Relación de Ejercicios de Apoyo

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

El problema del Aprendizaje

1. (B) Elementos de un problema de Aprendizaje Estadístico.
   a) Un problema de aprendizaje estadístico se nota formalmente por su vector de elementos. Considere el vector $\{\mathcal{P}, \mathcal{X}, \mathcal{Y}, \mathcal{D}, f, \mathcal{A}, \mathcal{L}, \mathcal{H}, g\}$ ¿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) Cómo se mide el error en cada punto
      6) ¿Que es el criterio ERM y cómo influye en la búsqueda de la solución?
   c) Dada, $h \in \mathcal{H}$ ¿Cómo se define el error de h dentro (E_{in}) y fuera (E_{out}) de la muestra ?

2. (B) Identificar, para cada una de los siguientes problemas, cúal es la tarea a realizar, que tipo de aprendizaje es el adecuado (supervisado, no supervisado, por refuerzo) y los elementos de aprendizaje $(\mathcal{X}, \mathcal{Y}, f)$ que deberíamos usar en cada caso. Si una tarea se ajusta a más de un tipo, explicar cómo 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) Programar que un dron capaz de rodear un obstáculo.

d) Dada una colección de fotos de perros, posiblemente de distintas razas, establecer cúantas razas distintas hay representadas en la colección.

e) Diagnóstico médico: Un paciente llega con su historia médica y algunos síntomas, y se desea identificar el problema.

f) Reconocimiento de dígitos manuscritos (p.e. para clasificación automática de códigos postales)

g) Determinar si un correo electrónico es spam o no.

h) Predecir como varía el consumo eléctrico con el coste, la temperatura y el día de la semana.

i) 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.

3. (B) ¿Cúales de los siguientes problemas son más adecuados para una aproximación por aprendizaje y cúales 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.

   f) Determinar los ingresos medios de una persona a partir de sus datos de nivel de educación, edad, experiencia y estatus social.

   g) Determinar la edad a la cúal se debería pasar un determinado examen médico.

   h) Clasificar números en primos y no primos.

   i) Detectar potenciales fraudes en cargos a tarjetas de crédito.

   j) Determinar el tiempo que tardará un objeto que cae en tocar el suelo.

4. (B) 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.

5. (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 más de un tipo, explicar como y describir los datos para cada tipo.

a) Construir perfiles de usuario en una tienda 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

MATRICES

1. (B) Verificar las siguientes propiedades de las matrices

   a) Dada una matriz X($N \times d$),$N > d$, de números reales las matrices $\mathrm{XX^T}$ y $\mathrm{X^TX}$ son simétricas.

   b) ¿Que representan los valores de traza($\mathrm{XX^T}$) y traza($\mathrm{X^TX}$)? (Ayuda: la traza de una matriz es la suma de los valores de su diagonal principal)

   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 $\mathrm{XX^T}$. Identifique alguna propiedad interesante de la SVD de matrices simétricas.

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

   e) ¿Establezca la conexión (∗∗) entre estos dos problemas?

$$\min_{\mathbf{w}}\{(\mathbf{y} - X\mathbf{w})^T(\mathbf{y} - X\mathbf{w})\}$$

$$\min_{\mathbf{w}}\{** ((\mathbf{y} - X\mathbf{w})(\mathbf{y} - X\mathbf{w})^T)\}$$

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

   g) Analizar el caso de matrices rectangulares. ¿Cómo sería la inversa de X?

2. (B) Sean $\mathbf{x}$ e $\mathbf{y}$ dos vectores de características. La covarianza de dos vectores 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)

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

   a) $E1 = \mathbf{1}\mathbf{1}^T\mathrm{X}$

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

   Track changes is on 22 ahora una matriz X cuyas columnas representan vectores de observaciones. La matriz de covarianzas asociada a la matriz X es el conjunto de covarianzas definidas por cada dos de sus vectores columnas. Defina la expresión matricial que muestra la matriz cov(X) como función de la matriz X (Ayuda: la traza de un matriz es la suma de los valores de su diagonal principal)

4. (B) Verifique matematicamente la validez de la expresión encontrada en el estudio de la función de perdida de Regresión.

$$\frac{d}{d\mathbf{w}}||(\mathbf{y} - X\mathbf{w}||^2 = 2X^T(X\mathbf{w} - \mathbf{y})$$

   Para ello calcule la derivada para un elemento genérico w_i y justifique la expresión matricial.

5. (B) Considere la matriz hat $\mathrm{H} = \mathrm{X}(\mathrm{X^TX})^{-1}\mathrm{X^T}$, donde X es una matriz $N \times (d+1)$, $N >> d$, y $\mathrm{X^TX}$ es invertible.

   a) Mostrar que H es simétrica

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

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

   d) Mostrar que traza(H) $= d + 1$, Ayuda: (considerar el uso de la SVD y que traza(AB)=traza(BA))

REGRESIÓN

1. (A) Consideremos que los datos están generados por una función f con ruido $y = \mathbf{w}^{*T}\mathbf{x} + \epsilon$, donde ϵ es un término de ruido con media cero y varianza σ^2, que está generado de forma independiente para cada muestra $(\mathbf{x}, y)$. Por tanto el error esperado del mejor ajuste posible a esta función es σ^2.

Supongamos una muestra de datos $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ donde el ruido en cada y_n se nota como ϵ_n y sea $\epsilon = [\epsilon_1, \epsilon_2, \ldots, \epsilon_N]^T$; asumimos que $\mathrm{X^T X}$ es invertible. Seguir los pasos que se muestran a continuación y mostrar que el error esperado (i.e. error esperado de entrenamiento) de regresión lineal respecto de $\mathcal{D}$ está dado por

$$\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(\mathbf{w}_{\text{lin}})] = \sigma^2 \left(1 - \frac{d+1}{N}\right)$$

a) Mostrar que la estimación de $\mathbf{y}$ está dada por $\hat{\mathbf{y}} = \mathrm{X}\mathbf{w}^* + \mathrm{H}\epsilon$

b) Mostrar que el vector de errores dentro de la muestra, $\hat{\mathbf{y}} - \mathbf{y}$, puede expresarse como el producto de una matriz por ϵ ¿Cual es la matriz?

c) Expresar $E_{\text{in}}(\mathbf{w}_{\text{lin}})$ en función de ϵ usando el apartado anterior y simplificar la expresión usando el ejercicio $5c$ de la sección de matrices.

d) Probar que $\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(\mathbf{w}_{\text{lin}})] = \sigma^2 \left(1 - \frac{d+1}{N}\right)$ usando el apartado anterior y la independencia de los errores, $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$. (Ayuda: Tener en cuenta la suma de los elementos de la diagonal de una matriz. Además el apartado $5d$ de la sección de matrices también es relevante)

e) Para analizar el error esperado fuera de la muestra, vamos a considerar un caso que es fácil de analizar. Consideremos un conjunto de test $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_1, y_1^{'}), \ldots, (\mathbf{x}_N, y_N^{'})\}$ que comparte los mismas entradas que $\mathcal{D}$ pero con términos de ruido de valor diferente. Notemos el ruido en $y_n^{'}$ como $\epsilon_n^{'}$ y sea $\epsilon^{'} = [\epsilon_1^{'}, \epsilon_2^{'}, \ldots, \epsilon_N^{'}]^T$. Definir $E_{\text{test}}(\mathbf{w}_{\text{lin}})$ como el error cuadrático medio sobre $\mathcal{D}_{\text{test}}$.

1) Probar que $\mathbb{E}_{\mathcal{D},\epsilon'}[E_{test}(\mathbf{w}_{lin})] = \sigma^2 \left(1 + \frac{d+1}{N}\right)$

   Este error de test especial, E_{test}, es un caso muy restrictivo del caso general de error-fuera-de-la-muestra.

2. (B) Consideremos las mismas condiciones generales del enunciado del ejercicio anterior. Considerar ahora $\sigma = 0{,}1$ y $d = 8$, ¿cual es el más pequeño tamaño muestral que resultará en un valor esperado de E_{in} mayor de 0,008?.

3. (M) 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.

4. (M) Suponiendo que $\mathrm{X}^T\mathrm{X}$ es invertible, mostrar que

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N}(\mathbf{w}^T\mathrm{X}^T\mathrm{X}\mathbf{w} - 2\mathbf{w}^T\mathrm{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y})$$

puede escribirse como

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\{(\mathbf{w} - \mathbf{w}_{lin})^T\mathrm{X}^T\mathrm{X}(\mathbf{w} - \mathbf{w}_{lin}) + \mathbf{y}^T(\mathrm{I} - \mathrm{H})\mathbf{y}\}$$

Extraer conclusiones sobre la contribución de cada término al error.

5. (A) Consideremos las condiciones del problema de regresión lineal establecidas en el ejercicio.1 de esta sección, donde los datos se generan a partir de una verdadera relación lineal más un ruido. El rudio se supone i.i.d. con media cero y varianza σ^2. Supongamos que la matriz de momentos de segundo orden $\Sigma = \mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^{\mathrm{T}}]$ es no singular. Seguir los pasos que se señalan para verificar que con alta probabilidad, el `error-fuera-de-la-muestra` en promedio vale

$$E_{\text{out}}(\mathbf{w}_{\text{lin}}) = \sigma^2\left(1 + \frac{d+1}{N} + o(\frac{1}{N})\right)$$

a) para un punto de test $\mathbf{x}$ mostrar que el error $y - g(\mathbf{x})$ es

$$\epsilon - \mathbf{x}^T(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^T\epsilon$$

donde ϵ es el valor del ruiso para el punto de test y $\boldsymbol{\epsilon}$ es el vector de valores de ruido en los datos.

b) Calcular la esperanza respecto del punto de test, i.e. $\mathbf{x}$ y ϵ, para obtener una expresión para E_{out}. Mostrar que

$$E_{\text{out}} = \sigma^2 + \text{traza}(\Sigma(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathbf{T}}\mathrm{X}^{\mathrm{T}}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}$$

(ayuda: traza(AB)=traza(BA) y esperanza y traza conmutan)

c) Que es $\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathbf{T}}]$?

d) Calcular la esperanza respecto a ϵ para ver que en promedio

$$E_{\text{out}} = \sigma^2 + \frac{\sigma^2}{\mathrm{N}}\text{traza}(\Sigma(\frac{1}{\mathrm{N}}\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1})$$

Note que $\frac{1}{\mathrm{N}}\mathrm{X}^{\mathrm{T}}\mathrm{X} = \frac{1}{\mathrm{N}}\sum_{\mathrm{n}=1}^{\mathrm{N}} \mathbf{x}_{\mathrm{n}}\mathbf{x}_{\mathrm{n}}^{\mathrm{T}}$ es una estimación de Σ con N muestras. Por tanto $\frac{1}{\mathrm{N}}\mathrm{X}^{\mathrm{T}}\mathrm{X} \approx \Sigma$. Si $\frac{1}{\mathrm{N}}\mathrm{X}^{\mathrm{T}}\mathrm{X} = \Sigma$ ¿cuanto vale E_{out} en promedio ?.

e) Mostrar que (después de calcular la esperanza respecto de todos los datos muestrales) con alta probabilidad

$$E_{\text{out}} = \sigma^2\left(1 + \frac{d+1}{N} + o(\frac{1}{N})\right)$$

(Ayuda: por la ley de los grandes números $\frac{1}{\mathrm{N}}\mathrm{X}^{\mathrm{T}}\mathrm{X}$ converge en probabilidad a Σ, y por continuidad de la inversa de Σ, $(\frac{1}{\mathrm{N}}\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}$ converge en probabilidad a Σ^{-1}.

f) (B) 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 estocástica binaria, $f = P(y|\mathbf{x})$, a partir de muestras etiquetadas con valores ± 1 y de funciones hipótesis que notamos por h. Escribir la verosimilitud de una muestra de tamaño 2. Comparar la expresiónes dadas por la maximización de la verosimilitud con la minimización de E_{in}.

MINIMIZACIÓN ITERATIVA: SGD

1. (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 $\mathbf{w} = -[1, 2, 3]^T$

2. La regla de adaptación de los pesos del Perceptron tiene la interesante propiedad de que mueve el vector de pesos en la dirección adecuada para clasificar $\mathbf{x}$ de froma correcta.
   a) Mostrar que $y(t)\mathbf{w}^T(t)\mathbf{x}(t) < 0$ (Suponer que $\mathbf{x}$ está mal clasificado por $\mathbf{w}$)
   b) Mostrar que $y(t)\mathbf{w}^T(t+1)\mathbf{x}(t) > y(t)\mathbf{w}^T(t)\mathbf{x}(t)$ (después de la adaptación)
   c) Basado en lo anterior, argumentar que el movimiento de $\mathbf{w}(t)$ hacia $\mathbf{w}(t+1)$ es un movimiento en la dirección correcta.

3. La regla de adaptación de los pesos del Perceptron tiene la interesante propiedad de alcanzar el separador óptimo en un número finitode pasos. Supongamos una muestra finita separable de N puntos etiquetados $\{+1, -1\}$. Por simplicidad supongamos $\mathbf{w}(0) = \mathbf{0}$. $\mathbf{w}^*$ representa la solución óptima. Probar los siguientes apartados:

   a) Sea $\rho = \text{mín}_{1 \leq n \leq N} \, y_n(\mathbf{w}^{*T}\mathbf{x}_n)$. Mostrar que $\rho > 0$

   b) Mostrar que $\mathbf{w}^T(t)\mathbf{w}^* \geq \mathbf{w}^T(t-1)\mathbf{w}^* + \rho$, y concluir que $\mathbf{w}^T(t)\mathbf{w}^* \geq t\rho$ (Ayuda: Usar inducción)

   c) Mostrar que $||\mathbf{w}(t)||^2 \leq ||\mathbf{w}(t-1)||^2 + ||\mathbf{x}(t-1)||^2$ (Ayuda: Usar $y(t-1) \cdot (\mathbf{w}^T(t-1)\mathbf{x}(t-1)) \leq 0$ porque $\mathbf{x}(t-1)$ fue mal clasificado por $\mathbf{w}(t-1)$

   d) Mostrar por inducción que $||\mathbf{w}(t)||^2 \leq tR^2$ donde $R = \text{máx}_{1 \leq n \leq N} \, ||\mathbf{x}||$

   e) Usando (b) y (d) mostrar que

$$\frac{\mathbf{w}^T(t)}{||\mathbf{w}(t)||}\mathbf{w}^* \geq \sqrt(t)\frac{\rho}{R}$$

   y por tanto probar que

$$t \leq \frac{R^2||\mathbf{w}^*||^2}{\rho^2}$$

   (Ayuda: $\mathbf{w}^T(t)\mathbf{w}^* < ||\mathbf{w}(t)|| \cdot ||\mathbf{w}^*||$)

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

   a) 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 $\nu = 1$.

   b) 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)

5. (M) Considerar $\mathbf{e}_n(\mathbf{w}) = \max(0, 1 - y_n\mathbf{w}^T\mathbf{x}_n)$.

   a) Mostrar que $\mathbf{e}_n(\mathbf{w})$ es continua y diferenciable excepto cuando $y_n = \mathbf{w}^T\mathbf{x}_n$.

   b) Mostrar que $\mathbf{e}_n(\mathbf{w})$ es una cota superior para $[\![\text{sign}(\mathbf{w}^T\mathbf{x}_n) \neq y_n]\!]$. Por tanto $\frac{1}{N}\sum_{n=1}^{N} \mathbf{e}_n(\mathbf{w})$ es una cota superior para $E_{\text{in}}(\mathbf{w})$, el error de entrenamiento de clasificación.

   c) Aplicar gradiente descendente estocástico (SGD) sobre $\frac{1}{N}\sum_{n=1}^{N} \mathbf{e}_n(\mathbf{w})$ (ignorando la singularidad de $y_n = \mathbf{w}^T\mathbf{x}_n$) y derivar un nuevo algoritmo perceptron (i.e. nueva regla de actualización de pesos)

6. (M) 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 $\mathbf{w}_{new} = \mathbf{w}_{old} + \eta(y_n - \mathbf{w}^T\mathbf{x}_n) \cdot \mathbf{x}_n$. Ahora vamos a interpretar ADALINE como un proceso de optimización. Considerar la función de error $E_n(\mathbf{w})) = (\max(0, 1 - y_n\mathbf{w}^T\mathbf{x}_n))^2$.
   a) Mostrar que $E_n(\mathbf{w})$ es continua y diferenciable. Escribir el gradiente $\nabla_{\mathbf{w}} E_n(\mathbf{w})$.
   b) Mostrar que $E_n(\mathbf{w})$ es una cota superior para $[\![\text{sign}(\mathbf{w}^T\mathbf{x}_n) \neq y_n]\!]$. Por tanto $\frac{1}{N}\sum_{n=1}^{N} E_n(\mathbf{w})$ es una cota superior para $E_{\text{in}}(\mathbf{w})$, el error de entrenamiento de clasificación.

REGRESION LOGISTICA

1. (B) Considerar la función
$$\tanh(s) = \frac{e^s - e^{-s}}{e^s + e^{-s}}$$
   a) ¿Como está relacionada esta función con la función logística o sigmoidal $\sigma(s)$?
   b) 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.
   c) Dibujar la función y compararla con la función $g(s) = -1$ para $s < 0$ y $g(s) = +1$ para $s \geq 0$.

2. Mostrar que la frontera de decisión del clasificador de regresión logística es siempre una recta. (Partir de $P(y = +1|x) = P(y = -1|x)$)

3. (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.

4. (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. Escribir la verosimilitud de una muestra de tamaño 2. Comparar la expresiónes dadas por la maximización de la verosimilitud con la minimización de E_{in}.
   a) Mostrar que la estimación de Máxima Verosimilitud 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)

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

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

Nota: dadas dos distribuciones de probabilidad $\{p, 1-p\}$ y $\{q, 1-q\}$ de variables aleatorias binarias, la entropía cruzada para estas distribuciones se define en teoría de la información por la expresión

$$p \log \frac{1}{q} + (1-p) \log \frac{1}{1-q}$$

El error de la muestra en el apartado.4*a* corresponde a una medida de error de entropía cruzada de los datos $(\mathbf{x}_n, y_n)$ con $p = [\![y_n = +1]\!]$ y $q = h(\mathbf{x}_n)$.

5. (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.

a) Definir la función de costo(aceptar) como el costo esperado si se acepta la persona. Definir de forma similar el costo(rechazo). Mostrar que

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

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

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

c) 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.

6. (B) Considerar la expresión de E_{in} dada en $4a$ de esta sección. Nuestro interés es derivar la expresión de la regla de adaptación de parámetros que genera el uso del método de optimización de Newton. La expresión general de la regla de adaptación de Newton es:

$$\theta^{(t+1)} = \theta^{(t)} - \mathrm{H}^{-1}\nabla_\theta(E_{\text{in}})$$

Para ello

   a) Calcular la expresión analítica en forma vectorial del gradiente $\nabla_\theta(E_{\text{in}})$

   b) Calcular la expresión analítica en forma vectorial para la matriz Hessiana H (recordar que la matriz Hessiana es la matriz de las derivadas segundas)

7. (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 $s = \mathbf{w}^T\mathbf{x}$.

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

   b) 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.

   c) 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.

8. (B) Este problema investiga como afecta la medida del error al proceso de aprendizaje de un valor/función. Suponga que dispone de N valores $y_1 \leq \cdots \leq y_n$ y desea estimar un valor representativo/resumen de los mismos.

   a) Si su algoritmo busca la hipótesis h que minimiza la suma de los cuadrados dentro de la muestra

$$E_{in}(h) = \sum_{n=1}^{N}(h - y_n)^2$$

   entonces mostrar que la estimacion será

$$h_{mean} = \frac{1}{N}\sum_{n=1}^{N} y_n$$

   b) Si su algoritmo busca la hipótesis h que minimiza la suma de los valores absolutos de la muestra,

$$E_{in}(h) = \sum_{n=1}^{N} |h - y_n|$$

entonces mostrar que la estimación será la mediana de la muestra, h_{med} (cualquier valor que deje el 50 % de valores muestrales a su derecha y el 50 % a su izquierda)

c) Suponga que y_N es modificado como $y_N + \epsilon$, donde $\epsilon \to \infty$. Obviamente el valor de y_N se convierte en un punto muy alejado de su valor original. ¿Como afecta esto a los estimadores anteriores?

OPTIMIZACIÓN CON RESTRICCIONES

1. (M) (Multiplicadores de Lagrange) Para resolver problemas de optimización con restricciones Lagrange propuso una técnica que permite convertir la optimización con restricciones en una optimización sin restricciones. Consideremos para el caso de 2 variables el siguiente problema de optimización con restricciones

$$\begin{gathered}\text{Optimizar}_{x,y} g(x,y) \\ \text{Sujeto a } f(x,y) = 0\end{gathered}$$

Es decir, queremos buscar los óptimos de la función g en un recinto del plano xy definido por los valores nulos de la función f. La solución es transformar este problema de optimización con restricciones en un problema de optimización sin restricciones y resolver este último derivando e igualando a cero. Para ello se construye una nueva función denominada Lagrangiana que se define como

$$\mathcal{L}(x,y,\lambda) = g(x,y) - \lambda f(x,y)$$

siendo λ una nueva variable denominada "multiplicador de Lagrange" (Nota: el signo $-$ delante de las restricciones también puede ser un signo positivo, ver ejercicios). Puede probarse que los valores de x, y que son la solución óptima de $\mathcal{L}$ coinciden con los valores del óptimo del problema inicial con restricciones. La condición necesaria para obtener una solución es que se verifique $\nabla_{x,y,\lambda}\mathcal{L}(x,y,\lambda) = \mathbf{0}$. En el caso de que exista más de una restricción en igualdad, cada una de ellas se añade a la Lagrangiana de la misma manera pero con un λ diferente.

$$\mathcal{L}(x,y,\lambda_1,\cdots,\lambda_n) = g(x,y) - \sum_{i=1}^{n} \lambda_i f_i(x,y)$$

La condición suficiente esta asociada al valor de la matriz Hessiana H de la función Lagragiana en los puntos estacionarios $\mathbf{x}_{est}$. Si $H(\mathbf{x}_{ets})$ es definida positiva (en dos dimensiones det(H)>0 y $H_{11} > 0$) tendriamos un mínimo condicionado, si $H(\mathbf{x}_{est})$ es definida negativa (en dos dimensiones det(H)>0 y $H_{11} < 0$) tendriamos un máximo condicionado. Si $H(\mathbf{x}_{est})$ es indefinida, es decir $det(H)) = 0$, entonces no se sabe nada. En algunos casos simples, p.e. con una única restricción, es posible caraterizar movimientos alrededor de los puntos estacionarios a partir de las ecuación $(\delta x, \delta y)^T \nabla f(\mathbf{x}_{est} = \mathbf{0})$ en los cuales si es posible evaluar el valor de la Hessiana y verificar si son máximos o mínimos condicionados.

Formular y resolver los siguientes problemas como problemas de optimización con restricciones:

a) Escribir dos formulaciones equivalentes de la Lagrangiana del problema: Minimizar $g(x, y)$ Sujeto a $f_1(x, y) = c_1$ y $f_2(x, y) = c_2$ que difieran en el signo de los multiplicadores.

b) Minimizar $g(x, y) = \mathrm{x}^2 + y^2$, sujeto a $x + y = 3$. Verificar que se obtiene un mínimo. Dibujar el problema.

c) Optimar $g(x, y) = xy$ sujeto a $x+y = 1$. ¿Qué tipo de óptimo aparece? Dibujar el espacio de soluciones.

d) Hallar los extremos de la función $g(x, y, z) = x+y+z$ condicionada a $x^2+y^2 = 1$ y $z = 1$.

e) Encontrar el punto (x_0, y_0) sobre la línea $ax + by + d = 0$ que está más cerca del punto (x_1, y_1).

f) La distancia entre dos curvas en el plano está dada por el mínimo de la expresión $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ donde (x_1, y_1) está sobre una de las curvas y (x_2, y_2) está sobre la otra. Calcular la distancia entre la línea $x + y = 4$ y la elipse $x^2 + 2y^2 = 1$.

En el caso de que algunas condiciones de restricción estén definidas en términos de desigualdad ($<$, $\leq$, etc), entonces las condiciones para que la solución del problema sin restricción coincida con la solución del problema con restricciones cambian respecto del caso Lagrangiano antes expuesto. Dichas condiciones se denominan las condiciones de Karush-Kuhn-Tucker y se estudiarán más adelante.

2. Explicar la relación que existe entre la constante C y el multiplicador de Lagrange β en la solución del problema: $\mathrm{Min}_{\mathbf{w}}||\mathbf{y} - Z\mathbf{w}||^2$ sujeto a $\mathbf{w}^T\mathbf{w} = C$. Justificar que el conjunto de soluciones que se generan variando β es el mismo que el del problema no restringuido $\mathrm{Min}_{\mathbf{w}}||\mathbf{y} - Z\mathbf{w}||^2 + \lambda_C||\mathbf{w}^T\mathbf{w}||$

3. En este ejercio vamos definir el ajuste de un modelo lineal de clasificación (PLA) como un problema de programación lineal. La programación líneal es una técnica de optimización que busca el óptimo (máximo o mínimo) de una función lineal de un vector de variables en una región de valores definida por un sistema de ecuaciones en desigualdad. En concreto se formula como,

$$\mathrm{Min}_{\mathbf{z}}\mathbf{c}^{\mathrm{T}}\mathbf{z}$$
$$\text{Sujeto a } \mathrm{A}\mathbf{z} \leq \mathbf{b}$$

donde ($\mathbf{z}$) es el vector de variables, $\mathbf{c}$ y $\mathbf{b}$ son vectores de números y A es una matriz de números. Para establecer nuestro resultado vamos a seguir los siguientes pasos:

a) (B) Para un conjunto de datos linealmente separable mostrar que para algún $\mathbf{w}$ se debe de verificar la condición $\mathrm{y_n}\mathbf{w}^{\mathrm{T}}\mathbf{x}_{\mathrm{n}} > 0$ para todo $(\mathbf{x}_{\mathrm{n}}, \mathrm{y_n})$ del conjunto.

b) (B) Formular el problema de programación lineal que resuelva el problema de la búsqueda del hiperplano separador. Es decir, identifique quienes son A, $\mathbf{z}$, $\mathbf{b}$ y $\mathbf{c}$ para este caso. (Ayuda: observe que solo necesita que se verifiquen las restricciones)

c) (M) Si los datos no son separables, la condición del primer apartado no se verifica para todos los puntos. Por tanto introducimos la variable holgura $\xi_n > 0$ para capturar la cantidad de violación que permitimos para el ejemplo $\mathbf{x}_n$. Por tanto, para $n = 1, \ldots, N$, tenemos

$$\begin{aligned} y_n(\mathbf{w}^T\mathbf{x}_n &\geq 1 - \xi_n \\ \xi_n &\geq 0 \end{aligned}$$

Logicamente nos gustaría minimizar la cantidad de violación. Una aproximación intuitiva es minimizar $\sum_{n=1}^{N} \xi_n$, i.e deseamos encontrar $\mathbf{w}$ que resuelve

$$\begin{aligned} \min_{\mathbf{w},\xi_n} \quad & \sum_{n=1}^{N} \xi_n \\ \text{sujeto a} \quad & y_n(\mathbf{w}^T\mathbf{x}_n) \geq 1 - \xi_n \\ & \xi_n \geq 0 \end{aligned}$$

donde las desigualdades deben de verificarse para $n = 1, \ldots, N$. Formular este problema como un problema de programción lineal

SOBREAJUSTE

1. (M) Mostrar que en el caso general de funciones con ruido se verifica que $\mathbb{E}_{\mathcal{D}}[E_{\text{out}}] = \sigma^2 + \texttt{bias} + \texttt{var}$ (ver transparencias de clase)

2. (B) El ruido determinista depende de $\mathcal{H}$, ya que algunos modelos aproximan mejor f que otros.

   a) Suponer que $\mathcal{H}$ es fija y que incrementamos la complejidad de f.

   b) Suponer que f es fija y decrementamos la complejidad de $\mathcal{H}$

   Contestar para ambos escenarios: ¿En general subirá o bajará el ruido determinista? ¿La tendencia a sobrejaustar será mayor o menor? (Ayuda: analizar los errores que influencian el sobreajuste)

3. (B) Si $\lambda < 0$ en el error aumentado $E_{\text{aug}}(\mathbf{w}) = E_{\text{in}}(\mathbf{w}) + \lambda\mathbf{w}^{\text{T}}\mathbf{w}$, ¿a que orden de restricción "suave" corresponde? (Ayuda: analizar que tipos de pesos favorece $\lambda < 0$).

REGULARIZACIÓN Y VALIDACIÓN DE MODELOS

1. (B) La técnica de regularización de Tikhonov generaliza de forma natural a weight-decay al usar funciones lineales de los pesos $\Gamma\mathbf{w}$ para definir regularizadores,

$$\mathbf{w}^{\text{T}}\Gamma^{\text{T}}\Gamma\mathbf{w} \leq C$$

   (La matriz Γ se denomina regularizador de Tikhonov)

   a) Calcular Γ cuando $\sum_{q=0}^{Q} w_q^2 \leq C$

   b) Calcular Γ cuando $(\sum_{q=0}^{Q} w_q)^2 \leq C$

2. (M) Fijar g^- (aprendida a partir de los datos de aprendizaje $\mathcal{D}_{\text{train}}$), y definir $\sigma^2_{\text{val}} \stackrel{\text{def}}{=} \text{Var}_{\mathcal{D}_{\text{val}}}[E_{\text{val}}(g^-)]$. Veamos como σ^2_{val} depende de K. Sea

$$\sigma^2(g^-) = \text{Var}_{\mathbf{x}}[\mathbf{e}(g^-(\mathbf{x}), y)]$$

la varianza del `error-fuera-de-la-muestra` de g^-.

   a) Mostrar que $\sigma^2_{\text{val}} = \frac{1}{K}\sigma^2(g^-)$.

   b) En un problema de clasificación donde $\mathbf{e}(g^-(\mathbf{x}), y) = [\![g^-(\mathbf{x}) \neq y]\!]$, expresar σ^2_{val} en términos de $\mathbb{P}[g^-(\mathbf{x}) \neq y]$.

   c) Mostrar que para cualquier g^- en un problema de clasificación, $\sigma^2_{\text{val}} = \frac{1}{4K}$.

3. (M) Suponga que tenemos M modelos $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_M$ que entrenamos con el conjunto de entrenamiento $\mathcal{D}_{\text{train}}$ para obtener una hipótesis g^-_m para cada modelo. Evaluamos cada una de estas hipótesis con el conjunto de evaluación para obterne sus errores de evaluación $E_1, E_2, \ldots, E_M$, donde $E_m = E_{\text{val}}(g^-_m)]$, para $m = 1, 2, \ldots, M$. ¿Es E_m un estimador insesgado del `error-fuera-de-la-muestra` $E_{\text{out}}(g^-_m)$?

4. (M) Ver la figura 4.12 del pdf(Learning from data(overfitting)).

   a) $\mathbb{E}[E_{\text{out}}(g^-_{m^*})]$ es inicialmente decreciente. ¿Como puede ser si $\mathbb{E}[E_{\text{out}}(g^-_m)]$ es creciente en K para cada m?

   b) $\mathbb{E}[E_{\text{out}}(g_{m^*})]$ es inicialmente decreciente y entonces comienza a crecer. ¿Cuáles son las posibles razones para ello?

   c) Cuando $K = 1$, $\mathbb{E}[E_{\text{out}}(g^-_{m^*})] < \mathbb{E}[E_{\text{out}}(g_{m^*})]$. ¿Cómo puede ser si las curvas de aprendizaje para ambos modelos son decrecientes ?

5. (A) Considere el caso de minimización del error aumentado con $\Gamma = \text{I}$ y $\lambda > 0$:

   a) Mostrar que $||\mathbf{w}_{\text{reg}}|| < ||\mathbf{w}_{\text{lin}}||$, justificando el término **‘‘`weight decay`’’** (Ayuda: comenzar suponiendo $||\mathbf{w}_{\text{reg}}|| > ||\mathbf{w}_{\text{lin}}||$ y derivar una contradicción).

   b) Suponer que E_{in} es diferenciable y usar gradiente descendente para minimizar E_{aug}:

$$\mathbf{w}(t+1) \leftarrow \mathbf{w}(t) - \eta \nabla E_{\text{aug}}(\mathbf{w}(t))$$

   Ver que esta regla de adaptación es la misma que

$$\mathbf{w}(t+1) \leftarrow (1 - 2\lambda\eta)\mathbf{w}(t) - \eta \nabla E_{\text{in}}(\mathbf{w}(t))$$

   Nota: esta es la razón del nombre **‘‘`weight decay`’’**: $\mathbf{w}(t)$ decae antes de ser adaptado por el gradiente de E_{in}

6. (M) Para modelos lineales y el regularizador general de Tikhonov Γ con término de penalización $\frac{\lambda}{\text{N}}\mathbf{w}^{\text{T}}\Gamma^{\text{T}}\Gamma\mathbf{w}$ en el error aumentado, mostrar que

$$\mathbf{w}_{\text{reg}} = (\text{Z}^{\text{T}}\text{Z} + \lambda\Gamma^{\text{T}}\Gamma)^{-1}\text{Z}^{\text{T}}\text{y}$$

   donde Z es la matriz de características.

a) Mostrar que las predicciones dentro de la muestra son

$$\hat{\mathbf{y}} = \mathrm{H}(\lambda)\mathbf{y}$$

donde $\mathrm{H}(\lambda) = \mathrm{Z}(\mathrm{Z}^{\mathrm{T}}\mathrm{Z} + \lambda\Gamma^{\mathrm{T}}\Gamma)^{-1}\mathrm{Z}^{\mathrm{T}}$

b) Simplificar esta expresión en el caso $\Gamma = \mathrm{Z}$ y obtener $\mathbf{w}_{\mathrm{reg}}$ en términos de $\mathbf{w}_{\mathrm{lin}}$. Esto se denomina uniforme "`weight decay`".

Factibilidad del Aprendizaje

1. (M) Consideremos una función objetivo booleana sobre un espacio 3D, $\mathcal{X} = \{0,1\}^3$ (supondremos que los valores de salida de la función son $\mathcal{Y} = \{0,1\}$). Disponemos de un conjunto de datos $\mathcal{D}$ compuesto por 5 ejemplos representados en la tabla adjunta, donde $y_n = f(x_n)$ para $n = 1,2,3,4,5$

x_n			y_n
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1

Notemos que en este caso Booleano simple podemos enumerar el espacio de entrada completo (solo hay 8 posibles vectores de entrada) y podemos enumerar todos las posibles funciones de salida (solo hay $2^8 = 256$ posible funciones para una función con 3 argumentos binarios y salida binaria). Analizamos el problema de como aprender f. Ya que f es deconocida excepto dentro de $\mathcal{D}$, cualquier función que coincida con $\mathcal{D}$ potencialmente podría ser f. Solo hay tres puntos de $\mathcal{X}$ fuera de $\mathcal{D}$, por tanto solo hay 8 posibles funciones distintas. Los argumentos de la función que no están en $\mathcal{D}$ son 101,110 y 111. El objetivo es determinar la hipótesis que más coincide con la función objetivo (i.e. los valores más probables para los tres argumentos que faltan). Para ello necesitamos cuantificar el error de cada una de las posibles propuestas. Como no conocemos f mediremos el error de cada hipotesis contando el número de coincidencias con cada una de las 8 posibles hipótesis.

Criterio-de-Bondad = (# de funciones que coinciden con la hipótesis en tres valores)$\times$ 3 + (# de funciones que coinciden con la hipótesis en dos valoress)$\times$ 2 + (# de funciones que coinciden con la hipótesis en un valor)$\times$ 1

Supongamos que usamos las siguientes clases de funciones:

a) $\mathcal{H}$ solo tiene dos hipótesis, una que devuelve siempre 0 y otra que devuelve siempre 1. El algoritmo de aprendizaje elige siempre aquella que más se ajusta a los datos

b) Igual que antes pero ahora suponemos que el algoritmo de aprendizaje elige la hipótesis que menos se ajusta a los datos.

c) $\mathcal{H} = \{\text{XOR}\}$, donde $\text{XOR}(x) = 0$ si el número de unos en x es impar, y $\text{XOR}(x) = 1$ si el número de unos en x es par.

d) $\mathcal{H}$ contiene todas las posibles hipótesis de 8 argumentos y el algoritmo de aprendizaje elige aquellas que mas coinciden con los datos de entrenamiento ($\mathcal{D}$) pero que menos coinciden con XOR.

¿Cuál de las hipotesis g es la que más coincide con las posibles funciones objetivos según el anterior criterio-de-bondad ?

2. (M) Supongamos que tenemos un conjunto de datos $\mathcal{D}$ de 25 ejemplos extraidos de una funcion desconocida $f : \mathcal{X} \to \mathcal{Y}$, donde $\mathcal{X} = \mathbb{R}$ e $\mathcal{Y} = \{-1, +1\}$. Para aprender f usamos un conjunto simple de hipótesis $\mathcal{H} = \{h_1, h_2\}$ donde h_1 es la función constante igual a $+1$ y h_2 la función constante igual a -1.

   Consideramos dos algoritmos de aprendizaje, S(smart) y C(crazy). S elige la hipótesis que mejor ajusta los datos y C elige deliberadamente la otra hipótesis. Vamos a estudiar como se comportan estos dos algoritmos fuera de la muestra desde un punto de vista determinístico y probabilístico. Suponga en el caso probabilístico que hay una distribución de probabilidad sobre $\mathcal{X}$, y sea $P[f(x) = +1] = p$

   a) ¿Puede S producir una hipótesis que garantice mejor comportamiento que la aleatoria sobre cualquier punto fuera de la muestra? (Ayuda: evaluar en cada caso que información me proporciona la hipótesis seleccionada sobre lo que ocurre fuera de la muestra)

   b) Asumir desde ahora que todos los ejemplos en $\mathcal{D}$ tienen $y_n = +1$. ¿Es posible que la hipotesis que produce C sea mejor que la hipótesis que produce S ? (Ayuda: No olvide que conoce lo que hace cada algoritmo)

   c) Si $p = 0{,}9$ ¿Cual es la probabilidad de que S produzca una hipótesis mejor que C?

   d) ¿Existe un valor de p para el cual es más probable que C produzca una hipótesis mejor que S?

Composición de Muestras

1. (B) En un recipiente con bolas blancas y negras se conoce que la probabilidad de bola blanca es $\mu = 0{,}9$ ¿ Cual es la probabilidad de que en una muestra de 10 bolas la frecuencia relativa de bolas blancas sea $\nu \leq 0{,}7$ (Ayuda: distribución binomial)

2. (B)Si $\mu = 0{,}9$ usar la desigualdad de Hoeffding para acotar la probabilidad de que una muestra de 10 bolas tenga $\nu \leq 0{,}1$ y comparar la contestación con el ejercico anterior.

3. (B) Supongamos 2 bolsas opacas cada una conteniendo 2 bolas. Una bolsa tiene 2 bolas negras y la otra tiene una blanca y una negra. Suponga que elige una bolsa al azar y entonces elige una bola de dicha bolsa también al azar. Cuando mira la bola elegida ve que es negra. Entonces saca la segunda bola de la bolsa. ¿Cual es la

probabilidad de que la segunda bola también sea negra? (Ayuda: usar Teorema de Bayes)

4. (M) Considerar una muestra como la extracción independiente de 10 bolas de un recipiente con bolas rojas y verdes. La probabilidad de bola rojas es μ. Para $\mu = 0{,}005$, $\mu = 0{,}5$ y $\mu = 0{,}8$, calcular la probabilidad de los siguientes sucesos:
   a) Extraemos una única muestra. Calcular la probabilida de que $\nu = 0$
   b) Extraemos 1.000 muestras independientes. Calcular la probabilidad de que al menos una de las muestras tenga $\nu = 0$
   c) repetir (b) para 1.000.000 de muestras independientes

   ¿Que conclusiones puede extraer que son útiles para el problema del aprendizaje?

Desigualdad de Hoeffding

1. (B) Identifique las hipótesis mínimas para que se cumpla la desigualdad de Hoeffding.

2. (B) ¿Por qué la desigualdad de Hoeffding no es aplicable de forma directa cuando el número de hipótesis es mayor de 1?

3. (M) Considere la cota alcanzada para la probabilidad del conjunto de muestras de error de la hipótesis solución g de un problema de aprendizaje, a partir de la desigualdad de Hoeffding, sobre una clase finita de hipótesis

$$P(\mathcal{D} : |E_{\text{out}}(g) - E_{\text{in}}(g)| > \epsilon) < 2|\mathcal{H}|e^{-2\epsilon^2 N}$$

   a) ¿Cuál es el algoritmo de aprendizaje que se usa para elegir g?
   b) Si elegimos g de forma aleatoria ¿seguiría verificando la desigualdad?
   c) ¿Depende g del algoritmo usado?.
   d) Es una cota ajustada o una cota laxa? (Ayuda: Razonar sobre las muestras que producen el suceso de error con cada función de $\mathcal{H}$.)

4. (B) Alguien se acerca a Ud con un problema de aprendizaje y le dice que la función objetivo que busca es *completamente* deconocida, pero le dice que tiene 4.000 datos de la misma. Esa persona está dispuesta a pagarle por resolver su problema aproximando la función desconocida f con una función g. ¿Que es lo mejor que le puede prometer de entre lo siguiente?
   a) Después de aprender le proporcionará una g, con garantía de que aproxima bien a f fuera de la muestra.
   b) Después de aprender le proporcionará un g, y con alta probabilidad la g que le da aproximará a f bien fuera de la muestra.
   c) Una de dos cosas sucederá:
      - Obtendrá una hipótesis g

- Declarará que ha fallado

Si devuelve una hipótesis g, entonces con alta probabilidad la g que ha encontrado aproximará f bien fuera de la muestra.

5. (M) Una muestra de caras y cruces es el resultado de lanzar una moneda un número de veces de forma independiente. Suponga que tenemos varias monedas que generan diferentes muestras de forma independiente. Para una moneda dada, sea μ la probabilidad de cara (probabilidad de error). La probabilidad de obtener k caras en N lanzamientos de esta moneda es dada por la distribución binomial:

$$P[k|N,\mu] = \binom{N}{k}\mu^k(1-\mu)^{1-k}$$

recordar que el error de entrenamiento ν es $\frac{k}{N}$.

a) Suponga que realizamos muestras de tamaño $N = 10$. Si todas las monedas tienen $\mu = 0{,}05$ calcular la probabilidad de que al menos una muestra tenga $\nu = 0$ en el caso de lanzar 1 moneda, lanzar 1.000 monedas y lanzar 1.000.000 monedas. Repetir para el caso de $\mu = 0{,}8$. ¿Que conclusiones saca que sean útiles para el problema del aprendizaje?

b) Para el caso de $N = 6$ y $N = 2$ con $\mu = 0{,}5$ en ambos casos, dibujar la probabilidad

$$P[\max_i |\nu_i - \mu_i| > \epsilon]$$

para ϵ en el rango [0,1] (el máximo es sobre las monedas). En la misma gráfica mostrar que la cota que se obtendría usando la desigualdad de Hoeffding. Recordar que para una única moneda la desigualdad es

$$P[|\nu - \mu| > \epsilon] \leq 2e^{2N\epsilon^2}$$

(Ayuda: usar $P(A \cup B) = P(A) + P(B) - P(A)P(B)$ en el máximo)

No-Free-Lunch y la Regla de Bayes

1. (B)¿Qué aporta la regla de Bayes al problema de clasificación?
2. (B)¿Por qué no usar siempre la regla de Bayes? Identifique pros y contras.
3. (B) Identifique la consecuencia más relevante sobre la teoría del aprendizaje del "Teorema de No-Free-Lunch"

Función de crecimiento y Punto de Ruptura

1. (B) Suponga que tenemos un modelo de aprendizaje simple cuya función de crecimiento es $m_{\mathcal{H}}(N) = N + 1$, por tanto $d_{VC} = 1$. Suponga que dispone de 100 muestras. Usar la cota de generalización para dar una cota de E_{out} con confianza del 90 %. Repetir para N=10.000

2. (M) Si $m_{\mathcal{H}}(k) < 2^k$ para algún k, entonces para todo N

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

Verificar esta cota en los siguientes casos:

- Flechas con signo: $\mathcal{H}$ consiste en todas las funciones de la forma $h(x) = \text{sign}(x - a)$.
- Intervalos positivos. $\mathcal{H}$ consiste en todas las funciones de una dimensión que toman valor $+1$ dentro de un intervalo arbitrario y -1 fuera de él.
- Funciones convexas: $\mathcal{H}$ consiste en todas las funciones de dos dimensiones que toman valor $+1$ dentro de un conjunto convexo arbitrario y -1 fuera de él.

3. (B) Dibujar las cotas para $m_{\mathcal{H}}(N)$ dadas por las expresiones

$$N^{d_{VC}} + 1 \quad \text{y} \quad \left(\frac{eN}{d_{VC}}\right)^{d_{VC}}$$

para $d_{VC} = 2$ y $d_{VC} = 5$ ¿Cuando prefiere una cota sobre la otra?

4. (B) Cuales de las siguientes expresiones son posibles funciones de crecimiento $m_{\mathcal{H}}(N)$ para algún conjunto de hipótesis

$$1 + N \ ; \ 1 + N + \frac{N(N-1)}{2} \ ; \ 2^N \ ; \ 2^{\lfloor \sqrt{N} \rfloor} \ ; \ 2^{\lfloor N/2 \rfloor} \ ; \ 1 + N + \frac{N(N-1)(N-2)}{6}$$

5. (B) En clase se mostró que el punto de ruptura más pequeño, para el modelo perceptron en el caso 2D es 4 puntos ¿Cual es el punto de ruptura más pequeño para el caso 3D (i.e., en lugar de rectas consideramos planos.)

6. (M) Consideremos el problema de aprendizaje definido por la clase "2-intervalos"donde $h \to \{-1, +1\}$, y $h(x) = +1$ si el punto está dentro de cualquiera de dos intervalos arbitrariamente elegidos y -1 en otro caso. ¿Cual es el punto de ruptura más pequeño para este conjunto de hipótesis?

7. Mostrar que $m_{\mathcal{H}}(2N) < m_{\mathcal{H}}(N)^2$

Error de Generalización

1. (B) Supongamos un conjunto de datos con 600 ejemplos. Para verificar adecuadamente la hipótesis final apartamos una muestra aleatoria de 200 ejemplos que nunca será usada en la fase de entrenamiento (conjunto de test). Suponga que usamos un modelo con 1.000 hipótesis y seleccionamos la hipótesis final a partir de las 400 ejemplos de entrenamiento. Nuestro interés es estimar $E_{out}(g)$. Tenemos acceso a dos estimaciones $E_{in}(g)$, el error de g con los 400 datos de entrenamiento y $E_{test}(g)$ el error de g con los 200 datos de test.

a) Usando un 5 % de error de tolerancia ($\delta = 0{,}05$) ¿que estimación tiene mayor error de generalización?

b) ¿Existe alguna razón por la que no deberíamos reservar tantos ejemplos para el test?

2. (M) Para funciones objetivo binarias, mostrar que $\mathbb{P}[h(x) \neq f(x)]$ puede escribirse como un valor esperado de una medida de error cuadrático medio en el siguiente caso:

   a) Cuando usamos $\{0, 1\}$ como los valores de una función binaria.

   b) Cuando usamos $\{-1, +1\}$ como los valores de una función binaria

   (Ayuda: La diferencia es solo una escala)

3. (B) Mostrar que si $\mathcal{H}$ es cerrada bajo combinaciones lineales (cualquier combinación lineal de funciones de $\mathcal{H}$ es una función de $\mathcal{H}$), entonces la función promedio $\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}] \equiv \overline{g} \in \mathcal{H}$.

   a) Dar un modelo para el cual la función promedio $\overline{g}$ no este en el conjunto de hipótesis (Ayuda: pensar en algo muy simple)

   b) En clasificación binaria, ¿Es esperable que $\overline{g}$ sea una función binaria?

4. (B) Usando la ecuación

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$$

   a) Para M=1, cuantos ejemplos necesitaremos para hacer $\epsilon \leq 0{,}05$

   b) Para M=100, cuantos ejemplos necesitaremos para hacer $\epsilon \leq 0{,}05$

   c) Para M=10.000, cuantos ejemplos necesitaremos para hacer $\epsilon \leq 0{,}05$

5. (B) Suponga $m_{\mathcal{H}}(N) = N + 1$, por tanto $d_{VC} = 1$ y que tenemos 100 ejemplos. Usar la cota de generalización para dar una cota para E_{out} con confianza del 90 %. Repetir para n=10.000.

6. Para un conjunto $\mathcal{H}$ con $d_{VC} = 10$, ¿que tamaño muestral se necesita (según la cota de generalización) para tener un 95 % de confianza de que el error de generalización es como mucho 0.05?

7. La desigualdad de Hoeffding modificada nos da una forma de caracterizar el error de generalización con una cota probabilística

$$\mathbb{P}[|E_{out}(g) - E_{in}(g)| > \epsilon] \leq 2Me^{-2N^2\epsilon}$$

   para cualquier $\epsilon > 0$. Si fijamos $\epsilon = 0{,}05$ y queremos que la cota probabilistica $2Me^{-2N^2\epsilon}$ sea como máximo 0.03 ¿cual será el valor más pequeño de N que verifique estas condiciones si $M = 1$?. Repetir para $M = 10$ y para $M = 100$

8. (B) Existen múltiples cotas sobre el error de generalización ϵ, todas verificandose con probabilidad $1-\delta$. Fijar $d_{VC}=50$ y $\delta=0{,}05$ y dibujar estas cotas como función de N. ¿Que cota es la más pequeña para N suficientemente grande, digamos N=10.000?
   a) VC: $\epsilon \leq \sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}}$
   b) Rademacher (Penalizada): $\epsilon \leq \sqrt{\frac{2\ln(2Nm_{\mathcal{H}}(N))}{N}}+\sqrt{\frac{2}{N}\ln\frac{1}{\delta}}+\frac{1}{N}$
   c) Parrondo and Van den Broek: $\epsilon \leq \sqrt{\frac{1}{N}(2\epsilon+\ln\frac{6m_{\mathcal{H}}(2N)}{\delta})}$
   d) Devroye: $\epsilon \leq \sqrt{\frac{1}{2N}(4\epsilon(1+\epsilon)+\ln\frac{4m_{\mathcal{H}}(N^2)}{\delta})}$

   Nota: Observar que algunas cotas están definidas de forma implicita.

9. (B) Con los mismos valores de $d_{VC}=50$ y $\delta=0{,}05$ del ejercicio anterior, pero para N pequeño digamos $N=5$, ¿Que cota sería la más pequeña?

Dimensión-VC

1. (M) Sea $\mathcal{H}=\{h_1,h_2,\cdots,h_M\}$ para algún número finito M. Probar que $d_{VC}(\mathcal{H}_k)\leq \log_2 M$

2. (A) Supongamos que $q\geq 1$ es un entero y sea $m_{\mathcal{H}}(1)=2$. ¿Cual es la dimensión-VC de un conjunto de hipótesis cuya función de crecimiento satisface: $m_{\mathcal{H}}(N+1)=2m_{\mathcal{H}}(N)-\binom{N}{q}$ ($\binom{M}{m}=0$ para $m>M$)

3. Sea $\mathcal{H}_1$,$\mathcal{H}_2$,...,$\mathcal{H}_K$, con dimensión-VC finita y positiva d_{vc}. Sea $\mathcal{H}=\mathcal{H}_1\cup\mathcal{H}_2\cup\cdots\cup\mathcal{H}_K$ la unión de estos modelos.
   - Mostrar $d_{vc}(\mathcal{H})<K(d_{vc}+1)$
   - Suponer que l satisface $2^l>2Kl^{d_{vc}}$. Mostrar que $d_{vc}(\mathcal{H})\leq l$

4. (M) Para el conjunto de hipótesis $\mathcal{H}_1$,$\mathcal{H}_2$,...,$\mathcal{H}_K$, con dimensión-VC finita y positiva $d_{VC}(\mathcal{H}_k)$, algunas de las siguientes cotas son correctas y algunas no. De entre las correctas, cual es la cota más ajustada (de menor rango de valores) para la dimensión-VC de las intersección de los conjuntos: $d_{VC}(\cap_{k=1}^K\mathcal{H}_k)$ (la dimensión-VC del conjunto vacio o de un "singleton"se toma como cero).
   - $0\leq d_{VC}(\cap_{k=1}^K\mathcal{H}_k)\leq\sum_{k=1}^K d_{VC}(\mathcal{H}_k)$
   - $0\leq d_{VC}(\cap_{k=1}^K\mathcal{H}_k)\leq \min_{1\leq k\leq K} d_{VC}(\mathcal{H}_k)$
   - $0\leq d_{VC}(\cap_{k=1}^K\mathcal{H}_k)\leq \max_{1\leq k\leq K} d_{VC}(\mathcal{H}_k)$
   - $\min_{1\leq k\leq K} d_{VC}(\mathcal{H}_k)\leq d_{VC}(\cap_{k=1}^K\mathcal{H}_k)\leq \max_{1\leq k\leq K} d_{VC}(\mathcal{H}_k)$
   - $\min_{1\leq k\leq K} d_{VC}(\mathcal{H}_k)\leq d_{VC}(\cap_{k=1}^K\mathcal{H}_k)\leq \sum_{k=1}^K d_{VC}(\mathcal{H}_k)$

5. (M) Bajo las mismas condiciones del ejercicio anterior pero considerando ahora la unión de los conjuntos, $d_{VC}(\cup_{k=1}^{K}\mathcal{H}_k)$
   - $0 \leq d_{VC}(\cup_{k=1}^{K}\mathcal{H}_k) \leq \sum_{k=1}^{K} d_{VC}(\mathcal{H}_k)$
   - $0 \leq d_{VC}(\cup_{k=1}^{K}\mathcal{H}_k) \leq K-1+\sum_{k=1}^{K} d_{VC}(\mathcal{H}_k)$
   - $\min\limits_{1\leq k\leq K} d_{VC}(\mathcal{H}_k) \leq d_{VC}(\cup_{k=1}^{K}\mathcal{H}_k) \leq \sum_{k=1}^{K} d_{VC}(\mathcal{H}_k)$
   - $\max\limits_{1\leq k\leq K} d_{VC}(\mathcal{H}_k) \leq d_{VC}(\cup_{k=1}^{K}\mathcal{H}_k) \leq \sum_{k=1}^{K} d_{VC}(\mathcal{H}_k)$
   - $\max\limits_{1\leq k\leq K} d_{VC}(\mathcal{H}_k) \leq d_{VC}(\cup_{k=1}^{K}\mathcal{H}_k) \leq K-1+\sum_{k=1}^{K} d_{VC}(\mathcal{H}_k)$
6. Suponga un conjunto de datos de 100 muestras de los cúales separa el 25 % para validación. Ud. entrena 100 modelos con VC-dimensión 10 y elige aquel con menor error de validación 0,25. De una cota del error fuera-de-la-muestra para la funciónn seleccionada. Suponga a continuación que entrena todos los modelos con todos los datos y elige aquel con menor error de ajuste que es 0.15. Dar una cota para el error fuera de la muestra en este caso.

Neural Networks

1. (B) Las funciones booleanas AND y OR de 2 entradas, pueden generalizarse a M entradas $\text{OR}(x_1, x_2, \ldots, x_M) = +1$ si alguna de las M entradas es igual a $+1$ y $\text{AND}(x_1, x_2, \ldots, x_M) = +1$ igual si todas las entradas son igual a $+1$
   - Establecer los grafos de representación de $\text{OR}(x_1, x_2, \ldots, x_M)$ y $\text{AND}(x_1, x_2, \ldots, x_M)$.
   - Dar el grafo de representación del perceptron $h(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x}))$
   - Dar el grafo de representación de $\text{OR}(x_1, \overline{x_2}, x_3)$
2. (B) Sea $f = h_1\overline{h_2} + \overline{h_1}h_2$ una función booleana a partir de las valores booleanos dados por los perceptrones h_1 y h_2. Usar los grafos de las funciones para mostrar que la función f se puede expresar como $f(x) = \text{sign}(\text{sign}(h_1(x) - h_2(x) - \frac{3}{2}) - \text{sign}(h_1(x) - h_2(x) + \frac{3}{2}) + \frac{3}{2})$ donde $h_i(x) = \text{sign}(\mathbf{w}_i^T\mathbf{x})$, $i = 1, 2$.
3. (M) Considerar la función
$$f = \overline{h_1}h_2h_3 + h_1\overline{h_2}h_3 + h_1h_2\overline{h_3}$$
composición de tres perceptrones, h_1, h_2, h_3. Construir el grafo del modelo MLP asociado, así como la expresión algebraica asociada.
4. (M) Dado $\mathbf{w}_1$ y $\epsilon > 0$, encontrar $\mathbf{w}_2$ tal que $|\text{sign}(\mathbf{w}_1^T\mathbf{x}) - \tanh(\mathbf{w}_2^T\mathbf{w})| \leq \epsilon$ para $\mathbf{x} \in \mathcal{D}$. (Ayuda: Para $\mathbf{x}$ grande $\text{sign}(\mathbf{x}) \approx \tanh(\alpha\mathbf{x})$).
5. (M) Sea V y Q el número de nodos y pesos en una red neuronal,
$$V = \sum_{l=0}^{L} d^{(l)}, \quad Q = \sum_{l=1}^{L} d^{(l)}(d^{(l+1)} + 1)$$

En términos de V y Q ¿cuantos operaciones se realizan en un pase hacia adelante (sumas, multiplicaciones y evaluaciones de θ)? (Ayuda: analizar la complejidad en términos de V y Q)

6. Considere las siguientes matrices de pesos

$$W^{(1)} = \begin{bmatrix} 0{,}1 & 0{,}2 \\ 0{,}3 & 0{,}4 \end{bmatrix}; W^{(2)} = \begin{bmatrix} 0{,}2 \\ 1 \\ -3 \end{bmatrix}; W^{(3)} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

como las correspondientes a un modelo MLP de clasificación de tres capas cuya entrada esta dada por vectores $[1, x]^T$ y tanh como función no lineal en todas las capas. Escribir el grafo del modelo. Propagar el vector $[1, 2]^T$ con etiqueta 1. Calcular los valores de entrada $\mathbf{s}$, salida $\mathbf{x}$ y sensibilidades δ en cada capa. Como consecuencia calcular las derivadas del error respecto de los parametros W.

7. (M) No es sorpresa que añadir nuevas unidades a las capas ocultas de una red neuronal significa un incremento de su capacidad de aproximación ya que añadimos nuevos parámetros. Cuantos parámetros contiene un modelo de una red neuronal especificada por el vector $\mathbf{d} = [d^{(0)}, d^{(1)}, \cdots, d^{(L)}]$ que especifica el número de unidades en cada capa ? Evaluar la fórmula encontrada para un modelo con 2 capas ocultas y 10 unidades en cada capa.

8. (M) Para el perceptrom sigmoidal $h(x) = \tanh(\mathbf{x}^T\mathbf{w})$, sea el error de ajuste $E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(\tanh(\mathbf{x}_n^T\mathbf{w}) - y_n)^2$. Mostrar que

$$\nabla E_{\text{in}}(\mathbf{w}) = \frac{2}{N}\sum_{n=1}^{N}(\tanh(\mathbf{x}_n^T\mathbf{w}) - y_n)(1 - \tanh(\mathbf{x}_n^T\mathbf{w})^2)\mathbf{x}_n$$

si $\mathbf{w} \to \infty$ ¿que le sucede al gradiente? ¿Cómo se relaciona esto con la dificultad de optimizar el perceptron multicapa?

9. (M) Consideremos la función de error $Q(\mathbf{w}) = (\mathbf{w} - \mathbf{w}^*)^T\mathrm{Q}(\mathbf{w} - \mathbf{w}^*)$ donde Q es una matriz definida positiva arbitraria. Poner $\mathbf{w} = 0$.
   - Mostrar que el gradiente $\nabla E(\mathbf{w}) = -\mathrm{Q}\mathbf{w}^*$. Que pesos minimiza $E(\mathbf{w})$.
   - ¿Se mueve el gradiente descendente en la dirección de estos pesos óptimos?
   - Reconciliar esta contestación con la encontrada con anterioridad y que nos dice que el gradiente es la mejor dirección en la que realizar el paso de avance (Ayuda: ¿Cuanto de grande era el paso?)

SVM

1. (B) Considerar los datos siguientes y un hiperplano $(b, \mathbf{w})$ que separa los datos (i.e. $y_n(\mathbf{w}^T\mathbf{x}_n + b) > 1$ $n = 1, 2, ..$)

$$\mathrm{X}=\begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \end{bmatrix} \quad \mathbf{y}=\begin{bmatrix} -1 \\ -1 \\ +1 \end{bmatrix} \quad \mathbf{w}=\begin{bmatrix} 1{,}2 \\ -3{,}2 \end{bmatrix} \quad \text{b=-0.5}$$

- Calcular $\rho = \text{mín}_{n=1,,N}\, y_n(\mathbf{w}^T\mathbf{x}_n + b)$
- Calcular los nuevos pesos $(b^*, \mathbf{w}^*) = \frac{1}{\rho}(b, \mathbf{w})$ y mostrar que satisfacen $\text{mín}_{n=1,,N}\, y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$
- Dibujar ambos hiperplanos para mostrar que son el mismo separador.

2. (B) Para datos separables que contienen ejemplos positivos y negativos y un hiperplano separador h, definir el margen positivo $\rho_+(h)$ como la distancia entre h y el punto más cercano de clase $+1$. De igual manera, definir el margen del lado negativo como la distancia entre h y el punto más cercano de clase -1. Argumentar que si h es el hiperplano óptimo entonces $\rho_+(h) = \rho_-(h)$. Es decir el ancho por cada parte debe ser el mismo.

3. (B) Derivar el problema de optimización cuadrática asociado al problema Lineal-Hard-Margin-SVM a partir de su formulación primal del problema de optimización.

4. (B) Sea Y una matriz diagonal $(N \times N)$ con valores $Y_{nn} = y_n$ (donde y_n son los valores de un vector). Se X la matriz de datos aumentada con una columna de 1s. Mostrar que $A = YX$ siendo A la matriz de problema-QP del ejercicio anterior.

5. (A) Mostrar que la matriz Q asociada al problema-QP del ejercicio anterior es semidefinida positiva. Es decir, $\mathbf{u}^T Q\mathbf{u} \geq 0$ para todo $\mathbf{u}$.

6. (M) Construir el problema-QP asociado al ejemplo del ejercicio 1.

7. (M) Ejercicio 8.8 del pdf de SVM.

8. (B) Ejercicio 8.9 del pdf de SVM.

9. (B) Ejercicio 8.10 del pdf de SVM.

10. (B) Ejercicio 8.11 del pdf de SVM.

11. (M) Resolver el problema: Minimizar $u_1^2 + u_2^2$ sujeto a $u_1 + 2u_2 \geq 2$, $u_1, u_2 \geq 0$.

12. (M) Derivar las restricciones asociadas al problema SVM-Hard-Dual debidas a las condiciones de KKT.

13. (A) Derivar las restriciones asociadas al problema SVM-Soft-Dual debidas a las condiciones de KKT.

14. (B) Si todos los datos en una muestra tienen la misma etiqueta, entonces $\alpha_n^* = 0$ para $n = 1, \ldots, N$. ¿Que valores tomarían en ese caso $\mathbf{w}*$ y b^* ? (Ayuda: mirar la solución dual)

15. (M) La condición de complementariedad de holguras dada por KKT, dice que si $\alpha_n^* > 0$, entonces el punto (x_n, y_n) está sobre la frontera del hiperplano óptimo y $y_n(\mathbf{w}^T\mathbf{x}_n + b^*) = 1$. Mostrar que la inversa no es verdad. Es decir, que es posible que $\alpha_n^* = 0$ y aún (x_n, y_n) siga sobre la frontera satisfaciendo $y_n(\mathbf{w}^T\mathbf{x}_n + b^*) = 1$.

(Ayuda: considerar un caso simple con dos ejemplos positivos en $(0,0)$ y $(1,0)$ y otro ejemplo negativo).

16. (B) Repasar el ejemplo 8.8 del pdf de SVM

17. (B) Consideremos dos transformaciones de dimensión finita Φ_1 y Φ_2 y sus correspondientes núcleos K_1 y K_2. Definir $\Phi(\mathbf{x}) = (\Phi_1(\mathbf{x}), \Phi_2(\mathbf{x}))$.

    a) Expresar el correspondiente kernel de Φ en términos de K_1 y K_2.

    b) Consider la matriz $\Phi_1(\mathbf{x}) \cdot \Phi_2(\mathbf{x})^T$. Expresar el correspondiente núcleo en términos de K_1 y K_2.

    c) Por tanto mostrar que si K_1 y K_2 son núcleos , entonces también lo son K_1+K_2 y K_1K_2

    El resultado anterior puede usarse para construir núcleos polinómicos generales y (cuando se extienden a transformaciones infinito dimensionales) para construir núcleos Gaussian-RBF generales.

18. (B) Ejercicio 8.16 del pdf de SVM.