Cuestionario 2. Aprendizaje Automático

José Carlos Martínez Velázquez

14 de Mayo de 2016

1. Sean x e y dos vectores de observaciones de tamaño N . Sea

$$cov(x, y) = \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

la covarianza de dichos vectores, donde $\bar{z}$ representa el valor medio de los elementos de z. Considere ahora una matriz X cuyas columnas representan vectores de observaciones. La matriz de covarianzas asociada a la matriz X es el conjunto de covarianzas definidas por cada dos de sus vectores columnas. Defina la expresión matricial que expresa la matriz $cov(X)$ en función de la matriz X

La respuesta es:

$$cov(X) = \begin{pmatrix} \frac{1}{N}\sum_{i=0}^{n}(x_{i,1} - \bar{x_1})(x_{i,1} - \bar{x_1}) & \dots & \frac{1}{N}\sum_{i=0}^{n}(x_{i,1} - \bar{x_1})(x_{i,ncol} - \bar{x_{ncol}}) \\ \frac{1}{N}\sum_{i=0}^{n}(x_{i,2} - \bar{x_2})(x_{i,1} - \bar{x_1}) & \dots & \frac{1}{N}\sum_{i=0}^{n}(x_{i,1} - \bar{x_1})(x_{i,ncol} - \bar{x_{ncol}}) \\ \vdots & \ddots & \vdots \\ \frac{1}{N}\sum_{i=0}^{n}(x_{i,ncol} - \bar{x_{ncol}})(x_{i,1} - \bar{x_1}) & \dots & \frac{1}{N}\sum_{i=0}^{n}(x_{i,ncol} - \bar{x_{ncol}})(x_{i,ncol} - \bar{x_{ncol}}) \end{pmatrix}$$

Veamos por qué.

Supongamos que tenemos la matriz X:

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,ncol} \\ x_{2,1} & x_{2,2} & \dots & x_{2,ncol} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,ncol} \end{pmatrix}$$

Tal que cada columna de X es un vector de observación y *ncol* es el número de columnas o vectores de observación que tiene X. Si quisiéramos obtener la covarianza entre el primer vector de observación y el segundo, recurriríamos a la expresión:

$$cov(X)_{1,2} = \frac{1}{N}\sum_{i=0}^{n}(x_{i,1} - \bar{x_1})(x_{i,2} - \bar{x_2})$$

Debemos generalizar, pues queremos la covarianza de cada dos columnas (vectores de observación) cualesquiera de X. La covarianza, pues, de dos vectores de observación j, k cualesquiera de la matriz X será:

$$cov(X)_{j,k} = \frac{1}{N}\sum_{i=0}^{n}(x_{i,j} - \bar{x_j})(x_{i,k} - \bar{x_k})$$

Siendo esto así, la matriz de covarianza resultante de una matriz X, será una matriz simétrica (pues es lo mismo $\frac{1}{N}\sum_{i=0}^{n}(x_{i,j} - \bar{x_j})(x_{i,k} - \bar{x_k})$ que $\frac{1}{N}\sum_{i=0}^{n}(x_{i,k} - \bar{x_k})(x_{i,j} - \bar{x_j})$), de dimensión *ncol* x *ncol*, donde cada posición $cov(X)_{j,k}$ es la covarianza del vector (columna) j con respecto al vector (columna) k:

$$cov(X) = \begin{pmatrix} \frac{1}{N}\sum_{i=0}^{n}(x_{i,1} - \bar{x_1})(x_{i,1} - \bar{x_1}) & \dots & \frac{1}{N}\sum_{i=0}^{n}(x_{i,1} - \bar{x_1})(x_{i,ncol} - \bar{x_{ncol}}) \\ \frac{1}{N}\sum_{i=0}^{n}(x_{i,2} - \bar{x_2})(x_{i,1} - \bar{x_1}) & \dots & \frac{1}{N}\sum_{i=0}^{n}(x_{i,1} - \bar{x_1})(x_{i,ncol} - \bar{x_{ncol}}) \\ \vdots & \ddots & \vdots \\ \frac{1}{N}\sum_{i=0}^{n}(x_{i,ncol} - \bar{x_{ncol}})(x_{i,1} - \bar{x_1}) & \dots & \frac{1}{N}\sum_{i=0}^{n}(x_{i,ncol} - \bar{x_{ncol}})(x_{i,ncol} - \bar{x_{ncol}}) \end{pmatrix}$$

o lo que es lo mismo:

$$cov(X) = \begin{pmatrix} cov(X)_{1,1} & cov(X)_{1,2} & \dots & cov(X)_{1,ncol} \\ cov(X)_{2,1} & cov(X)_{2,2} & \dots & cov(X)_{2,ncol} \\ \vdots & \vdots & \ddots & \vdots \\ cov(X)_{ncol,1} & cov(X)_{ncol,2} & \dots & cov(X)_{ncol,ncol} \end{pmatrix}$$

Con esta matriz, la covarianza del vector j con respecto al vector k es simplemente el elemento $cov(X)_{j,k}$.

2. Considerar la matriz hat definida en regresión, $H = X(X^TX)^{-1}X^T$, donde X es una matriz $Nx(d+1)$ y X^TX es invertible.

a)Mostrar que H es simétrica.

Paso 1

Lo primero que hay que hacer es demostrar que X^TX es simétrica, lo cual quiere decir demostrar que $X^TX = (X^TX)^T$.

Aplicando la propiedad del producto traspuesto de matrices lo tenemos:

$$(X^TX)^T = (X)^T(X^T)^T = X^TX$$

Paso 2

Como la matriz (X^TX) aparece invertida en la expresión de H, es decir, $(X^TX)^{-1}$, hay que demostrar que $(X^TX)^{-1}$ es también simétrica.

Por comodidad, renombramos $S = X^TX$. Lo que queremos demostrar ahora es que S^{-1} es simétrica, es decir, $S^{-1} = S^{-T}$.

Como hemos dicho que $X^TX = S$ es invertible, entonces $SS^{-1} = I$. Dado que $I = I^T$, tenemos que $SS^{-1} = (SS^{-1})^T$.

Aplicamos de nuevo la propiedad del producto traspuesto de matrices y nos queda que:

$$SS^{-1} = (SS^{-1})^T \Rightarrow SS^{-1} = S^{-T}S^T$$

Como S es simétrica, tenemos que $S = S^T$, entonces tenemos:

$$SS^{-1} = S^{-T}S$$

Como hemos dicho antes, $SS^{-1} = I$, entonces lo que tenemos es:

$$I = S^{-T}S$$

Dado que $S^{-T}S = I$, sabemos que S^{-T} tiene que ser la inversa de S. Sabemos que la matriz inversa es única (como la madre), tenemos entonces que:

$$S^{-T} = S^{-1} \Rightarrow (X^TX)^{-T} = (X^TX)^{-1}$$

O lo que es lo mismo, que la matriz inversa de una simétrica también es simétrica.

Paso 3 (último)

Sabiendo todo lo anterior, vamos a ver qué ocurre si operamos la traspuesta de H:

$$H^T = (X(X^TX)^{-1}X^T)^T = (X^T)^T(X^TX)^{-T}X^T = X(X^TX)^{-1}X^T = H$$

Con esto, vemos que H es simétrica, es decir, $H = H^T$.

b)Mostrar que $H^K = H$ para cualquier entero K positivo

Para demostrarlo, comencemos por ver qué pasa con H^2:

$$H^2 = HH = (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T) = X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T =$$
$$= XI(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$$

Ahora que sabemos que $H^2 = H$, podemos ver cómo, por ejemplo, $H^3 = H^2H = HH = H^2 = H$. Veamos entonces los casos posibles:

- Cualquier número entero positivo K **par** podría ser expresado como $2 + 2 + ... + 2 = 2n \ \forall n \in \mathbb{N} \geq 1$, lo que implicaría que $H^{2n} = H^2 * ... * H^2 = H * ... * H = H$.
- De forma similar, cualquier entero positivo K **impar**, podría ser expresado como $2 + 2 + ... + 2 + 1 = 2n - 1 \ \forall n \in \mathbb{N} \geq 1$, lo que implicaría que $H^{2n+1} = H^2 * ... * H^2 * H = H * ... * H * H = H$.

De este modo vemos que sea K cualquier entero positivo, se cumple que $H^K = H$.

3. Resolver el siguiente problema: Encontrar el punto (x_0, y_0) sobre la línea $ax + by + c = 0$ que esté más cerca del punto (x_1, y_1).

La respuesta es:

- Si $a \neq 0$

$$x_0 = \frac{b^2 x_1 - aby_1 - ad}{a^2 + b^2}$$
$$y_0 = \frac{b(b^2 x_1 - aby_1 - ad)}{a(a^2 + b^2)} - \frac{b}{a} x_1 + y_1$$

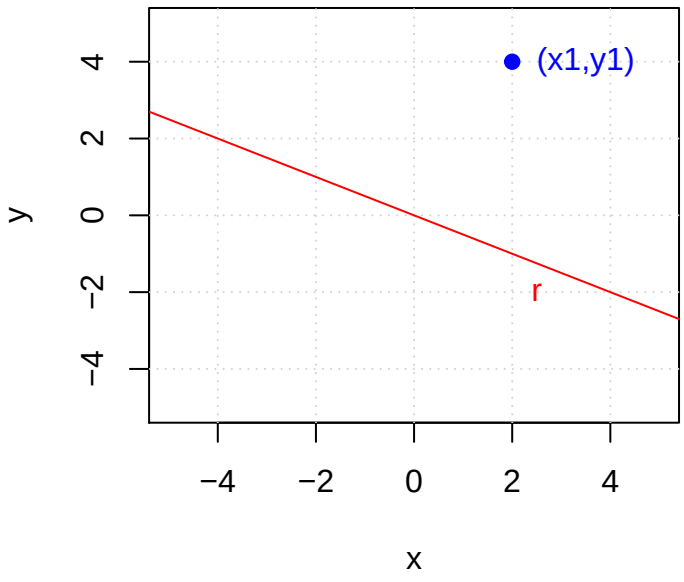
- Si $a = 0$:

$$x_0 = x_1$$
$$y_0 = \frac{-d}{b}$$

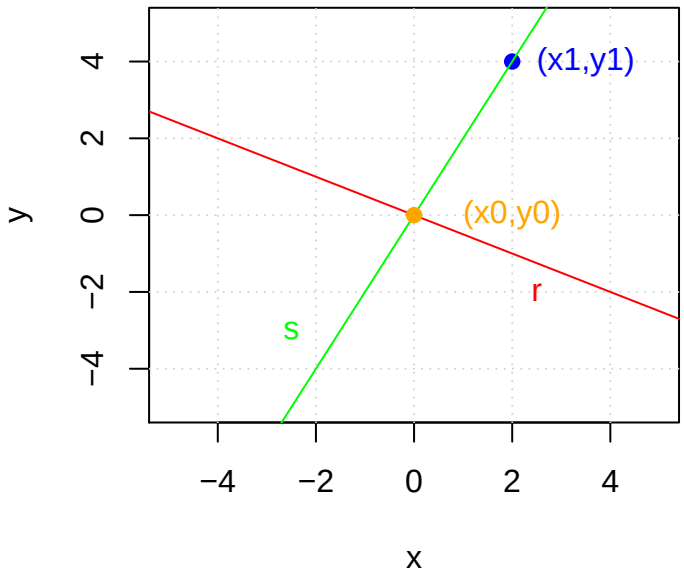
Veamos por qué.

El razonamiento lógico es que el camino más corto entre dos puntos es una línea recta. Lo que tenemos entre manos es un punto (x_1, y_1) y una recta r. Queremos hallar el punto (x_0, y_0) perteneciente a la recta r que minimice la distancia. Automáticamente, sabemos lo que necesitamos. Lo primero es una recta s que pase por el punto (x_1, y_1) y lo segundo es hacer que s sea perpendicular a r.

En un principio, nosotros tenemos lo siguiente:



Una vez conseguimos la recta perpendicular a r y que pase por el punto (x_1, y_1), es decir, la recta s, lo que tenemos es lo siguiente:



Donde (x_0, y_0) es el punto que buscamos.

Situados en el problema, pasemos a las matemáticas. Partimos de la recta $r \equiv ax + by + d = 0$, cuya pendiente es $m_r = -\frac{a}{b}$. Calcular la pendiente de s es tan fácil como $m_s = -\frac{1}{m_r} = \frac{b}{a}$. Una vez conocida la pendiente de s, calculamos su ecuación:

$$s \equiv y - y_1 = m_s(x - x_1) \Rightarrow s \equiv \frac{b}{a}(x - x_1) - y + y_1 = 0$$

Conocidas r y s, tenemos que hallar el punto (x_0, y_0), que no es más que el punto donde se cortan ambas rectas, para ello, resolvemos el sistema siguiente:

$$\left.\begin{array}{c} r \equiv ax_0 + by_0 + d = 0 \\ s \equiv \frac{b}{a}(x_0 - x_1) - y_0 + y_1 = 0 \end{array}\right\}$$

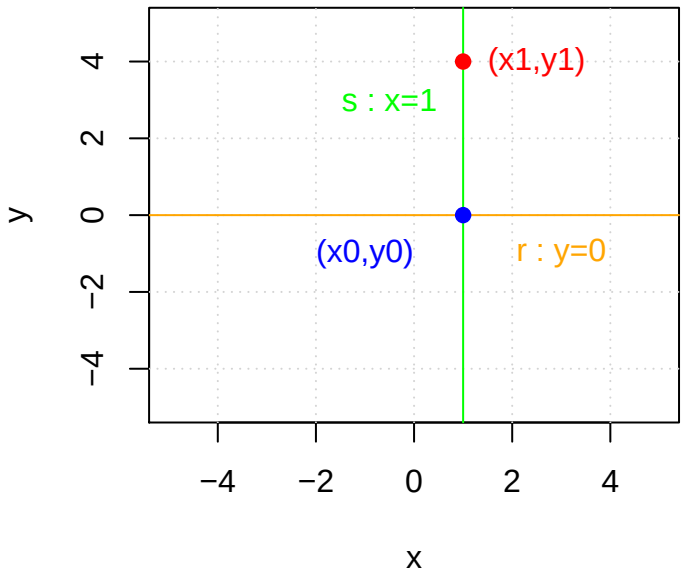
Resolviendo el sistema, (despejo primero y_0 de la ecuación de s) nos queda que:

$$x_0 = \frac{b^2x_1 - aby_1 - ad}{a^2 + b^2}$$

$$y_0 = \frac{b(b^2x_1 - aby_1 - ad)}{a(a^2 + b^2)} - \frac{b}{a}x_1 + y_1$$

Pero cuidado, hay un caso particular. ¿Qué pasa si $a = 0$? Cuando $a = 0$, la recta r dada es totalmente horizontal, por lo que para el mismo valor de x_0, serían válidas infinitas y_0. Eso no es lo que buscamos, por eso aislamos el caso.

Veamos un ejemplo de lo que quiero expresar:



¿Cuál sería el valor de y para $x = 1$ en s, por ejemplo?. Para $x = 1$, tendriamos que $x_0 = x_1 = 1$ pero $y_0 = -\infty, ..., y_1, ..., \infty$, lo cual no debería ocurrir en funciones de este tipo.

Para dar solución a este caso particular, vemos cómo la coordenada x_0 que buscamos coincide con la coordenada x_1 dada. Para encontrar y_0 recurrimos a la ecuación de la recta dada que tendrá la forma $r \equiv by + d = 0$, es sencillo ver que la propia ecuación restringe un único valor de y. Si forzamos que y_0 debe estar en la recta, obtenemos directamente su valor: $by_0 + d = 0 \Rightarrow y_0 = \frac{-d}{b}$.

En resumen, en el caso particular donde $a = 0$, la solución es:

$$x_0 = x_1$$
$$y_0 = \frac{-d}{b}$$

4. Consideremos el problema de optimización lineal con restricciones definido por

$$Min_z c^T z$$
$$Sujeto\ a\ Az \leq b$$

donde c y b son vectores y A es una matriz.

a) Para un conjunto de datos linealmente separable mostrar que para algún w se debe verificar la condición $y_n w^T x_n > 0$ para todo (x__n,y__n) del conjunto.

b) Formular un problema de programación lineal que resuelva el problema de la búsqueda del hiperplano separador. Es decir, idenfique quienes son A, z, b y c para este caso.

5. Probar que en el caso general de funciones con ruido se verifica que $\mathbb{E}_D[E_{out}] = \sigma^2 + bias + var$ (ver transparencias de clase)

En la transparencia 7 de la sesión 5 llegamos a la conclusión de que:

$$\mathbb{E}_D[E_{out}(g^{(D)})] = \mathbb{E}_x[\underbrace{\mathbb{E}_D[(g^{(D)}(x) - \bar{g}(x))^2]}_{var(x)} + \underbrace{(\bar{g}(x) - f_{ruido}(x))^2}_{bias(x)}]$$

Que podría reescribirse como:

$$\mathbb{E}_D[E_{out}(g^{(D)})] = \underbrace{\mathbb{E}_x[\mathbb{E}_D[(g^{(D)}(x) - \bar{g}(x))^2]]}_{var(x)} + \underbrace{\mathbb{E}_x[(\bar{g}(x) - f_{ruido}(x))^2]}_{bias(x)}$$

Si entendemos que $f_{ruido}(x)$ contiene ruido, entonces $f_{ruido}(x) = f(x) + \epsilon$, donde ϵ = ruido. Vamos a operar en el término bias(x), sustituyendo $f_{ruido}(x)$ por $f(x) + \epsilon$.

$$\begin{aligned}(\bar{g}(x) - f_{ruido}(x))^2 &= \bar{g}(x)^2 + f_{ruido}(x)^2 - 2\bar{g}(x)f_{ruido}(x) = \\ &= \bar{g}(x)^2 + (f(x) + \epsilon)^2 - 2\bar{g}(x)(f(x) + \epsilon) = \\ &= \bar{g}(x)^2 + f(x)^2 + \epsilon^2 + 2f(x)\epsilon - 2\bar{g}(x)f(x) - 2\bar{g}(x)\epsilon\end{aligned}$$

Vamos a reagrupar esto en la expresión que tenía en cuenta la esperanza matemática:

$$\mathbb{E}_D[E_{out}(g^{(D)})] = \underbrace{\mathbb{E}_x[\mathbb{E}_D[(g^{(D)}(x) - \bar{g}(x))^2]]}_{var(x)} + \underbrace{\mathbb{E}_x[\bar{g}(x)^2 + f(x)^2 + \epsilon^2 + 2f(x)\epsilon - 2\bar{g}(x)f(x) - 2\bar{g}(x)\epsilon]}_{bias(x)}$$

Agrupemos los términos que tienen ruido (ϵ) y los que no cada uno en su propia esperanza:

$$\mathbb{E}_D[E_{out}(g^{(D)})] = \underbrace{\mathbb{E}_x[\mathbb{E}_D[(g^{(D)}(x) - \bar{g}(x))^2]]}_{var(x)} + \underbrace{\mathbb{E}_x[\bar{g}(x)^2 + f(x)^2 - 2\bar{g}(x)f(x)]}_{bias(x)} + \underbrace{\mathbb{E}_x[\epsilon^2 + 2f(x)\epsilon - 2\bar{g}(x)\epsilon]}_{\sigma^2(x)}$$

Si nos fijamos en el término $bias(x)$, podemos agruparlo en una diferencia de cuadrados, volviendo a tener lo que en la segunda expresión en dicho término, esta vez aislado el ruido:

$$\mathbb{E}_D[E_{out}(g^{(D)})] = \underbrace{\mathbb{E}_x[\mathbb{E}_D[(g^{(D)}(x) - \bar{g}(x))^2]]}_{var(x)} + \underbrace{\mathbb{E}_x[(\bar{g}(x) - f(x))^2]}_{bias(x)} + \underbrace{\mathbb{E}_x[\epsilon^2 + 2f(x)\epsilon - 2\bar{g}(x)\epsilon]}_{\sigma^2(x)}$$

Reagrupando todo:

$$\mathbb{E}_D[E_{out}(g^{(D)})] = \mathbb{E}_x[\underbrace{\mathbb{E}_D[(g^{(D)}(x) - \bar{g}(x))^2]}_{var(x)} + \underbrace{(\bar{g}(x) - f(x))^2}_{bias(x)} + \underbrace{\epsilon^2 + 2f(x)\epsilon - 2\bar{g}(x)\epsilon]}_{\sigma^2(x)}$$

Si comparamos la primera expresión del ejercicio con esta última, vemos claramente cómo si una función f tiene ruido (f_{ruido}), entonces $\mathbb{E}_D[E_{out}(g^{(D)})] = var + bias + \sigma^2$

6. Consideremos las mismas condiciones generales del enunciado del Ejercicio.2 del apartado de Regresión de la relación de ejercicios.2. Considerar ahora $\sigma = 0,1$ y $d = 8$, ¿cual es el más pequeño tamaño muestral que resultará en un valor esperado de E_{in} mayor de 0,008?.

La respuesta es 46.

Veamos por qué.

El ejercicio al que nos referimos nos proporciona la siguiente expresión para estimar el error dentro de la muestra esperado.

$$\mathbb{E}_D[E_{in}(w_{lin})] = \sigma^2 \left(1 - \frac{d+1}{N}\right)$$

Sustituyendo los datos que nos dan en esta expresión tenemos:

$$0.008 = (0.1)^2 \left(1 - \frac{8+1}{N}\right)$$

Despejando N, tenemos que:

$$N = \frac{-9}{\frac{0.008}{(0.1)^2} - 1} = 45$$

Con una muestra de tamaño 45, esperamos obtener un error dentro de la muestra de 0.008. Ahora bien, si el tamaño de la muestra fuese incrementado en 1, es decir, $45 + 1 = 46$, el error esperado dentro de la muestra es:

$$\mathbb{E}_D[E_{in}(w_{lin})] = (0.1)^2 \left(1 - \frac{9}{46}\right) = 0.0080435$$

Lo cual es mayor a 0.008. Lógicamente, cuantas más muestras tengamos más dificil será acertar en cada una.

7. En regresión logística mostrar que:

$$\nabla E_{in}(w) = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n x_n}{1+e^{y_n w^T x_n}} = \frac{1}{N}\sum_{n=1}^{N} -y_n x_n \sigma(-y_n w^T x_n)$$

Argumentar que un ejemplo mal clasificado contribuye al gradiente más que un ejemplo bien clasificado.

Para mostrar lo que pide el enunciado, lo primero que debemos hacer es definir la función $\sigma(x)$ o función sigmoide. Dicha función es la siguiente:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

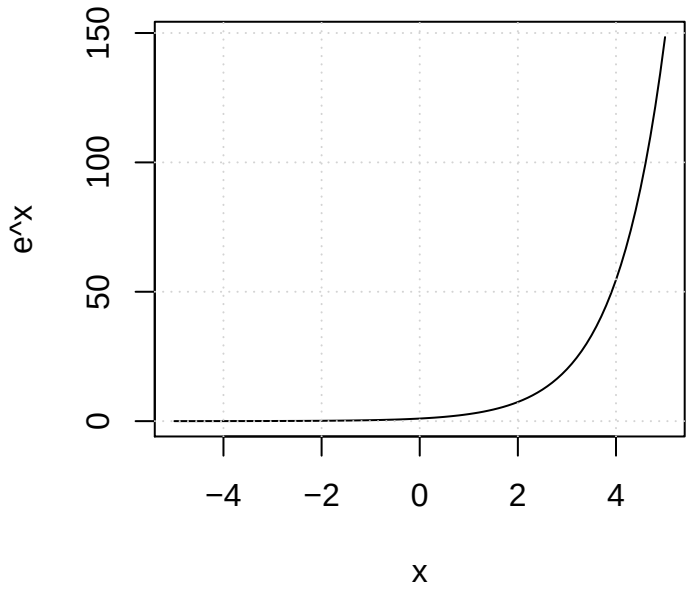
Sabiendo esto, vamos a ir de la expresión final a la inicial:

$$\nabla E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N} -y_n x_n \sigma(-y_n w^T x_n) = \frac{1}{N}\sum_{n=1}^{N} -y_n x_n \frac{1}{1+e^{-(-y_n w^T x_n)}} =$$

$$= \frac{1}{N}\sum_{n=1}^{N}(-1) y_n x_n \frac{1}{1+e^{y_n w^T x_n}} = (-1)\frac{1}{N}\sum_{n=1}^{N} y_n x_n \frac{1}{1+e^{y_n w^T x_n}} =$$

$$= -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n x_n}{1+e^{y_n w^T x_n}}$$

Para argumentar por qué un ejemplo mal clasificado contribuye al gradiente más que un ejemplo mal clasificado, vamos a echar un vistazo a la función $f(x) = e^x$.



- Si un ejemplo está bien clasificado, quiere decir que tanto y_n como $w^T x_n$ tienen el mismo signo, por lo que el producto $y_n w^T x_n > 0$. Si esto ocurre, el término $1 + e^{y_n w^T x_n}$ tenderá a ser muy grande, entonces la expresión $\frac{y_n x_n}{1+e^{y_n w^T x_n}}$ tenderá a 0, por lo que apenas aportará al gradiente.
- Si un ejemplo está, por contra, mal clasificado, quiere decir que y_n tiene distinto signo de $w^T x_n$, por lo que el producto $y_n w^T x_n < 0$. Si esto ocurre, el término $1 + e^{y_n w^T x_n}$ tenderá a ser 1, entonces la expresión $\frac{y_n x_n}{1+e^{y_n w^T x_n}}$ tenderá a $y_n x_n$, por lo que aportará al gradiente mucho más que un ejemplo bien clasificado.

8. Definamos el error en un punto (x_n, y_n) por

$$e_n(x) = max(0, -y_i w^T x_i)$$

Argumentar que el algoritmo PLA puede interpretarse como SGD sobre e_n con tasa de aprendizaje $\eta = 1$.

Lo primero que tenemos que tener en cuenta es que PLA pretende resolver un problema de clasificación, mientras que SGD de minimización. En PLA, hemos cometido los siguientes errores de clasificación para un punto (x_i, y_i):

$$error(x_i, y_i) = \begin{cases} 0 & si \quad y_i w^T x_i \geq 0 \\ \\ y_i x_i & si \quad y_i w^T x_i < 0 \end{cases}$$

Si hemos cometido un error en un punto (x_i, y_i), la actualización de pesos en PLA se hace del siguiente modo:

$$w^{(t+1)} = w^{(t)} + y_i x_i$$

Es decir, hay que compensar el error cometido. Aunque no hace falta, vemos que $w^{(t+1)} = w^{(t)} + 0$ si no cometemos errores.

Vayámonos ahora a SGD. La regla de aprendizaje en SGD nos dice que:

$$w^{(t+1)} = w^{(t)} + \eta - \nabla E_{in}(w^{(t)})$$

Sabemos que $E_{in}(w^{(t)}) = \frac{1}{N} \sum_{i=1}^{N} e_n$. Como en PLA se corrige un sólo punto cada vez, lo que tenemos es $E_{in}(w^{(t)}) = \frac{1}{1} \sum_{i=1}^{1} e_n = e_n$ es decir, $E_{in}(w^{(t)}) = max(0, -y_i w^T x_i)$. Entonces, lo que tenemos es lo siguiente:

$$w^{(t+1)} = w^{(t)} + \eta - \nabla max(0, -y_i w^T x_i)$$

Si fijamos la tasa de aprendizaje $\eta = 1$, lo que tenemos es lo siguiente:

$$w^{(t+1)} = w^{(t)} + (-\nabla max(0, -y_i w^T x_i))$$

Si nos fijamos bien, cuando cometemos un error, $-y_i w^T x_i > 0$, entonces $max(0, -y_i w^T x_i) = -y_i w^T x_i$. El gradiente $\nabla(E_{in}(w))$ es la derivada de $E_{in}(w)$, que nos indica hacia dónde se debe mover el vector $w^{(t)}$ en la siguiente iteración. Esa modificación permitirá corregir el punto que se está considerando en el instante t. Entonces, lo que obtenemos es $\nabla(-y_i w^T x_i) = -y_i x_i$.

$$w^{(t+1)} = w^{(t)} + (-(-y_i x_i)) \Rightarrow w^{(t+1)} = w^{(t)} + y_i x_i$$

Si, por el contrario, no hemos cometido errores, lo que tenemos es que $max(0, -y_i w^T x_i) = 0$, por lo que tenemos que $\nabla 0 = 0$:

$$w^{(t+1)} = w^{(t)} + 0$$

Ambos casos, son exactamente lo mismo que obtendríamos con el algoritmo PLA, por lo que podríamos concluir que PLA no es más que un caso particular de SGD.

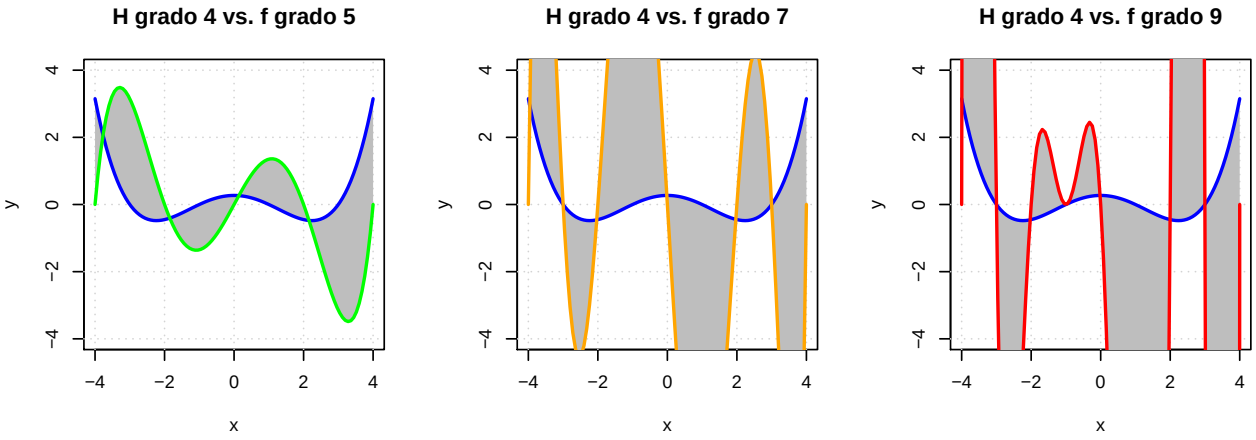
9. El ruido determinista depende de H, ya que algunos modelos aproximan mejor f que otros.

a) Suponer que H es fija y que incrementamos la complejidad de f.

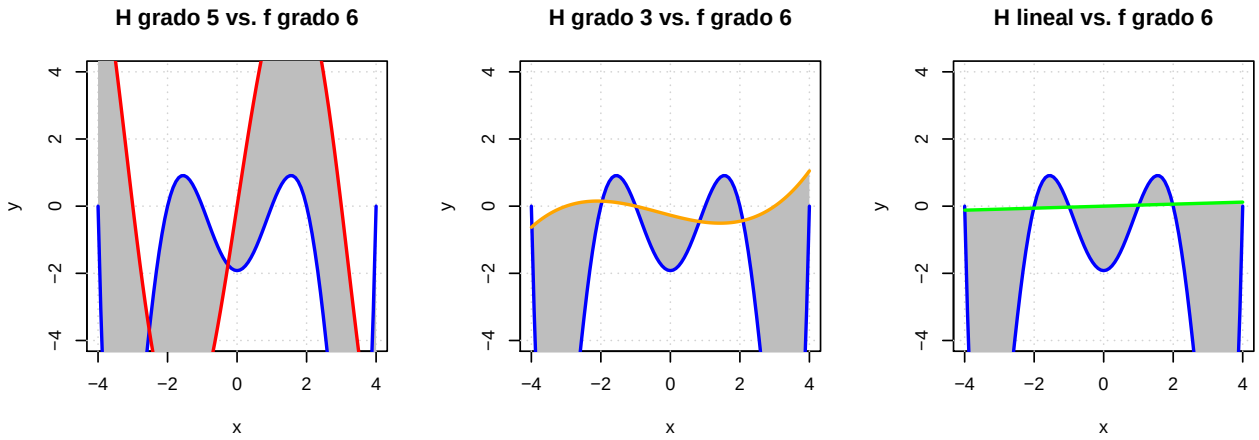
b) Suponer que f es fija y decrementamos la complejidad de H

Contestar para ambos escenarios: ¿En general subirá o bajará el ruido determinista? ¿La tendencia a sobrejaustar será mayor o menor? (Ayuda: analizar los detalles que influencian el sobreajuste)

En el **apartado a**, teóricamente, si mantenemos fija la hipótesis H y aumentamos la complejidad de f, la **varianza entre f y el mejor estimador de H debería aumentar**, por lo que **el ruido determinístico debería ser cada vez mayor**. Al aumentar la varianza, **el sesgo disminuiría**, por lo que el estimador se mantendría más cerca de los valores esperados de f, pero dado que la complejidad del estimador de H cada vez estaría a más distancia por debajo de la complejidad de f, **el estimador de H no puede sobreajustar, o sobreajustaría cada vez menos**, pues no llega a la complejidad de f. Veamos algunos ejemplos de cómo aumenta el ruido determinístico (área gris) de un estimador de H fijo mientras f aumenta su complejidad.



En el **apartado b**, si suponemos que f es fija y decrementamos la complejidad de H, **la varianza debería ser cada vez menor** y por ende **el sesgo cada vez mayor**. A medida que el estimador H reduce su complejidad, el ruido determinísico que se produce tiende a ser sólo a causa de f (el máximo valor del estimador de H está por debajo del máximo valor de f y el mínimo valor del estimador de H está por encima del mínimo valor de f), luego, teóricamente **el ruido determinístico debería ir bajando**. Dado que f es fija y se decrementa la complejidad de H, **la tendencia a sobreajustar sería cada vez menor**, ya que la complejidad de H se iría alejando cada vez más (no llegaría) de la complejidad de f. Veamos algunos ejemplos de cómo el ruido determinístico va disminuyendo y cómo el sesgo (sobre todo en los casos de H naranja y verde), se dirige a la parte superior de f.


H grado 5 vs. f grado 6
H grado 3 vs. f grado 6
H lineal vs. f grado 6
y
x
−4
−2
0
2
4

10. La técnica de regularización de Tikhonov es bastante general al usar la condición

$$w^T\Gamma^T\Gamma w \leq C$$

que define relaciones entre las w_i (La matriz Γ_i se demonina regularizador de Tikhonov)

a) Calcular Γ cuando $\sum_{q=0}^{Q} w_q^2 \leq C$

b) Calcular Γ cuando $(\sum_{q=0}^{Q} w_q)^2 \leq C$

Argumentar si el estudio de los regularizadores de Tikhonov puede hacerse a través de las propiedades algebráicas de las matrices Γ

Apartado a)

La respuesta es $\Gamma = I$, la matriz identidad. Veamos por qué

En el primer apartado, se nos restringe a que $\sum_{q=0}^{Q} w_q^2 \leq C$, entonces, nada nos impide reescribir la expresión así:

$$w^T\Gamma^T\Gamma w \leq \sum_{q=0}^{Q} w_q^2$$

Vamos a desarrollar la parte izquierda y derecha para entender mejor qué se está pidiendo:

$$\underbrace{\begin{pmatrix} w_1 & w_2 & \dots & w_Q \end{pmatrix} \begin{pmatrix} \Gamma_{11} & \Gamma_{21} & \dots & \Gamma_{m1} \\ \Gamma_{12} & \Gamma_{22} & \dots & \Gamma_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{1Q} & \Gamma_{2Q} & \dots & \Gamma_{mQ} \end{pmatrix}}_{\begin{pmatrix} w_1 & w_2 & \dots & w_Q \end{pmatrix}} \underbrace{\begin{pmatrix} \Gamma_{11} & \Gamma_{12} & \dots & \Gamma_{1m} \\ \Gamma_{21} & \Gamma_{22} & \dots & \Gamma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{Q1} & \Gamma_{Q2} & \dots & \Gamma_{Qm} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_Q \end{pmatrix}}_{\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_Q \end{pmatrix}} \leq w_1^2 + w_2^2 + ... + w_Q^2$$

Es decir, vemos cómo para que se cumpla, $w^T\Gamma^T = w^T$ y $\Gamma w = w$. La única forma de conseguirlo es que $w^T\Gamma^T\Gamma w = w^T I w$, es decir, que $\Gamma^T = \Gamma = I$.

Apartado b)

La respuesta es $\Gamma = \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix}$, es decir, una fila de Q unos. Veamos por qué:

En este caso, a diferencia del anterior, estamos buscando el cuadrado de una suma, es decir, lo que buscamos es que $w^T\Gamma^T = w_1 + w_2 + ... + w_Q$ y que $\Gamma w = w_1 + w_2 + ... + w_Q$. Si lo vemos como multiplicación de matrices, lo que buscamos es esto:

$$\underbrace{\begin{pmatrix} w_1 & w_2 & \dots & w_Q \end{pmatrix} \begin{pmatrix} \Gamma_{11} & \Gamma_{21} & \dots & \Gamma_{m1} \\ \Gamma_{12} & \Gamma_{22} & \dots & \Gamma_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{1Q} & \Gamma_{2Q} & \dots & \Gamma_{mQ} \end{pmatrix}}_{w_1+w_2+...+w_Q} \underbrace{\begin{pmatrix} \Gamma_{11} & \Gamma_{12} & \dots & \Gamma_{1m} \\ \Gamma_{21} & \Gamma_{22} & \dots & \Gamma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{Q1} & \Gamma_{Q2} & \dots & \Gamma_{Qm} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_Q \end{pmatrix}}_{w_1+w_2+...+w_Q} \leq (w_1 + w_2 + ... + w_Q)^2$$

La única forma de conseguirlo es que $\Gamma = \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix}$, ya que:

$$\begin{pmatrix} w_1 & w_2 & \dots & w_Q \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_Q \end{pmatrix} = (w_1+w_2+...+w_Q)(w_1+w_2+...+w_Q) \leq (w_1+w_2+...+w_Q)^2$$

BONUS

B1. Considerar la matriz $H = X(X^TX)^{-1}X^T$. Sea X una matriz Nx$(d+1)$, y X^TX invertible. Mostrar que $traza(H) = d+1$, donde traza significa la suma de los elementos de la diagonal principal.

En [1] se nos proporciona información adicional para realizar esta demostración y consiste en utilizar la propiedad $traza(AB) = traza(BA)$.

Dicha propiedad no siempre es cierta, pues para cumplirse, debe ocurrir que $dimension(A) = dimension(B^T)$. En otras palabras, si la propiedad se cumple y $dimension(A) = m \times n$ entonces forzosamente $dimension(B) = n \times m$,

La demostración es sencilla y se puede encontrar en [2].

Paso 1

Sabiendo ya cierta la propiedad anteriormente citada, tras revisar la demostración, vamos a ver si estamos en las condiciones de dimensionalidad para aplicarla. Tomemos $A = X$ y $B = (X^TX)^{-1}X^T$. Entonces:

$$A = \underbrace{X}_{N\times(d+1)}$$

$$B = \underbrace{\underbrace{(X^TX)^{-1}}_{(d+1)\times(d+1)} \underbrace{X^T}_{(d+1)\times N}}_{(d+1)\times N}$$

Como podemos comprobar, $dimension(A) = dimension(B^T)$ y estamos en las condiciones de aplicar la propiedad $traza(AB) = traza(BA)$.

Paso 2

Sustituímos los valores de A y B por sus valores originales para aplicar la propiedad anteriormente citada.

$$traza(AB) = traza(BA) \Rightarrow traza(X(X^TX)^{-1}X^T) = traza((X^TX)^{-1}X^TX)$$

Si nos fijamos, en la parte derecha tenemos una matriz por su inversa, lo que se convierte en la identidad, entonces tenemos:

$$traza(AB) = traza(BA) \Rightarrow traza(X(X^TX)^{-1}X^T) = traza(I_{d+1})$$

Como sabemos, la traza de la matriz identidad es igual al número de filas (o al de columnas, ya que es cuadrada). Dado que $dimension(I_{d+1}) = (d+1) \times (d+1)$, tenemos que $traza(I_{d+1}) = d+1$. Entonces:

$$traza(X(X^TX)^{-1}X^T) = d+1 \Rightarrow traza(H) = d+1$$

.

Referencias

[1] Learning from data. Mostafa, Ismail, Lin. 2012. Página 87.

[2] Linear Algebra Theory and Applications. Kenneth Kuttler. 2012. Página 180.
http://www.saylor.org/site/wp-content/uploads/2012/02/Linear-Algebra-Kuttler-1-30-11-OTC.pdf