Trabajo.1: Relación de Ejercicios de Apoyo

Grado de dificultad: Bajo(B), Medio(M), Alto(A)

Estos ejercicios están diseñados como elementos de ayuda a la comprensión de la teoría.

Intentar resolver el máximo número posible

El problema del Aprendizaje

1. (B) Expresar cada una de las siguientes tareas en el escenario de "learning from data", especificando el espacio de entrada $\mathcal{X}$, el espacio de salida $\mathcal{Y}$, la función objetivo $f : \mathcal{X} \to \mathcal{Y}$ y detalles específicos de los datos que usaremos para aprender
   a) Diagnostico médico: Un paciente llega con su historia médica y algunos sintomas, y se desea identificar el problema.
   b) Reconocimiento de digitos manuscritos (p.e. para clasificación automática de códigos postales)
   c) Determinar si un correo electrónico es spam o no.
   d) Predecir como varía el consumo eléctrico con el coste, la temperatura y el día de la semana.
   e) Suponga que tiene un problema para el que no conoce una solución analítica pero dispone de datos a partir de los cuales construir una solución empírica.
2. (B) ¿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 la edad a la cual se debería pasar un determinado examen médico.
   b) Clasificar números en primos y no primos.
   c) Detectar potenciales fraudes en cargos a tarjetas de crédito.
   d) Determinar el tiempo que tardará un objeto que cae en tocar el suelo.
   e) Determinar el ciclo óptimo para las luces de los semáforos en un cruce con mucho tráfico.

3. (B) Identificar, para cada una de las siguientes tareas, que tipo de aprendizaje es el adecuado (supervisado, no supervisado, por refuerzo) y los datos de aprendizaje que deberiamos usar. Si una tarea se ajusta a mas de un tipo, explicar como y describir los datos para cada tipo.
   a) Recomendar un libro a un usuario en una librería on-line.
   b) Jugar al tres en raya
   c) Categorizar películas entre diferentes tipos/categorias
   d) Aprender a tocar un instrumento musical
   e) Decidir el máximo crédito permitido para cada cliente de un banco
4. (B) Un problema de aprendizaje estadístico se nota formalmente por su vector de elementos
   a) Considere el vector $\{\mathcal{P}, \mathcal{X}, \mathcal{Y}, \mathcal{D}, f, \mathcal{A}, \mathcal{H}, g\}$ ¿Que significan cada uno de los elementos del vector? ¿Hay alguna propiedad que deba de cumplir $\mathcal{D}$?
   b) Identifique los elementos del vector que representan:
      1) La entrada al aprendizaje
      2) La salida del aprendizaje
      3) El clase de funciones usada
      4) El algoritmo de búsqueda usado
      5) ¿Que denota g?

MATRICES Y VECTORES

En esta sección aprenderá algunas propiedades interesantes de las matrices de datos asociadas a los datos de vectores de características. recuerde que por defecto los vectores de características son las filas de la matriz de datos X

1. (B) Verificar para matrices A 2×2 que $\frac{\partial}{\partial \mathrm{x}} \mathrm{x}^T A\mathrm{x} = 2A\mathrm{x}$ si la matriz A es simétrica e igual a $(A + A^T)\mathrm{x}$ si no lo es.
2. (B) Verificar las siguientes propiedades de las matrices de datos:
   a) Dada una matriz X($N \times d$),$N > d$, de números reales las matrices $\mathrm{XX^T}$ y $\mathrm{X^TX}$ son simétricas. (ver que los elementos con indices (ij) y (ji) son siempre iguales para valores cualesquiera de i y j)
   b) Verificar que representan los valores de traza($\mathrm{XX^T}$) y traza($\mathrm{X^TX}$) (La traza de una matriz es la suma de los valores de su diagonal principal) (Ayuda: hacerlo con tamaño 2x2 y generalizar)
   c) Sea X una matriz de números reales, Sea $\mathrm{X = UDV^T}$ su descomposición en valores singulares (SVD). Calcular la SVD de $\mathrm{X^TX}$ y $\mathrm{XX^T}$. ¿que diferencia observa entre ambas?

d) Establezca una relación entre los valores singulares de las matrices $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ y $\mathrm{X}\mathrm{X}^{\mathrm{T}}$ y los valores singulares de X. (Ayuda: los valores singulares son los valores de la matriz D)

e) Verificar que si una matriz cuadrada X tiene inversa, entonces $\mathrm{X}^{-1} = \mathrm{V}\mathrm{D}^{-1}\mathrm{U}^{\mathrm{T}}$ si $\mathrm{SVD}(\mathrm{X}) = \mathrm{U}\mathrm{D}\mathrm{V}^{\mathrm{T}}$ ¿Cómo sería la inversa si además X es simétrica?

3. (B) Sean $\mathbf{x}$ e $\mathbf{y}$ dos vectores de características. La covarianza de dos vectores es un número que mide la dependencia estadística que existe entre ellos. Covarianza cero indica que no existe dependencia estadística entre ellos. Su expresión está definida por

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

donde $\bar{\mathbf{x}}$ e $\bar{\mathbf{y}}$ representan la media de los vectores $\mathbf{x}$ e $\mathbf{y}$ respectivamente. Verificar:

a) Que la covarianza de dos vectores se puede escribir como un producto escalar de vectores.

b) Que usando los dos vectores del producto escalar es posible definir una matriz cuya traza coincide con el valor de la covarianza. (Ayuda: La traza de una matriz es la suma de los valores de su diagonal principal)

4. 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,

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

Sabemos que $\mathrm{cov}(\mathbf{x}, \mathbf{y})$ tiene la forma de un producto escalar de vectores. Usando este hecho en combinación con la forma de cov(X), transforme cov(X) en una expresión que dependa de $\mathrm{X}^T\mathrm{X}$.

5. (B) Considere la matriz hat del modelo de regresión que define los pesos de regresión(predicción) para los datos de aprendizaje X, $\hat{\mathrm{H}} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$, donde X es una matriz $N \times (d+1)$, $N >> d$, y $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ es invertible.

a) Mostrar que H es simétrica (usar SVD)

b) Mostrar que $\mathrm{H}^{\mathrm{K}} = \mathrm{H}$ para cualquier entero K (Ayuda: probar primero $\mathrm{H}^2 = \mathrm{H}$)

c) Si I es la matriz identidad de tamaño N, mostrar que $(\mathrm{I} - \mathrm{H})^{\mathrm{K}} = \mathrm{I} - \mathrm{H}$ para cualquier entero positivo K

6. En regresión lineal con ruido independiente en las etiquetas, el `error fuera de la muestra para una h dada` esta dado por

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

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

Donde $\mathbb{E}_{\mathbf{x},y}$ significa el valor medio del error de $h(\mathbf{x})$ medido sobre todas los posibles muestras $(\mathbf{x}, y)$

a) Dar una interpretación al valor de la integral entre parentesis. (ayuda: usar el teorema de Gauss-Markov)

b) Desarrollar la expresión anterior y mostrar que

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

c) ¿Cual es la interpretación de $\int y p(y|\mathbf{x}) dy$?

d) ¿Que consecuencias tendría elegir $h(\mathbf{x}) = \int y p(y|\mathbf{x}) dy$.

El Algoritmo Perceptron Lineal (PLA)

1. (B) Suponga que usamos un Perceptron para detectar correos electrónicos spam. Suponga que el vector de características asociado a cada correo está dado por un histograma que mide la frecuencia de ocurrencia de un conjunto de palabras previamente seleccionado. Los mensajes spam son etiquetados con $+1$ y los no spam con -1

   a) ¿Podría sugerir algunas palabras a las que el perceptrón terminará asignando un peso positivo grande? (Ayuda: de acuerdo con la clase de funciones perceptron el peso de cada palabra empuja hacia una de las clases)

   b) Lo mismo que antes pero ahora con peso negativo.

   c) ¿Que parámetro del perceptron afecta de forma directa a la cantidad de mensajes indecisos que acaben siendo clasificados como spam? (Ayuda: los mensajes indecisos son aquellos que no podemos decidir usando solo el histogrma de palabras)

2. (B) Considere el perceptron en dos dimensiones: $h(\mathbf{x}) = \text{sign}(\mathbf{w}^T \mathbf{x})$ donde $\mathbf{w} = [w_0, w_1, w_2]^T$ y $\mathbf{x} = [1, x_1, x_2]^T$.

   a) Mostrar que la regiones del plano donde $h(x) = +1$ y $h(x) = -1$ están separadas por una línea.

   b) Si expresamos esta línea por la ecuación $x_2 = ax_1 + b$, ¿cuales son las expresiones de a y b en términos de w_0, w_1, w_2?

   c) Dibujar un gráfico para los casos $\mathbf{w} = [1, 2, 3]^T$ y $\mathbf{w} = -[1, 2, 3]^T$

3. (M) 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 adapta en la dirección adecuada para clasificar $\mathbf{x}$ de forma correcta.

a) Mostrar que cuando en la iteración t, $\mathbf{x}(t)$ está mal clasificado por $\mathbf{w}(t)$ se verifica que $y(t)\mathbf{w}^T(t)\mathbf{x}(t) < 0$

b) Mostrar que tras adaptar los pesos en la iteración t con el punto $\mathbf{x}(t)$, se verifica $y(t)\mathbf{w}^T(t+1)\mathbf{x}(t) > y(t)\mathbf{w}^T(t)\mathbf{x}(t)$. Interprete este resultado.

TRANSFORMACIONES NO LINEALES

a) (B) Consideremos la siguiente transformación de características $\Phi(\mathbf{x}) = (1, x_1^2, x_2^2)$ ¿Que clase de curva genera en $\mathcal{X}$ el hiperplanos $\hat{\mathbf{w}}$ de $\mathcal{Z}$ correspondiente a los siguientes casos

1) $\tilde{w}_1 > 0$, $\tilde{w}_2 < 0$
2) $\tilde{w}_1 > 0$, $\tilde{w}_2 = 0$
3) $\tilde{w}_1 > 0$, $\tilde{w}_2 > 0$, $\tilde{w}_0 \leq 0$
4) $\tilde{w}_1 > 0$, $\tilde{w}_2 > 0$, $\tilde{w}_0 > 0$

b) (B) Considerar ahora la transformación $\Phi_2(\mathbf{x}) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)$. ¿Que hiperplanos $\hat{\mathbf{w}}$ de $\mathcal{Z}$ representan a las siguientes curvas frontera de $\mathcal{X}$?

1) La parábola $(x_1 - 3)^2 + x_2 = 1$
2) El círculo $(x_1 - 3)^2 + (x_2 - 4)^2 = 1$
3) La elipse $2(x_1 - 3)^2 + (x_2 - 4)^2 = 1$
4) La hipérbola $(x - 1 - 3)^2 - (x_2 - 4)^2 = 1$
5) La elipse $2(x_1 + x_2 - 3)^2 + (x_1 - x_2 - 4)^2 = 1$
6) La línea $2x_1 + x_2 = 1$

c) (B) Considerar la transformación polinomial de Q-ésimo orden, Φ_Q para $\mathcal{X} = \mathbb{R}^d$. ¿Cuál es la dimensión del espacio de características $\mathcal{Z}$ (excluyendo la coordenada fija $z_0 = 1$)? Evaluar su resultado para $d \in \{2, 3, 5, 10\}$ y $Q \in \{2, 3, 5, 10\}$

REGRESION LOGISTICA

a) (B) Considerar la función

$$\tanh(s) = \frac{e^s - e^{-s}}{e^s + e^{-s}}$$

1) ¿Como está relacionada esta función con la función logística o sigmoidal $\sigma(s)$?
2) Mostrar que $\tanh(s)$ converge a un valor asintótico finito para valores de $|s|$ grandes y no converge a ningún valor para valores de $|s|$ pequeños.

3) Dibujar la función y compararla con la función $g(s) = -1$ para $s < 0$ y $g(s) = +1$ para $s \geq 0$.

b) (B) 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.

c) (M) Este ejercicio conecta la regla de inducción de Máxima Verosimilitud, con la regla ERM en el caso de variables aleatorias binarias. Para ello supongamos que queremos predecir una función objetivo estocástica binaria, $f = P(y|\mathbf{x})$, a partir de muestras etiquetadas con valores ± 1 y de funciones hipótesis que notamos por h (observese que es un caso particular de regresión logística).

1) Escribir la verosimilitud de una muestra de tamaño 2.
2) Escribir la expresiíon de E_{in} de una muestra de tamaño 2
3) Comparar la expresiónes dadas por la maximización de la verosimilitud con la minimización de E_{in}.
4) Mostrar que la estimación de Máxima Verosimilitud para una nmuestra de tamaño N se reduce a la tarea de encontrar la función h que minimiza

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

(Ayuda: recordar que E_{in} debe minimizar en lugar de maximizar)

5) 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

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

d) (M) Consideremos el caso de la verificación de la huella digital (ver transparencias de clase). Tras aprender con un modelo de regresión logística a partir de datos obtenemos una hipótesis final

$$g(x) = \mathbb{P}[y = +1|\mathbf{x}]$$

que representa la estimación de la probabilidad de que $y = +1$. Suponga que la matriz de coste está dada por

		Verdadera Clasificación	
		+1 (persona correcta)	-1 (intruso)
decisión	+1	0	c_a
decisión	-1	c_r	0

Para una nueva persona con huella digital $\mathbf{x}$, calculamos $g(\mathbf{x})$ y tenemos que decidir si aceptar o rechazar a la persona (i.e. tenemos que usar una decisión 1/0). Por tanto aceptaremos si $g(\mathbf{x}) \geq \kappa$, donde κ es un umbral.

1) Verificar que la funciones de coste de aceptar definida como el coste promedio de las aceptaciones, y de forma semejante para el coste de rechazo, son.

$$\begin{array}{rcl} \text{costo(acceptar)} & = & (1 - g(\mathbf{x}))c_a \\ \text{costo(rechazar)} & = & g(\mathbf{x})c_r \end{array}$$

2) Usar el apartado anterior para derivar una condición sobre $g(x)$ para aceptar la persona y mostrar que

$$\kappa = \frac{c_a}{c_a + c_r}$$

3) Usar las matrices de costo para la aplicación del supermercado y la CIA (transparencias de clase) para calcular el umbral κ para cada una de las dos clases. Dar alguna interpretación del umbral obtenido.

e) (B) En regresión logística mostrar que

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

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

f) (M) En este ejercicio mostramos la relación entre distintas funciones de error que aparecen en los modelos lineales estudiados y como su relación nos permite obtener soluciones iniciales al problema de clasificación a partir de las soluciones del problema de regresión o del de regresión logística. Considerar las siguientes medidas puntuales de error, $\mathbf{eclass}(s, y) = [\![y \neq \text{sign}(s)]\!]$ (clasificación), $\mathbf{esq}(s, y) = (y - s)^2$ (regresión), y $\mathbf{elog}(s, y) = \ln(1 + \exp(-ys))$ (regresión logística), donde $s = \mathbf{w}^T \mathbf{x}$.

1) Para $y = +1$, dibujar las curvas $\mathbf{eclass}, \mathbf{esq}$ y $\frac{1}{\ln 2}\mathbf{elog}$ versus s en los mismos ejes.

2) Mostrar que $\mathbf{eclass}(s, y) \leq \mathbf{esq}(s, y)$, y por tanto que el error de clasificación esta acotado superiormente por el error cuadrático.

3) Mostrar que $\mathbf{eclass}(s, y) \leq \frac{1}{\ln 2}\mathbf{elog}(s, y)$, y como en el apartado anterior obtener una cota superior (salvo una constante) usando el error de regresión logística.

Nota: Estas cotas indican que minimizar el error cuadrático o el error de regresión logística deberían hacer decrecer también el error de clasificación.

g) (M) En este ejercicio establecemos una conexión entre el algoritmo PLA y la regla de inducción SGD

1) Definamos el error en un punto $(\mathbf{x}_n, y_n)$ respecto de un modelo $\mathbf{w}$ como

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

Mostrar que el algoritmo PLA puede interpretarse como un modelo SGD de minimización iterativa sobre $\mathbf{e}_n$ con tasa de aprendizaje $\nu = 1$.

2) Mostrar que en regresión logística si el vector de pesos $\mathbf{w}$ es muy grande, minimizar E_{in} usando SGD es similar a PLA. (Otra indicación de que los pesos de regresión logística pueden ser usados como buena aproximación en clasificación)