IIC2433 — Minería de Datos

Apuntes Minería: Gradiente, Logit

1. Descenso del gradiente

A pesar de que con las regresiones lineales podemos encontrar los parámetros de forma analítica, eso muchas veces no es el caso. E incluso aunque pudiéramos encontrarlo siempre de forma analítica, esto puede que sea costoso (calcular $(X^T X)^-1$ toma tiempo cúbico cuando se programa con herramientas usuales). De cualquier forma, cuando no queremos o no podemos maximizar la verosimilitud de forma "elegante", mediante un desarrollo analítico, podemos usar el método del descenso del gradiente.

La idea del gradiente es simple: ir modificando los parámetros de la verosimilitud (en el caso de la regresión, los β) de a poquito, en la dirección en que van maximizando la verosimilitud. Es importate hacer notar que el método del gradiente es un método general para encontrar mínimos y máximos de funciones. En nuestro caso, por ejemplo, vamos a encontrarnos con estimaciones donde incluso la función de verosimilitud es muy difícil de estimar. En este caso, vamos a tratar de ir minimizando el Mean Squared Error (MSE), o en su defecto, alguna noción de error.

Tenemos entonces tres directrices principales de diseño a la hora de implementar un descenso del gradiente:

- Qué función maximizar o minimizar. Cuando la verosimilitud sea fácil de manejar vamos a optar por esa, pero cuando no, siempre podemos ir por el MSE. Nota que para el caso de la regresión estas dós métricas coinciden: la fórmulación para minimizar la log-verosimilitud es mín $\sum_{1\leq i\leq m}(y_i - \beta\bar{x}^i)^2$ y la fórmula para minimizar el MSE es mín $\frac{1}{n}\sum_{1\leq i\leq m}(y_i - \beta\bar{x}^i)^2$. Claramente el multiplicador por $\frac{1}{n}$ no importa en la minimización.

  Una vez que se elige la función, podemos saber en qué dirección movernos. Esta dirección normalmente está dada por el gradiente de la función evaluada en ese punto. Para el caso de la regresión,

$$\frac{\delta MSE}{\delta \beta_j} = \frac{2}{m}\sum_{1\leq i\leq m}(\beta\bar{x}^i - y_i)x_i,$$

$$\nabla_{\bar{\beta}} = \frac{2}{m}X^T(X\bar{\beta} - \bar{y})$$

- Cuanto avanzar. Evaluar el gradiente en un punto $\bar{\beta}^*$ dado nos da la dirección haca donde queremos minimizar la función. Pero la pregunta es ¿Cuánto avanzamos? Esto

se conoce como el *learning rate*, y es un hiperparametro que normalmente es fijado a priori. **Ejercicio:** piensa en las ventajas y desvetajas de un learning rate alto o bajo.

- Como saber cuando llegamos al óptimo? Es imposible saber, pero de cualquier forma tenemos que tener una condición para cuando parar el algoritmo. Algunas alternativas posibles son: parar cuando de un paso a otro la reducción de la verosimilitud/MSE es más baja que cierto parámetro, parar luego de un número de iteraciones, o una combinación de ambas métricas.

  Un aspecto relacionado es el tema de los óptimos locales vs óptimos globales: puede ser que en el descenso del gradiente nos detengamos en un óptimo local en donde el gradiente tiende a cero, pero que existe un optimo global en otro lado de la función. A veces podemos probar matemáticamente que la función tiene un optimo global (como en el caso de la regresión), pero no siempre. Existe una gran batería de técnicas que intentan refinar el método del descenso de gradiente para lograr sacarlo de óptimos locales, pero están fuera del alcance de este curso.

2. Regresión Logística o modelo Logit

Imaginemos ahora que queremos *clasificar* un conjunto de entidades en dos grupos distintos. Por ejemplo, la gente con estatura mayor a $1,8$ y la menor o igual a $1,8$. O las personas que votaron apruebo en el plebiscito del 2020 y las que votaron rechazo. O las observaciones astronómicas que corresponden a hoyos negros, y los que no. Los otoños que pasan a inviernos lluviosos, y los que no. Si hablamos de *clasificar*, la regresión lineal nos dice los valores de y, y podemos usarla para clasificar cuando las categorías tienen que ver con que el y sea mayor o igual a un threshold T: si $y > T$ entonces es una categoría, si $y <= T$ es otra.

Pero ahora: qué pasa cuando NO tenemos los valores de y para entrenar, pero solo sus categorías? Ahí la regresión no es un buen modelo (y se puede explicar gráficamente de manera simple), y una mejor alternativa es el modelo Logit.

Al igual que en regresión lineal, tenemos entidades $\bar{x} = x_1, \ldots, x_n$, agrupadas en una matriz X. Para cada una de estas entidades $\bar{x}$, sabemos si pertenece a una categoría u otra, y codificamos eso de forma binaria, con un 1 si pertenece a una categoría y un 0 a otra. El modelo Logit usa la hipótesis de que este valor, que nuevamente llamamos y, es un valor entre 0 y 1, distribuye de acuerdo a una distribución de bernoulli con probabilidad p. Esta probabilidad p va a estar dada por evaluar el polinomio $\beta\bar{x}$ en la función sigmoide $\sigma(t) = \frac{1}{1+e^{-t}}$, es decir,

$$p = \frac{1}{1 + e^{-\beta\bar{x}}}$$

2.1. De donde sale usar esto

La función sigmoide termina siendo bastante común en computación, porque tiene la gracia que mapea los reales al intervalo $[0, 1]$. Esto es justo lo que queremos, por que buscamos

una probabilidad que se compute a partir de datos números que no necesariamente son acotados. Pero hay otra razón intuitiva de usar el sigmoide, que explicamos brevemente a continuación. Tomemos una distribución bernoulli con probabilidad p, y pensemos en los *odds*, como $\frac{p}{1-p}$, la razón entre un éxito y un fracaso. Este número va entre 0 e infinito, y es una medida de cuan verosímil es un evento. Para suavizar estos cambios nos concentramos en los *log-odds*, o

$$\ln(\frac{p}{1-p}).$$

Podemos pensar en el modelo logit como un modelo en donde las variaciones en el vector $\bar{x}$, que representa a las entidades, influye de forma lineal en las log-odds:

$$\begin{aligned} \beta\bar{x} &= \ln(\frac{p}{1-p}) \\ e^{\beta\bar{x}} &= \frac{p}{1-p} \\ e^{\beta\bar{x}} - pe^{\beta\bar{x}} &= p \\ e^{\beta\bar{x}} &= p(1+e^{\beta\bar{x}}) \\ p &= \frac{e^{\beta\bar{x}}}{1+e^{\beta\bar{x}}} = \frac{1}{1+e^{-\beta\bar{x}}} \end{aligned}$$

2.2. Estimando el modelo a partir de máxima verosimilitud

Veamos el valor de $\mathcal{L}(\beta \mid y)$ para este modelo. Recordemos que los y distribuyen Bernoulli, con probabilidad $p = p(\beta\bar{x}) = \frac{1}{1+e^{-\beta\bar{x}}}$, y por tanto

$$\mathcal{L}(\beta \mid y) = \big(p(\beta\bar{x})\big)^y \big(1 - p(\beta\bar{x})\big)^{1-y}$$

y reordenando y usando el hecho que $\ln(\frac{p}{1-p}) = \beta\bar{x}$ visto arriba, y por lo tanto $\frac{p}{1-p} = e^{\beta\bar{x}}$, tenemos:

$$\begin{aligned} \mathcal{L}(\beta \mid y) &= \big(\frac{p(\beta\bar{x})}{1-p(\beta\bar{x})}\big)^y (1 - p(\beta\bar{x})) && (1) \\ &= \big(e^{\beta\bar{x}}\big)^y (1 - p(\beta\bar{x})) && (2) \\ &= \big(e^{\beta\bar{x}}\big)^y (\frac{1}{1+e^{\beta\bar{x}}}), && (3) \end{aligned}$$

En donde la tercera ecuación sale de la definición de $p(\beta\bar{x})$.

Como decíamos, asumamos que contamos con m entidades, y para cada entidad $\bar{x}j$ conocemos su valor final y_j. Obtenemos que $\mathcal{L}(\beta \mid \bar{y}) = \Pi_{1\leq i\leq m}\mathcal{L}(\beta \mid y_j)$. Pasamos de nuevo a la log-verosimilitud $\ell(\beta \mid \bar{y})$ y planteamos nuestro problema:

$$\text{maximizar } \ell(\beta \mid \bar{y}) = \sum_{1\leq j\leq m} \ln\left(\big(e^{\beta\bar{x}^j}\big)^{y_j} (\frac{1}{1+e^{\beta\bar{x}^j}})\right)$$

y reordenando

$$\ell(\beta \mid \bar{y}) \quad = \quad \sum_{1 \leq j \leq m} \big(y_j(\beta \bar{x}^j)\big) - \sum_{1 \leq j \leq m} \ln\big(1 + e^{\beta \bar{x}^j}\big)$$

Podemos demostrar que no existe una solución analítica para este problema. Pero siempre podemos usar el método del gradiente!

la derivada de $\ell(\beta \mid \bar{y})$ con respecto a algún β_i es:

$$\frac{\delta \ell(\beta \mid \bar{y})}{\delta \beta_i} \quad = \quad \sum_{1 \leq j \leq m} (y_j x_i^j) - \sum_{1 \leq j \leq m} \frac{x_i^j}{1 + e^{-\beta \bar{x}^j}} \tag{4}$$

$$\frac{\delta \ell(\beta \mid \bar{y})}{\delta \beta_i} \quad = \quad \sum_{1 \leq j \leq m} x_i^j \big(y_j - \frac{1}{1 + e^{-\beta \bar{x}^j}}\big) \tag{5}$$