x) \ln g(x) \mu(dx) - \int g(x) \ln h(x) \mu(dx) \end{aligned} \tag{1.1}$$

Es un método iterativo en el cual, primero se generan un conjunto aleatorio de valores para el valor que queremos estimar y después de actualizan los prámetros para poder generar unos valores *mejores*, más aproximados en el sentido de Kullback-Leibler, en la siguiente iteración. La ventaja de este método es que se puede aplicar tanto a problemas *determinísticos* [1] como a problemas con *ruido*.

Veremos cómo se ha aplicado ya con éxito en problemas de optimización, convirtiendo estos problemas *deteminísticos* en otro *estocástico* equivalente y luego aplicando la técnica de simulación de rare-events parecida a la usada por Rubinstein en 1997. Veremos cómo se puede aplicar el método de la entropía cruzada a optimización y a clasificación con K-means, de [18]. Finalmente usaremos el método CE para resolver problemas concretos de regresión y clasificación, estudiando los resultados obtenidos y comparándo éstos con los resultados de que se obtienen con los métodos tradicionales.

[1] Con determinístico, nos referimos a problemas que produce invariablemente las mismas salidas para las mismas entradas.

PARTE 2

Qué es el método de la entropía cruzada.

En este apartado haremos un breve resumen de cómo funciona el método de la entropía cruzada, extraído de [1], libro que Rubinstein publicó sobre este método en 2004. En sus comienzos fué una solución a un problema de estimación de probabilidades rare-event como alternativa al muestreo por importancia. Después resultó una eficaz herramienta para todo tipo de problemas complejos de optimización. Se ha aplicado en muchas más situaciones pero nos centraremos en el problema de clasificación mediante el método de K-mean, al final de esta sección, para dar un ejemplo de cómo se aplica este método.

2.1. Muestreo por importancia

Vamos a desarrollar la teoría tal y cómo lo hacen en [1]. Para ello primero repasemos en qué consiste el muestro por importancia, ver [17]. Supongamos que queremos estimar l dada de la siguiente forma:

$$l = \mathbb{E}_f H(X) = \mathbb{E}_f \varphi(S(X);\gamma) = \int \varphi(S(X);\gamma) f(x) \mu(dx), \tag{2.1}$$

donde S es un estadístico muestral, $\varphi(\cdot;\gamma)$ es una función real de transformación de la muestra, que depende de γ, f es la función de densidad de X con respecto a la medida μ, es decir $X \sim f(\cdot;\mu)$. Algunos ejemplos de $\varphi(S(X);\gamma)$ son la función indicadora $\varphi(S(X);\gamma) = I_{\{S(X)\geq\gamma\}}$ o las funciones Boltzmann $\varphi(S(X);\gamma) = exp(-S(X)/\gamma)$.

Sea g otra función de densidad tal que Hf esta *dominada* por g . Esto es, $g(x) = 0 \Rightarrow H(x)f(x) = 0$. Usando g podemos reescribir l como:

$$l = \int H(x)\frac{f(x)}{g(x)} g(x)\mu(dx) = \mathbb{E}_g H(X)\frac{f(x)}{g(x)}, \tag{2.2}$$

donde ahora vemos que la esperanza es con respecto a g, la cual se denomina densidad de *muestro por importancia (IS)*,ver [13]. Un estimador estándar de l es

$$\hat{l} = \frac{1}{N}\sum_{i=1}^{N} H(X_i)W(X_i), \tag{2.3}$$

donde $\hat{l}$ se conoce como estimador de muestreo por importancia o *estimador del ratio de verosimilitud*, $W(x) = f(x)/g(x)$ se llama *ratio de versimilitud*, véase [3]. Sea $X_1, X_2, ..., X_N$ un conjunto aleatorio de vectores de densidad g. En el caso particular en el que no hay *cambio de medida*, es decir $g = f$

con lo que $W = 1$, y (2.3) se convierte en el siguiente estimador que es conocido con el nombre de Montecarlo simple, ver [13], $\hat{l} = \frac{1}{N}\sum_{i=1}^{N} H(X_i)$.

La elección de la función de densidad g es crucial en la varianza del parámetro $\hat{l}$. Abordemos el problema de minimizar la varianza de $\hat{l}$ con respecto de g,

$$\min_{g} Var_g \left\{ H(X)\frac{f(X)}{g(X)} \right\}. \tag{2.4}$$

La *solución* del problema (2.4) es

$$g^*(x) = \frac{|H(x)|f(x)}{\int |H(x)|f(x)\mu(dx)}. \tag{2.5}$$

Si $H(x) \geq 0$, entonces

$$g^*(x) = \frac{H(x)f(x)}{l}$$

y

$$Var_{g^*}\left(\hat{l}\right) = Var_{g^*}\left(H(X)W(X)\right) = Var_{g^*}(l) = 0.$$

A esta nueva función de densidad g^* la denominamos *densidad óptima de muestreo por importancia*. El principal problema que surge al intentar hallar g^* es que depende de l, que es precisamente lo que queremos estimar. En la mayoría de los casos hay un problema añadido, el desconocimiento previo de la expresión de H. Para solventar este problema se pueden generar a su vez $H(X_1), H(X_2), ..., H(X_N)$ para estimar g^*. Estos hechos nos hacen sospechar que construir esta densidad es muy complicado y lleva mucho tiempo, más aún si g es de una dimensión grande.

Para solventar el primero de estos problemas vamos a asumir a partir de ahora que f pertenezca a una determinada familia paramétrica $\mathcal{F} = \{f(\cdot; v), v \in \mathcal{V}\}$. Sea $f(\cdot; u)$ la función de densidad del vector aleatorio X de (2.1), para algún parámetro fijo $u \in \mathcal{V}$. Además restringiremos la elección de la densidad IS, g, a aquellas que pertenzcan a la misma familia paramétrica $\mathcal{F}$; así g sólo difiere de la densidad original $f(\cdot; u)$ en un parámetro v, que llamaremos *parámetro de referencia*. Entonces podremos reescribir el ratio de versimilitud con $g(x) = f(x; v)$ como

$$W(X; u, v) = \frac{f(X; u)}{f(X; v)} \tag{2.6}$$

y el estimador (2.3) como

$$\hat{l} = \frac{1}{N}\sum_{i=1}^{N} H(X_i)W(X_i; u, v), \tag{2.7}$$

donde $X_1, X_2, ..., X_N$ es una muestra aleatoria de $f(\cdot; v)$. Llamaremos a este último *estimador estándar del ratio de verosimilitud (SLR)*; que no tiene solución analítica, ver [3]. Se resuelve por diversos métodos que aquí no tratamos ya que lo que haremos es ofrecer una alternativa a este estimador.

2.2. La distancia Kullback-Leibler

Una alternativa a cómo se minimiza la varianza de la estimación en (2.7) esta basada en la *entropía cruzada CE* de Kullback-Leibler (1.1), que define la distancia entre dos funciones de densidad de probabilidad g y h y se puede escribir como

$$\begin{aligned}\mathcal{D}(g,h) &= \int g(x) \ln \frac{g(x)}{h(x)} \mu(dx) \\ &= \int g(x) \ln g(x) \mu(dx) - \int g(x) \ln h(x) \mu(dx).\end{aligned}$$

La idea del método de la entropía cruzada es coger la función de densidad IS h tal que la distancia entre la densidad óptima IS g^* de (2.5) y h sea mínima; esto es, que la función CE de densidad óptima IS h^* es la solución del siguiente problema de optimización *funcional*

$$\min_h \mathcal{D}(g^*, h).$$

De la condición de que $\mathcal{D}(g^*, h) \geq 0$ se deduce que $h^* = g^*$, las soluciones de minimizar la varianza de $\hat{l}$ con respecto de g y la que se obtiene como CE de densidad óptima, coinciden.

Vamos a usar el estimador SLR, la clase de las densidades se restringen a la familia $\mathcal{F} = \{f(\cdot; v), v \in \mathcal{V}\}$ que a su vez contiene a la densidad $f(\cdot; u)$. Entonces el método de CE pretende resolver el problema de optimización *paramétrica*

$$\min_h \mathcal{D}(g^*, f(\cdot; v)),$$

con

$$g^*(x) = \frac{|H(x)| f(x; u)}{\int |H(x)| f(x; u) \mu(dx)}.$$

Como el primer término en (1.1) de la derecha no depende de v (de h), minimizar la distancia Kullback-Leibler entre g^* y $f(\cdot; v)$ es equivalente a maximizar, con respecto a v

$$\int |H(x)| f(x; u) \ln f(x; v) \mu(dx) = \mathbb{E}_u |H(X)| \ln f(X; v).$$

Vamos a suponer para eliminar valores absolutos que $H(X) \geq 0$. Entonces el parámetro óptimo de referencia, con respecto a la distnacia de Kullback-Leibler es la solución de

$$\max_v = \mathcal{D}(v) = \max_v \mathbb{E}_u H(X) \ln f(X; v). \tag{2.8}$$

Lo que es equivalente al siguiente programa

$$\max_v \mathcal{D}(v) = \max_v \mathbb{E}_w H(X) W(X; u, w) \ln f(X; v), \tag{2.9}$$

donde $W(X; u, w)$ es el ratio de verosimilitud, como en (2.6). Sea v^* el vector de parámetros que minimiza (2.9), lo llamaremos vector de parámetros óptimo de referencia CE. Análogo a (2.7) podemos estimar la solución óptima como resultado de

$$\max_v \hat{\mathcal{D}}(v) = \max_v \frac{1}{N} \sum_{i=1}^{N} H(X_i) W(X_i; u, w) \ln f(X_i; v), \tag{2.10}$$

donde $X_1, X_2, ..., X_N$ es una muestra aleatoria de $f(\cdot; w)$. En las aplicaciones usuales la función $\hat{D}$ es *cóncava y diferenciable* con respecto a v (véase [14]) y se obtiene resolviendo el siguiente sistema de ecuaciones

$$\frac{1}{N} \sum_{i=1}^{N} H(X_i) W(X_i; u, w) \bigtriangledown \ln f(X_i; v) = 0. \tag{2.11}$$

2.3. Estimación de probabilidades *rare-event*

Sea S una transformación de la muestra X y supongamos que queremos hallar la probabilidad de que $S(X)$ sea mayor o igual que cierto número real γ bajo $f(\cdot; u)$. Es decir, que queremos estimar

$$l = \mathbb{P}_u(S(X) \geq \gamma) = \mathbb{E}_u I_{\{S(X) \geq \gamma\}}.$$

LLamamos *rare-event* a probabilidades de este tipo que sean menores que 10^{-5}. Este es un caso particular de (2.1) con $H(X) = I_{\{S(X) \geq \gamma\}}$. Supondermos que X tiene densidad $f(\cdot; u)$ en alguna familia $\{f(\cdot; v)\}$. Como ya hemos visto podemos estimar l usando el estimador SLR

$$\hat{l} = \frac{1}{N} \sum_{i=1}^{N} I_{\{S(X_i) \geq \gamma\}} W(X_i; u, w). \tag{2.12}$$

El cambio de medida óptimo (de varianza cero) en este caso tiene una interpretación sencilla. De (2.5) con $\mathcal{A} = \{x : S(x) \geq \gamma\}$ tenemos que

$$g^* = \begin{cases} f(x; u) / \int_{\mathcal{A}} f(x; u) \mu(dx) & \text{si} \quad S(x) \geq \gamma \\ 0 & \text{si} \quad S(x) < \gamma \end{cases} \tag{2.13}$$

Vemos que g^* es la densidad condicionada de $X \sim (\cdot; u)$ dado que el evento $I_{\{S(X_i)\geq\gamma\}}$ ocurre.

Observamos que el programa de simulación CE no es muy útil para *rare events*, ya que los indicadores $H(X) = I_{\{S(X)\geq\gamma\}}$ son la mayoría cero. Para estos problemas hacemos un procedimiento CE en dos fases, en los que ambos, el parámetro de referencia v y el nivel γ, son actualizados, evitando así que demasiados indicadores sean cero. Creamos pues una secuencia de pares $\{(v_t, \gamma_t)\}$ para hallar la estimación del parámetro óptimo de referencia, v^*.

Empezamos con un $\hat{v}_0$ y un ϱ (*rarety* parámetro) no muy pequeño, $\varrho = 0.01$, las dos fases son:

1. Actualización adaptativa de γ_t***:*** Para un v_{t-1}, sea γ_t un $(1-\varrho)$-cuantil de $S(X)$ bajo v_{t-1}. Esto es, γ_t satisface

$$\mathbb{P}_{v_{t-1}}(S(X) \geq \gamma_t) \geq \varrho \tag{2.14}$$
$$\mathbb{P}_{v_{t-1}}(S(X) \leq \gamma_t) \geq 1 - \varrho \tag{2.15}$$

donde $X \sim f(\cdot; v_{t-1})$. Un estimador simple de γ_t es el estadístico ordenado. Es decir, para cada $X_1, X_2, ..., X_N$ de $f(\cdot; v_{t-1})$ evaluamos $S(X_i)$ para cada i y los ordenamos $S_{(1)} \leq S_{(2)} \leq ... \leq S_{(N)}$ y finalmente

$$\hat{\gamma}_t = S_{\lceil (1-\varrho)N \rceil}. \tag{2.16}$$

Así escogemos $\hat{\gamma}_t$ de manera que el *rare event* $\{S(X) \geq \gamma_t\}$ no es tan raro ya que tiene probabilidad ϱ. Por lo que actualizar el parámetro de referencia tiene sentido.

2. Actualización adaptativa de v_t***:*** Dados γ_t y v_{t-1}, obtenemos v_t como solución del siguiente programa de CE:

$$\max_v \mathcal{D}(v) = \max_v \mathbb{E}_{v_{t-1}} I_{\{S(X)\geq\gamma_t\}} W(X; u, v_{t-1}) \ln f(X; v). \tag{2.17}$$

Equivalentemente, para unos $\hat{\gamma}_t$ y $\hat{v}_{t-1}$ dados, hallamos $\hat{v}_t$ como solución de

$$\max_v \hat{\mathcal{D}}(v) = \max_v \frac{1}{N} \sum_{i=1}^{N} I_{\{S(X_i)\geq\hat{\gamma}_t\}} W(X_i; u, \hat{v}_{t-1}) \ln f(X_i; v). \tag{2.18}$$

Normalmente la solución óptima se puede obtener analíticamente, en particular, si $f(x; v)$ es una familia exponencial natural o es una función de masa de soporte finito, ver [1]. Ya podemos exponer un algoritmo para este tipo de simulación.

Algoritmo principal de CE para simulación rare-events:

1. Definir $\hat{v}_0 = u$ y $t = 1$ que será el contador de nivel
2. Generar $X_1, ..., X_N$ de la densidad $f(\cdot; v_{t-1})$ y hallar el $(1-\varrho)$-cuantil, $\hat{\gamma}_t$, de S de acuerdo con (2.16), si $\hat{\gamma}_t$ es menor que γ si no poner $\hat{\gamma}_t = \gamma$.
3. Utilizar la **misma** muestra $X_1, ..., X_N$ para resolver el programa (2.18) llamando a la solución $\hat{v}_t$.
4. Si $\hat{\gamma}_t < \gamma$, poner $t = t + 1$ y volver al paso 2, si no vamos al paso 5.
5. Estimar la probabilidad rare-event, l, usando el estimador SLR, $\hat{l}$ de (2.12) con v reemplazado por v_T, donde T es el número final de iteraciones.

Versión determinista del algoritmo:

1. Definir $v_0 = u$ y $t = 1$
2. Calcular γ_t como

$$\gamma_t = \text{máx}\{s : \mathbb{P}_{v_{t-1}}(S(X) \geq s) \geq \varrho\},$$

   si $\gamma_t < \gamma$, sino poner $\gamma_t = \gamma$.
3. Hallar v_t como solución de

$$v_t = \arg\max_v \mathbb{E}_{v_{t-1}} I_{\{S(X)\geq\gamma_t\}} W(X; u, v_{t-1}) \ln f(X; v).$$

4. Si $\gamma_t = \gamma$, entonces PARAR, si no $t = t + 1$ y volver al paso 2.

2.4. CE para optimización

El objetivo del método CE para optimzación es hallar el máximo, que llamaremos γ^*, de una función $S(x)$ para $x \in \mathcal{X}$ con $\mathcal{X}$ finito,

$$\gamma^* = S(x^*) = \max_{x\in\mathcal{X}} S(x). \tag{2.19}$$

Lo vamos a convertir en lo que llamaremos el *problema estocástico asociado, ASP*, para ello primero definimos la familia de funciones de densidad $\{f(\cdot; v), v \in \mathcal{V}\}$ en $\mathcal{X}$ con lo que (2.19) se transforma en

$$l(\gamma) = \mathbb{P}_u(S(X) \geq \gamma) = \mathbb{E}_u I_{\{S(X)\geq\gamma\}}, \tag{2.20}$$

donde X es un vector aleatorio con función de densidad de probabilidad $f(\cdot; u)$ para algún $u \in \mathcal{V}$. A (2.20) se le pueden asociar 2 problemas de estimación; dado l hallar γ ó dado γ hallar l. Nosotros resolveremos el segundo caso y estimaremos l para un cierto γ cercano a γ^*. Lo que convierte a $\{S(X) \geq \gamma\}$ en un rare-event. Usaremos el algoritmo de CE para rare-events del apartado anterior 2.3, haciendo cambios adaptativos en las funciones de densidad de probabilidad acorde con la CE de Kullback-Leibler, creando así una secuencia $f(\cdot; u), f(\cdot; v_1), f(\cdot; v_2), ...$ de funciones de densidad de probabilidad que se va aproximando a la densidad óptima. La siguiente proposición demuestra que a menudo la densidad CE óptima es la densidad degenerada en x^*.

Proposición 2.1 *Sea γ^* el valor máximo de una función real S en $\mathcal{X}$ finito. Supongamos que dicho máximo es único, x^*, y que la clase de funciones de densidad $\{f(\cdot; v)\}$ usados en el programa CE contiene a la densidad degenerada en x^*;*

$$\delta_{x^*}(x) = \begin{cases} 1 & si \quad x = x^* \\ 0 & si \quad x \neq x^* \end{cases}$$

Entonces las soluciones del programa VM (mínima varianza) y CE para estimar $\mathbb{P}(S(X) \geq \gamma^*)$ coinciden y se corresponden con δ_{x^*}.

Demostración: Sea ${}_*v$ tal que $f(\cdot;{}_* v) = \delta_{x^*}(\cdot)$

En el caso del estimador por VM, esta claro que como bajo ${}_*v$ la varinza de $\hat{l}$ de (2.3) es cero, $Var(\hat{l}) = 0$, tenenmos que ${}_*v$ es el parámetro óptimo de referencia con $H(X) = I_{\{S(X) \geq \gamma^*\}}$. Del hecho de que $f(\cdot;{}_* v) \subset \{f(\cdot; v)\}$ es también inmediato que $\mathcal{D}(\delta_{x^*}, f(\cdot; v^*)) = 0$ para $v^* =$, es decir, la solución de ambos programas coincide. □

Ahora, un vez definido el ASP, queremos generar una secuencia de pares $\{(\hat{\gamma}_t, \hat{v}_t)\}$ que converja a un entorno pequeño del par óptimo (γ^*, v^*). Basándonos en el método CE para probabilidades rare-event, inicializamos $v_0 = u$ y $\varrho = 0.01$. Procedemos en 2 pasos:

1. Actualización adaptativa de γ_t: Fijado v_{t-1} sea γ_t el $(1 - \varrho)$-cuantil de $S(X)$ bajo v_{t-1}. Esto es, γ_t satisface (2.14) y (2.15). Análogo al procedimiento con rare-events, escogemos el estimador de γ_t como estadístico ordenado. Es decir, para cada $X_1, X_2, ..., X_N$ de $f(\cdot; v_{t-1})$ evaluamos $S(X_i)$ para cada i y los ordenamos $S_{(1)(2)} \leq ...{}_{(N)}$ y finalmente

$$\hat{\gamma}_t = S_{\lceil (1-\varrho)N \rceil}. \tag{2.21}$$

2. Actualización adaptativa de v_t: Fijados γ_t y v_{t-1}, obtenemos v_t como solución del siguiente programa de CE:

$$\max_v \mathcal{D}(v) = \max_v \mathbb{E}_{v_{t-1}} I_{\{S(X) \geq \gamma_t\}} \ln f(X; v). \tag{2.22}$$

Equivalentemente, para unos $\hat{\gamma}_t$ y $\hat{v}_{t-1}$ dados, hallamos $\hat{v}_t$ como solución de

$$\max_v \hat{\mathcal{D}}(v) = \max_v \frac{1}{N} \sum_{i=1}^{N} I_{\{S(X_i) \geq \hat{\gamma}_t\}} \ln f(X_i; v). \tag{2.23}$$

En este caso, en comparación con (2.18), no usamos el término del ratio de verosimilitud W. Esto se debe a que en la estimación CE de probabilidades rare-event el parámetro u se especifica al inicio y es una parte esencial del problema. Aquí, el parámetro de referncia u en el ASP es bastante arbitrario. De hecho es bueno cambiar el ASP cuando avancemos en las iteraciones. Eliminando el término W estimamos eficientemente en cada iteración el parámetro de referencia v_t para la probabilidad rare-event $\mathbb{P}_{v_t}(S(X) \geq \gamma_t) \geq \mathbb{P}_{v_{t-1}}(S(X) \geq \gamma_t)$. Se podría incluir el término W pero experimentos númericos sugieren que esto sólo introduce ruido en las estimaciones de v_t y v^*.

Otra observación que debemos hacer sobre este método de CE, es que hay que suavizar los parámetros en cada iteración. En vez de calcular en cada iteración v directamente de (2.23), hallamos su versión suavizada

$$\hat{v}_t = \alpha \tilde{v}_t + (1 - \alpha)\hat{v}_{t-1}, \tag{2.24}$$

donde $\tilde{v}_t$ es la solución de (2.23) y α al que llamaremos *parámetro de suavidad*, que suele estar entre $0.7 < \alpha \leq 1$. Suavizamos el parámetro por dos motivos; el primero es que suavizamos los valores de v y la segunda es que reducimos la probabilidad de que alguna componente $\hat{v}_{t,i}$ de $\hat{v}_t$ sea 0 ó 1 en las primeras iteraciones. Eto es importante cuando $\hat{v}_t$ es un vector de probabilidades. Si $0 < \alpha < 1$ entonces para todo i tenemos que $\hat{v}_{t,i} > 0$, y si $\alpha = 1$ puede que $\hat{v}_{t,i} = 0$ ó $\hat{v}_{t,i} = 1$ y no convegería a la solución. Ahora ya podemos resumir el principal algoritmo CE para optimización, suavizando los parámetros.

Algoritmo principal de CE para optimización:

1. Escoger $\hat{v}_0$ y $t = 1$ que será el contador de nivel.
2. Generar $X_1, ..., X_N$ de la densidad $f(\cdot; v_{t-1})$ y hallar el $(1 - \varrho)$-cuantil, $\hat{\gamma}_t$, de S de acuerdo con (2.21).
3. Utilizar la **misma** muestra $X_1, ..., X_N$ para resolver el programa (2.23) llamando a la solución $\tilde{v}_t$.
4. Aplicar (2.24) para suaviazar y hallar $\hat{v}_t$.

5. Si para algún $t \geq d$, por ejemplo $d = 5$

$$\hat{\gamma}_t = \hat{\gamma}_{t-1} = ... = \hat{\gamma}_{t-d}$$

entonces PARAR, si no $t = t + 1$ y nos vamos al paso 2.

Versión determinista del algoritmo:

1. Escoger un v_0 y $t = 1$
2. Calcular γ_t como

$$\gamma_t = \text{máx}\{s : \mathbb{P}_{v_{t-1}}(S(X) \geq s) \geq \varrho\}.$$

3. Hallar v_t como solución de

$$v_t = \arg \max_v \mathbb{E}_{v_{t-1}} I_{\{S(X) \geq \gamma_t\}} \ln f(X; v).$$

4. Si para algún $t \geq d$, por ejemplo $d = 5$

$$\gamma_t = \gamma_{t-1} = ... = \gamma_{t-d}$$

entonces PARAR, si no $t = t + 1$ y nos vamos al paso 2.

2.5. Ejemplo de CE aplicado, clasificación con K-means.

Dado $\mathcal{Z} = \{z_1, ..., z_n\}$ conjunto de puntos de un espacio euclídeo de dimensión d. Queremos hacer una partición de ese conjunto en K clases $R_1, ..., R_K$, con $R_i \bigcap R_j \neq \emptyset$ y $\bigcup_j R_j = \mathcal{Z}$; y tal que minimicemos una determinada función de pérdida. Una función de pérdida usual es

$$\sum_{j=1}^{K} \sum_{z \in R_j} \| z - c_j \|^2, \tag{2.25}$$

donde $c_j = \frac{1}{|R_j|} \sum_{z \in R_j} z$ representa el centro de la clase o *centroide* de la clase R_j.

Llamemos $x = (x_1, ..., x_n)$ con $x_i = j$ cuando $z_i \in R_j$; y $z_{ij} = I_{\{x_i = j\}} z_i$. Entonces podemos escribir (2.25) como

$$\sum_{j=1}^{K} \sum_{i=1}^{n} I_{\{x_i = j\}} \| z_{ij} - c_j \|^2,$$

donde los centroides se pueden escribir como $c_j = \frac{1}{n_j}\sum_{i=1}^{n} z_{ij}$ con $n_j = \sum_{i=1}^{n} I_{\{x_i=j\}}$, que es el número de puntos de la j-ésima clase.

Queremos hacer una partición de $\mathcal{Z}$ en K clases, no necesariamente del mismo tamaño, tal que minimicemos (2.25). En otras palabras, queremos encontrar un vector de centroides $(c_1, ..., c_K)$ y las correspondientes particiones R_j que minimicen (2.25). Para poder aplicar el método de la entropía cruzada aquí, tenemos que ver este problema de clasificación como un problema de optimización *continua de multiextremos*, donde los centroides $c_1, ..., c_K$ son las variables decisivas. Es decir,

$$\min_{c_1,...,c_K} S(c_1, ..., c_K) = \min_{c_1,...,c_K} \sum_{j=1}^{K} \sum_{z \in R_j} \| z - c_j \|^2, \tag{2.26}$$

donde $R_j = \{z : \| z - c_j \| < \| z - c_k \|, k \neq j\}$, es decir que R_j es el conjunto de puntos que están más cerca de c_j que de cualquier otro centroide.

Para ver (2.26) un poco más sencillo cogemos sólo, $K = 2$ clases, usaremos funciones de densidad normales para actualizar los centroides c_j, $j = 1, 2$ y suponemos que cada $z_i \in \mathbb{R}^2$. Asociamos a (2.26) dos distribuciones normales[2] de dimensión 2, $\mathcal{N}(\mu_1, \Sigma_1)$ y $\mathcal{N}(\mu_2, \Sigma_2)$, donde Σ_1 y Σ_2 son las matrices de covarianzas respectivamente. Ponemos que las matrices iniciales de Σ_1 y Σ_2 sean diagonales con grandes varianzas en las diagonales, y procedemos:

1. Elegir determinística o aleatoriamente los vectores iniciales μ_1 y μ_2.
2. Generar $K = 2$ secuencias de centroides, para la clase 1 y 2 respectivamente, $Y_{11}, ..., Y_{1N}$ y $Y_{21}, ..., Y_{2N}$; con $Y_{jk} \sim \mathcal{N}(\mu_j, \Sigma_j)$ independientes $j = 1, 2$. Para cada $k = 1, ..., N$ calculamos la función objetivo como en (2.26), con c_j reemplazado por Y_{ji} con $j = 1, 2$.
3. Aplicar el algoritmo de CE para optimización, con por ejemplo $\varrho = 0.01$ y $\alpha = 0.7$, y actualizamos (μ_1, μ_2) y (Σ_1, Σ_2).
4. Cuando haya terminado el algoritmo cogemos los resultados de las medias (μ_{1T}, μ_{2T}) como estimación de la solución óptima (c_1^*, c_2^*) de (2.27).

[2] *Esta elección la estudiaremos en detalle en la sección 4.2.*

El siguiente ejemplo está expuesto y desarrollado en [1] y el software esta disponible en [2]. Se trata de aplicar el método CE a clasificación de K-means y compararlo con el método de K-means, el conjunto de datos es el *Banana data set* que son 200 observaciones de 2 dimensiones cada, estos puntos se han generado artificialmente usando una mezcla de distribuciones gaussianas.

Hay que especificar los siguientes parámetros: el número de clases K, el tamaño de las muestras que generamos en cada iteración N, el percentil que tomamos para la minimizar ϱ y el parámetro para suavizar α. Los valores iniciales de las medias se han escogido uniformemente dentro de los valores de los datos. También se pueden escoger los valores iniciales como se hace para el método de K-means. Repetimos cada algoritmo 10 veces, y obtenemos: T que es la media del total de iteraciones, $\bar{\gamma}_T$ que es la media de las soluciones, γ es la mejor solución conocida, $\bar{\varepsilon}$ es el error y CPU es la media del tiempo que tarda en segundos.

Vamos a poner los siguientes parámetro iguales para todas las estimaciones: $N = 800$, $\varrho = 0.025$ y $\alpha = 0.7$. Lo que variaremos será el número de clases en las que queremos que nos clasifique.

Para $K = 5$ clases

Método	T	$\bar{\gamma}_T$	γ	$\bar{\varepsilon}$	CPU
CE	49.6	288.49	288.11	0.00	26.67
K-means	9.3	294.31	288.11	0.02	0.09

Para $K = 10$ clases

Método	T	$\bar{\gamma}_T$	γ	$\bar{\varepsilon}$	CPU
CE	75.8	197.22	195.87	0.01	64.25
K-means	9	221.49	195.87	0.13	0.20

Para $K = 20$ clases

Método	T	$\bar{\gamma}_T$	γ	$\bar{\varepsilon}$	CPU
CE	142.1	138.06	135.80	0.02	261.92
K-means	10.1	169.03	135.80	0.24	1.20

Se puede ver que pese a ser significativamente más lento, el método de CE es más exacto y consistente que el de K-means. En ambos casos, el error aumenta conforme lo hace el número de clases, esto se debe a que le pedimos que halle más mínimos. Pero aún así el del método de CE es significativanmente mejor cuantas más clases tengamos. Pese a tener los datos ruido, el método CE obtiene errores aceptables.

PARTE 3

Regresión mediante entropía cruzada.

3.1. El método CE para regresión lineal.

Pongamos el caso más simple de regresión; queremos ajustar una nube de puntos

$$X = (X_1, X_2, ..., X_n)$$

mediante la recta $y = ax + b$.

Para ello, hay que hallar a y b que minimicen, la función objetivo

$$S(X) = (y - (ax + b))^2 \quad \Rightarrow \quad S(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

es decir,

$$\min_{a,b} S(X) = \min_{a,b} \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

equivalentemente

$$\max_{a,b} -S(X) = \max_{a,b} -\sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

Para aplicar el método de la entropía cruzada aquí y ver mejor los pasos que tenemos que hacer y cómo lo vamos a dividir en las siguientes partes:

1. Tenemos que elegir determinística o aleatoriamente los valores iniciales de las distribuciones de los parámetros. Podemos suponer sin problemas que sigan distibuciones normales: $a_0 \sim \mathcal{N}(\mu_{1,0}, \sigma_{1,0})$ y $b_0 \sim \mathcal{N}(\mu_{2,0}, \sigma_{2,0})$.

2. Generamos 2 muestras de tamaño N de los parámetros que queremos hallar. Es decir, que simulamos

$$a_1, a_2, ..., a_N \text{ donde } a_k \sim \mathcal{N}(\mu_{1,t}, \sigma_{1,t}) \text{ con } k = 1, ..., N$$

$$b_1, b_2, ..., b_N \text{ donde } b_k \sim \mathcal{N}(\mu_{2,t}, \sigma_{2,t}) \text{ con } k = 1, ..., N$$

3. Evaluamos la función objetivo para cada par de parámetros (a_k, b_k), $k = 1, ..., N$

$$S_k = \sum_{i=1}^{n} (y_i - a_k x_i - b_k)^2.$$

4. Aplicamos el método de la entropía cruzada para optimización.

5. Ponemos $a = \mu_{1,T}$ y $b = \mu_{2,T}$ siendo T la última iteración.

Veamos más en detalle cómo se hace el paso 4. en este caso concreto.

- Ponemos $t = 1$, será el número de iteración.
- Hallamos $\hat{\gamma}_t$, que es el $(1 - \varrho)$-cuantil de S con $\varrho = 0.01$.
- Hallamos los nuevos parámetros $(\tilde{\mu}_{1,t}, \tilde{\sigma}_{1,t})$ y $(\tilde{\mu}_{2,t}, \tilde{\sigma}_{2,t})$.

  Para ello utilizamos la distancia de Kullback-Leibler, resolvemos el sistema (2.23) que recordemos es

$$\max_v \mathcal{D}(v) = \max_v \frac{1}{N} \sum_{i=1}^{N} I_{\{S(X_i) \geq \hat{\gamma}_t\}} \ln f(X_i; v) \text{ aquí } X_i = a_i, b_i.$$

  Como en nuestro caso $v_{i,t-1} = (\mu_{i,t-1}, \sigma_{i,t-1})$ con $i = 1, 2$, $a_k \sim \mathcal{N}(\mu_{1,t-1}, \sigma_{1,t-1})$ y $b_k \sim \mathcal{N}(\mu_{2,t}, \sigma_{2,t})$ con $k = 1, ..., N$ y suponiendo que estos parámetros son independientes; tenemos que $f(X_i, v_t) = f(a_i, v_{1,t}) \cdot f(b_i, v_{2,t})$. Es decir, que queremos

$$\begin{aligned}
\max_v \mathcal{D}(v) &= \max_v \frac{1}{N} \sum_{i=1}^{N} \ln \left(f(a_i, v_1) \cdot f(b_i, v_2) \right) \\
\max_v \mathcal{D}(v) &= \max_v \frac{1}{N} \sum_{i=1}^{N} \ln \frac{1}{\sigma_1 \sqrt{2\pi}} \exp\left(-\frac{(a_i - \mu_1)^2}{2\sigma_1^2} \right) \\
&\quad \cdot \frac{1}{\sigma_2 \sqrt{2\pi}} \exp\left(-\frac{(b_i - \mu_2)^2}{2\sigma_2^2} \right) \\
\max_v \mathcal{D}(v) &= \max_v \frac{1}{N} \sum_{i=1}^{N} - \ln \sigma_1 \sqrt{2\pi} - \frac{(a_i - \mu_1)^2}{2\sigma_1^2} \\
&\quad - \ln \sigma_2 \sqrt{2\pi} - \frac{(b_i - \mu_2)^2}{2\sigma_2^2}.
\end{aligned}$$

Para obtener el máximo hacemos

$$* \frac{\partial \mathcal{D}}{\mu_1} = 0 \Rightarrow \frac{1}{N} \sum_{i=1}^{N} \left(\frac{a_i - \mu_1}{\sigma_1^2} \right) = 0 \Rightarrow \sum_{i=1}^{N} (a_i - \mu_1) = 0$$

$$\Rightarrow \sum_{i=1}^{N} a_i = \sum_{i=1}^{N} \mu_1 \Rightarrow \sum_{i=1}^{N} a_i = N \cdot \mu_1 \Rightarrow \mu_1 = \frac{\sum_{i=1}^{N} a_i}{N}.$$

$$* \frac{\partial \mathcal{D}}{\mu_2} = 0 \text{ desarrollando como antes tenemos } \mu_2 = \frac{\sum_{i=1}^{N} b_i}{N}.$$

$$* \frac{\partial \mathcal{D}}{\sigma_1} = 0 \Rightarrow \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{2\sigma_1} - \frac{(a_i - \mu_1)^2}{2\sigma_1^3} \right) = 0 \Rightarrow$$

$$\Rightarrow \frac{1}{N} \sum_{i=1}^{N} \left(\sigma_1^2 - (a_i - \mu_1)^2 \right)) = 0 \Rightarrow \sum_{i=1}^{N} (\sigma_1^2) = \sum_{i=1}^{N} (a_i - \mu_1)^2 \Rightarrow$$

$$\Rightarrow N\sigma_1^2 = \sum_{i=1}^{N} (a_i - \mu_1)^2 \Rightarrow \sigma_1 = \sqrt{\frac{\sum_{i=1}^{N} (a_i - \mu_1)^2}{N}}.$$

$$* \frac{\partial \mathcal{D}}{\sigma_2} = 0 \text{ desarrollando como antes } \sigma_2 = \sqrt{\frac{\sum_{i=1}^{N} (b_i - \mu_2)^2}{N}}.$$

Es decir que para los $a_{i,t}$ y $b_{i,t}$ cuyo $S_i \geq \hat{\gamma}_t$

$$\tilde{\mu}_{1,t} = \frac{\sum_{i=1}^{N} a_{i,t}}{N} \qquad \tilde{\sigma}_{1,t} = \sqrt{\frac{\sum_{i=1}^{N} (a_{i,t} - \mu_{1,t-1})^2}{N}}$$

$$\tilde{\mu}_{2,t} = \frac{\sum_{i=1}^{N} b_{i,t}}{N} \qquad \tilde{\sigma}_{2,t} = \sqrt{\frac{\sum_{i=1}^{N} (b_{i,t} - \mu_{2,t-1})^2}{N}}$$

Básicamente los nuevos valores de media y desviación se calculan haciendo la media y la desviación típica de los valores que hacen que la función objetivo esté por encima de $\hat{\gamma}_t$, los que la maximizan.

- Hay que suavizar los parámetros con undeterminado valor de $\alpha \geq 0.7$

$$\hat{\mu}_{1,t} = \alpha\tilde{\mu}_{1,t} + (1-\alpha)\hat{\mu}_{1,t-1}$$

$$\hat{\sigma}_{1,t} = \alpha\tilde{\sigma}_{1,t} + (1-\alpha)\hat{\sigma}_{1,t-1}$$

$$\hat{\mu}_{2,t} = \alpha\tilde{\mu}_{2,t} + (1-\alpha)\hat{\mu}_{2,t-1}$$

$$\hat{\sigma}_{2,t} = \alpha\tilde{\sigma}_{2,t} + (1-\alpha)\hat{\sigma}_{2,t-1}$$

- Paramos si para algún $t \geq d$, por ejemplo $d = 5$ iteraciones

$$\hat{\gamma}_t = \hat{\gamma}_{t-1} = ... = \hat{\gamma}_{t-d}$$

  es decir que ya no varíe el cuantil. Y si no se cumple volvemos al paso 2.

3.2. Ejemplos de CE aplicado a regresión lineal.

Veamos algunos ejemplos del método CE aplicado a regresión lineal. Empezamos por el caso más simple de una recta de regresión.

Ejemplo 3.1:

Se trata del famoso conjunto de datos "Iris"de Edgar Anderson [6], vamos a ver si se puede hacer un ajuste lineal con la longitud y el ancho del pétalo, sin tener en cuenta la especie de Iris a las que pertenece.

$$\text{Ancho de pétalo} = a \cdot \text{largo del pétalo} + b$$

En un primer análisis gráfico parece que si pueden tener una buena relación lineal:

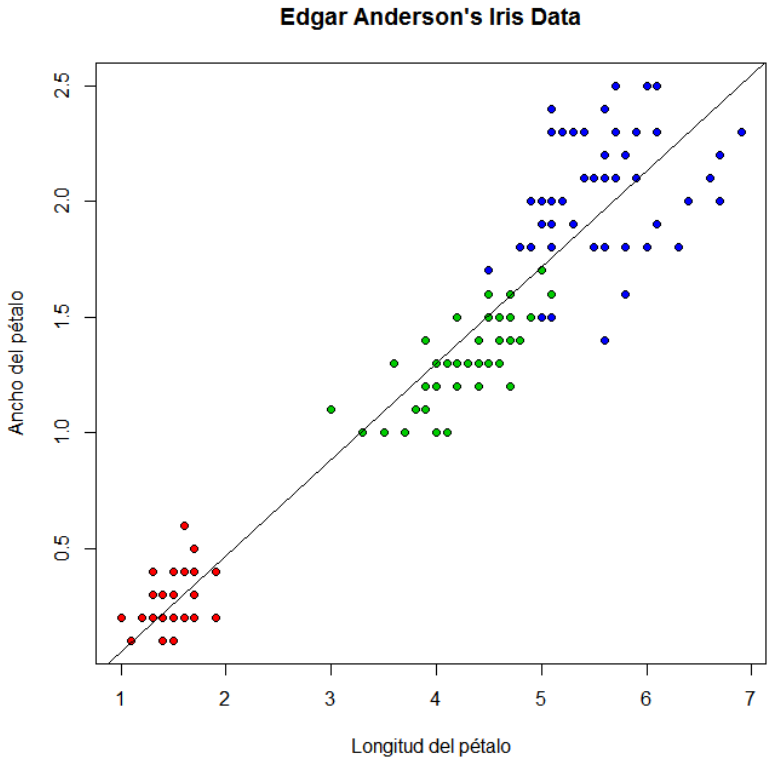


Figura 1: Parecen que hay buena relación lineal independientemente de la especie a la que pertenezcan, que se diferencian por el color.

Aplicando el método de mínimos cuadrados llegamos a que

$$\text{Ancho de pétalo} = 0.416 \cdot \text{largo del pétalo} - 0.363$$

con un error de 0.2065. En los resultados del análisis completo de este método se ve que se trata de un buen ajuste ya que tiene unos p-valores muy bajos, la correlación alta y un error no muy alto. Veamos el resultado hallando el ajuste con el método CE para optimización. Ponemos $a_{k,1} \sim \mathcal{N}(0, 10)$ y $b_{k,1} \sim \mathcal{N}(0, 100)$ como valores iniciales de los parámetros; y $\alpha = 0.7$, $\varrho = 0.01$, $d = 5$ y $N = 500$. En 10 repeticiones obtenemos, en una media de 20 iteraciones:

$$\text{Ancho de pétalo} = 0.416 \cdot \text{largo del pétalo} - 0.363$$

con un error de 0.2064. Es decir, que no hay diferencias significativas con respecto al método de mínimos cuadrados, ni en los parámetros ni en el error. Tarda más tiempo y mejora en muy poco el error. Podemos ver la evolución de los parámetros en la Figura 2:

t	a_t	b_t	$\hat{\gamma}_t$	$S_{t,1}$
"1"	0.00000	0.00000	-7017.53456	-1452224.47842
"2"	-1.03509	3.82417	-1006.46147	-101956.79538
"3"	-0.17107	0.66877	-130.47604	-39871.99123
"4"	0.31805	-0.55798	-13.15159	-2189.67234
"5"	0.38891	-0.47265	-12.57090	-3883.45833
"6"	0.44691	-0.57342	-7.42778	-287.50584
"7"	0.42445	-0.42699	-6.39710	-17.28937
"8"	0.41666	-0.37385	-6.35949	-21.12850
"9"	0.41729	-0.36557	-6.31208	-22.31597
"10"	0.41662	-0.36545	-6.31059	-6.32865
"11"	0.41606	-0.36402	-6.31020	-6.33238
"12"	0.41583	-0.36324	-6.31011	-6.31259
"13"	0.41575	-0.36299	-6.31010	-6.31131
"14"	0.41575	-0.36304	-6.31010	-6.31011
"15"	0.41575	-0.36305	-6.31010	-6.31010
"16"	0.41575	-0.36307	-6.31010	-6.31010
"17"	0.41575	-0.36307	-6.31010	-6.31010
"18"	0.41576	-0.36308	-6.31010	-6.31010
"19"	0.41576	-0.36308	-6.31010	-6.31010

Figura 2: Evolución de los parámetros principales en una de las pruebas.

Ejemplo 3.2:

Hacemos otra prueba con los datos simulados a partir de $y = 3x + 18$, generamos $n = 300$ valores de x y hallamos los valores de y de la siguiente expresión $y = 3x + 18 + \varepsilon$, donde ε son n valores aleatorios de uns $\mathcal{N}(0, 0.1)$.

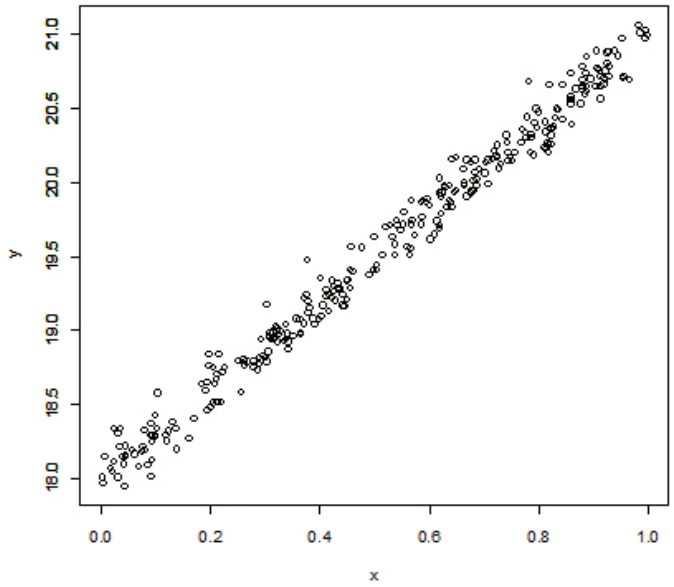


Figura 3: Representación de la nube de puntos generada artificialmente.

Aplicando el método de mínimos cuadrados llegamos a que

$$y = 3.002x + 18.004$$

Se trata de un buen ajuste con un error 0.1006.

Veamos el resultado hallando el ajuste con el método de la entropía cruzada para optimización. Viendo la gráfica podríamos poner $a_{k,1} \sim \mathcal{N}(0, 10)$ y $b_{k,1} \sim \mathcal{N}(0, 100)$ como punto de partida; y $\alpha = 0.7$, $\varrho = 0.01$, $d = 5$ y $N = 500$. Obtenemos el modelo ajustado

$$y = 3.002x + 18.004$$

en una media de 28 iteraciones con un error de 0.1006. Es decir, que tampoco no hay diferencias significativas con respecto al método de mínimos cuadrados, en lo que se refiere a los parámetros a estimar. Tarda mucho más tiempo y mejora en muy poco el error (función objetivo). Podemos ver la evolución de los parámetros más importantes en la siguiente tabla:

t	a_t	b_t	$\hat{\gamma}_t$	$S_{t,1}$
"1"	0.00000	0.00000	-2266.22383	-1227812.18296
"2"	0.76721	13.69927	-915.83691	-3228.37503
"3"	2.88088	16.88135	-161.53884	-250344.75723
"4"	2.53349	18.08552	-34.69823	-1060.91204
"5"	2.94117	18.05282	-8.44754	-1177.72234
"6"	2.94978	18.04805	-3.42107	-264.91840
"7"	2.93777	18.04361	-3.10666	-5.42283
"8"	2.98286	18.01784	-3.04793	-3.22116
"9"	2.99554	18.00759	-3.01968	-3.03150
"10"	2.99887	18.00590	-3.01772	-3.02229
"11"	3.00002	18.00459	-3.01707	-3.02927
"12"	3.00194	18.00413	-3.01701	-3.01837
"13"	3.00240	18.00411	-3.01699	-3.01849
"14"	3.00242	18.00413	-3.01699	-3.01703
"15"	3.00246	18.00412	-3.01699	-3.01700
"16"	3.00244	18.00413	-3.01699	-3.01699
"17"	3.00245	18.00412	-3.01699	-3.01699
"18"	3.00245	18.00412	-3.01699	-3.01699
"19"	3.00245	18.00412	-3.01699	-3.01699
"20"	3.00245	18.00412	-3.01699	-3.01699

Figura 4: Evolución de los parámetros principales en una prueba en concreto.

Ejemplo 3.3:

Vemos otro ejemplo, en este caso de regresión lineal múltiple. Los datos en este caso son los de la library(faraway) y se llaman data(gala) [8]. Contiene el número de especies de tortugas en 30 islas diferentes del archipiélago de las Galápagos entre otras variables. En total contiene 7 variables para cada una de las 30 islas:
Species = nº de especies diferentes de tortugas encontradas en esa isla

Endemics = el área de la isla, (km^2)
Elevation = la altura máxima de la isla, (m)
Nearest = distancia a la isla más cercana, (km)
Scruz = distancia a la isla de Santa Cruz, (km^2)
Adjacent = área de la isla adyacente, (km^2)

Vamos a hacer el siguiente ajuste lineal múltiple

$$Species = a \cdot Area + b \cdot Elevation + c \cdot Nearest + d \cdot Scruz + e \cdot Adjacent + f$$

Aplicando el método de mínimos cuadrados llegamos a que, con un error de 60.98:

$$Species = -0.024 \cdot Area + 0.32 \cdot Elev. + 0.01 \cdot Nearest - 0.24 \cdot Scruz - 0.07 \cdot Adj. + 7.07$$

Veamos el resultado hallando el ajuste con el método de la entropía cruzada para optimización modificado para regresión múltiple. Ponemos $a_{k,1}, b_{k,1}, c_{k,1}, d_{k,1}, e_{k,1}, f_{k,1} \sim \mathcal{N}(0, 10)$ como punto de partida además de $\alpha = 0.7$, $\varrho = 0.01$, $d = 5$ y $N = 300$. Obtenemos

$$Species = -0.024 \cdot Area + 0.32 \cdot Elev. + 0.01 \cdot Nearest - 0.24 \cdot Scruz - 0.07 \cdot Adj. + 7.07$$

con un error de 60.9. Es decir, que no hay diferencias significativas con respecto al método de mínimos cuadrados, en lo que se refiere a los parámetros a estimar. Tarda más tiempo y mejora en muy poco el error (función objetivo). Podemos ver la evolución de los parámetros más importantes:

t	a_t	b_t	c_t	d_t	e_t	f_t	$S_{t,1}$
"1"	0	0	0	0	0	0	120606389.48
"2"	0.57	0	2.63	-5.29	0.38	-0.33	45871324.18
"3"	0.64	-0.43	9.91	-4.33	0.31	-11.76	8255420.64
"4"	0.29	-0.22	13.3	-4.24	0.33	-1.85	2423246.53
"5"	0.24	-0.25	14.04	-2.07	0.18	-2.29	1553967.1
"6"	0.15	-0.07	10.12	-1.44	0.07	-8.45	708472.47
"7"	0.06	0.12	6.37	-0.94	-0.02	-17.18	361606.38
"8"	0.04	0.16	3.47	-0.74	-0.03	-11.6	212833.31
"9"	0	0.28	1.94	-0.52	-0.04	-19.76	169137.77
"10"	0	0.3	1	-0.36	-0.06	-16.04	118650.56
"11"	-0.02	0.34	0.4	-0.27	-0.08	-13.66	102637.75
"12"	-0.03	0.35	-0.01	-0.3	-0.07	-1.51	95093.96
"13"	-0.04	0.34	0	-0.28	-0.07	2.17	91520.19
"14"	-0.03	0.34	0.11	-0.24	-0.07	-0.06	90524.17
"15"	-0.03	0.33	0	-0.22	-0.08	1.4	89747.12
"16"	-0.03	0.33	0.08	-0.25	-0.08	5.37	89528.03
"17"	-0.03	0.33	0.02	-0.25	-0.08	6.97	89423.04
"18"	-0.03	0.32	-0.02	-0.24	-0.08	6.35	89305.31
"19"	-0.03	0.32	0	-0.24	-0.08	6.24	89253.35
"20"	-0.03	0.32	0	-0.24	-0.08	6.83	89247.05
"21"	-0.02	0.32	-0.01	-0.24	-0.08	7.01	89239.47
"22"	-0.02	0.32	0	-0.24	-0.08	7.11	89236.61
"23"	-0.02	0.32	0	-0.24	-0.08	7.03	89233.4
"24"	-0.02	0.32	0	-0.24	-0.08	7.03	89232.36
"25"	-0.02	0.32	0	-0.24	-0.07	6.99	89231.82
"26"	-0.02	0.32	0.01	-0.24	-0.07	7.01	89231.62
"27"	-0.02	0.32	0.01	-0.24	-0.07	7.02	89231.48
"28"	-0.02	0.32	0.01	-0.24	-0.07	7.03	89231.4
"29"	-0.02	0.32	0.01	-0.24	-0.07	7.04	89231.39
"30"	-0.02	0.32	0.01	-0.24	-0.07	7.05	89231.38
"31"	-0.02	0.32	0.01	-0.24	-0.07	7.05	89231.37
"32"	-0.02	0.32	0.01	-0.24	-0.07	7.05	89231.37
"33"	-0.02	0.32	0.01	-0.24	-0.07	7.06	89231.37
"34"	-0.02	0.32	0.01	-0.24	-0.07	7.06	89231.37
"35"	-0.02	0.32	0.01	-0.24	-0.07	7.06	89231.37
"36"	-0.02	0.32	0.01	-0.24	-0.07	7.07	89231.37

Figura 5: Evolución de los parámetros principales en una prueba en concreto.

PARTE 4

El método CE para clasificación Naïve Bayes.

En esta parte trataremos el problema de clasificación que cada vez aparece más y en más diversos ámbitos. Empezaremos hablando del clasificador Naïve Bayes; veremos en qué se basa y cómo funciona, ver también [10]. Después plantearemos cómo se puede aplicar el método CE a este tipo de clasificación y trataremos en detalle el problema de la elección del tipo de distribución con el que generar las muestras del método CE. Terminaremos con algunos ejemplos de experimentos y viendo la mejora que se produce gracias a la entropía cruzada.

4.1. Clasificación Naïve Bayes.

Se trata sin duda de un método muy simple para clasificación mediante redes bayesianas, ver [3]. Que supone que los datos tienen una estructura de red fija (ver Figura 6.) y que sólo tenemos que aprender los parámetros. A pesar de tener una larga tradición en la comunidad de reconocimiento de patrones (Duda y Hart, 1973, en [10]) el clasificador Naïve Bayes aparece por primera vez en la literatura del aprendizaje automático a finales de los ochenta principios de los noventa (Langley, 1992, en [11]) con el objetivo de comparar su capacidad predictiva con la de métodos más sofisticados. De manera gradual los investigadores de esta comunidad de aprendizaje automático se han dado cuenta de su potencialidad y robustez en problemas de clasificación.

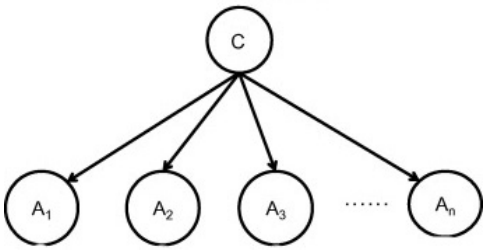


Figura 6: Estructura de la red bayesiana que presenta el método Naïve Bayes.

Lo que queremos es clasificar con el llamado *método MAP* (maximum a posteriori) un conjunto de datos en K clases, ver [3]. Sea C la variable aleatoria que tiene los K valores posibles $\Omega_C = C_1, C_2, ..., C_k$. Este método se basa en:

1. El teorema de Bayes:

$$P(C|A_1, A_2, ..., A_n) = \frac{P(A_1, A_2, ..., A_n|C) \cdot P(C)}{P(A_1, A_2, ..., A_n)},$$

   donde $A_1, A_2, ..., A_n$ son los atributos con los que clasificaremos la observación.

2. La hipótesis MAP; la hipótesis más plausible es la que tiene la máxima probabilidad a posteriori dados los atributos. Es decir que para cada clase tenemos que hallar,

$$C_{MAP} = \arg \max_{C \in \Omega_C} P(C|A_1, ..., A_n) = \arg \max_{C \in \Omega_C} \frac{P(A_1, ..., A_n|C) \cdot P(C)}{P(A_1, ..., A_n)}.$$

   Se puede simplificar y

$$C_{MAP} = \arg \max_{C \in \Omega_C} P(A_1, A_2, ..., A_n|C) \cdot P(C).$$

El método Naïve Bayes se basa en la suposición de que todos los atributos son independientes conocido el valor de la variable clase. Por lo que

$$C_{MAP} = \arg \max_{C \in \Omega_C} P(A_1, A_2, ..., A_n|C) \cdot P(C) = \arg \max_{C \in \Omega_C} \prod_{i=1}^{n} P(A_i|C) \cdot P(C).$$

Tenemos que estimar $P(A_i|C)$ para cada A_i además de la probabilidad a priori de la variable clase $P(C)$. Estas probabilidades las estimaremos con el método CE para optimización, minimizando la función de fallos. Es decir, que tendremos un $C_{MAP,i}$ para cada observación X_i; y queremos que $P(A_i|C)$ y $P(C)$ minimicen la función objetivo

$$\min_{C_{MAP}} \sum_{i=1}^{n} S(X_i) \quad \text{donde} \quad S(X_i) = \begin{cases} 0 & \text{si} \quad C_i = C_{MAP,i} \\ 1 & \text{si} \quad C_i \neq C_{MAP,i} \end{cases}$$

Ahora bien tenemos que distinguir entre

1. **Atributos que son continuos.** En este caso el clsificador Naïve Bayes supone que el atributo sigue una distribución normal $P(A_i|C) \sim \mathcal{N}(\mu, \sigma)$; y nosotros estimaremos los parámetros de esas normales también con distribuciones normales, tal y como lo hemos hecho hasta ahora.

2. **Atributos discretos.** En caso de ser los atributos discretos, los estimaremos igual que las probabilidades a priori $P(C)$ que veremos en detalle en las siguiente sección.

4.2. Elección de la distribución a partir de la que generar los parámetros.

En el caso de queramos generar muestras aleatorias para un parámetro, un número concreto, como por ejemplo el valor de a, la pendiente en el caso de la regresión lineal; parece más que lógico que se escoja la opción de la distribución normal ya que esta distribución nos va a generar valores aleatorios entorno a su media y se va a ir *desplazando* y *apuntando* hasta alcanzar el valor óptimo. Es por esto que conviene siempre elegir valores iniciales grandes para la varianza, que después se va reduciendo considerablemente. Esto se puede ver en cualquiera de los ejemplos ya vistos, cojamos el Ejemplo 3.1 de regresión lineal simple con los datos de "Iris". Teníamos 2 parámetros a estimar a y b. Veamos cómo evolucionan los valores de las normales con las que generamos las secuencias de los parámetros, Figura 7.

t	$\mu_{1,t}$	$\mu_{2,t}$	$\sigma_{1,t}$	$\sigma_{2,t}$
1	0	0	10	100
2	-10.4670	56.7140	4.30155	36.51913
3	-0.4406	31.6926	2.01770	14.2367
4	0.0638	11.8543	1.13517	6.39215
5	0.2158	0.3039	0.52333	2.94459
6	0.3372	-0.0792	0.28163	1.29357
7	0.4032	-0.2876	0.15246	0.61861
8	0.4125	-0.3197	0.05602	0.22803
9	0.4138	-0.3508	0.02239	0.10691
10	0.4159	-0.3632	0.00969	0.04772
11	0.4159	-0.3627	0.00360	0.01728
12	0.4156	-0.3621	0.00169	0.00772
13	0.4156	-0.3625	0.00061	0.00298
14	0.4157	-0.3629	0.00024	0.00136
15	0.4157	-0.3630	0.00009	0.00049

Figura 7: Evolución de los parámetros de las distribuciones normales donde $a \sim \mathcal{N}(\mu_{1,t}, \sigma_{1,t})$ y $b \sim \mathcal{N}(\mu_{2,t}, \sigma_{2,t})$.

En el método CE Naïve Bayes con atributos continuos, para hallar probabilidades $P(A_i|C) \sim \mathcal{N}(\mu, \sigma)$ necesitamos saber μ y σ que serán los parámetros a estimar. Como se trata de valores concretos sin condición alguna los generaremos también a partir de distribuciones normales.

El problema surge cuando queremos generar valores para $P(C_i)$ o $P(A_i|C)$ siendo estas variables aleatorias discretas. No se pueden generar a partir de distribucione normales ya que se podrían generar valores fuera del intervalo $[0, 1]$, y descartar o manipular estos casos, por pocos que sean , distorsionaría mucho en método poniendo en riesgo la convergencia. Hay que descartar también todas las distribuciones que no tengan una forma apuntada, como la distribución Uniforme o la distribución Gamma.

Una de las distribuciones que podría valernos es la distribución Beta, en concreto, $Be(\alpha, \alpha)$ por la forma que tienen, ver Figura 8.

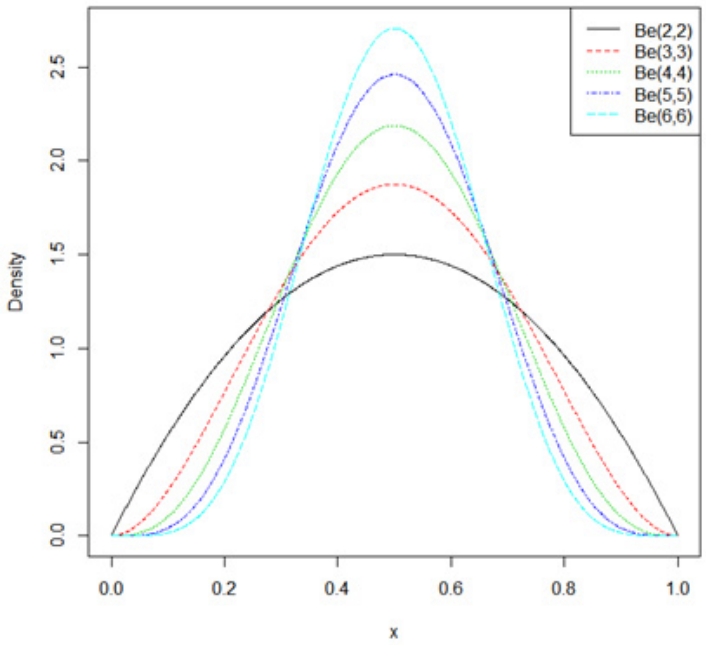
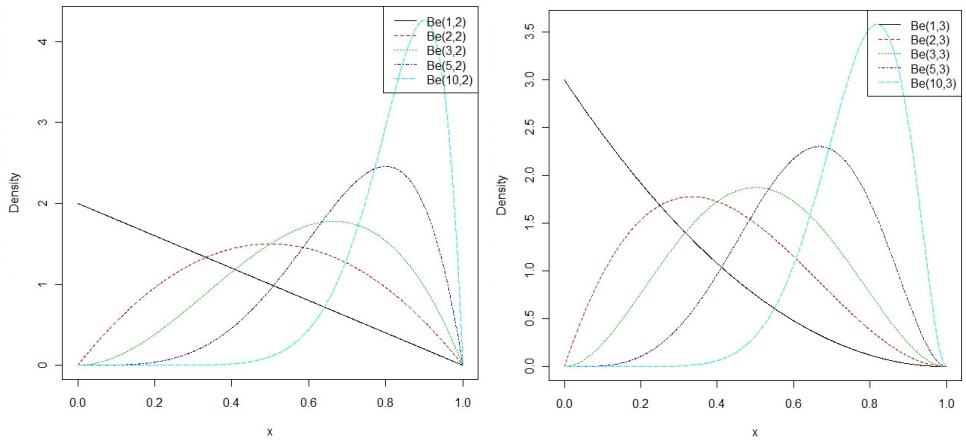


Figura 8: Por su forma apuntada y por estar en el intevalo $[0, 1]$, $Be(\alpha, \alpha)$ parece un buen candidato.

No nos valen este tipo de distribuciones ya que no se puede hallar el parámetro óptimo de referencia con respecto a la distancia de Kullback-Leibler, ver Anexo 1. Hay otra posibilidad sin cambiar de distribución, que es la de considerar las distribuciones Beta con uno de los dos parámetros fijos, por ejemplo $Be(\alpha, 1)$, $Be(\alpha, 2)$ o $Be(\alpha, 3)$. La primera de las opciones la podemos decartar por no tener forma apuntada en ningún caso; y de las otras dos parece mejor $Be(\alpha, 3)$ en cuanto a apuntamiento, ver Figura 9.



(a) Distribuciones $B(\alpha, 2)$ (b) Distribuciones $B(\alpha, 3)$

Figura 9: Dos densidades candidatas para simular valores de probabilidades.

Pero las distribuciones $Be(\alpha, 3)$ no nos valen ya tampoco se puede hallar el parámetro óptimo de referencia con respecto a la distancia de Kullback-Leibler, ver Anexo 2. Sin embargo las distribuciones $Be(\alpha, 3)$ si nos valen y el parámetro óptimo de referencia con respecto a la distancia de Kullback-Leibler, es

$$\alpha = \frac{-N}{\sum_{i=1}^{N} \ln P_i} - 1,$$

el desarrollo está en el Anexo 3.
Y por último esta la opción de generar estas probabilidades a partir de distribuciones log-normales. Esta última es la opción que hemos elegido para nuestros experimentos ya que pese a generar algunos valores por encima de 1, no se altera la convergencia si ajustamos a 1 esos pocos valores. Elegimos esta opción por la forma de las distribuciones, que es muy apuntada para valores por debajo de 0.5 (ver Figura 10); y porque se parece bastante a la normal en lo que se refiere a la estimación de los parámetros. Vamos a ver cuál es el parámetro óptimo de referencia en este caso.

Para ello utilizamos la distancia de Kullback-Leibler, resolvemos el sistema (2.23) que recordemos es

$$\max_{v} \mathcal{D}(v) = \max_{v} \frac{1}{N} \sum_{i=1}^{N} I_{\{S(X_i) \geq \hat{\gamma}_t\}} \ln f(X_i; v) \text{ aquí } X_i = P_i \geq 0.$$

Como en nuestro caso $v_{=}(\mu, \sigma)$, $P_k \sim log\mathcal{N}(\mu, \sigma)$ con $k = 1, ..., N$, entonces queremos

$$\max_v \mathcal{D}(v) = \max_v \frac{1}{N} \sum_{i=1}^{N} \ln \left(\frac{1}{P_i \sigma \sqrt{2\pi}} \exp \left(- \frac{(\ln P_i - \mu)^2}{2\sigma^2} \right) \right)$$

$$\max_v \mathcal{D}(v) = \max_v \frac{1}{N} \sum_{i=1}^{N} \ln \left(- \ln(P_i \sqrt{2\pi}) - \ln(\sigma) - \ln \left(\frac{(\ln P_i - \mu)^2 \cdot \sigma^{-2}}{2} \right) \right).$$

Para obtener el máximo hacemos

$$* \ \frac{\partial \mathcal{D}}{\mu} = 0 \ \Rightarrow \ \frac{1}{N} \sum_{i=1}^{N} - \left(\frac{-2(\ln P_i - \mu)}{2\sigma_2} \right) = 0 \ \Rightarrow \ \sum_{i=1}^{N} (\ln P_i - \mu) = 0$$

$$\Rightarrow \ \sum_{i=1}^{N} \ln P_i = \sum_{i=1}^{N} \mu \ \Rightarrow \ \sum_{i=1}^{N} \ln P_i = N \cdot \mu \ \Rightarrow \ \mu = \frac{\sum_{i=1}^{N} \ln P_i}{N}.$$

$$* \ \frac{\partial \mathcal{D}}{\sigma} = 0 \Rightarrow \frac{1}{N} \sum_{i=1}^{N} - \left(\frac{1}{\sigma} - \frac{(\ln P_i - \mu)^2}{2\sigma^3} \right) = 0 \Rightarrow$$

$$\Rightarrow \frac{1}{N} \sum_{i=1}^{N} \left(\sigma^2 - (\ln P_i - \mu)^2 \right) = 0 \ \Rightarrow \sum_{i=1}^{N} (\sigma^2) = \sum_{i=1}^{N} (\ln P_i - \mu)^2 \Rightarrow$$

$$\Rightarrow N\sigma^2 = \sum_{i=1}^{N} (\ln P_i - \mu)^2 \ \Rightarrow \sigma_1 = \sqrt{\frac{\sum_{i=1}^{N} (\ln P_i - \mu)^2}{N}}.$$

Básicamente los nuevos valores de media y desviación se calculan haciendo la media y la desviación típica del logaritmo de los que se han mantenido por encima del cuantil establecido.

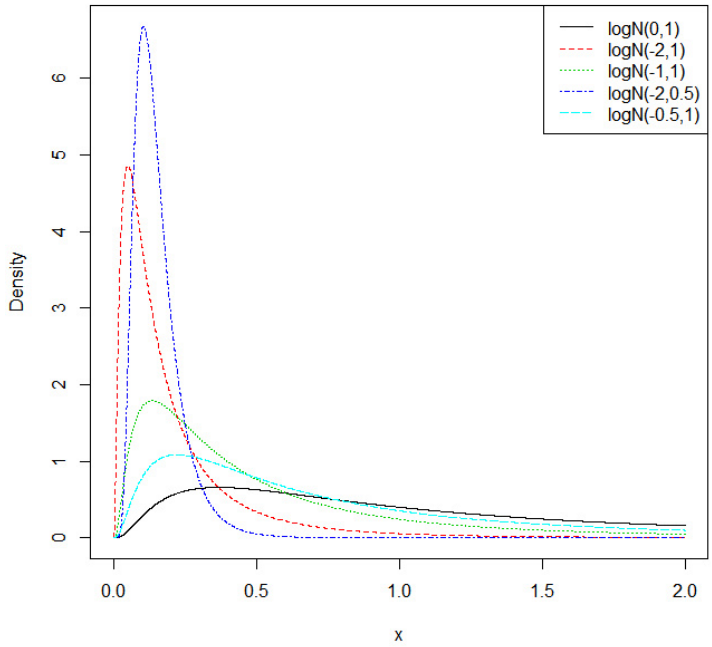


Figura 10: Distribuciones lognormales.

4.3. Algunos ejemplos.

Ejemplo 4.1:

Hacemos la prueba con los datos de la flores Iris de Edgar Anderson, más en concreto vamos a clasificar las observaciones en las distintas especies a partir de los atributos de longitud y ancho de pétalo y sépalo. Las distintas especies de Iris que serán nuestras tres clases son: setosa, versicolor y virginica. Compararemos los resultados obtenidos con el método CE, con los del método Naïve Bayes que tiene *R*, ver [5]. Tenemos que clasificar en $C_1 = setosa$, $C_2 = versicolor$ o $C_3 = virginica$ las 150 observaciones que componen el conjunto de datos Iris, en el que hay exáctamente 50 de cada clase. Para ello tenemos 4 atributos continuos: A_1 =longitud de sépalo, A_2 =ancho de sépalo, A_3 =longitud de pétalo y A_4 =ancho de pétalo.

Usando el método Naïve Bayes de *R* obtenemos los siguientes resultados:

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	47	3
virginica	0	3	47

Se trata de una tabla que refleja los aciertos de las predicciones hechas con el método. Se han cometido 6 fallos clasificando las 150 observaciones. Veamos cómo se aplica el método CE para este tipo de clasificación. Tendremos que estimar $2 \cdot 12 + 3 - 1$ parámetros, ya que al ser continuos los 4 atributos tenemos que hallar los parámetros de cada una de las siguientes probabilidades:

$P(A_1\|C)$	
$C = C_1$	$\mathrm{N}(\mu_{11}, \sigma_{11})$
$C = C_2$	$\mathrm{N}(\mu_{12}, \sigma_{12})$
$C = C_3$	$\mathrm{N}(\mu_{13}, \sigma_{13})$

$P(A_2\|C)$	
$C = C_1$	$\mathrm{N}(\mu_{21}, \sigma_{21})$
$C = C_2$	$\mathrm{N}(\mu_{22}, \sigma_{22})$
$C = C_3$	$\mathrm{N}(\mu_{23}, \sigma_{23})$

$P(A_3\|C)$	
$C = C_1$	$\mathrm{N}(\mu_{31}, \sigma_{31})$
$C = C_2$	$\mathrm{N}(\mu_{32}, \sigma_{32})$
$C = C_3$	$\mathrm{N}(\mu_{33}, \sigma_{33})$

$P(A_4\|C)$	
$C = C_1$	$\mathrm{N}(\mu_{41}, \sigma_{41})$
$C = C_2$	$\mathrm{N}(\mu_{42}, \sigma_{42})$
$C = C_3$	$\mathrm{N}(\mu_{43}, \sigma_{43})$

Para cada parámetro ponemos, para simplificar, que inicialmente siguen una $\mathcal{N}(0,1)$. Pero además tenemos que estimar las probabilidades a priori de las tres clases; $P(C_1)$,$P(C_2)$ y $P(C_3)$. En realidad sólo tenemos que estimar 2 ya que $P(C_3) = 1 - P(C_1) - P(C_2)$. Estas dos probabilidades se estiman, como hemos visto en el apartado anterior, con distribuciones log-normales, $P(C_1) \sim log\mathcal{N}(\mu_1, \sigma_1)$ y $P(C_2) \sim log\mathcal{N}(\mu_2, \sigma_2)$. La siguiente tabla recoge los resultados de distintos experimentos en función de la elección inicial de estos 4 parámetros.

N	$\mu_{1,0}$	$\mu_{2,0}$	$\sigma_{1,0}$	$\sigma_{2,0}$	Nº de iteraciones	Nº de fallos	Tasa de error
500	-3	-3	0.5	0.5	8	6	0.04
500	-3	-3	1	1	16	5	0.033
500	-2	-2	1	1	10	6	0.04
500	-1	-1	1	1	10	6	0.04
500	-5	-5	1	1	11	6	0.04
1000	-5	-5	1	1	16	5	0.033

Parece que la segunda opción es bastante buena veámos la correspondiente tabla de aciertos. Mejoramos en 1 el número de errores y parece que las 5 observaciones restantes están dentro de las 6 que el método Naïve Bayes normal comete.

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	47	4
virginica	0	1	47

Ejemplo 4.2:

Vamos a ver otro ejemplo de claficación con todos los atributos numéricos. Se trata de los datos, de [7], "Haberman's Survival data"de 1999, que contiene 306 observaciones, 3 atributos y dos clases. Es decir, A_1 =edad del paciente al operarse, A_3 =año de la operación y A_3 =número de nodos detectados; y las dos clases son $C_1 = 1$ si el paciente sobrevivió 5 años o más y $C_2 = 2$ si el paciente murió en menos de 5 años. Usando el método Naïve Bayes de R obtenemos los siguientes resultados:

	1	2
1	212	65
2	13	16

Se han cometido 78 fallos clasificando las 306 observaciones.

Veamos cómo se aplica el método CE para este tipo de clasificación. Tendremos que estimar $2 \cdot 6 + 2 - 1$ parámetros, ya que al ser continuos los 3 atributos tenemos que estimarlos a partir de distribuciones normales, igual que antes. Para cada parámetro ponemos, para simplificar, que inicialmente siguen una $\mathcal{N}(0, 1)$. Pero además tenemos que estimar las probabilidades a priori de las dos clases; $P(C_1)$ y $P(C_2)$. En realidad sólo tenemos que estimar 1 ya que $P(C_2) = 1 - P(C_1)$. Estas probabilidades se estiman, como hemos visto en el apartado anterior, con distribuciones log-normales, $P(C_1) \sim log\mathcal{N}(\mu_1, \sigma_1)$. La siguiente tabla recoge los resultados de distintos experimentos en función de la elección inicial de estos 2 parámetros, de la log-normal.

N	me_P_1inicial	sd_P_1inicial	Iteraciones	Nº de fallos
1000	-1	0.1	12	70
500	-1	0.25	10	70
500	-0.7	0.5	19	72

Parece que la segunda opción es bastante buena veamos la correspondiente tabla de aciertos:

	1	2
1	210	55
2	15	26

Mejoramos en 8 el número de observaciones correctamente clasificadas con respecto al método Naïve Bayes normal, lo que supone una mejora del 2.6 % .

Ejemplo 4.3:

Vamos a ver otro ejemplo esta vez con atributos continuos y discretos. Se trata de las datos, de [7], "Teaching Assistant Evaluation"de 1997, que contiene 151 observaciones, 5 atributos y tres clases. Donde $A_1 = 1 \vee 2$ dependiendo de si el alumno tiene ó no el inglés como lengua materna, A_2 =un número asociado a cada profesor, A_3 =número de curso, $A_4 = 1 \vee 2$ dependiendo de si se trata de un semestre normal o de verano y A_5 =tamaño de la clase; y las tres clases son las calificaciones obtenidas por cada alumno $C_1 = 1$ baja, $C_2 = 2$ media y $C_3 = 3$ alta. Usando el método Naïve Bayes de R obtenemos los siguientes resultados:

	1	2	3
1	38	25	12
2	6	14	13
3	5	11	27

Se han cometido 72 fallos clasificando las 151 observaciones.

Hay que mencionar que R aplica el método Naïve Bayes discretizando todas las variables, independientemente de si son discretas, continuas o cualitativas. Este hecho junto con el de que este método es un clasificador muy simple, hacen que las tasas de error sean elevadas. La ventaja que tendremos con el método CE, es que sí distinguiremos entre los distintos tipos que pueden ser las variables.

Los atributos A_2, A_3 y A_5 los trataremos cómo si fuesen continuos. Por lo que hallaremos sus probabilidades condicionadas a partir de distribuciones normales.

$P(A_2\|C)$		$P(A_3\|C)$		$P(A_5\|C)$	
$C = C_1$	$\mathrm{N}(\mu_{21}, \sigma_{21})$	$C = C_1$	$\mathrm{N}(\mu_{31}, \sigma_{31})$	$C = C_1$	$\mathrm{N}(\mu_{51}, \sigma_{51})$
$C = C_2$	$\mathrm{N}(\mu_{22}, \sigma_{22})$	$C = C_2$	$\mathrm{N}(\mu_{32}, \sigma_{32})$	$C = C_2$	$\mathrm{N}(\mu_{52}, \sigma_{52})$
$C = C_3$	$\mathrm{N}(\mu_{23}, \sigma_{23})$	$C = C_3$	$\mathrm{N}(\mu_{33}, \sigma_{33})$	$C = C_3$	$\mathrm{N}(\mu_{53}, \sigma_{53})$

Los atributos A_1 y A_4 son discretos, con soporte finito $\{1, 2\}$, por lo que los generamos a partir de distribuciones log-normales.

$P(A_1\|C)$	$x=1$	$x=2$
$C=C_1$	$P1_{11}$	$P2_{12}$
$C=C_2$	$P1_{12}$	$P2_{12}$
$C=C_3$	$P1_{13}$	$P2_{13}$

$P(A_4\|C)$	$x=1$	$x=2$
$C=C_1$	$P1_{41}$	$P2_{42}$
$C=C_2$	$P1_{42}$	$P2_{42}$
$C=C_3$	$P1_{43}$	$P2_{43}$

Donde $P2 = 1-P1$, es decir que sólo tendremos que generar $P1$ mediante log-normales. Resumiendo, tenemos $18 + 6 \times 2 + 2$ parámetros que estimar. Si ponemos como valores iniciales para todas las distribuciones log-normales $log\mathcal{N}(-2, 0.25)$ y para las normales $\mathcal{N}(0, 50)$ obtenemos para $N = 500$, que en converge en 20 iteraciones con 63 fallos:

	1	2	3
1	28	10	6
2	18	35	21
3	3	5	25

Si ponemos $N = 1000$, en 14 iteraciones, obtenemos

	1	2	3
1	30	10	9
2	16	35	18
3	3	5	25

En este caso cometemos 61 fallos, esto es una mejora de 11 fallos con respecto al método de Naïve Bayes normal, es lo supone algo más de 7 % de mejora del error. Los resultados mediante CE son:

$P(A_2\|C)$	
$C=C_1$	N(0.68,0.46)
$C=C_2$	N(-0.62,0.22)
$C=C_3$	N(1.36,0.36)

$P(A_3\|C)$	
$C=C_1$	N(-0.65,0.31)
$C=C_2$	N(2.38,0.40)
$C=C_3$	N(-0.04,0.24)

$P(A_5\|C)$	
$C=C_1$	N(-0.02,0.33)
$C=C_2$	N(-0.46,0.42)
$C=C_3$	N(-0.13,0.24)

$P(A_1\|C)$	$x=1$	$x=2$
$C=C_1$	0.2479	0.7521
$C=C_2$	0.3423	0.6576
$C=C_3$	0.4012	0.5987

$P(A_4\|C)$	$x=1$	$x=2$
$C=C_1$	0.0061	0.9938
$C=C_2$	0.2094	0.7905
$C=C_3$	0.4136	0.5863

Y las probabilidades a priori con CE son

$$P(C_1) = 0.2657 \quad P(C_2) = 0.4091 \quad P(C_3) = 0.3252.$$

las del método Naïve Bayes de R son

$$P(C_1) = 0.3245 \quad P(C_2) = 0.3311 \quad P(C_3) = 0.3443.$$

Con estos ejemplos se puede comprobar cómo va aumentando la diferencia de error, entre el usar el método tradicional Naïve Bayes o el de CE, conforme aumenta la dificultad del problema de clasificación.

PARTE 5

Conclusiones

Como hemos visto, el método de la entropía cruzada tiene una estructura sencilla que se puede aplicar a muchos problemas. Funciona para problemas simples, como la regresión y también para problemas más complejos, como el de clasificación. En el caso de regresión de no produce mejoras significativas pero esto se debe a que son problemas sencillos que ya están optimizados al máximo con los métodos tradicionales. Sin embargo, en clasificación se obtienen resultados mejores que con las soluciones clásicas. Además, cuanto más complejo sea el problema (más clases, más variables y diferentes tipos de variables) mayor es la mejora del error con este método.

Como defectos de este método de la CE, sólo tenemos que detacar uno, que es el tiempo de ejecución, que en parte se puede reducir algo optimizando los programas. Pero hay problemas que hemos resuelto que no hemos incluido en el trabajo pese a sus buenos resultados en términos de error, ya tardan demasiado tiempo en converger. Por lo demás es bastante sencillo de implementar, porque es simple e intuitivo; y todo el software sobre él esta disponible en [2], el de los experimentos realizados por Rubinstein.

Merece la pena seguir indagando en este tema ya que se puede hacer un mejor estudio de los valores iniciales según el caso en el que los estemos aplicando. Además, aquí no lo hemos hecho por falta de tiempo, se puede aplicar a métodos más sofisticados de clasificación, como los que aparecen en [3], y ver qué mejoras se obtienen. Lo que es interesante ya que los problemas de clasificación aparecen cada vez más en muchos ámbitos. Otra aplicación interesante de este método puede ser a regresión con variables que no sean continuas, por ejemplo con varibles cualitativas; y también será interesante ver cómo hacer regresiones no lineales, que no se puedan estimar mediante mínimios cuadrados, como son las mixturas de exponenciales, que hasta ahora se han resuelto con métodos muy complejos.

PARTE 6

Anexos

Anexo 1

Una de las distribuciones que podría valernos para estimar probabilidades a priori es la distribución Beta, en concreto, $Be(\alpha, \alpha)$ por la forma que tienen, pero no nos valen este tipo de distribuciones ya que no se puede hallar el parámetro óptimo de referencia con respecto a la distancia de Kullback-Leibler. El problema se presenta al resolver el sistema (2.23) que recordemos es

$$\max_v \mathcal{D}(v) = \max_v \frac{1}{N} \sum_{i=1}^{N} I_{\{S(X_i) \geq \hat{\gamma}_t\}} \ln f(X_i; v)$$

Aquí $X_i = P_i \geq 0$ y $v = \alpha$. Es decir, que queremos maximizar,

$$\max_v \mathcal{D}(v) = \max_v \frac{1}{N} \sum_{i=1}^{N} \ln \left(\frac{\left(P_i - P_i^2\right)^{\alpha-1}}{\beta(\alpha, \alpha)} \right)$$

$$\max_v \mathcal{D}(v) = \max_v \frac{1}{N} \sum_{i=1}^{N} \ln \left((\alpha - 1) \ln \left(P_i - P_i^2\right) - \ln(\beta(\alpha, \alpha)) \right)$$

donde $\beta(\alpha, \alpha) = \int_0^1 \left(P_i - P_i^2\right)^{\alpha-1} dP_i$. Y es esta integral la causa de los problemas para poder hallar este máximo.

Anexo 2

Una de las distribuciones que podría valernos para estimar probabilidades a priori es la distribución Beta, en concreto, $Be(\alpha, cte = 3)$ por la forma que tienen, pero no nos valen este tipo de distribuciones ya que no se puede hallar el parámetro óptimo de referencia con respecto a la distancia de Kullback-Leibler. El problema se presenta al resolver el sistema (2.23) que recordemos es

$$\max_v \mathcal{D}(v) = \max_v \frac{1}{N} \sum_{i=1}^{N} I_{\{S(X_i) \geq \hat{\gamma}_t\}} \ln f(X_i; v)$$

Aquí $X_i = P_i \geq 0$ y $v = \alpha$. Es decir, que queremos maximizar,

$$\max_v \mathcal{D}(v) = \max_v \frac{1}{N} \sum_{i=1}^{N} \ln \left(\frac{P_i^{\alpha-1} (1 - P_i)^2}{\beta(\alpha, 3)} \right)$$

$$\max_v \mathcal{D}(v) = \max_v \frac{1}{N} \sum_{i=1}^{N} \ln \left((\alpha - 1) \ln(P_i) + 2 \ln(1 - P_i) - \ln(\beta(\alpha, 3)) \right)$$

donde

$$\beta(\alpha, 3) = \int_0^1 P_i^{\alpha-1} (1 - P_i)^2 dP_i = \frac{2}{\alpha(\alpha + 1)(\alpha + 2)}.$$

Entonces,

$$\max_v \mathcal{D}(v) = \max_v \frac{1}{N} \sum_{i=1}^{N} \left((\alpha - 1) \ln(P_i) + 2 \ln(1 - P_i) - \ln \left(\frac{2}{\alpha(\alpha + 1)(\alpha + 2)} \right) \right)$$

$$\max_v \mathcal{D}(v) = \max_v \frac{1}{N} \sum_{i=1}^{N} \left((\alpha - 1) \ln(P_i) + 2 \ln(1 - P_i) - \ln 2 - \ln(\alpha(\alpha + 1)(\alpha + 2)) \right)$$

Para hallar el máximos hacemos

$$\frac{\partial \mathcal{D}}{\partial \alpha} = \frac{1}{N} \sum_{i=1}^{N} \left(\ln(P_i) + \frac{3\alpha^2 + 6\alpha + 2}{\alpha(\alpha + 1)(\alpha + 2)} \right)$$

$$N \left(\frac{3\alpha^2 + 6\alpha + 2}{\alpha(\alpha + 1)(\alpha + 2)} \right) = - \sum_{i=1}^{N} \ln(P_i)$$

Y en este caso es esta ecuación la que no impide hallar un estimador de α.

Anexo 3

Una de las distribuciones que podría valernos para estimar probabilidades a priori es la distribución Beta, en concreto, $Be(\alpha, cte = 2)$ por la forma que tienen, en este caso si se puede hallar el parámetro óptimo de referencia con respecto a la distancia de Kullback-Leibler. Es decir, que al igual que antes queremos maximizar,

$$\max_{v} \mathcal{D}(v) = \max_{v} \frac{1}{N} \sum_{i=1}^{N} \ln \left(\frac{P_i^{\alpha-1} (1 - P_i)}{\beta(\alpha, 2)} \right)$$

$$\max_{v} \mathcal{D}(v) = \max_{v} \frac{1}{N} \sum_{i=1}^{N} \ln \left((\alpha - 1) \ln(P_i) + \ln(1 - P_i) - \ln(\beta(\alpha, 2)) \right)$$

donde

$$\beta(\alpha, 2) = \int_0^1 P_i^{\alpha-1} (1 - P_i) dP_i = \frac{2}{\alpha + 1}.$$

Entonces,

$$\max_{v} \mathcal{D}(v) = \max_{v} \frac{1}{N} \sum_{i=1}^{N} \left((\alpha - 1) \ln(P_i) + \ln(1 - P_i) - \ln \left(\frac{}{\alpha + 1} \right) \right)$$

$$\max_{v} \mathcal{D}(v) = \max_{v} \frac{1}{N} \sum_{i=1}^{N} \left((\alpha - 1) \ln(P_i) + \ln(1 - P_i) + \ln(\alpha + 1) \right)$$

Para hallar el máximos hacemos

$$\frac{\partial \mathcal{D}}{\partial \alpha} = \frac{1}{N} \sum_{i=1}^{N} \left(\ln(P_i) + \frac{1}{\alpha + 1} \right) \Rightarrow \frac{\partial \mathcal{D}}{\partial \alpha} = 0 \Rightarrow \sum_{i=1}^{N} \left(\ln(P_i) + \frac{1}{\alpha + 1} \right) = 0$$

$$\sum_{i=1}^{N} (\ln(P_i)) + \frac{N}{\alpha + 1} = 0 \Rightarrow (\alpha + 1) \sum_{i=1}^{N} (\ln(P_i)) + N = 0$$

Y finalmente obtenemos

$$\alpha = \frac{-N}{\sum_{i=1}^{N} (\ln(P_i))} - 1.$$

PARTE 7

Bibliografía

- [**1**] D. P. Kroese and R. Y. Rubinstein. The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. Springer, 2004.
- [**2**] http://iew3.technion.ac.il/CE/
- [**3**] J.Hernández Orallo, M.J. Ramírez Quintana, C. Ferri Ramírez. Introducción a laMinería de Datos.
- [**4**] P.-T. de Boer, D.P. Kroese, S. Mannor and R.Y. Rubinstein. A Tutorial on the Cross-Entropy Method. 2005
- [**5**] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel and F. Leisch. Pachkage é1071', 2013.
- [**6**] http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/iris.html
- [**7**] http://archive.ics.uci.edu/ml/datasets.html
- [**8**] http://poincare.matf.bg.ac.rs/ marcone//LSM/faraway.pdf
- [**9**] B. Cascales Salinas, P.L. Salorín, J.M. Mira Ros, A. Pallarés Ruiz y S. Sánchez-Pedreño Guillén. LATEX una imprenta en sus manos.
- [**10**] R. O. Duda, P. E. Hart. Pattern Classification and Scene Analysis, 1973.
- [**11**] P. Langley, W. Iba, K. Thompson. An analysis of Bayesian classifiers, 1992.
- [**12**] R. Y. Rubinstein. Optimization of Computer simulation Models with Rare Events, European Journal of Operations Research, 99, 89-112.
- [**13**] D. MacKay. Information Theory, Inference, and Learning Algorithms. 2003
- [**14**] R. Y. Rubinstein and A. Shapiro. Discrete event systems: Sensitivity analysis and stochastic optimization by the score function method. 1993
- [**15**] M. J. Crawley. The R book. 2008
- [**16**] S. Kullback and R. Leibler. On information and sufficiency, Annals of Mathematical Statistics Pages 76-86. 1951.

- [**17**] R. Y. Rubinstein. Simulation and the Monte Carlo Method. 1981.
- [**18**] J. B. MacQueen. Some Methods for classification and Analysis of Multivariate Observations. 1967

Agradecimientos

Este trabajo no podría haberse realizado sin los conocimientos adquiridos en el presente Máster; aportándome todas sus asignaturas las nociones necesarias para llevar a cabo este estudio.

La realización del presente trabajo final del Máster es fruto de las orientaciones, sugerencias y estímulo del profesor D. Antonio Salmerón Cerdán, quien me ha conducido durante estos meses con un talante abierto y generoso, guiándome sin ser directivo y mostrando en cada momento una inmejorable disposición ante las dudas que durante la realización del mismo me surgieron, aportando valiosas observaciones que en todo momento guiaron este trabajo.

Y, por supuesto a mi pareja y mis familiares.