Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Segundo Semestre 2024
Fundamentos Matemáticos de Ciencia de Datos IMT3120

Tarea 1

Instrucciones. Puede usar las conclusiones de una pregunta para responder las preguntas que siguen. Por ejemplo: puede usar las conclusiones obtenidas de la Pregunta 1 para responder la Pregunta 2.
Puede entregar soluciones con constantes más grandes que las dadas por el enunciado (esto no afectará su puntaje). Si obtiene constantes mejores que las del enunciado, por favor enfatícelo en su respuesta.

Pregunta 1. (a) Sea $X \sim \mathcal{N}(0,1)$. Pruebe que para $\lambda < 1/2$, $\Lambda_{X^2}(\lambda) = -\frac{1}{2}\ln(1-2\lambda)$. Qué pasa si $\lambda \geq 1/2$?

(b) Una v.a. $Z \sim \chi^2_k$ (chi-cuadrado con k grados de libertad) si existen $X_1, \ldots, X_k \overset{iid}{\sim} \mathcal{N}(0,1)$ y $Z = X_1^2 + \ldots X_k^2$. Note que en tal caso $\mathbb{E}[Z] = k$. Pruebe que para todo $t > 0$,

$$\mathbb{P}[Z \geq k + 2\sqrt{tk} + 2t] \leq \exp(-t)$$
$$\mathbb{P}[Z \leq k - 2\sqrt{tk}] \leq \exp(-t).$$

Indicación. Calcule cotas genéricas para la fgml cuando $\lambda > 0$ y $\lambda < 0$.

(c) Concluya que si $0 < \delta < 1$ y $k > \ln(2/\delta)$ entonces para $\varepsilon = 3\sqrt{\frac{\ln(2/\delta)}{k}}$,

$$\mathbb{P}[(1-\varepsilon)k < Z < (1+\varepsilon)k] \geq 1 - \delta.$$

Pregunta 2. Sea A una matriz $k \times d$ con coeficientes iid $a_{ij} \sim \mathcal{N}(0,1)$

(a) Sea $x \in \mathbb{R}^d$, $0 < \delta < 1$ y $k > \ln(2/\delta)$ entonces para $\varepsilon = 3\sqrt{\frac{\ln(2/\delta)}{k}}$,

$$\mathbb{P}\left[(1-\varepsilon)\|x\|_2^2 \leq \frac{1}{k}\|Ax\|_2^2 \leq (1+\varepsilon)\|x\|_2^2\right] \geq 1 - \delta.$$

(b) Pruebe el siguiente resultado:

Teorema 1 (Lema de Johnson-Lindenstrauss)**.** Sea $\mathcal{Q} \subseteq \mathbb{R}^d$ finito y $\varepsilon > 0$. Entonces existe una matriz $A \in \mathbb{R}^{k \times d}$ con $k = \left\lfloor \frac{9\ln(4|\mathcal{Q}|^2)}{\varepsilon^2} \right\rfloor$ tal que

$$(1-\varepsilon)\|x-y\|_2^2 \leq \|Ax - Ay\|_2^2 \leq (1+\varepsilon)\|x-y\|_2^2 \quad (\forall x, y \in \mathcal{Q}).$$

Observación. En este caso, se dice que A es una *inmersión de baja distorsión*, ya que transforma un conjunto de datos en alta dimensión a uno de baja dimensión, preservando aproximadamente las distancias.

(c) Sea ahora $\mathcal{Q} \subseteq \mathcal{B}_2^d = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ un conjunto finito. Pruebe que una inmersión de baja distorsión A sobre $\tilde{\mathcal{Q}} := \mathcal{Q} \cup (-\mathcal{Q})$ (donde $-\mathcal{Q} = \{-x : x \in \mathcal{Q}\}$) posee la siguiente propiedad:

$$|\langle Ax, Ay\rangle - \langle x, y\rangle| \leq \varepsilon \quad (\forall x, y \in \mathcal{Q}).$$

Indicación. Use (y pruebe si no la conoce) la identidad de polarización:

$$\langle x, y\rangle = \frac{1}{4}\Big[\|x+y\|_2^2 - \|x-y\|_2^2\Big].$$

Pregunta 3. Considere un problema de clasificación binaria. $S = \{(X_i, Y_i)\}_{i\in[n]} \overset{iid}{\sim} \mathcal{D}$, con $\text{supp}(\mathcal{D}) \subseteq \mathcal{B}_2^d \times \{-1, +1\}$. Definimos el margen de S como

$$\rho := \sup_{w\in\mathbb{R}^d} \min_{i\in[n]} \frac{Y_i\langle w, X_i\rangle}{\|w\|_2}.$$

(a) Pruebe que el margen es estrictamente positivo si y sólo los datos observados son estrictamente separables; es decir, existe $w \in \mathbb{R}^d$ tal que

$$\langle w, X_i\rangle Y_i > 0 \quad (\forall i \in [n]).$$

Pruebe además que en este caso, existe $\hat{w}$ con $\|\hat{w}\|_2 = 1$ tal que la distancia entre $\{Y_iX_i : i \in [n]\}$ y el semiespacio $\mathcal{H} = \{z \in \mathbb{R}^d : \langle \hat{w}, z\rangle \leq 0\}$ es igual a ρ. Interprete geométricamente este resultado.

(b) Pruebe que el $\hat{w}$ mencionado en la parte (a) satisface $\hat{w} = \frac{w^*}{\|w^*\|_2}$, donde w^* es la única solución óptima del siguiente problema de optimización (debe justificar la existencia y unicidad de soluciones para este problema)

$$\text{(Hard-SVM)} \qquad \min_{w\in\mathbb{R}^d} \Big\{\frac{1}{2}\|w\|_2^2 : Y_i\langle w, X_i\rangle \geq 1\,(\forall i \in [n])\Big\}.$$

Pruebe además que $\rho = 1/\|w^*\|_2$.

(c) Pruebe que el dual Lagrangeano de (Hard-SVM) es el siguiente

$$\text{(Hard-SVM)}^* \qquad \max_{\alpha\in\mathbb{R}^n} \Big\{\sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n\sum_{k=1}^n \alpha_i\alpha_k Y_iY_k\langle X_i, X_k\rangle : \alpha \geq 0\Big\}.$$

Pruebe además que si el primal es factible, entonces tanto el primal como el dual poseen soluciones óptimas, w^*, α^*, y que están relacionadas por

$$w^* = \sum_{i=1}^n \alpha_i^* Y_i X_i.$$

Se dice que los vectores X_i con $\alpha_i^* > 0$ son *vectores de soporte*. De una interpretación geométrica de estos vectores.
Indicación. Use el Teorema de Dualidad Fuerte.

Pregunta 4. Sea $S = \{(X_i, Y_i)\}_{i\in[n]} \subseteq \mathcal{B}_2^d \times \{-1, +1\}$ un conjunto de datos linealmente separable con margen $\rho > 0$. Pruebe que si $k = \left\lfloor \frac{81\ln(16|S|^2)}{\rho^2} \right\rfloor$, entonces existe $A \in \mathbb{R}^{k\times d}$ tal que el conjunto de datos $\tilde{S} = \{(AX_i, Y_i)\}_{i\in[n]}$ es linealmente separable con margen $\rho/3$.

Utilice este resultado para proponer un algoritmo (aleatorizado) de clasificación lineal que opere con vectores $v \in \mathbb{R}^k$.

Parte Computacional.

El objetivo de esta pregunta es implementar métodos para resolver problemas de clasificación lineal en el caso realizable y no-realizable.

Pregunta 5. En el caso que los datos no son linealmente separables, podemos considerar la *violación de restricciones* como penalización en la función objetivo. Para esto se introducen variables de holgura $(\xi_i)_{i\in[n]}$, un parámetro de regularización $\lambda > 0$ y se formula el problema penalizado

$$\text{(Soft-SVM)} \qquad \begin{array}{lrcll} \min_{w\in\mathbb{R}^d,\xi\in\mathbb{R}^n} & \frac{\lambda}{2}\|w\|_2^2 + \sum_{i=1}^n \xi_i & & & \\ & Y_i\langle w, X_i\rangle & \geq & 1-\xi_i & \forall i \in [n] \\ & \xi_i & \geq & 0 & \forall i \in [n]. \end{array}$$

(a) Implemente un solver para los problemas (Hard-SVM) y (Soft-SVM). Para esto puede optar por cualquiera de las siguientes opciones:

(i) Diseñar su propio algoritmo de optimización convexa en Python. Si escoge esta opción, puede usar el hecho de que podemos eliminar las restricciones haciendo la sustitución

$$\xi_i = [1 - \langle w, Y_i X_i\rangle]_+.$$

(ii) Formular un problema convexo cuadrático y ocupar alguna librería de Python (ver algunos ejemplos en `https://scaron.info/blog/quadratic-programming-in-python.html`)

Importante.

- Los datos entregados en los ejemplos no son linealmente separables, pero (potencialmente) afínmente separables; es decir, podrían existir $w \in \mathbb{R}^d$ y $b \in \mathbb{R}$ tales que

$$Y_i(\langle w, X_i\rangle + b) \geq 1 \quad (\forall i \in [n]).$$

  Para pre-procesar estos datos, considere el espacio extendido $\mathbb{R}^{d+1}$ con datos $(X_i', Y_i) = ([X_i; 1], Y_i)$. Ahora existe un separador lineal en $\mathbb{R}^{d+1}$ (más precisamente, $[w; b]$).
- Su método debe escoger automaticamente el parámetro de regularización λ. Para esto se aconseja separar la muestra en conjuntos (disjuntos y aleatorios) de entrenamiento (80 %), validación (10 %) y prueba (10 %). Sobre el conjunto de entrenamiento resuelva (Soft-SVM) para valores de $\lambda \in \Lambda := \{2^{-10}, 2^{-9}, \ldots, 2^{-1}, 2^0\}$; esto da lugar a modelos $(\hat{w}_\lambda)_{\lambda\in\Lambda}$. Luego, escoja el modelo $\hat{w}_\lambda$ que minimiza la función objetivo sobre el conjunto de validación. Finalmente, estime el riesgo de población del modelo seleccionado evaluando su error de clasificación (pérdida 0-1) sobre el conjunto de prueba.

Aplique su algoritmo a las distintas instancias entregadas, y determine si los conjuntos son linealmente separables o no. Determine también la escalabilidad de su algoritmo (es decir, en qué dimensión o tamaño de muestra deja de funcionar).

(b) Pre-procese sus datos utilizando una matriz $\frac{1}{\sqrt{k}}A$ aleatoria con coeficientes $a_{ij} \sim \mathcal{N}(0,1)$ (como se sugiere en la Pregunta 4). Evalúe la habilidad de su algoritmo para detectar la condición de separabilidad, así como la escalabilidad de su nuevo algoritmo para el primal de (Soft-SVM) y (Hard-SVM), en las mismas instancias de arriba.

(c) Calcule el dual Lagrangeano de (Soft-SVM). Además implemente un solver para los problemas duales de (Hard-SVM) y (Soft-SVM). Determine la escalabilidad de sus algoritmos, probándolos en las mismas instancias de la parte (a).