MDS7104 Aprendizaje de Máquinas
Profesor: Francisco Vásquez L.
Auxiliares: Catalina Lizana G., Álvaro Márquez S., Diego Olguín W.
Fecha: 15 de Octubre de 2024.

Auxiliar 9: Repaso C2 y PCA

P1. Considere las siguientes situaciones:

(a) Una empresa necesita su ayuda. Él necesita clasificar postulaciones de trabajo entre las categorías buena/mala, y detectar postulantes que mientan en su postulación usando estimación de densidad para detectar outliers. Para esta situación ¿Recomienda usar un modelo discriminativo o uno generativo? Por qué?

(b) La misma empresa también quiere clasificar aplicaciones de software para detectar cuales tienen más tendencia a presentar bugs usando características del código fuente. Este proyecto solo dispone de algunas aplicaciones para usarlos de data de entrenamiento. Para crear el clasificador más preciso ¿Recomienda usar un modelo discriminativo o uno generativo? Por qué?

(c) Por último, la misma empresa quiere clasificar otras compañías para decidir cuales comprar. Este proyecto tiene mucha data de entrenamiento basada en décadas de investigación. Para crear el clasificador más preciso ¿Recomienda usar un modelo discriminativo o uno generativo? Por qué?

P2. Para distinguir entre billetes falsos y verdaderos, se hicieron mediciones del largo y la diagonal de la imagen presente en estos. Para 1000 billetes (500 verdaderos y 500 falsos) se obtuvieron los siguientes valores para el promedio y la matriz de covarianza (usando estimadores insesgados), donde la primera coordenada corresponde al largo:

$$\text{Billetes verdaderos:} \quad \bar{x}_V = \begin{bmatrix} 214.97 \\ 141.52 \end{bmatrix} \text{ y } \hat{\Sigma}_V = \begin{bmatrix} 0.1502 & 0.0055 \\ 0.0055 & 0.1998 \end{bmatrix}$$

$$\text{Billetes falsos:} \quad \bar{x}_F = \begin{bmatrix} 214.82 \\ 139.45 \end{bmatrix} \text{ y } \hat{\Sigma}_F = \begin{bmatrix} 0.1240 & 0.0116 \\ 0.0116 & 0.3112 \end{bmatrix}$$

(a) Asuma que las verdaderas matrices de covarianza de los billetes verdaderos y falsos son iguales. ¿Cómo podría estimar la matriz de covarianza en común?

(b) Explique las suposiciones hechas para hacer uso de LDA para clasificar un nuevo dato entre billete verdadero y billete falso. Escriba la regla de clasificación.

(c) Use el método anterior para determinar si un billete de largo 214.0 y diagonal 140.4 es verdadero o falso.

P3. Considere los siguientes datos:

Obs.	X_1	X_2	Y
1	0.8	0.4	Rojo
2	0.4	0.8	Rojo
3	0.4	0.2	Rojo
4	0.2	0.4	Rojo
5	0.4	0.4	Azul
6	0.8	0.8	Azul

(a) Considere el siguiente kernel $K(x,z) = \frac{z^T x}{\|x\|\|z\|}$. Muestre que cumple ser un Mercel kernel en $\mathbb{R}^2_+ \times R^2_+$ y encuentre el vector de características $\phi(x)$ correspondiente al kernel.

(b) Mapee los datos anteriores al nuevo espacio de características definido implicitamente por $K(x,z)$. Son los nuevos datos linealmente separables?

(c) Dibuje el plano de decisión para SVM de margen maximal en el espacio de características implícito.

(d) Dibuje el plano de decisión en el espacio de características original resultante del kernel $K(x,z)$.

P4. Dado un dataset de N datos x_i, se desea realizar kernel PCA (KPCA). Para esto, primero se debe mapear x_i a una característica no lineal $\phi(x_i)$ en un espacio RKHS $\mathcal{H}$ correspondiente a un kernel $k(x,y) = \langle \phi(x), \phi(y) \rangle$. Luego se deben centrar los datos en el nuevo espacio de características $\tilde{\phi}(x_i)$ y definir el operador de covarianza C.

(a) Muestre que cada vector propio de C puede ser expresado como una combinación lineal de las características $\tilde{\phi}(x_i)$.

(b) Implemente KPCA para algunos dataset de juguete (presentes en el colab asociado a este auxiliar).