Aprendizaje de Máquinas Probabilístico: Tarea 1

Autor: Lerko Araya Hernández
Profesor: Felipe Tobar
Auxiliares: Alejandro Cuevas
Alejandro Veragua
Fecha: 27 de abril de 2017

Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Matemática
Universidad de Chile
MA5203 Aprendizaje de Máquinas Probabilístico

The laws of probability, so true in general,
so fallacious in particular.
– Edward Gibbon

---

1. Máxima Verosimilitud

1.1. a)

Para calcular los estimadores de máxima verosimilitud, basta calcular la función de log-verosimilitud que está definida por:

$$\ell(x_1, ..., x_n|\mu, \sigma) = \log p(x_1, ...x_n|\mu, \sigma)$$

Como las variables x_i son i.i.d. $\forall i = 1, ..., n$. Entonces se tiene:

$$\begin{aligned}\ell(x_1, ..., x_n|\mu, \sigma) &= \log \prod_{i=1}^{n} p(x_i|\mu, \sigma) \\ &= \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x_i - \mu)^2}{2\sigma^2}\end{aligned}$$

Finalmente, derivando e igualando a cero para maximizar:

$$\begin{aligned}\frac{d\ell(x_1, ..., x_n|\mu, \sigma)}{d\mu} &= 0 \\ \Rightarrow \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \frac{d\ell(x_1, ..., x_n|\mu, \sigma)}{d\sigma^2} &= 0 \\ \Rightarrow \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\end{aligned}$$

Si el conjunto es solo una muestra, $\hat{\mu} = 1$ y $\hat{\sigma}^2 = 0$. Este caso $\hat{\sigma}^2$ no representa el parámetro, dado que es un estimador sesgado, esto quiere decir que la esperanza del estimador es distinta del parámetro estimado. Por ende, se debe considerar el estimador insesgado y de esta manera $\hat{\sigma}^2_{inses} \to \infty$ para una muestra.

1.2. b)

Para el caso multivariado la función de log-verosimilitud está dada por:

$$\begin{aligned}\ell(x_1, ..., x_n|\mu, \Sigma) &= \log \prod_{i=1}^{n} p(x_i|\mu, \sigma) \\ &= \frac{N}{2}\log(|\Sigma^{-1}|) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu)\end{aligned}$$

Para calcular el estimador de la esperanza basta derivar con respecto a μ e igualar a cero:

$$\begin{aligned}\frac{d\ell(x_1, ..., x_n|\mu, \Sigma)}{d\mu} &= 0 \\ \Rightarrow \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i\end{aligned}$$

Por otro lado, para calcular el estimador para la matriz de covarianza se debe reescribir la expresión y usar que $\frac{\partial}{\partial A} log|A| = A^{-T}$ y $\frac{\partial}{\partial A}\text{Tr}[AB] = B^T$:

$$\begin{aligned}\ell(x_1, ..., x_n|\mu, \Sigma) &= \frac{N}{2}\log(|\Sigma^{-1}|) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu) \\ &= \frac{N}{2}\log(|\Sigma^{-1}|) - \frac{1}{2}\text{Tr}[\Sigma^{-1}\sum_{i=1}^{n}(x_i - \mu)^T(x_i - \mu)]\end{aligned}$$

En este caso, conviene derivar respecto a Σ^{-1}:

$$\begin{aligned}\frac{d\ell(x_1, ..., x_n|\mu, \Sigma)}{d\Sigma^{-1}} &= 0 \\ \Rightarrow \hat{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^T(x_i - \hat{\mu})\end{aligned}$$

1.3. c)

En este caso, nuestro estimador de máxima verosimilitud para la esperanza cambia. Por ende:

$$\begin{aligned}\ell(x_1, ..., x_n|\mu, \sigma_i) &= \log \prod_{i=1}^{n} p(x_i|\mu, \sigma_i) \\ &= \sum_{i=1}^{n} log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\right) - \frac{(x_i - \mu)^2}{2\sigma_i^2} \quad / \frac{\partial(\cdot)}{\partial\mu} = 0 \\ \Rightarrow \hat{\mu} &= \frac{\sum_{i=1}^{n}\frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}\end{aligned}$$

Es decir, las muestras se ponderan por el reciproco de la varianza, lo cual implica que a mayor varianza, la ponderación en menor. Por ende, a los científicos A y B, si tienen una mayor varianza, el estimador les cree menos.

2. Selección de modelos, máxima verosimilitud y optimización

2.1. a)

El modelo para calcular el MAP está dado por:

$$\underset{\theta}{\text{argmax}}\ p(\theta|y,x) = p(y|\theta,x)p(\theta|x)$$

$$p(\theta|y,x) = \frac{1}{\sqrt{2\pi\sigma_\eta^2}} exp\left\{-\frac{(y-f^n(x))^2}{2\sigma_\eta}\right\} \cdot (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)\right\}$$

Donde:

$$f^1(x) = \theta_3 + \theta_2 x$$
$$f^2(x) = \theta_3 + \theta_2 x + \theta_1 x^2$$
$$f^3(x) = \theta_3 + \theta_2 x + \theta_1 x^2 + \theta_0 x^3$$

Optimizando en python con la libraría `pymc3` con $n = 1, 2, 3$ se obtiene:

$$\theta^1 = [2{,}498; 1{,}678; 97{,}100]$$
$$\text{con } l(x_{1:n}) = 561{,}359$$
$$\theta^2 = [1{,}149 \cdot 10^{-2}; 1{,}269; 118{,}780]$$
$$\text{con } l(x_{1:n}) = 556{,}749$$
$$\theta^3 = [5{,}996 \cdot 10^{-5}; 1{,}881 \cdot 10^{-3}; 1{,}678; 115{,}243]$$
$$\text{con } l(x_{1:n}) = 556{,}650$$

Con un $MSE^1_{test} = 4975{,}955$, $MSE^2_{test} = 5461{,}675$ y $MSE^3_{test} = 6889{,}322$.
Considerando estas realizaciones y además el gráfico superior de la figura 1, se puede observar que la componente lineal es la predominante y es la más probable en terminos de verosimilitud, por otro lado, la componente cuadrática realiza un aporte, aunque sea pequeño, esto se puede observar en dado que es la que permite que la función vaya por el promedio de la señal original. Por otro lado, se puede observar que la componente cúbica no realiza mayor aporte, dado que esta es 3 ordenes de magnitud menor que la componente cuadrática del polinomio cuadrático. Con lo anterior, de aquí en adelante se considerará la el polinómio cuadrático para seguir.

2.2. b)

La función de log-verosimilitud está dada por la misma expresión:

$$p(y|\theta,x) = \frac{1}{\sqrt{2\pi\sigma_\eta^2}} exp\left\{-\frac{(y-f(x))^2}{2\sigma_\eta}\right)$$

Solo que esta vez la función $f(\cdot)$ está definida por la expresión del enunciado. Para implementar esta parte se utilizó la librería `mle`, la cual por defecto utiliza BFGS para optimizar.

Así, los parámetros obtenidos por máxima verosimilitud son los siguientes y la forma de la función que inducen se puede observar en el gráfico central de la figura 1:

$$\theta_1 = 11{,}866; \theta_2 = 57{,}070; \theta_3 = 11{,}326; \theta_4 = 0{,}017$$

Con un $MSE_{test} = 1939{,}814$.
Como primera observación, se puede notar que el parámetros θ_4 se encuentra cercano a la condición inicial sugerida en el enunciado, esto muestra que la condición inicial para ese parámetro era buena y el optimizador no tuvo que variar mucho desde ese punto.

2.3. c)

Finalmente, los parámetros obtenidos por máxima verosimilitud obtenidos para esta modelo son los que siguen:

$$\theta_1 = 8{,}972; \theta_2 = 1000{,}080; \theta_3 = -12{,}129; \theta_4 = 0{,}012$$

Con un $MSE_{test} = 1232{,}467$.
Se puede notar que el parámetro θ_2 que está asociado a la frecuencia de la forma de onda, es el doble del modelo anterior, lo cual concuerda con los datos, dado que los datos reales muestran como hay una onda principal y otras ondas más rápidas montadas sobre ella con aproximadamente la mitad del periodo.

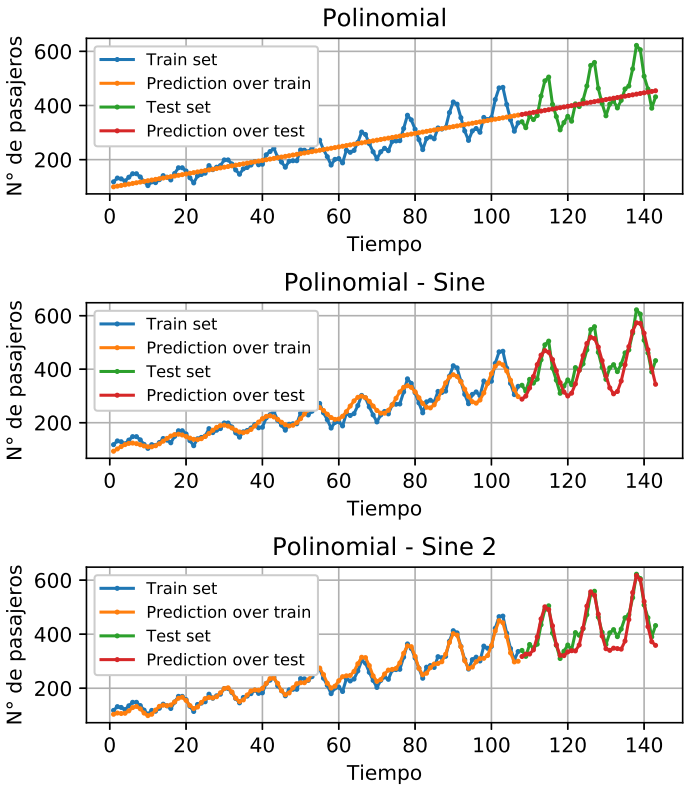

Figura 1: Resultados de la predicción. Superior: Parte (a); Medio: Parte (b); Inferior parte (c).

En este punto, resulta interesante analizar las variaciones de las condiciones iniciales al maximizar la verosimilitud. Para

este modelo, si fijamos $\theta_1 = 10$, $\theta_2 = 10$, $\theta_3 = 10$ y $\theta_4 = 0{,}01$, entonces los resultados obtenidos son:

$$\theta_1 = 391{,}97; \theta_2 = 12{,}501; \theta_3 = 9{,}405; \theta_4 = -0{,}969$$

Con un $MSE_{test} = 4135{,}725$.

Notemos que en este caso, el MSE es mayor que el anterior e idéntido al de la parte (b), esto debido a que θ_4 es negativo, por ende, la amplitud se va a cero y el ajuste no es percibido. De este modo, se puede concluir que la solución encontrada varía dependiendo de las condiciones iniciales. Esto dado que el optimizador encuentra sub-óptimos.

Finalmente, es interesante notar que si bien es posible realizar esta regresión toda de una vez, en la práctica es altamente complejo. Lo anterior, dado que cada parámetro debe ajustarse completamente a los datos, lo cual implica tener buenas condiciones iniciales para que el optimizador no se quede atrapado en sub-óptimos.

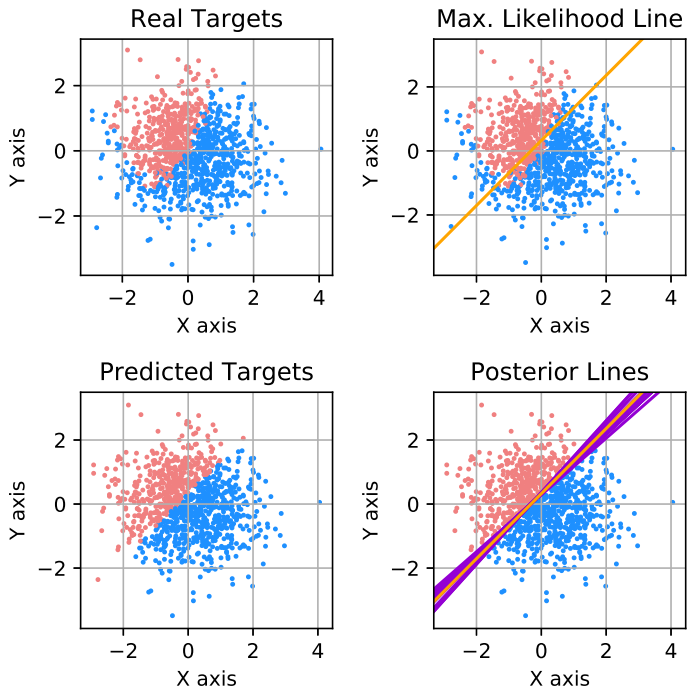

Figura 2: De izquierda a derecha, de arriba hacia abajo. Datos etiquetados. Recta separadora. Etiquetas predichas 20 rectas muestreadas de la posterior.

3. Regresión Logística, clasificación y Metrópolis-Hastings

3.1. a)

Los datos etiquetados se pueden ver en la figura 2. El modelo utilizado para utilizar encontrar los parámetros de la regresión logística consiste en modelar las etiquetas como variables aleatorias con distribución Bernoulli de parámetro sigmoide(x_1, x_2). Formalmente, sea una tripleta (x_1^i, x_2^i, y^i) entonces, la variable $y_i \sim Ber(p(x_1^i, x_2^i))$. Donde:

$$p(x_1^i, x_2^i) = \frac{1}{1 + \exp\left(-\beta_0 - \beta_1 x_1^i - \beta_2 x_2^i\right)}$$

De esta manera, la función de log-verosimilitud está dada por:

$$\ell(y_1, ..., y_n|\beta) = \sum_{i=1}^{n} y \cdot \log p(x_1^i, x_2^i) + (1 - y) \cdot \log(1 - p(x_1^i, x_2^i))$$

Con lo anterior, utilizando la librería `mle`, se tiene que los parámetros óptimos están dados por:

$$\beta_0 = 0{,}717; \beta_1 = 2{,}232; \beta_2 = -2{,}196$$

Es decir, el plano separador está dado por la ecuación 3.1.

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0 \tag{3.1}$$

3.2. b)

Los datos clasificados por la regresión logística se pueden observar en el gráfico derecho de la figura 2. Por otro lado, este sistema obtiene una tasa de clasificación de 87,1 %. Además, la matriz de confusión se puede observar en la figura 3, en donde se puede notar que los datos de la clase 1 (azul) tienen un mejor desempeño.

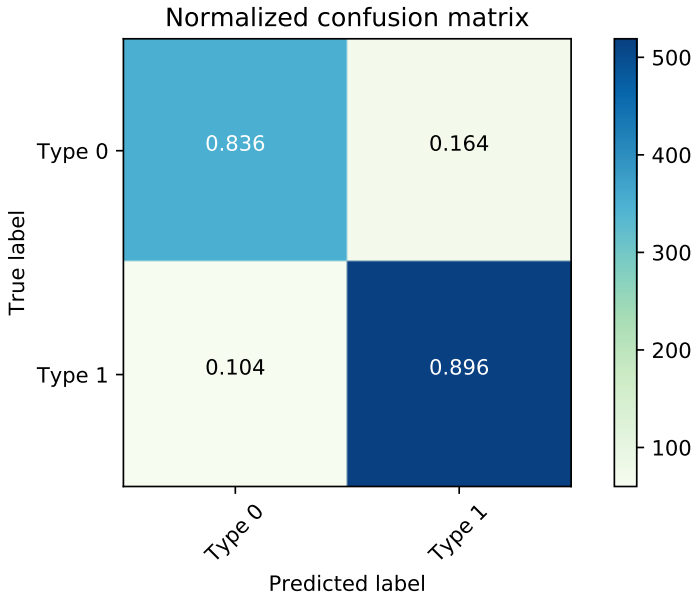

Figura 3: Matriz de confusión para regresión logística sobre los datos.

3.3. c)

Para poder realizar MAP, se debe fijar *prior* sobre los parámetros β_0, β_1 y β_2. De esta manera se asumirá gaussianidad e independencia sobre estos los parámetros. Es decir:

$$\beta_0 \sim N(\beta_0^{init}, \sigma_0^2)$$
$$\beta_1 \sim N(\beta_1^{init}, \sigma_1^2)$$
$$\beta_2 \sim N(\beta_2^{init}, \sigma_2^2)$$

Donde β_0^{init}, β_1^{init} y β_2^{init} son los valores encontrados en la sección (a).

Al utilizar el algoritmo de metrópolis implementado por `pymc3`. Se encuentra que las distribuciones posteriores tienen como máximo un valor cercano a los parámetros iniciales y el *sampler*, al partir en condiciones iniciales con alto valor de la densidad de probabilidad, este se queda muestreando alrededor de la condición inicial (ver figura 4). Sin embargo, si la función se le perturba ligeramente los parámetros iniciales (+10 unidades a β_0), entonces el *sampler* parte desde la condición dada, pero luego vuelve a converger a los parámetros iniciales, ya que estos representan zonas de alta probabilidad, tal como se muestra en la figura 5.

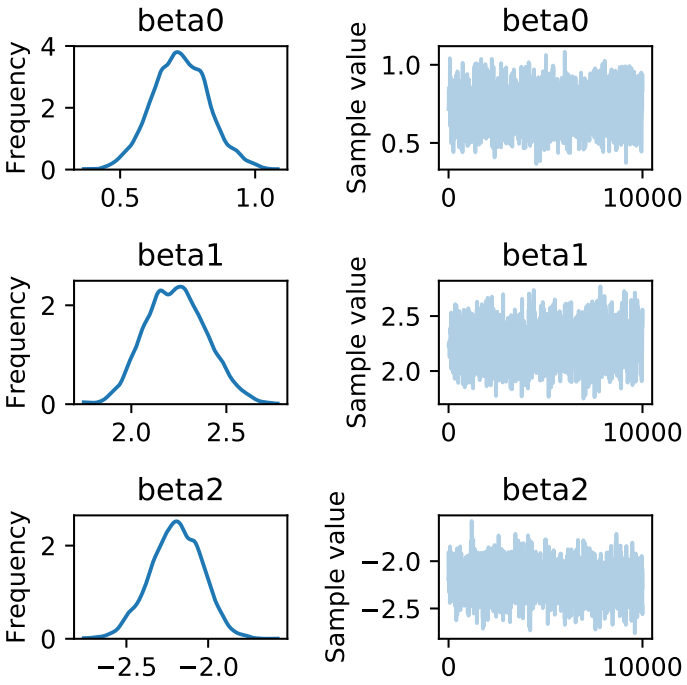

Figura 4: Primera realización de metrópolis. Parámetros iniciales sin perturbar.

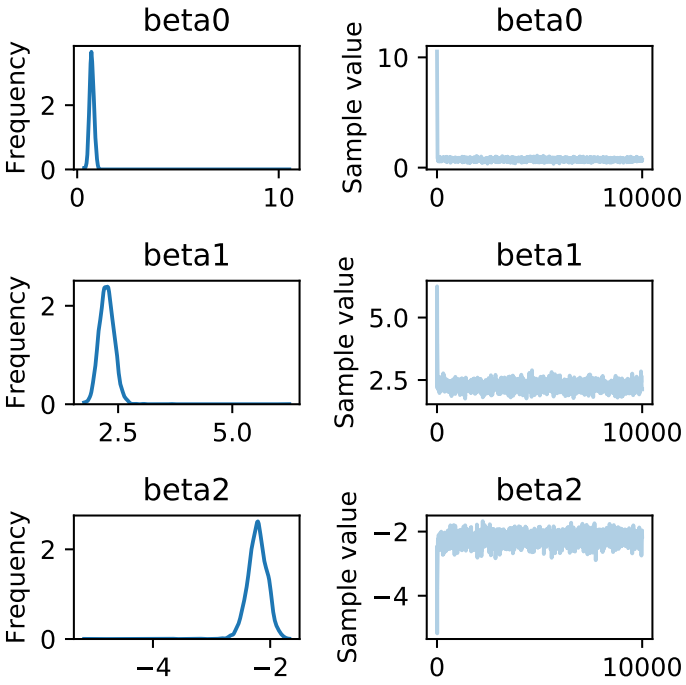

Figura 5: Segunda realización de metrópolis. Parámetros iniciales perturbados.

Finalmente, al analizar visualmente las rectas que muestreadas de la posterior, se puede notar que todas pasan por un mismo punto. Además, se tiene que las rectas están en torno a la recta de verosimilitud, lo cual es reflejo de que la posterior tiene como valor máximo los valores de la verosimilitud.