Trabajo 1: Cuestiones de Teoría

Ignacio Vellido Expósito

PREGUNTAS

Todas las preguntas tienen el mismo valor

1. Identificar, para cada una de las siguientes tareas, cual es el prioblema, que tipo de aprendizaje es el adecuado (supervisado, no supervisado, por refuerzo) y los elementos de aprendizaje $(\mathcal{X}, f, \mathcal{Y})$ que deberíamos usar en cada caso. Si una tarea se ajusta a más de un tipo, explicar como y describir los elementos para cada tipo.
   a) Clasificación automática de cartas por distrito postal.
   b) Decidir si un determinado índice del mercado de valores subirá o bajará dentro de un periodo de tiempo fijado.
   c) Hacer que un dron sea capaz de rodear un obstaculo.
   d) Dada una colección de fotos de perros, posiblemente de distintas razas, establecer cuantas razas distintas hay representadas en la colección.

   a **Problema**: Clasificar cartas en distrito según su código postal.
   Tipo de aprendizaje: Supervisado, pues debemos saber a qué distrito pertenece cada calle.
   X: Direcciones de las cartas.
   f: Clasificación perfecta de una carta a un distrito en base a su dirección.
   Y: Distritos postales.

   b **Problema**: Predecir el estado del índice del mercado de valores en un instante dado. La comparativa con el valor actual no tiene por qué hacerla la máquina, podemos comparar el valor/rango devuelto con el actual.
   Tipo de aprendizaje: Supervisado, ya que el sistema debe comprender en qué forma afecta los valores de entrada en el mercado.
   X: Información sobre el mercado. No solo tendría por qué ser el valor de las acciones de cada empresa, también se podría incluir información sobre cada una.
   f: Predicción exacta del rango o valor del índice del mercado.
   Y: Valor o rango del índice.

   c **Problema**: Decidir la serie de movimientos a realizar para que el dron evite el obstáculo.
   Tipo de aprendizaje: Por refuerzo, la serie de movimientos puede medirse con un posible beneficio que le ayuda a realizar la maniobra.
   X: Datos devueltos por los sensores del dron.
   f: Indica los pasos correctos para poder rodear el obstáculo.
   Y: Movimientos para esquivar el obstáculo.

   d **Problema**: Agrupar los perros de una foto en distintas razas, sin tener que identificar cada una.
   Tipo de aprendizaje: No supervisado, ya que solo es necesario que el sistema sea capaz de diferenciar las distintas razas.
   X: Fotos de perros, por ejemplo, en forma de matriz de píxeles.
   f: Devuelve el número de razas de perros diferentes en una foto.
   Y: Número de razas encontradas en una foto.

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) Determinar si un vertebrado es mamífero, reptil, ave, anfibio o pez.
   b) Determinar si se debe aplicar una campaña de vacunación contra una enfermedad.
   c) Determinar perfiles de consumidor en una cadena de supermercados.
   d) Determinar el estado anímico de una persona a partir de una foto de su cara.
   e) Determinar el ciclo óptimo para las luces de los semáforos en un cruce con mucho tráfico.

   a Al ser las diferencias grandes entre las posibles clases, una aproximación por diseño resulta suficiente para este problema. Por ejemplo, para determinar si es un ave nos vale con determinar si el vertebrado tiene plumas y un pico sin dientes. Si estuviésemos interesados en clasificar los animales dentro de cada clase una aproximación por aprendizaje nos resultaría más eficiente.

   b Averiguando datos sobre la enfermedad (peligrosidad, facilidad de contagio, etc.) debe ser suficiente para determinar si se debe aplicar una campaña de vacunación, por lo que no necesitamos aprendizaje, una aproximación por diseño es suficiente.

   c Por aprendizaje, ya que puede haber relaciones ocultas entre los datos del perfil de un consumidor que ayuden a agruparlos en diferentes clases, consiguiendo más potencia de clasificación que una aproximación por diseño.

   d Claramente se necesita una aproximación por aprendizaje, pues no tenemos una regla clara para la distinción en estados de ánimo. Recogiendo imágenes y aplicando un aprendizaje supervisado podríamos aprender a clasificar los estados anímicos.

   e Una aproximación por diseño nos vale, pues se puede determinar el ciclo óptimo simplemente con la cantidad y los horarios del tráfico.

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.

   El problema consiste en identificar una fruta según sus características, teniendo como posibles clases mangos, papayas y guayabas.

   X: Datos sobre las frutas, se proporciona un vector de características por cada entrada con valores como: color, tacto, forma, tamaño...
   D: Pares (x, y) con $x \in X$ e $y \in Y$
   f: Función que clasifica perfectamente una fruta en función de sus características.
   Y: Una de las tres posibles clases.

   Es inevitable y evidente que en una situación real las etiquetas van a contener ruido. Siempre existe una alta probabilidad de que algún dato venga con algún error, pues estamos recopilando grandes cantidades de datos.

Nota: Le pido disculpas por el formato, tuve problemas con Latex y el resto del trabajo he tenido que hacerlo en Microsoft Office.

4. 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.

Aplicando la descomposición en valores singulares:

$$X = UDV^T$$
$$X^T = VDU^T$$

U y V son ortogonales, por tanto:

$$X^T X = VDDV^T$$
$$A = X^T X$$

Por ser A matriz cuadrada y V ortogonal, la descomposición anterior también se puede considerar como una en valores singulares, teniendo:

$$A = X^T X = VDDV^T = U_2 D_2 V^T{}_2$$

$$V = U_2$$
$$D^2 = D_2$$
$$V^T = V^T{}_2$$

Por la segunda igualdad podemos ver que los valores singulares de A son los cuadrados de los de X.

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

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

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 $\mathrm{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$ es el conjunto de covarianzas definidas por cada dos de sus vectores columnas. Es decir,

$$\mathrm{cov(X)} = \begin{pmatrix} \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_N) \\ \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_N) \\ \cdots & \cdots & \cdots & \cdots \\ \mathrm{cov}(\mathbf{x}_N, \mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_N, \mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_N, \mathbf{x}_N) \end{pmatrix} \tag{0.1}$$

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

a) $E1 = 11^T \mathrm{X}$
b) $E2 = (\mathrm{X} - \frac{1}{M} E1)^T (\mathrm{X} - \frac{1}{M} E1)$

a)

$$E1 = 11^T X$$

Desplegamos:

$$E1 = \begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \end{pmatrix} (1,1,\dots,1)\ X$$

Y realizamos la multiplicación:

$$E1 = \left(\sum_1^N x_{j_1}, \sum_1^N x_{j_2}, \dots, \sum_1^N x_{j_N} \right)$$

Lo que equivale a la suma de los valores de cada característica de entrada

b)

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

A partir de lo calculado en el apartado anterior, tenemos que:

$$\frac{1}{M} E1 = \left(\widetilde{x_{j_1}}, \widetilde{x_{j_2}}, \dots, \widetilde{x_{j_N}} \right)$$

$$X - \frac{1}{M} E1 = \left(x_{j_1} - \widetilde{x_{j_1}}, x_{j_2} - \widetilde{x_{j_2}}, \dots, x_{j_N} - \widetilde{x_{j_N}} \right)$$

Por lo que, sustituyendo en la ecuación anterior nos queda:

$$E2 = M * cov(X)$$

6. Considerar la matriz **hat** definida en regresión, $\hat{\mathrm{H}} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$, donde X es la matriz de observaciones de dimensión $N \times (d+1)$, y $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ es invertible.
   a) ¿Que representa la matriz $\hat{\mathrm{H}}$ en un modelo de regresión?
   b) Identifique la propiedad más relevante de dicha matriz en relación con regresión lineal.

Justificar las respuestas.

a) $\widehat{H}$ es una matriz que indica la estimación de y como combinación de las columnas de X:

$$\hat{\boldsymbol{y}} = \mathrm{X}\mathbf{w}_{lin} = \mathrm{X}\,(X^T X)^{-1} X^T \boldsymbol{y} = \widehat{\mathrm{H}}\boldsymbol{y}$$

También es llamada *matriz de proyección* pues proyecta el vector de observaciones $\boldsymbol{y}$ en el de predicciones $\hat{\boldsymbol{y}}$.

Esta matriz además se utiliza para calcular el *leverage*, que es una medida del efecto de una muestra concreta en la predicción realizada.

a) Esta matriz es idempotente ($\widehat{H}^k = \widehat{H} \quad k > 0$)
Esto indica que no importa el número de veces que apliquemos la regresión, que para el mismo conjunto de datos $\widehat{H}$ siempre será igual.

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 mueve el vector de pesos 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 matematicamente que el movimiento de la regla de adaptación de pesos siempre produce un movimiento de $\mathbf{w}$ en la dirección correcta para clasificar bien $\mathbf{x}(t)$.

Por lo visto en clase:

$$\text{LEARNING RULE :}$$
$$\begin{cases} w_{updated} = w_{current} + y \cdot \boldsymbol{x} \;\; if \quad sign(\boldsymbol{w}^T \boldsymbol{x}_i) \neq y_i \\ w_{updated} = w_{current} \;\; if \quad sign(\boldsymbol{w}^T \boldsymbol{x}_i) = y_i \end{cases}$$

tenemos que el Perceptron solo actualiza los pesos cuando existe un dato mal clasificado, aplicando la regla:

$$w_{new} = w_{old} + yx$$

O de igual manera:

$$w(t+1) = w(t) + y(t)\, x(t)$$

Multiplicando en ambos lados por $x(t)$:

$$x(t)\, w(t+1) = x(t)w(t) + y(t)x(t)x(t)$$

$$Siendo\ x(t)x(t) \geq 0\ por\ definición\ del\ producto\ escalar$$

Al estar el dato mal clasificado, tenemos que:

$$sign\left(w^T(t)\, x(t)\right) \neq y(t)$$
$$y(t)\, w^T(t)\, x(t) < 0$$

Y dos opciones:

$$Si\ sign\left(w^T(t)\, x(t)\right) > 0 \;\rightarrow\; y(t) < 0$$
$$y\ \ w_{new}\ aumenta\ por\ la\ tercera\ ecuación$$
$$Si\ sign\left(w^T(t)\, x(t)\right) < 0 \;\rightarrow\; y(t) > 0$$
$$y\ w_{new}\ disminuye\ por\ la\ tercera\ ecuación$$

En cualquiera de las dos opciones w_{new} se mueve en la opción correcta de manera que:

$$y\, {w_{new}}^T\, x > 0$$

8. Sea un problema probabilistico de clasificación binaria con etiquetas {0,1}, es decir $\mathrm{P}(\mathrm{Y}=1)=h(\mathbf{x})$ y $\mathrm{P}(\mathrm{Y}=0)=1-h(\mathbf{x})$, para una función $h()$ dependiente de la muestra

   a) Considere una muestra i.i.d. de tamaño N $(\mathbf{x}_1,\cdots,\mathbf{x}_N)$. Mostrar que la función h que maximiza la verosimilitud de la muestra es la misma que minimiza

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

   donde $[\![\cdot]\!]$ vale 1 o 0 según que sea verdad o falso respectivamente la expresión en su interior.

   b) 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

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

a) La función que maximiza la verosimilitud es:

$$\frac{1}{N}\sum_{1}^{N}ln(\frac{1}{P(y_n|x_n)})$$

Y por el enunciado:

$$P(y_n|x_n)=\begin{cases}h(x) & y_n=1\\ 1-h(x) & y_n=0\end{cases}$$

Podemos substituir obteniendo:

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

En términos de optimización ambas ecuaciones (esta última y la del enunciado) son iguales puesto que 1/N no afecta a la hora de encontrar el mínimo.

b) Partimos de:

$$h(x)=\sigma(w^Tx)=\frac{1}{1+e^{w^Tx}}\qquad 1-h(x)=\sigma(-w^Tx)=\frac{1}{1+e^{-w^Tx}}$$

Sustituyendo obtenemos:

$$E_{in}(w)=\sum_{1}^{N}[[y_n=1]]\ln(1+e^{w^Tx})+[[y_n=0]]\ln(1+e^{-w^Tx})$$

Como en lo único que afecta y_n es el signo del exponente de *e*:

$$E_{in}(w) = \sum_{1}^{N} \ln(1 + e^{(y_n - 1)w^T x})$$

No es posible obtener la ecuación del enunciado si utilizamos las etiquetas {0,1}. En caso de que fueran {-1,1} el término (y_n - 1) en nuestra ecuación pasaría a ser (y_n).

9. Derivar el error E_{in} para mostrar que en regresión logística se verifica:

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

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

Partimos de:

$$E_{in}(w) = \frac{1}{N}\sum_{1}^{n} ln(1 + e^{y_i w^T x_i})$$

$$\nabla E_{in}(w) = \frac{\partial E_{in}}{\partial w} = -\frac{1}{N}\sum_{1}^{n}\frac{y_i x_i}{1 + e^{y_i w^T x_i}}$$

Como:

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

Entonces:

$$\frac{1}{1 + e^{y_i w^T x_i}} = \sigma(-y_i w^T x_i)$$

Sustituyendo tenemos lo que pretendíamos demostrar:

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

Un ejemplo mal clasificado hace el signo del exponente de *e* negativo, por lo que *e* tiende a cero y el denominador de la división a uno, por lo que no se reduce el numerador. Esto hace que influyan más los ejemplos mal clasificados.

10. Definamos el error en un punto $(\mathbf{x}_n, y_n)$ 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 $\nu = 1$.

Nos preguntan si la actualización de los pesos es igual a:

$$w_{new} = w_{old} + yx$$

SGD ahora se definiría como:

$$w_{new} = w_{old} - \nabla e_n(w)$$

Y tenemos los siguientes casos:

$$Si \quad -y(t)\, w^T(t)\, x(t) < 0$$

x *está bien clasificado y por tanto PLA no actualizaría los pesos, y de igual manera nuestro nuevo SGD tampoco, ya que*:

$$\nabla e_n(w) = 0$$

$$Si \quad -y(t)\, w^T(t)\, x(t) > 0$$

x *está mal clasificado y PLA añadiría a los pesos* $\boldsymbol{yx}$.

Nuestro SGD tendría:

$$e_n(w) = y(t)\, w^T(t)\, x(t)$$
$$\nabla e_n(w) = y(t)\, x(t)$$

Entonces actualizaría de la misma manera que PLA.

De forma que SGD no tendría diferencia con PLA.

BONUS

Los BONUS solo serán tenidos en cuenta si en el cuestionario obligatorio se ha conseguido al menos un 75 % de los puntos totales.

1. (2 puntos) En regresión lineal con ruido en las etiquetas, el `error fuera de la muestra para una h dada` puede expresarse como

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

   a) Desarrollar la expresión y mostrar que

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

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

$$\int (h(\mathbf{x}) - y)^2 p(y|\mathbf{x}) dy$$

   ¿Que mide este término para una h dada?.

   c) El objetivo que se persigue en Regression Lineal es encontrar la función $h \in \mathcal{H}$ que minimiza $E_{\text{out}}(h)$. Verificar que si la distribución de probabilida $p(\mathbf{x}, y)$ con la que extraemos las muestras es conocida, entonces la hipótesis óptima h^* que minimiza $E_{\text{out}}(h)$ está dada por

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

   d) ¿Cuál es el valor de $E_{\text{out}}(h^*)$?

   e) Dar una interpretación, en términos de una muestra de datos, de la definición de la hipótesis óptima.

Pregunta no contestada.

2. (1 punto) Una modificación del algoritmo perceptron denominada ADALINE, incorpora en la regla de adaptación una poderación sobre la cantidad de movimiento necesaria. En PLA se aplica $\mathbf{w}_{new} = \mathbf{w}_{old} + y_n \mathbf{x}_n$ y en ADALINE se aplica la regla $\mathbf{w}_{new} = \mathbf{w}_{old} + \eta(y_n - \mathbf{w}^T \mathbf{x}_n)\mathbf{x}_n$. Considerar la función de error $E_n(\mathbf{w})) = (\max(0, 1 - y_n \mathbf{w}^T \mathbf{x}_n))^2$. Argumentar que la regla de adaptación de ADALINE es equivalente a gradiente descendente estocástico (SGD) sobre $\frac{1}{N} \sum_{n=1}^{N} E_n(\mathbf{w})$.

Pregunta no contestada.