Aprendizaje Automático (2016-2017)
Grado en Ingeniería Informática
Universidad de Granada

Cuestionario de teoría 2

David Criado Ramón

1 Sea X una matriz de números reales de dimensiones N x d, N >d. Sea $X = UDV^T$ su descomposición en valores singulares (SVD). Calcular la SVD de X^TX y XX^T en función de la SVD de X. ¿Qué propiedades tienen estas nuevas matrices que no están presente en X? ¿Qué representa la suma de diagonal principal de cada una de las matrices producto?

Empecemos por X^TX

$$X^TX = (UDV^T)^TUDV^T \tag{1.1}$$

Aplicando la transpuesta del producto de matrices.

$$X^TX = VD^TU^TUDV^T \tag{1.2}$$

Puesto que U es una matriz ortogonal, entonces $U^TU = I$, por tanto:

$$X^TX = VD^TDV^T \tag{1.3}$$

Como D es una matriz diagonal coincide con su traspuesta por lo que $D^TD = D^2$

$$X^TX = VD^2V^T \tag{1.4}$$

Vamos ahora a por XX^T

$$XX^T = UDV^T(UDV^T)^T \tag{1.5}$$

Aplicamos la traspuesta del producto de matrices

$$XX^T = UDV^TVD^TU^T \tag{1.6}$$

Como V es una matriz ortogonal, $V^TV = I$.

$$XX^T = UDD^TU^T \tag{1.7}$$

Como D es una matriz diagonal coincide con sus traspuesta por lo que

$$XX^T = UD^2U^T \tag{1.8}$$

Propiedades:

- U es una matriz ortogonal de dimensión N x N.
- V es otra matriz ortogonal de dimensión d x d. En este caso, como en el anterior, por ser ortogonales el producto de la matriz con su traspuesta (o viceversa) resulta en la matriz identidad de mismas dimensiones.
- D es una matriz diagonal de dimensión N x d. Todos sus elementos son 0 menos los de la diagonal principal. D es igual a su traspuesta.

La diagonal principal de D representa a los valores singulares de X en orden decreciente.

2 Suponga una matriz cuadrada A que admita la descomposición $A = X^T X$ para alguna matriz X de números reales. Establezca una relación entre los valores singulares de las matriz A y los valores singulares de X.

Partiendo del ejercicio anterior $A = VD^2V^T$. Vamos a poner subíndices para indicar si la descomposición es respecto a matriz A o a la X.
Partiendo de lo anterior y aplicando SVD en el miembro izquierdo.

$$U_A D_A V_A^T = V_X D_X^2 V_X^T \tag{2.1}$$

Por tanto $U_A = V_X$, $D_A = D_X^2$, $V_A^T = V_X^T$ y los valores singulares de A (que se encuentran en la diagonal principal de D) son el cuadrado de los valores singulares de X.

3 Definir el problema con restricciones para encontrar los valores extremos de $f(x, y) = ax + by$ sujeto a la restricción $x^2 + y^2 = r^2$. Definir la Langraniana y calcular los valores de x, y y $f(x, y)$ en el óptimo. Discutir las distintas soluciones que se pueden presentar en función de los valores de a, b y r

Dadas estas restricciones, definimos la Langraniana del problema a estudiar como:

$$\mathcal{L}(x, y, \lambda) = ax + by - \lambda(x^2 + y^2 - r^2) \tag{3.1}$$

Hacemos las derivadas parciales.

$$\frac{\delta \mathcal{L}(x, y, \lambda)}{\delta x} = a - 2\lambda x \tag{3.2}$$

$$\frac{\delta \mathcal{L}(x, y, \lambda)}{\delta y} = b - 2\lambda y \tag{3.3}$$

$$\frac{\delta \mathcal{L}(x, y, \lambda)}{\delta \lambda} = -x^2 - y^2 + r^2 \tag{3.4}$$

Igualando a 0.

$$a - 2\lambda x = 0 \tag{3.5}$$

$$b - 2\lambda y = 0 \tag{3.6}$$

$$-x^2 - y^2 + r^2 = 0 \tag{3.7}$$

Despejamos x e y de las dos primeras ecuaciones (3.5 y 3.6).

$$x = \frac{a}{2\lambda} \tag{3.8}$$

$$y = \frac{b}{2\lambda} \tag{3.9}$$

Ahora sustituyendo x e y en la última ecuación (3.7).

$$-\left(\frac{a}{2\lambda}\right)^2 - \left(\frac{b}{2\lambda}\right)^2 + r^2 = 0 \tag{3.10}$$

Despejamos lambda ahora del resultado.

$$\frac{-a^2 - b^2}{4\lambda^2} = -r^2 \tag{3.11}$$

$$r^2 = \frac{a^2 + b^2}{4\lambda^2} \tag{3.12}$$

$$\lambda^2 = \frac{a^2 + b^2}{4r^2} \tag{3.13}$$

$$\lambda = \sqrt{\frac{a^2 + b^2}{4r^2}} \tag{3.14}$$

Con lambda despejado sustituimos en lambda en las ecuaciones de x e y (3.8 y 3.9) para obtener la x e y del extremo.

$$x = \frac{a}{2\sqrt{\frac{a^2+b^2}{4r^2}}} \tag{3.15}$$

$$y = \frac{b}{2\sqrt{\frac{a^2+b^2}{4r^2}}} \tag{3.16}$$

Para terminar el valor de la función en dicho punto es de

$$f(x,y) = ax + by = \frac{a^2}{2\sqrt{\frac{a^2+b^2}{4r^2}}} + \frac{b^2}{2\sqrt{\frac{a^2+b^2}{4r^2}}} = \frac{a^2+b^2}{2\sqrt{\frac{a^2+b^2}{4r^2}}} = \frac{2(a^2+b^2)|r|}{2\sqrt{a^2+b^2}} = \frac{(a^2+b^2)|r|}{\sqrt{a^2+b^2}} \tag{3.17}$$

Por tanto, si $r = 0$ ó a y b son ambas 0 a la vez la función no tiene extremos que cumplan las restricciones

4 En regresión lineal con ruido en las etiquetas, el error fuera de la muestra viene dado por $E_{out}(h) = \mathbb{E}_{x,y}[(h(x) - y)^2] = \int\int (h(x) - y)^2\, p(x,y)\, dxdy$. Mostrar que de todas las posibles hipótesis la que minimiza E_{out} está dada por $h^*(x) = \mathbb{E}_y[y|x] = \int y \cdot p(y|x)\, dy$

5 Escribir la función de Máxima Verosimilitud de una muestra de un tamaño N para un problema de clasificación binaria. Además

Siendo

$$P(y|x) = \begin{cases} h(x) & si\ y = +1 \\ 1 - h(x) & si\ y = -1 \end{cases} \tag{5.1}$$

La función de máxima verosimilitud viene dada por

$$argmax(\mathcal{L}) = argmax\left(\prod_{n=1}^{N} P(y_n|x_n)\right) \tag{5.2}$$

5.a) Mostrar que la estimación de Máxima Verosimilitud se reduce a la tarea de encontrar la función h que minimiza $E_{in}(w) = \sum_{n=1}^{N}[[y = +1]]\ ln\frac{1}{h(x_n)}\ [[y = -1]]\ ln\frac{1}{1-h(x_n)}$

Para empezar resulta evidente que si intentamos maximizar una función podemos minimizar uno partido dicha función puesto que cuanto más grande es el denominador más decrece 1 partido la función, por tanto:

$$argmax(\mathcal{L}) = argmin\left(\frac{1}{N}\mathcal{L}\right) \tag{5.3}$$

$$argmax(\mathcal{L}) = argmin\left(\prod_{n=1}^{N} \frac{1}{P(y_n|x_n)}\right) \tag{5.4}$$

Además, la función de máxima verosimilitud podemos cambiarla por su versión con logaritmos, cambiaremos la escala pero no el crecimiento de la función . Por tanto, ahora la función de máxima verosimilitud logarítmica viene dada por:

$$argmin(\mathcal{L}) = argmin(ln\ \mathcal{L}) \tag{5.5}$$

$$argmax(\mathcal{L}) = argmin\left(\sum_{n=1}^{N} ln\ \frac{1}{P(y_n|x_n)}\right) \tag{5.6}$$

De igual manera si dividiésemos la suma por N volvemos a hacer la escala más pequeña pero no cambiamos los argumentos que minimizan el resultado de la función.

$$argmin(\mathcal{L}) = argmin(\frac{1}{N}\mathcal{L}) \tag{5.7}$$

$$argmax(\mathcal{L}) = argmin\left(\frac{1}{N}\sum_{n=1}^{N} ln\ \frac{1}{P(y_n|x_n)}\right) \tag{5.8}$$

Expandiendo $P(y|x)$ en la última ecuación obtenemos el resultado que queríamos demostrar:

$$argmax(\mathcal{L}) = argmin\left(\frac{1}{N}\sum_{n=1}^{N}[[y_n = +1]]\ ln\frac{1}{h(x_n)} + [[y_n = -1]]\ ln\frac{1}{1-h(x_n)}\right) \tag{5.9}$$

5.b) Para el caso $h = \sigma(w^T x)$ mostrar que minimizar el error de la muestra en el apartado anterior es equivalente a minimizar el error muestral $E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N} ln(1 + e^{-y_n w^T x_n})$

Podemos observar que en este caso no se diferencia entre la hipótesis de que y sea +1 ó -1, por tanto, vamos a demostrar primero que los siguiente ocurre

$$1 - h(z_n) = h(-z_n) \tag{5.10}$$

Sustituimos $h(z_n)$ con la función proporcionada

$$1 - \sigma(w^T x_n) = \sigma(-w^T x_n) \tag{5.11}$$

Aplicamos que $\sigma(z) = \frac{e^z}{1+e^z}$ en la ecuación

$$1 - \frac{e^{w^T x_n}}{1 + e^{w^T x_n}} = \frac{e^{-w^T x_n}}{1 + e^{-w^T x_n}} \tag{5.12}$$

Expandimos el miembro de la izquierda

$$\frac{1}{1 + e^{w^T x_n}} = \frac{e^{-w^T x_n}}{1 + e^{-w^T x_n}} \tag{5.13}$$

Aplicando que $z^{-b} = \frac{1}{z^b}$

$$\frac{1}{1 + e^{w^T x_n}} = \frac{\frac{1}{e^{w^T x_n}}}{1 + \frac{1}{e^{w^T x_n}}} \tag{5.14}$$

Expandimos el denominador del miembro de la derecha

$$\frac{1}{1 + e^{w^T x_n}} = \frac{\frac{1}{e^{w^T x_n}}}{1 + \frac{1}{e^{w^T x_n}}} \tag{5.15}$$

$$\frac{1}{1 + e^{w^T x_n}} = \frac{\frac{1}{e^{w^T x_n}}}{\frac{1+e^{w^T x_n}}{e^{w^T x_n}}} \tag{5.16}$$

Realizando la división de la derecha queda demostrado que son iguales

$$\frac{1}{1 + e^{w^T x_n}} = \frac{1}{1 + e^{w^T x_n}} \tag{5.17}$$

Puesto que para $y = 1$ tenemos que calcular $h(x_n)$ y para $y = -1$ tenemos que calcular $h(-x_n)$ y lo que acabamos de calcular es $h(-x_n)$ multiplicamos el h original por la etiqueta correspondiente cambiada de signo, así pues:

$$h = \frac{1}{1 + e^{-y w^T x_n}} \tag{5.18}$$

Si sustituimos $h(x_n)$ por nuestra nueva h en la ecuación (5.9) obtenemos que

$$argmin\left(\frac{1}{N}\sum_{n=1}^{N}[[y_n=+1]]ln\frac{1}{h(x_n)}+[[y_n=-1]]ln\frac{1}{1-h(x_n)}\right)=argmin\left(\frac{1}{N}\sum_{n=1}^{N}1+e^{-yw^Tx_n}\right) \tag{5.19}$$

6 Definamos el error en el punto (x_n, y_n) por $e_n(w) = max(0, -y_n w^T x_n)$. Argumentar si con esta función el algoritmo PLA puede interpretarse como SGD sobre e_n con tasa de aprendizaje $\nu = 1$.

El algoritmo PLA va cogiendo datos de la muestra en orden aleatorio y si el dato **está mal clasificado** ajusta con respecto a la regla $w_{nuevo} = w_{anterior} + y_i x_i$.

SGD recorre los datos en orden aleatorio de la muestra y siempre cambia el ajuste aplicando $w_{nuevo} = w_{anterior} - \nu \cdot \nabla e_n(w)$.

Para que sean iguales han de ocurrir 2 cosas.

- Si el dato está bien clasificado, SGD no ha de realizar ningún cambio.
- Si el dato está mal clasificado, SGD y PLA han de tener la misma regla de actualización para el ajuste de pesos.

Empecemos por si el dato está mal clasificado. Como $\nu = 1$

$$\triangle w_{PLA} = \triangle w_{SGD} \tag{6.1}$$

$$w_{anterior} + y_n x_n = w_{anterior} - \nabla e_n(w) \tag{6.2}$$

Por lo que hemos de demostrar que $\nabla e_n(w) = y_n x_n$. Para ello hemos de observar que ocurre si el datos está bien clasificado o mal clasificado qué ocurre con el máximo de la función de error. Si el dato está bien clasificado el signo de la etiqueta original (y_n) coincide con el signo de la predicción hecha por el modelo $(w^T x_n)$ siendo el producto de ellos positivo y por tanto al tener un signo menos delante se quedaría negativo y al aplicar el máximo nos quedaríamos con 0, por lo que **si el dato está bien clasificado** el SGD con esta función de error hace lo mismo que PLA, es decir, **no modifica el ajuste**. En caso de que el dato esté mal clasificado, el signo de la etiqueta original difiere del signo de la etiqueta predicha, por lo que, su producto es negativo y como hay un símbolo menos delante queda positivo, y, por tanto siempre será el máximo al compararse con 0. Sabiendo ya qué es lo que ocurre cuando el dato está mal clasificado, calculemos el gradiente.

$$\nabla e_n(w) = \frac{\delta e_n(w)}{\delta w} = -y_n x_n \tag{6.3}$$

Sustituyendo el gradiente en (6.2) queda demostrado que **tienen la misma regla de actualización cuando los datos están mal clasificados.**

$$w_{anterior} + y_n x_n = w_{anterior} + y_n x_n \tag{6.4}$$

7 Considerar la función de error $E_n(w) = (max(0, 1 - y_n w^T x_n))^2$. Argumentar que el algoritmo ADALINE con la regla de adaptación $w_{new} = w_{old} + \eta(y_n - w^T x_n) \cdot x_n$ es equivalente a gradiente descendente estocástico (SGD) sobre $\frac{1}{N}\sum_{n=1}^{N} E_n(w)$

Veamos qué ocurre en el caso de un **dato bien clasificado** en ambos algoritmos.

- **ADALINE** - $w^T x_n = y_n$ puesto que está bien clasificado, por tanto, $y_n - w^T x_n = 0$, por lo que **no modifica ningún peso** si está bien clasificado.
- **SGD** Como $w^T x_n = y_n$ quedándonos 1 - $y_n \cdot w^T x_n$ que en problemas de clasificación, como el que estamos evaluando va a ser 0 porque $y_n \cdot w^T x_n = 1$. También podemos observar que, si está mal clasificado, el dato del error calculado será 1 o mayor, por tanto, se aplicará la regla de actualización de pesos.

He podido quitar $y_n y_n$ puesto que ya sea $y_n = 1$ ó $y_n = -1$ el producto $y_n y_n$ siempre va a dar 1.

Para ver que se comportan igual si hay un error de clasificación, de forma similar al ejercicio anterior, comparamos las funciones de actualización de pesos para que sean la misma. Primero vamos a calcular el gradiente de $(1 - y_n w^T x_n)^2$.

$$\begin{aligned}\nabla E_n(w) = \frac{\delta E_n(w)}{\delta w} = 2 \cdot (1 - y_n w^T x_n) \cdot (-y_n x_n) = 2 \cdot (-y_n x_n + y_n w^T x_n y_n x_n) = \\ = 2 \cdot (-y_n + y_n y_n w^T x_n) \cdot x_n = -2 \cdot (y_n - w^T x_n) \cdot x_n\end{aligned} \tag{7.1}$$

Comparemos ambas reglas de actualización de pesos.

$$\triangle w_{ADALINE} = \triangle w_{SGD} \tag{7.2}$$

$$w_{anterior} + \eta(y_n - w^T x_n) \cdot x_n = w_{anterior} - \nu \cdot \nabla E_n(w) \tag{7.3}$$

Sustituimos el gradiente que hemos calculado.

$$w_{anterior} + \eta(y_n - w^T x_n) \cdot x_n = w_{anterior} + 2\nu(y_n - w^T x_n) \cdot x_n \tag{7.4}$$

Por tanto, podemos concluir, que en un problema de clasificación ADALINE es igual a SGD si la tasa de aprendizaje $\nu = \frac{\eta}{2}$, donde η es el parámetro del algoritmo ADALINE.

BONUS

1 Sea X una matriz $NxM, N > M$ de números reales. ¿Cómo son los valores singulares de las matrices X, X^TX y XX^T y qué relación existe entre ellos?

Resolvemos este problema de manera similar a los ejercicios 1 y 2 de este cuestionario. En el ejercicio 2 ya se encuentra la relación entre X y X^TX. Procedamos ahora a ver qué ocurre con X y XX^T. Partiendo de lo que veíamos en (1.8)

$$XX^T = UD^2U^T \tag{B.1.1}$$

Resulta fácil ver que las equivalencias de $SVD(XX^T)$ son:

- $U_{XX^T} = U_X$
- $D_{XX^T} = D_X^2$
- $V_{XX^T}^T = U_X^T$

Podemos observar que los valores singulares de X^TX y XX^T son los mismos y son el cuadrado de los valores singulares de X.