

# Trabajo 1 - Teoría

## Aprendizaje Automático

Francisco Solano López Rodríguez

March 26, 2018

1. Identificar, para cada una de las siguientes tareas, que tipo de aprendizaje es el adecuado (supervisado, no supervisado, por refuerzo) así como los datos de aprendizaje que deberíamos usar en su caso. Si una tarea se ajusta a más de un tipo, explicar como y describir los datos para cada tipo.
  - (a) Dada una colección de fotos de caras de personas de distintas razas establecer cuantas razas distintas hay representadas en la colección.  
No supervisado, ya que no disponemos de datos previos y queremos obtener el número de clases (que en este caso son razas) hay representadas en la colección
  - (b) Clasificación automática de cartas por distrito postal.  
Podría ser adecuado por aprendizaje supervisado, ya que necesitamos datos previos para entrenar al programa si no queremos que el programa nos agrupe las cartas estableciendo clases posiblemente no deseadas. Aunque también podría ser interesante si no tenemos ninguna información previa utilizar aprendizaje no supervisado y obtener agrupaciones automáticas.
  - (c) Decidir si un determinado índice del mercado de valores subirá o bajará dentro de un periodo de tiempo fijado.  
Se podría realizar mediante aprendizaje supervisado a partir de datos de por ejemplo meses anteriores y con esos utilizar un método de regresión para clasificar los datos actuales. También podría ser positivo realizar lo mediante refuerzo 'premiando' aquellos pesos que hayan dado buenos resultados.
  - (d) Aprender un algoritmo que permita a un robot rodear un obstáculo.  
La mejor opción sería mediante aprendizaje por refuerzo, guiar a robot en su tarea de rodear un obstáculo reforzando aquellas acciones que sean positivas.
2. ¿Cuales de los siguientes problemas son más adecuados para una aproximación por aprendizaje y cuales más adecuados para una aproximación por diseño? Justificar la decisión.
  - (a) Agrupar los animales vertebrados en mamíferos, reptiles, aves, anfibios y peces.  
Este problema es más adecuado para una aproximación por diseño, de hecho existen programas por diseño que resuelven este problema.
  - (b) Determinar si se debe aplicar una campaña de vacunación contra una enfermedad. Por aprendizaje tomando datos de años anteriores para decidir mediante un modelo de regresión si aplicar la campaña de vacunación.
  - (c) Determinar si un correo electrónico es de propaganda o no.  
Es más adecuado por aprendizaje, ya que un correo utiliza el lenguaje humano una de las cosas más complejas de clasificar, por lo que será mas adecuado entrenar al programa mediante muchos casos correos ya clasificados.
  - (d) Determinar el estado de ánimo de una persona a partir de una foto de su cara.  
Por aprendizaje, el reconocimiento facial en sí ya es un problema difícil que necesita del aprendizaje, no existe una definición rigurosa de lo que es un rostro facial, cada persona tiene unos rasgos únicos totalmente diferentes al de los demás. Si además tenemos que determinar el estado de ánimo a partir de la cara, aún más complejo será el problema.

- (e) Determinar el ciclo óptimo para las luces de los semáforos en un cruce con mucho tráfico.  
Este problema sería factible de determinar mediante aprendizaje, aunque posiblemente sería mejor opción mediante una metaheurística, ya que para el aprendizaje sería difícil obtener el conjunto de datos necesario para el entrenamiento.
3. Construir un problema de aprendizaje desde datos para un problema de clasificación de fruta en una explotación agraria que produce mangos, papayas y guayabas. Identificar los siguientes elementos formales  $\mathcal{X}, \mathcal{Y}, \mathcal{D}, f$  del problema. Dar una descripción de los mismos que pueda ser usada por un computador. ¿Considera que en este problema estamos ante un caso de etiquetas con ruido o sin ruido? Justificar las respuestas.
- $\mathcal{X}$ : espacio d-dimensional donde cada coordenada corresponde a una característica de las frutas a clasificar como podría ser el peso, tamaño, color, forma, etc...
  - $\mathcal{Y}$ : clases a las que puede pertenecer una fruta, podríamos usar por ejemplo 1, 2, 3 donde el 1 podría ser la clase asociada al mango, el 2 a la papaya y el 3 a la guayaba.
  - $\mathcal{D}$ : Conjunto de parejas características-etiqueta  $\mathcal{D} = \{(x_1, f(x_1)), \dots, (x_n, f(x_n))\}$ , donde  $f(x_i) = y_i \in \mathcal{Y}$  y que servirá de entrenamiento.
  - $f$ : es la función objetivo cuyo dominio es  $\mathcal{X}$  y toma valores en  $\mathcal{Y}$ .  $f: \mathcal{X} \rightarrow \mathcal{Y}$

Considero que pueden tener ruido ya que puede haber equivocaciones en la medida de las características ya sea por error humano o por inexactitud de precisión en la medida por ejemplo al tomar el peso de la fruta y además más aún teniendo en cuenta de que las frutas tienen características en las que son muy parecidas como puede ser el peso o el tamaño.

4. Sea  $X$  una matriz de números reales de dimensiones  $N \times d$ ,  $N > d$ . Sea  $X = UDV^T$  su descomposición en valores singulares (SVD). Calcular la SVD de  $X^T X$  y  $XX^T$  en función de la SVD de  $X$ . Identifique dos propiedades de estas nuevas matrices que no tiene  $X$ ?. ¿Qué valor representa la suma de la diagonal principal de cada una de las matrices producto?

**Solución:**

$$X^T X = (UDV^T)^T (UDV^T) = VD^T U^T U D V^T$$

Como  $U$  es ortogonal se tiene que  $UU^T = I$  donde  $I$  es la matriz identidad, luego tenemos:

$$X^T X = VD^T U^T U D V^T = VD^T D V^T = VD_1 V^{-1}$$

Donde  $D_1 = D^T D$ , es una matriz diagonal, y  $VD_1 V^{-1}$  es la descomposición en valores singulares de  $X^T X$ .

$$XX^T = (UDV^T)(UDV^T)^T = UDV^T VD^T U^T = UDD^T U^T = UD_1 U^{-1}$$

$UD_1 U^{-1}$  es la descomposición en valores singulares de  $XX^T$ .

Una propiedad que vemos directamente de  $X^T X$  y  $XX^T$  es que son diagonalizables ortogonalmente. También vemos que son matrices cuadradas simétricas.

La traza de  $X^T X$  y  $XX^T$  tiene el mismo valor. La suma de una corresponde a la raíz cuadrada del módulo de las columnas de  $X$  y la de la otra a la suma de la raíz cuadrada del módulo de las filas de  $X$ .

5. Sean  $\mathbf{x}$  y  $\mathbf{y}$  dos vectores de características de dimensión  $M \times 1$ . La expresión

$$\text{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{M} \sum_{i=1}^M (x_i - \bar{x})(y_i - \bar{y})$$

define la covarianza entre dichos vectores, donde  $\bar{z}$  representa el valor medio de los elementos de  $\mathbf{z}$ . Considere ahora una matriz  $X$  cuyas columnas representan vectores de características. La matriz de covarianzas asociada a la matriz  $X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$  es el conjunto de covarianzas definidas por cada dos de sus vectores columnas. Es decir,

$$\text{cov}(X) = \begin{pmatrix} \text{cov}(\mathbf{x}_1, \mathbf{x}_1) & \text{cov}(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \text{cov}(\mathbf{x}_1, \mathbf{x}_N) \\ \text{cov}(\mathbf{x}_2, \mathbf{x}_1) & \text{cov}(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \text{cov}(\mathbf{x}_2, \mathbf{x}_N) \\ \cdots & \cdots & \cdots & \cdots \\ \text{cov}(\mathbf{x}_N, \mathbf{x}_1) & \text{cov}(\mathbf{x}_N, \mathbf{x}_2) & \cdots & \text{cov}(\mathbf{x}_N, \mathbf{x}_N) \end{pmatrix}$$

Sea  $1_M^T = (1, 1, \dots, 1)$  un vector  $M \times 1$  de unos. Mostrar que representan las siguientes expresiones

(a)  $E1 = 11^T X$

(b)  $E2 = (X - \frac{1}{M} E1)^T (X - \frac{1}{M} E1)$

**Solución:**

(a)  $E1 = 11^T X$

$$E1 = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{N1} \\ x_{12} & x_{22} & \cdots & x_{N2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1M} & x_{2M} & \cdots & x_{NM} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^M x_{1i} & \sum_{i=1}^M x_{2i} & \cdots & \sum_{i=1}^M x_{Ni} \\ \sum_{i=1}^M x_{1i} & \sum_{i=1}^M x_{2i} & \cdots & \sum_{i=1}^M x_{Ni} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^M x_{1i} & \sum_{i=1}^M x_{2i} & \cdots & \sum_{i=1}^M x_{Ni} \end{pmatrix}$$

(b)  $E2 = (X - \frac{1}{M} E1)^T (X - \frac{1}{M} E1)$

$$\begin{aligned} (X - \frac{1}{M} E1) &= \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{N1} \\ x_{12} & x_{22} & \cdots & x_{N2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1M} & x_{2M} & \cdots & x_{NM} \end{pmatrix} - \begin{pmatrix} \frac{1}{M} \sum_{i=1}^M x_{1i} & \frac{1}{M} \sum_{i=1}^M x_{2i} & \cdots & \frac{1}{M} \sum_{i=1}^M x_{Ni} \\ \frac{1}{M} \sum_{i=1}^M x_{1i} & \frac{1}{M} \sum_{i=1}^M x_{2i} & \cdots & \frac{1}{M} \sum_{i=1}^M x_{Ni} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{M} \sum_{i=1}^M x_{1i} & \frac{1}{M} \sum_{i=1}^M x_{2i} & \cdots & \frac{1}{M} \sum_{i=1}^M x_{Ni} \end{pmatrix} = \\ &= \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{N1} \\ x_{12} & x_{22} & \cdots & x_{N2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1M} & x_{2M} & \cdots & x_{NM} \end{pmatrix} - \begin{pmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_N \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_N \\ \cdots & \cdots & \cdots & \cdots \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_N \end{pmatrix} = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \cdots & x_{N1} - \bar{x}_N \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{N2} - \bar{x}_N \\ \cdots & \cdots & \cdots & \cdots \\ x_{1M} - \bar{x}_1 & x_{2M} - \bar{x}_2 & \cdots & x_{NM} - \bar{x}_N \end{pmatrix} \\ E2 &= \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_1 & \cdots & x_{1M} - \bar{x}_1 \\ x_{21} - \bar{x}_2 & x_{22} - \bar{x}_2 & \cdots & x_{2M} - \bar{x}_2 \\ \cdots & \cdots & \cdots & \cdots \\ x_{N1} - \bar{x}_N & x_{N2} - \bar{x}_N & \cdots & x_{NM} - \bar{x}_N \end{pmatrix} \begin{pmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \cdots & x_{N1} - \bar{x}_N \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{N2} - \bar{x}_N \\ \cdots & \cdots & \cdots & \cdots \\ x_{1M} - \bar{x}_1 & x_{2M} - \bar{x}_2 & \cdots & x_{NM} - \bar{x}_N \end{pmatrix} = \\ &= \begin{pmatrix} \sum_{i=1}^M (x_{1i} - \bar{x}_1)(x_{1i} - \bar{x}_1) & \sum_{i=1}^M (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) & \cdots & \sum_{i=1}^M (x_{1i} - \bar{x}_1)(x_{Ni} - \bar{x}_N) \\ \sum_{i=1}^M (x_{2i} - \bar{x}_2)(x_{1i} - \bar{x}_1) & \sum_{i=1}^M (x_{2i} - \bar{x}_2)(x_{2i} - \bar{x}_2) & \cdots & \sum_{i=1}^M (x_{2i} - \bar{x}_2)(x_{Ni} - \bar{x}_N) \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^M (x_{Ni} - \bar{x}_N)(x_{1i} - \bar{x}_1) & \sum_{i=1}^M (x_{Ni} - \bar{x}_N)(x_{2i} - \bar{x}_2) & \cdots & \sum_{i=1}^M (x_{Ni} - \bar{x}_N)(x_{Ni} - \bar{x}_N) \end{pmatrix} \end{aligned}$$

$$\begin{pmatrix} M \cdot \text{cov}(\mathbf{x}_1, \mathbf{x}_1) & M \cdot \text{cov}(\mathbf{x}_1, \mathbf{x}_2) & \cdots & M \cdot \text{cov}(\mathbf{x}_1, \mathbf{x}_N) \\ M \cdot \text{cov}(\mathbf{x}_2, \mathbf{x}_1) & M \cdot \text{cov}(\mathbf{x}_2, \mathbf{x}_2) & \cdots & M \cdot \text{cov}(\mathbf{x}_2, \mathbf{x}_N) \\ \cdots & \cdots & \cdots & \cdots \\ M \cdot \text{cov}(\mathbf{x}_N, \mathbf{x}_1) & M \cdot \text{cov}(\mathbf{x}_N, \mathbf{x}_2) & \cdots & M \cdot \text{cov}(\mathbf{x}_N, \mathbf{x}_N) \end{pmatrix} = M \cdot \text{cov}(\mathbf{X})$$

Luego tenemos que:  $E_2 = M \cdot \text{cov}(\mathbf{X})$

6. Considerar la matriz **hat** definida en regresión,  $H = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ , donde  $\mathbf{X}$  es una matriz  $N \times (d+1)$ ,  $\mathbf{X}^T \mathbf{X}$  es invertible.

(a) Mostrar que  $H$  es simétrica.

**Solución:** Para ver que  $H$  es simétrica tenemos que comprobar si  $H = H^T$ . Para ello calculemos  $H^T$ .

$$\begin{aligned} H^T &= (X(X^T X)^{-1} X^T)^T = (X^T)^T ((X^T X)^{-1})^T X^T = \\ &= X((X^T X)^T)^{-1} X^T = X(X^T X)^{-1} X^T = H \end{aligned}$$

Luego queda demostrado que  $H$  es simétrica.

(b) Mostrar que es idempotente  $H^2 = H$

**Solución:**

$$\begin{aligned} H^2 &= (X(X^T X)^{-1} X^T)^2 = X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T = \\ &= X(X^T X)^{-1} (X^T X)(X^T X)^{-1} X^T = X(X^T X)^{-1} X^T = H \end{aligned}$$

(c) ¿Qué representa la matriz  $H$  en un modelo de regresión?

Representa a la matriz de transformación del vector de datos observados en un vector de estimaciones y define los pesos de las etiquetas de aprendizaje usando la matriz  $X$ .

7. La regla de adaptación de los pesos del Perceptron ( $\mathbf{w}_{new} = \mathbf{w}_{old} + y\mathbf{x}$ ) tiene la interesante propiedad de que los mueve en la dirección adecuada para clasificar  $\mathbf{x}$  de forma correcta. Suponga el vector de pesos  $\mathbf{w}$  de un modelo y un dato  $\mathbf{x}(t)$  mal clasificado respecto de dicho modelo. Probar que la regla de adaptación de pesos siempre produce un movimiento en la dirección correcta para clasificar bien  $\mathbf{x}(t)$ .

$$(\mathbf{w}_{new} = \mathbf{w}_{old} + y\mathbf{x}) \Rightarrow \mathbf{x} \cdot \mathbf{w}_{new} = \mathbf{x} \cdot \mathbf{w}_{old} + \mathbf{x} \cdot y\mathbf{x} = \mathbf{x} \cdot \mathbf{w}_{old} + y(\mathbf{x} \cdot \mathbf{x})$$

- Si  $\mathbf{x}(t)$  se clasificó correctamente, entonces el algoritmo no aplica la regla de actualización, por lo que nada cambia.
- Si  $\mathbf{x}(t)$  se clasificó incorrectamente como negativo, entonces  $y = 1$ , y se desplaza hacia el lugar correcto.
- Si  $\mathbf{x}(t)$  se clasificó incorrectamente como positivo, entonces  $y = -1$ , y se vuelve a desplazar hacia el lugar correcto.

8. Sea un problema probabilístico de clasificación binaria cuyas etiquetas son  $\{0,1\}$ , es decir  $P(Y = 1) = h(x)$  y  $P(Y = 0) = 1 - h(x)$

(a) Dar una expresión para  $P(Y)$  que sea válida tanto para  $Y=1$  como para  $Y=0$ .

$$P(Y) = h(x)^y (1 - h(x))^{1-y}$$

- (b) Considere una muestra  $N$  v.a. independientes. Escribir la función de Máxima Verosimilitud para dicha muestra.

$$L(Y|w_1, \dots, w_N) = \prod_{i=1}^N h(x_n)^{y_n} \prod_{i=1}^N (1 - h(x_n))^{1-y_n}$$

- (c) Mostrar que la función que maximiza la verosimilitud de la muestra es la misma que minimiza

$$E_{in}(\mathbf{w}) = \sum_{n=1}^N [y_n = 1] \ln \frac{1}{h(x_n)} + [y_n = 0] \ln \frac{1}{1 - h(x_n)}$$

donde  $[\cdot]$  vale 1 ó 0 según que sea verdad o falso respectivamente la expresión en su interior. Calculamos el menos logaritmo de la verosimilitud:

$$\begin{aligned} -\ln\left(\prod_{n=1}^N h(x_n)^{y_n} \prod_{n=1}^N (1 - h(x_n))^{1-y_n}\right) &= -\sum_{n=1}^N y_n \ln(h(x_n)) - \sum_{n=1}^N (1 - y_n) \ln(1 - h(x_n)) = \\ &= \sum_{n=1}^N y_n \ln\left(\frac{1}{h(x_n)}\right) + \sum_{n=1}^N (1 - y_n) \ln\left(\frac{1}{1 - h(x_n)}\right) \end{aligned}$$

Basta darse cuenta de que como el logaritmo es una función monótona estrictamente creciente la función que maximiza la verosimilitud es la misma que maximiza el logaritmo de la verosimilitud. Y por último ver que la función que maximiza el logaritmo de la verosimilitud es la misma que minimiza lo anterior si lo multiplicamos por -1, y dicha expresión es la dada en el enunciado.

- (d) Para el caso  $h(x) = \sigma(\mathbf{w}^T \mathbf{x})$  mostrar que minimizar el error de la muestra en el apartado anterior es equivalente a minimizar el error muestral

$$\begin{aligned} E_{in}(\mathbf{w}) &= \frac{1}{N} \sum_{n=1}^N \ln\left(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}\right) \\ \ln\left(\frac{1}{\sigma(\mathbf{w}^T \mathbf{x}_n)}\right) &= -\ln\left(\frac{e^{\mathbf{w}^T \mathbf{x}_n}}{1 + e^{\mathbf{w}^T \mathbf{x}_n}}\right) = \ln(1 + e^{\mathbf{w}^T \mathbf{x}_n}) - \mathbf{w}^T \mathbf{x}_n \\ \ln\left(\frac{1}{1 - \sigma(\mathbf{w}^T \mathbf{x}_n)}\right) &= -\ln\left(\frac{1}{1 + e^{\mathbf{w}^T \mathbf{x}_n}}\right) = \ln(1 + e^{\mathbf{w}^T \mathbf{x}_n}) \end{aligned}$$

$$E_{in}(\mathbf{w}) = \sum_{n=1}^N [y_n = 1] \ln \frac{1}{\sigma(\mathbf{w}^T \mathbf{x}_n)} + [y_n = 0] \ln \frac{1}{1 - \sigma(\mathbf{w}^T \mathbf{x}_n)} = \sum_{n=1}^N \ln(1 + e^{\mathbf{w}^T \mathbf{x}_n}) - \sum_{n=1}^N [y_n = 1] \mathbf{w}^T \mathbf{x}_n$$

9. Mostrar que en regresión logística se verifica:

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

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

$$\nabla E_{in}(\mathbf{w}) = \frac{\partial}{\partial \mathbf{w}} \left( \frac{1}{N} \sum_{n=1}^N \ln(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}) \right) = \frac{1}{N} \sum_{n=1}^N -y_n \mathbf{x}_n \frac{e^{-y_n \mathbf{w}^T \mathbf{x}_n}}{1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}}$$

Multiplicando en la fracción que aparece en la última sumatoria por  $e^{y_n \mathbf{w}^T \mathbf{x}_n}$  arriba y abajo obtenemos la primera igualdad:

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

La segunda igualdad es evidente teniendo en cuenta que:

$$\sigma(-y_n \mathbf{w}^T \mathbf{x}_n) = \frac{e^{-y_n \mathbf{w}^T \mathbf{x}_n}}{1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}}$$

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

$$\mathbf{e}_n(\mathbf{w}) = \max(0, -y_n \mathbf{w}^T \mathbf{x}_n)$$

Argumentar si con esta función de error el algoritmo PLA puede interpretarse como SGD sobre  $\mathbf{e}_n$  con tasa de aprendizaje  $v = 1$ .

El SGD usa la regla  $w_j = w_j - \eta \frac{\partial e_n(w)}{\partial w_j}$ , como  $\eta = 1$  la expresión queda como  $w_j = w_j - \frac{\partial e_n(w)}{\partial w_j}$ .

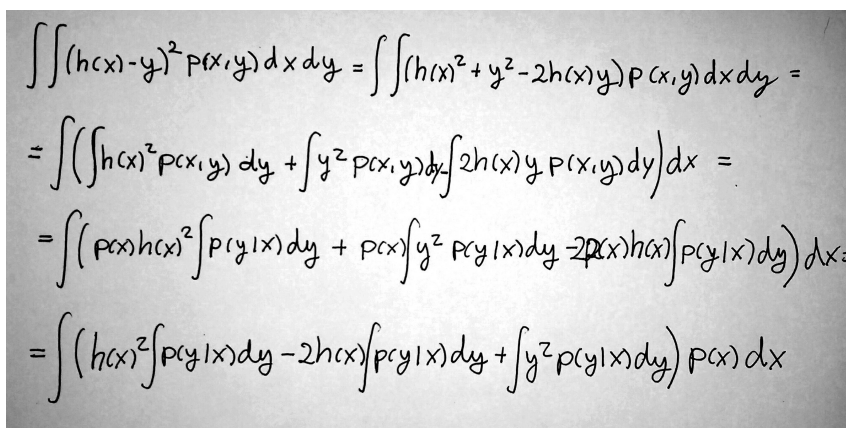
En este caso tenemos que  $\mathbf{e}_n(\mathbf{w}) = \max(0, -y_n \mathbf{w}^T \mathbf{x}_n)$ , derivamos la expresión  $-y_n \mathbf{w}^T \mathbf{x}_n = -y_n \sum w_i x_{ni}$  a la que voy a llamar  $h(w)$  y obtenemos  $\frac{\partial h(w)}{\partial w_j} = -y_n x_{nj}$ . Si  $x$  está bien clasificado entonces el  $\max(0, -y_n \mathbf{w}^T \mathbf{x}_n) = 0$ , luego el SGD se queda como estaba ( $w_j = w_j$ ) al igual que en el perceptron, si  $x$  está mal clasificado  $-y_n \mathbf{w}^T \mathbf{x}_n$  es positivo con lo que quedaría  $w_j = w_j + y_n x_{nj}$ , y queda comprobado de que en este caso el perceptron y el SGD son iguales.

1. (BONUS) En regresión lineal con ruido en las etiquetas, el error fuera de la muestra para una  $h$  dada puede expresarse como

$$E_{out}(h) = \mathbb{E}_{x,y}[(h(x) - y)^2] = \int \int (h(x) - y)^2 p(x, y) dx dy$$

(a) Desarrollar la expresión y mostrar que

$$E_{out}(h) = \int \left( h(x)^2 \int p(y|x) dy - 2h(x) \int y \cdot p(y|x) dy + \int y^2 p(y|x) dy \right) p(x) dx$$



$$\begin{aligned} \int \int (h(x) - y)^2 p(x, y) dx dy &= \int \int (h(x)^2 + y^2 - 2h(x)y) p(x, y) dx dy = \\ &= \int \left( \int h(x)^2 p(x, y) dy + \int y^2 p(x, y) dy - 2h(x) \int y p(x, y) dy \right) dx = \\ &= \int \left( p(x) h(x)^2 \int p(y|x) dy + p(x) \int y^2 p(y|x) dy - 2p(x) h(x) \int y p(y|x) dy \right) dx = \\ &= \int \left( h(x)^2 \int p(y|x) dy - 2h(x) \int y p(y|x) dy + \int y^2 p(y|x) dy \right) p(x) dx \end{aligned}$$

- (b) El término entre en paréntesis  $E_{out}$  corresponde al desarrollo de la expresión

$$\int (h(x) - y)^2 p(x, y) dy$$

¿Qué mide este término para una  $h$  dada?

Mide la media de los residuos a cuadrado, es decir a desviación de dicha variable respecto a su media, lo que es conocido como varianza.

- (c) Verificar que si la distribución de probabilidad con la que extraemos las muestras es conocida, entonces la hipótesis óptima  $h^*$  que minimiza  $E_{out}(h)$  está dada por

$$h^*(x) = \mathbb{E}_y[y|x] = \int y \cdot p(y|x) dy$$

Si minimiza el  $E_{out}$  ya que el valor que minimiza la varianza es la esperanza.

- (d) ¿Cuál es el valor de  $E_{out}(h^*)$  ?
- (e) Dar una interpretación, en términos de una muestra de datos, de la definición de la hipótesis óptima.
2. Una modificación del algoritmo perceptron denominada ADALINE, incorpora en la regla de adaptación una ponderación sobre la cantidad de movimiento necesaria. En PLA se aplica  $w_{new} = w_{old} + y_n x_n$  en ADALINE se aplica la regla  $w_{new} = w_{old} + \eta (y_n - w^T x_n) x_n$ .