# Tarea 1: Aprendizaje de Máquinas

Autor: Arturo Lazcano

Profesor: Felipe Tobar

Auxiliares: Catherine Benavides, Camila Bergasa, Víctor Caro, Camilo Carvajal, Diego Cortéz y Stefano Schiappacasse

## P1: Regresión Lineal Bayesiana

Sea $X$ una variable aleatoria en $\mathbb{R}^p$ e $Y$ variable aleatoria a valores $\mathbb{R}$. Consideremos el modelo de regresión lineal: $Y = \beta_0 + \beta^T X + \epsilon$, con $\epsilon \sim \mathcal{N}(0, \sigma^2)$, con $\beta = (\beta_1, \dots, \beta_p)^T$. Suponemos $\sigma^2$ conocido. Nos interesamos en la estimación Bayesiana de los parámetros $\theta = (\beta_0, \beta_1, \dots, \beta_p)$. Suponemos que la distribución $X$ no depende de $\theta$. Sean $(x_i, y_i)\ 1 \leq i \leq N$ datos.

### (a)

Demuestre que el estimador máximo a posteriori con un prior constante (que no depende de $\theta$) equivale al estimador de máxima verosimilitud.

**Demostración:**

Por definición de máximo a posteriori (MAP) tenemos que:

$$\theta_{MAP} = \underset{\Theta}{\operatorname{argmax}}\, p(\theta|\mathcal{D}) = \underset{\Theta}{\operatorname{argmax}}\, \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \underset{\Theta}{\operatorname{argmax}}\, p(\mathcal{D}|\theta)p(\theta)$$

Pues $p(\mathcal{D})$ no depende de $\theta$ y no juega un rol en el argmax. Luego,

$$\underset{\Theta}{\operatorname{argmax}}\, p(\mathcal{D}|\theta)p(\theta) = \underset{\Theta}{\operatorname{argmax}}\, \left[\log(p(\mathcal{D}|\theta)) + \log(p(\theta))\right]$$

Ya que la función $\log$ es una función monótona creciente por lo cual el problema de optimización es equivalente. Así, si escogemos un prior constante (independiente de $\theta$, su logaritmo tampoco) juega un rol en encontrar el argmax por lo que podemos quitarlo de la ecuación. Así, obtenemos:

$\theta_{MAP} = \underset{\Theta}{\operatorname{argmax}}\, \left[\log(p(\mathcal{D}|\theta))\right] = \underset{\Theta}{\operatorname{argmax}}\, \left[p(\mathcal{D}|\theta)\right] = \theta_{MLE}$ ya que podemos pensar en quitar el logaritmo sin afectar al problema de optimización.

Supongamos que usamos el prior siguiente:

$$p(\beta_i) = \frac{1}{\sqrt{2\pi\tau}} \cdot e^{-\frac{\beta_i^2}{2\tau^2}}$$

## (b)

Compute el estimador máximo a posteriori. Identifique a qué estimador conocido corresponde.

**Respuesta:**

Para obtener el máximo a posteriori, procedemos a calcular la expresión:

$$\max_{\Theta} \left[p(\theta|x,y)\right] = \max_{\Theta} \left[p(y|x,\theta)p(\theta)\right]$$

En la ecuación anterior, $\theta$ son nuestros parámetros, es decir, $\theta = (\beta_0, \ldots, \beta_p)$.
Luego, para calcular el primer término, podemos pensar en que al fijar el parámetro $\theta$, la probabilidad de un par de datos $(x, y)$ estará dada por una alrededor de la recta $\beta_0 + \beta^T x$. Así, podemos pensar en una distribución normal pero centrada en $\beta_0 + \beta^T x$ de varianza $\sigma^2$. Esto se refleja en la siguiente ecuación:

$$p(y|x,\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{y-(\beta_0+\beta^T x)}{\sigma}\right)^2}$$

Luego, por enunciado, $p(\beta_i) = \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{\beta_i^2}{2\tau^2}}$ por lo que procedemos a multiplicar estas dos expresiones:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{y-(\beta_0+\beta^T x)}{\sigma}\right)^2} \cdot \frac{1}{\sqrt{2\pi\tau}} \prod_{i=1}^{p} e^{-\frac{\beta_i^2}{2\tau^2}}$$
$$= \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(y-(\beta_0+\beta^T x))^2} \cdot \frac{1}{\sqrt{2\pi\tau}} \prod_{i=1}^{p} e^{-\frac{\beta_i^2}{2\tau^2}}$$
$$= \frac{1}{2\pi\sigma\tau} e^{\frac{-1}{2\sigma^2}(y-(\beta_0+\beta^T x))^2} \cdot \prod_{i=1}^{p} e^{-\frac{\beta_i^2}{2\tau^2}}$$

Por lo tanto, $\theta_{MAP} = \max_{\Theta} \left[\frac{1}{2\pi\sigma\tau} e^{\frac{-1}{2\sigma^2}(y-(\beta_0+\beta^T x))^2} \cdot \prod_{i=1}^{p} e^{-\frac{\beta_i^2}{2\tau^2}}\right]$
y podemos tirar $\log$ por ser una función monótona creciente y no afectar al cálcuo del máximo:

$$\theta_{MAP} = \max_{\Theta} \left[\log\left(\frac{1}{2\pi\sigma\tau} e^{\frac{-1}{2\sigma^2}(y-(\beta_0+\beta^T x))^2} \cdot \prod_{i=1}^{p} e^{-\frac{\beta_i^2}{2\tau^2}}\right)\right]$$
$$= \max_{\Theta} \left[\log\left(\frac{1}{2\pi\sigma\tau} e^{\frac{-1}{2\sigma^2}(y-(\beta_0+\beta^T x))^2}\right) + \log\left(\prod_{i=1}^{p} e^{-\frac{\beta_i^2}{2\tau^2}}\right)\right]$$
$$= \max_{\Theta} \left[\log\left(\frac{1}{2\pi\sigma\tau}\right) - \frac{1}{2\sigma^2}(y-(\beta_0+\beta^T x))^2\right) + \sum_{i=1}^{p} \log\left(e^{-\frac{\beta_i^2}{2\tau^2}}\right)\right]$$
$$= \max_{\Theta} \left[\log\left(\frac{1}{2\pi\sigma\tau}\right) - \frac{1}{2\sigma^2}(y-(\beta_0+\beta^T x))^2\right) + \sum_{i=1}^{p} -\frac{\beta_i^2}{2\tau^2}\right]$$
$$= \max_{\Theta} \left[-\frac{1}{2\sigma^2}(y-(\beta_0+\beta^T x))^2\right) + \sum_{i=1}^{p} -\frac{\beta_i^2}{2\tau^2}\right]$$

Ya que el término $\log\left(\frac{1}{2\pi\sigma\tau}\right)$ no depende de $\theta$ y como $máx f = -mín - f$ obtenemos:

$\theta_{MAP} = \min_{\Theta}\left[\frac{1}{2\sigma^2}(y-(\beta_0+\beta^T x))^2\right) + \sum_{i=1}^{p}\frac{\beta_i^2}{2\tau^2}\Big]$ y multiplicando por $2\sigma^2 \neq 0$ resulta:

$$\theta_{MAP} = \min_{\Theta}\left[(y-(\beta_0+\beta^T x))^2\right) + \frac{2\sigma^2}{2\tau^2}\sum_{i=1}^{p}\beta_i^2\Big]$$

$$= \min_{\Theta}\left[(y-(\beta_0+\beta^T x))^2\right) + \frac{\sigma^2}{\tau^2}||\theta||_2^2\Big]$$

Siendo este la solución del problema de regularización de Tikhonov (Ridge regression) donde $\frac{\sigma^2}{\tau^2} > 0$, y como se puede ver en el apunte, página 19, su solución es explícita y viene de calcular el gradiente con respecto a $\theta$ e imponer esto igual a 0. Así, si definimos la expresión anterior como $J$ y $\rho = \sigma^2/\tau^2$, queda:

$\nabla_\theta J = -2(Y-X\theta)^T X + 2\rho\theta^T = 0$
$\iff -Y^T X + \theta X^T X + \rho\theta^T = 0$
$\iff \theta^T = Y^T X(X^T X + \rho I_d)^{-1}$ donde $(X^T X + \rho I_d)$ es s.d.p por lo cual es invertible
$\iff \theta = (X^T X + \rho I_d)^{-1} X^T Y$

Finamente, concluímos que $\theta_{MAP} = (X^T X + \rho I_d)^{-1} X^T Y$ es el estimador de Ridge.

## (c)

Considere los casos

- $\tau \to \infty$
- $\tau \to 0$

Argumente qué sucede en cada caso, tanto en términos del estimador encontrado, como en términos “Bayesianos” (esto es, piense en la forma que tendrían los priors en ambos casos).

**Respuesta:**

- $\tau \to \infty$ : El estimador $\hat{\theta} = (X^T X + \frac{\sigma^2}{\tau^2} I_d)^{-1} X^T Y$ tiende a $\hat{\theta} = (X^T X)^{-1} X^T Y$, es decir, que el estimador tiende a ser el mínimos cuadrados. Esto también se puede ver en su formulación al problema de optimización, pues al hacer tender $\tau$ a $\infty$ lo que hago es quitar la restricción de regularización, recuperando el problema original (sin regularización).
  Con respecto al prior, la distribución normal tiende a una distribución uniforme que sólo toma valores iguales a 0, sin embargo, al ser una distribución de probabilidad, esta sigue integrando uno, pero a lo largo de toda la recta real. Esto se puede pensar como una distribución uniforme que se acerca lo que más queramos al 0 pero que no es justo igual a 0.

- $\tau \to 0$ : El estimador en este caso queda un término que tiende a infinito en la expresión $\hat{\theta} = (X^T X + \frac{\sigma^2}{\tau^2} I_d)^{-1} X^T Y$, por lo tanto, veamos mejor el problema de optimización correspondiente:
  $\hat{\theta} = \underset{\Theta}{\operatorname{argmin}} \left[ ||Y - X\theta||^2 + \rho ||\theta||_2^2 \right]$
  Acá, vemos que, como $\tau \to 0$ :, $\rho \to \infty$ : por lo que el único caso en el que el problema está bien definido es cuando los parámetros $\theta$ tienden a 0, a la misma (o más) rapidez que $\rho$. Es por esto que, a pesar de no conocer la forma explícita de $\hat{\theta}$ necesitamos que los $\theta$ tiendan a 0.
  Con respecto al prior, vemos que la forma de la normal se va haciendo cada vez más angosta, pues la varianza está tendiendo a 0. Así, al punto límite, vemos que la distribución original (Gaussiana) pasa a ser un delta de Dirac.
  Recordemos que esta "función" (distribución en el sentido matemático) también cumple la propiedad de integrar 1 sobre los reales. Notar que, como estamos haciendo tender la varianza a 0, el punto el cual valdrá 1 en el delta de Dirac es justamente la media, es decir, $0$.

## (d)

Ahora denote $\rho = \frac{\sigma^2}{\tau^2}$ que lo interpretaremos como un hiper-parámetro de nuestro modelo. Demuestre que el estimador resultante de la minimización es insesgado si y sólo si $\rho = 0$.

**Respuesta:**

Recordemos que la definición de estimador insesgado es que para un estimador $\hat{g}(X)$ de $g(\theta)$, se cumpla que $\mathbb{E}(\hat{g}(x)) = g(\theta)$.

Veamos que el estimador $\hat{\theta} = (X^T X + \rho I_d)^{-1} X^T Y$ se puede escribir de la siguiente forma:

$\hat{\theta} = (X^T X + \rho I_d)^{-1} X^T Y = (X^T X + \rho I_d)^{-1} X^T X (X^T X)^{-1} X^T Y$

Donde podemos identificar que los últimos términos corresponden al estimador de mínimos cuadrados ordinario, ie, $\hat{\theta}_{MC}$ por lo tanto obtenemos:

$\hat{\theta} = (X^T X + \rho I_d)^{-1} X^T Y = (X^T X + \rho I_d)^{-1} X^T X \, \hat{\theta}_{MC}$. Así,

$\mathbb{E}(\hat{\theta}|X) = (X^T X + \rho I_d)^{-1} X^T X \, \mathbb{E}(\hat{\theta}_{MC})$ donde sabemos que el estimador $\hat{\theta}_{MC}$ es insesgado, por lo cual su esperanza es $\mathbb{E}(\hat{\theta}_{MC}) = \theta$. Así, reemplazando en la ecuación anterior:

$\mathbb{E}(\hat{\theta}|X) = (X^T X + \rho I_d)^{-1} X^T X \, \theta.$

En la expresión anterior podemos ver que la única opción de que nuestro estimador sea insesgado es que la matriz $(X^T X + \rho I_d)^{-1} X^T X$ sea la identidad. Notemos que esto ocurre

si, y solo si, $\rho = 0$. La implicancia podemos pensar en su contrarecíproca y es trivial ver que, si $\rho \neq 0$, entonces $(X^T X + \rho I_d)^{-1} X^T X$ no es la matriz identidad. Probando el resultado pedido.

# Tarea 1: Aprendizaje de Máquinas

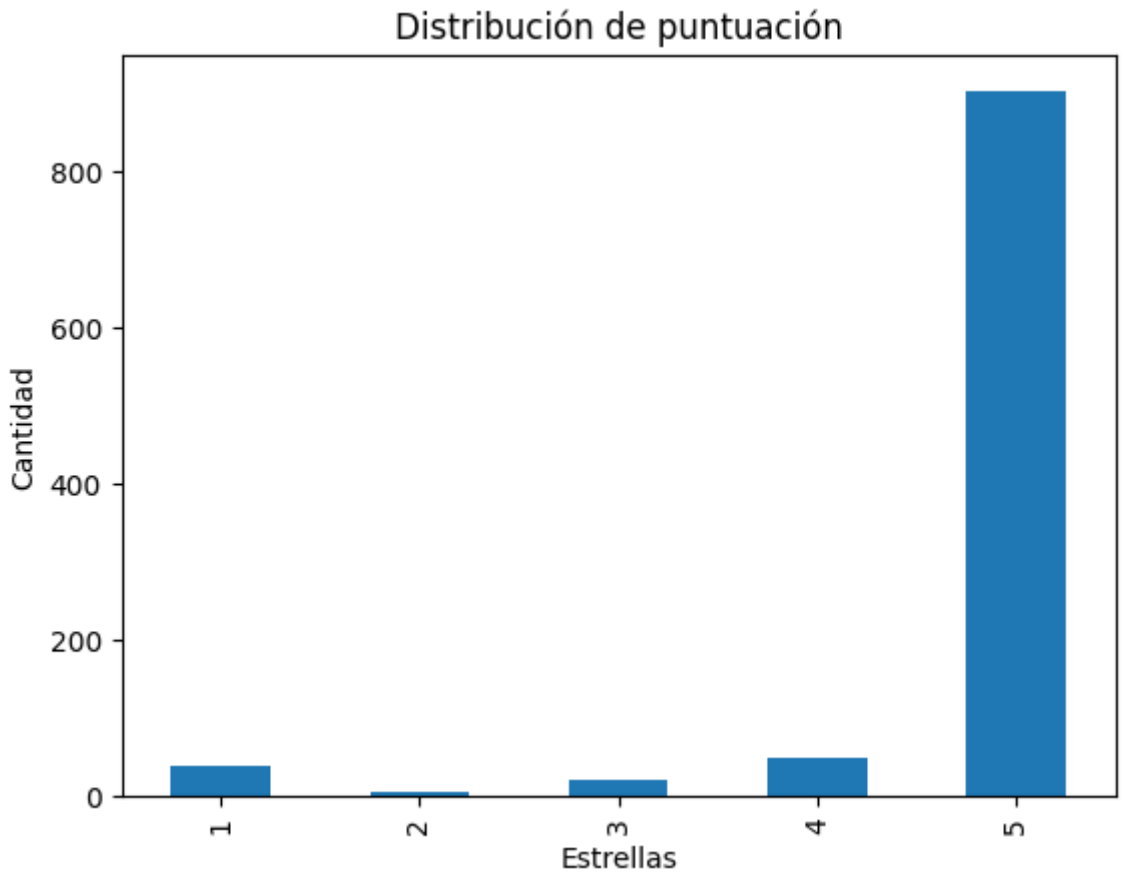
Autor: Arturo Lazcano

Profesor: Felipe Tobar

Auxiliares: Catherine Benavides, Camila Bergasa, Víctor Caro, Camilo Carvajal, Diego Cortéz y Stefano Schiappacasse

## P2

(a) La distribución de datos es como sigue:




Se puede apreciar lo desbalanceado que está el dataset, cargando la mayor parte de sus datos a las 5 estrellas.

(b) Coeficientes de la regresión lineal:

$\theta_0, \theta_1, \theta_2, \theta_3 = (4.647951826289638, 0.11141577759927594, -0.2177152630405732,$ -

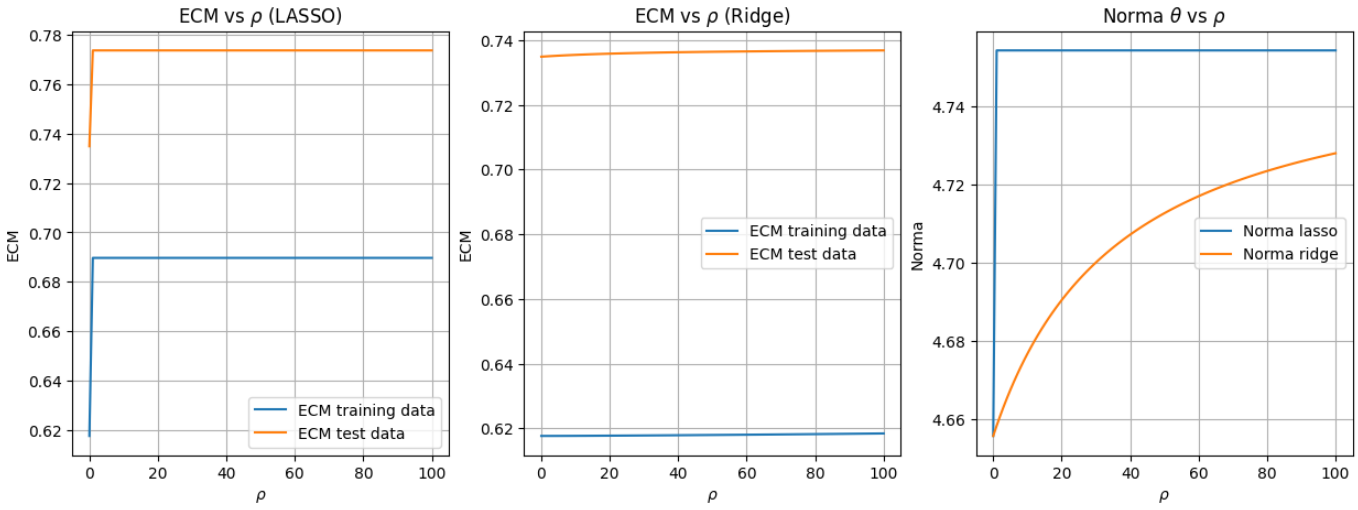
Error Cuadrático Medio (ECM):

Conjunto de entrenamiento: 0.6176515615339531
Conjunto de testeo: 0.7348324250616085

Estos parámetros $(\theta_0, \theta_1, \theta_2, \theta_3)$ son el peso que toma cada variable de entrada $(x)$ para predecir de mejor forma la variable target $(y)$ que en este caso es `star_rating`. Es decir, representan la relación que deben tener en un modelo lineal para ajustar de mejor forma los datos. En este caso, $\theta_0$ es el intercepto, es decir, el término independiente del modelo lineal.

Con respecto a ECM, como sabemos es una métrica de evaluación con rango en $[0, \infty]$, donde mientras más cercano al 0, significa que el modelo entrega un mejor resultado sobre sus predicciones con respecto al dato real.
En este caso los ECM en el conjunto de entrenamiento y testeo son similares y están alrededor de 0.6 - 0.7, lo cual es un número poco interpretable sin datos extras. Sin embargo, que los errores en los distintos conjuntos sean similares es una buena señal de que el modelo no tuvo overfitting.

(c) A continuación se ve el gráfico de lo pedido:




En el tercer gráfico sobre la norma de los parámetros podemos ver que, en magnitud, son similares para ambos modelos, sin embargo, que dependiendo del valor de $\rho$, la norma puede tener tendencias a crecer. En este caso, la norma de $\theta$ para el modelo de Tikhonov es siempre menor al de LASSO.

(d) La siguiente celda contiene un modelo polinómico de grado 2 donde, para una entrada tipo $[x, y]$, su formulación matemática es la siguiente:

$$[1, x, y, x^2, y^2, xy]$$

Es decir, agrega 6 columnas al modelo que son tomadas en cuenta para su entrenamiento (en la práctica, el método fit_transform agrega otra columna de 1's para ajustar dimensionalidad/bias)

Error Cuadrático Medio (ECM):

Conjunto de entrenamiento: 0.6041763828322753
Conjunto de testeo: 0.7503631013903158

(e) A modo de elegir un modelo que resuelva este problema, recomendaría la regresión de Ridge ya que posee menor ECM, lo cual es un buen indicador que el modelo ajusta de mejor forma, notando que en el conjunto de testeo este valor tampoco crece de forma abrupta. Sin embargo, tanto el modelo lineal como de LASSO se escapan con sus métricas, ya que para esta cantidad de datos y tipo de problema todos estos modelos se ajustan de forma similar.
Pensando en el principio de parsimonia, el modelo lineal es también una buena opción.

En caso de usar mayor cantidad de datos, habría que considerar el tiempo de ejecución que pudiésen tener estos algoritmos para tomar en cuenta este factor en la decisión final.

(f) Error cuadrático medio (ECM) Regresión Logística Multiclase:

Conjunto de entrenamiento: 0.7587719298245614
Conjunto de testeo: 1.0098039215686274

En la celda anterior se pueden ver los resultados de la regresión logistica construida desde 0. Con respecto al ECM, estos valores son un poco más altos que el de los modelos vistos anteriormente, del orden de 0.1 - 0.3.
Esto se puede deber al tipo de modelo que estamos imponiendo a resolver el problema como a la implementación no optimizada que tiene esta regresión logistica pues recordemos que, en los modelos anteriores, estos vienen de librerías expertas en la implementación y optimización.

Es por esto que, para este problema, seguiría recomendando uno de los modelos vistos en las preguntas anteriores.

(g) En la teoría, un modelo de regresión es más adecuado que uno de clasificación, meramente por el hecho de que el target en este dataset es `star_rating`, es decir, es un número del 1 al 5. Así, métodos de regresión interpretan fácilmente que un número del estilo 4.8 es más cercano a la predicción del valor 5 que del valor 1.
Recordemos que los modelos de clasificación suelen ser usados con datos categóricos o que posean "diferencias" similares, como puede ser clasificar, por ejemplo, si un animal es perro, gato o conejo, ya que no hay "un animal más cerca que otro".

Sin embargo, en la práctica, dependiendo de os datos, podemos encontrar escenarios muy distintos. Es así, como a pesar de que en la teoría un modelo es mejor que otro, siempre se deben pensar (y ojalá probar) en ambos métodos.