APRENDIZAJE AUTOMÁTICO

Trabajo.1: Cuestiones de Teoría

Fecha de entrega: 23 marzo 2018. Valor máximo: 12.5 puntos

NORMAS DE DESARROLLO Y ENTREGA DE TRABAJOS

Para este trabajo como para los demás es obligatorio presentar un informe escrito (hacerlo en pdf, MS Word). **Sin este informe se considera que el trabajo NO ha sido presentado.**

Normas para el desarrollo de los Trabajos: EL INCUMPLIMIENTO DE LAS NORMAS (*) SIGNIFICA PERDIDA DE 2 PUNTOS POR CADA INCUMPLIMIENTO.

- En su informe de contestación debe incluir todas las preguntas en el orden y tal y como se les formula en este documento. (*)
- Las contestaciones irán a continuación de cada pregunta, dejando en blanco las que no conteste.(*)
- Todas las contestaciones deben ser justificadas con argumentos. Sin argumentos la pregunta se considera no contestada.
- Todas las justificaciones matemáticas deben contener todos y cada uno de los pasos de la misma. En caso de duda la contestación no se considerará válida.
- Cualquier desarrollo matemático hecho a mano que no presente la claridad y calidad de un editor de ecuaciones (MS WORD, latex) no se considerará válida. Se recomienda vivamente usar el editor latex para estos casos.
- **Forma de entrega**: Subir el pdf a la web docente de CCIA.

PREGUNTAS

Todas las preguntas tienen el mismo valor

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 deberiamos 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.
   b) Clasificación automática de cartas por distrito postal
   c) Decidir si un determinado índice del mercado de valores subirá o bajará dentro de un periodo de tiempo fijado.
   d) Aprender un algoritmo que permita a un robot rodear un obstaculo.
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.
   b) Determinar si se debe aplicar una campaña de vacunación contra una enfermedad.
   c) Determinar si un correo electrónico es de propaganda o no.
   d) Determinar el estado de ánimo 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.
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.
4. Sea X una matriz de números reales de dimensiones $N \times d$, $N > d$. Sea $\mathrm{X} = \mathrm{UDV}^{\mathrm{T}}$ su descomposición en valores singulares (SVD). Calcular la SVD de $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ y XX^{T} en función de la SVD de X. Identifique dos propiedades de estás nuevas matrices que no tiene X?. ¿Qué valor representa la suma de la diagonal principal de cada una de las matrices producto?
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,

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

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

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

   a) Mostrar que H es simétrica

   b) Mostrar que es idempotente $\mathrm{H}^2 = \mathrm{H}$

   c) ¿Que representa la matriz H en un modelo de regresión?

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 $\mathbf{x}(t)$.

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

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

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

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

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

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

   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

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

9. Mostrar que en regresión logística se verifica:

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

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

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 $\nu = 1$.

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.

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 $\frac{1}{N} \sum_{n=1}^{N} E_n(\mathbf{w})$.