IIC2433 — Minería de Datos

Apuntes: Regresión

1. Regresión Lineal

La primera herramienta que vamos a ver es regresión lineal. Esta herramienta es quizás una de las más clásicas, pero y nos sirve para introducir muchos conceptos que van a ser interesantes en este curso. Por otro lado, por su simplicidad y sus usos cuando queremos explicar lo aprendido, aun se mantiene enormemente vigente en el mundo actual.

1.1. Lo básico

Tenemos una tabla con un conjunto de atributos $A_1, \ldots, A_n$, y un atributo B, todas numéricas. Como explicamos, estas tablas representan una función $f : \mathbb{R}^n \to \mathbb{R}$ que nos interesa aprender. Para el caso de una regresión lineal, vamos a limitarnos a aprender funciones del tipo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n.$$

En otras palabras, estamos suponiendo que nuestra función es una *transformación lineal* que toma vectores en $\mathbb{R}^n$ y entrega un número en $\mathbb{R}$. La transformación de arriba se le llama lineal por que corresponde a una simple multiplicación a nivel de vectores: si $\bar{x} = (x_1, \ldots, x_n)$ y $\beta = (\beta_1, \ldots, \beta_n)$, entonces

$$y = \beta_0 + \beta^T \cdot \bar{x}$$

1.2. Aprendiendo la función

¿Por que nos conviene una función de la forma $y = \beta_0 + \beta^T \cdot \bar{x}$? Primero, por que nos entrega una noción de qué cosas son importantes y cuáles no: el valor de y depende de una combinación lineal de $x_1, \ldots, x_n$ ponderada por los pesos $\beta_1, \ldots, \beta_n$, y por tanto (asumiendo que los valores de $x_1, \ldots, x_n$ tienen el mismo rango), mientras más grande el peso β_i, más grande la importancia del atributo A_i frente al resto.

Segundo, por que el aprender la función de puede hacer de forma analítica, y podemos encontrar una fórmula para eso, lo que es muy práctico. Vamos a ver que hay muchos modelos donde no es posible una función analítica, y hay que *entrar a picar* con métodos de gradiente, pero por ahora disfrutemos de lo simple que son las regresiones.

1.2.1. La verosimilitud

La técnica más importante que vamos a ver en este curso es la idea de máximizar la *verosimilitud* de que una función en particular sea la real. Partamos por un ejemplo.

Imaginen que se encuentran con un psíquico, y les pide que saquen una moneda de su bolsillo, y que la lancen 10 veces. El resultado de esa moneda es:

cara, cara, cara, cara, cara, cara, cara, sello, cara, cara

El psíquico les comenta que con sus habilidades estaba evitando que la moneda saliera en sello, pero que falló en el octavo lanzamiento por que se desconcentró. La hipótesis del psíquico es que la moneda está bajo su control, y que por tanto la probabilidad de que la moneda saque *cara* es, digamos, $0,9$, y la que saque *sello* es $0,1$. Ustedes no creen mucho en los fenómenos paranormales, y creen que las caras fueron pura coincidencia, que la probabilidad de cara es igual a la de sello: $0,5$.

A estas alturas ya sabemos como trabajar: estamos hablando de una función sobre entidades tipo lanzamiento. Cada lanzamiento es una entidad distinta, y la función vale c si la moneda sale cara y s si la moneda sale sello: podemos definir la función como $f : \mathbb{N} \to \{c, s\}$. El psíquico les quiere convencer de que la función real es una función aleatoria que asigna c en un $90\,\%$ de las veces. Ustedes, por otro lado, creen que la función es una función aleatoria donde c y s tienen igual probabilidad. ¿Cuál función podemos asignarle una mayor probabilidad de ser correcta?

- La probabilidad de que las monedas hayan salido como arriba, asumiendo que la moneda es controlada por el psíquico, es $(\frac{9}{10})^7 \cdot \frac{1}{10} \cdot (\frac{9}{10})^2 = 0{,}0387420489$
- La probabilidad de que las monedas hayan salido como arriba si la moneda es normal es $(\frac{1}{2})^{10} = 0{,}0009765625$

A pesar de que ambas probabilidades son bajas, es mucho, mucho más probable haber sacado el lanzamiento de arriba si la moneda la controla el psíquico. Por eso, deberíamos estar más inclinados a aprender esa función.

Veamoslo en términos formales. Tenemos un set de 10 lanzamientos. Lo que queremos es ver la verosimilitud de que esos 10 lanzamientos salieron como los vimos. Cada lanzamiento distribuye de acuerdo a una distribución Bernoulli con parámetro p, y para el caso del psíquico, $p = 0,9$ mientras que $p = 0,5$ para una moneda normal.

Podemos pensar en la verosimilitud como una función que depende de p y de el valor de esos 10 lanzamientos. Es decir, la verosimilitud toma un valor de p, toma una instancia con 10 lanzamientos, y nos dice cual es la probabilidad de haber sacado la secuencia de caras y sellos que vimos en esos lanzamientos. Entonces: la verosimilitud es una función $\mathcal{L}$ (por *likelyhood*, verosimilitud en inglés) definida como

$\mathcal{L}(p|ccccccscc) =$ probabilidad de haber sacado $ccccccscc$,
sabiendo que la prob. de una cara es p,

Lo primero que vamos a asumir que cada lanzamiento es independiente, y por tanto la probabilidad de que esos 10 lanzamientos salieron como los vimos es igual a la multiplicación de la probabilidad de que cada uno de esos lanzamientos salió como lo vimos.

$$\mathcal{L}(p|ccccccscc) = \Pi_{1\leq i\leq 7} L(p|c) \cdot L(p|c) \cdot \Pi_{1\leq i\leq 2} L(p|c)$$

Reemplazando por bernoulli, donde prob de cara es p y prob. de sello es $(1-p)$, tenemos

$$\mathcal{L}(p|x) = p^7(1-p)p^2 = p^9(1-p).$$

Notemos entonces que con un x fijo, la función $\mathcal{L}$ solo depende de p. Lo que nos interesa entonces para nuestro ejercicio es

Encontrar el p que maximiza a $\mathcal{L}(p \mid x)$, dado que $x = ccccccscc$.

En general, lo que buscamos es encontrar el o los parámetros que maximizan la verosimilitud $\mathcal{L}$. ¿Cuál será el parámetro p que maximiza la verosimilitud de arriba, sabiendo que $x = 9$? De acuerdo a la forma de esta función, es posible derivar e igualar a cero (esto no siempre es posible, vamos a discutirlo en el curso).

En nuestro caso, al derivar $p^9(1-p)$ en p tenemos $9(1-p)p^8 - p^9$. Esto tiene dos raíces, $p = 0$ y $p = 0,9$. ¡Justo la probabilidad del psíquico!

Concluímos que $p = 0,9$ es el estimador de la probabilidad de sacar cara que *maximiza la verosimilitud*.

1.2.2. Encontrando los parámetros de una regresión lineal

En una regresión lineal nuestras entidades son vectores $\bar{x} = x_1, \ldots, x_n$, y vamos a asumir que tenemos m de esas entidades, $\bar{x}^1, \ldots, \bar{x}^m$ para las cuales ya sabemos el valor de y, digamos $\bar{y} = y_1, \ldots, y_m$. Entonces, la función $\mathcal{L}$ depende de parámetros $\beta_0, \beta_1, \ldots, \beta_n$.

Lo primero que vamos a asumir es que cada una de las m entidades es independiente entre si[1]. Entonces

$$\mathcal{L} = \Pi_{1\leq j\leq m} \mathcal{L}(\bar{\beta} \mid y_j)$$

Afirmarse los lectores por que vamos a meter una serie de cañonazos, verlos en detalle nos demoraría varias clases.

Cañonazo 1. Queremos calcular $\mathcal{L}(\bar{\beta}|\ y_j)$. Pero cuál es la función de probabilidad de y_j? Conviene pensar en que nuestro modelo entrega valores en base a un error. Entonces la distribución de y es:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon$$

Donde ϵ es usado para la diferencia entre el valor que predice la regresión y el valor real (nadie espera que la regresión ajuste perfecto). Ahora, algunas suposiciones.

[1]Esto no siempre es cierto, hay formas de tratarlo, pero en este curso llegamos hasta acá.

- El error en realidad es un errorcito para cada término: $\epsilon = \epsilon_1 + \cdots + \epsilon_n$, donde estos son mutuamente independientes, cada ϵ_i solo depende de x_i, y estos distribuyen normal con media 0 y varianza σ^2.

- La matriz de $n \times m$ que resulta de tomar todas las entidades es de rango completo (en otras palabras, no hay un atributo que determine a otro).

Cañonazo 2. Con todo esto podemos probar que y_j distribuye como una normal con media $\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$ y varianza σ^2.

Perfecto! ya tenemos la distribución de probabilidad para el resultado de la función, tal como antes teníamos una función de probabilidad para cara y/o sello. Escribimos la fórmula de la verosimilitud para un y_j dado. Para acortar digamos que $\beta = \beta_0, \beta_1, \ldots, \beta_n$ y que $\beta^T \bar{x}^j = \beta_0 + \beta_1 x_1^j, \ldots, \beta_n x_n^j$. Entonces:

$$L(\bar{\beta}, \sigma^2 \mid y_j, \bar{x}^j) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{1}{2\sigma^2}(y_j - \beta^T \bar{x}^j)^2\right)}$$

ahora para escribir la verosimilitud dadas las m entidades $y_1, \ldots, y_m$, simplemente multiplicamos esto (tal como lo hicimos antes).

$$L(\bar{\beta}, \sigma^2 \mid y_1, \ldots, y_m, \bar{x}^1, \ldots, \bar{x}_m) = \frac{1}{\sigma\sqrt{2\pi}} \Pi_{1 \leq j \leq m} e^{-\left(\frac{1}{2\sigma^2}(y_j - \beta^T \bar{x}^j)^2\right)}.$$

Y listo! ahora solo tenemos que encontrar los valores de $\bar{\beta}$ y σ^2 que maximizan la función. Esos son los valores de nuestra regresión, los vamos a escribir como $\hat{\beta}_0, \ldots, \hat{\beta}_n$.

1.2.3. Derivando los parámetros óptimos

Tenemos que encontrar el valor que maximiza L. Vamos a hacer dos cosas. Vamos a hacer dos cosas. Primero, abandonamos el término $\frac{1}{\sigma\sqrt{2\pi}}$, por que no influye en el máximo. Lo segundo que vamos a hacer es calcular el logaritmo de la verosimilitud (llamado en inglés como log-likelyhood). Esto siempre tiene sentido cuando uno minimiza multiplicaciones de cosas mayores que cero: el mínimo es el mismo pero encontrarlo es usualmente más fácil. Buscamos entonces los valores de $\bar{\beta}$ y σ^2 que maximizan:

$$\sum_{1 \leq j \leq m} \log\left(e^{-\left(\frac{1}{2\sigma^2}(y_j - \beta^T \bar{x}^j)^2\right)}\right) = \sum_{1 \leq j \leq m} -\left(\frac{1}{2\sigma^2}(y_j - \beta^T \bar{x}^j)^2\right)$$

Ahora observamos que en realidad la varianza σ^2 no tiene nada que ver con los parámetros que estamos tratando de ajustar (los $\bar{\beta}$) y por lo tanto podemos sacarlo también. Finalmente, tomando el mínimo en vez del máximo multiplicamos todo por -1, para que quede algo positivo, y llegamos a la siguiente formulación:

$$\text{minimizar} \sum_{1 \leq j \leq m} (y_j - \beta^T \bar{x}^j)^2$$

A veces puede que veas esto de acuerdo a una notación matricial. En este caso llamémosle X a la matriz que tiene a cada entidad $\bar{x}^j$ como una fila. Supongamos además que añadimos a X una columna de unos, de forma de poder escribir $\bar{y} = \beta^T X$ (si no le agrega estos 1's no podemos agregar el coeficiente β_0 y que quede como forma matricial). Escribimos nuestro problema como

$$\text{minimizar}||\bar{y} - \beta^T X||^2$$

Al derivar esto, igualar a cero y despejar los β, encontramos los $\hat{\beta}$ que minimizan esa función, y corresponde a

$$\hat{\beta} = (X^T X)^{-}1 X^T \bar{y}.$$

Ejercicio. Escribe esto con sumas, no en forma matricial.

Para profundizar. Existe una derivación alternativa de los factores $\bar{\beta}$ que tiene que ver con minimizar el cuadrado del error entre el y_j real y el y_j que corresponde a $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_n x_n$. Puedes averiguar en internet o en alguno de los libros de la bibliografía como se obtienen estos parámetros. Para una regresión como la que vimos, y con estos supuestos, ambos métodos coinciden, y entregan exactamente los mismos coeficientes $\hat{\beta}$.