Análisis Multivariado: Tarea 5

Aprendizaje supervisado y no supervisado

Fecha de entrega: 5 de junio.

Naive Bayes

1. (2 puntos) El archivo *SMSSpams.txt* contiene una colección de mensajes de texto clasificados como *spam* y *ham*. El objetivo de este ejercicio es crear un filtro de *spam* usando el 80% de los datos como muestra de entrenamiento y probando el algoritmo en el 20% restante. Para este ejercicio algunas posibles consideraciones a tomar en cuenta son las siguientes:

   (a) Signos de puntuación, símbolos y caracteres numéricos pueden no incluirse en el análisis.

   (b) La capitalización no es relevante para el proceso de aprendizaje.

   (c) Cada mensaje puede caracterizarse por la presencia de las palabras.

   (d) Palabras poco frecuentes o que no aparecen en la muestra de entrenamiento pueden generar problemas en el aprendizaje. Para palabras poco frecuentes se puede utilizar un suavizamiento, i.e., para un parámetro $\alpha > 0$ (usualmente $\alpha = 1$) obtenemos, por ejemplo que

   $$\mathbb{P}(w_i|S) = \frac{N_{w_i|S} + \alpha}{N_S + V\alpha},$$

   donde $N_{w_i|S}$ es el número de veces que aparece la palabra w_i en el *spam*, N_S el número de *spams* y V el tamaño del vocabulario. Las palabras que no aparecen pueden tratarse igual o quitarlas al momento de clasificar un nuevo mensaje y solo trabajar con las palabras existentes.

   (e) Para una palabra w_i el teorema de Bayes garantiza

   $$\mathbb{P}(S|w_i) = \frac{\mathbb{P}(w_i|S)\mathbb{P}(S)}{\mathbb{P}(w_i|S)\mathbb{P}(S) + \mathbb{P}(w_i|H)\mathbb{P}(H)},$$

   donde $\mathbb{P}(S)$ y $\mathbb{P}(H)$ pueden ser estimadas con las frecuencias de la muestra de entrenamiento o en algunas ocasiones se asumen iguales, i.e., $\mathbb{P}(S) = \mathbb{P}(H) = .5$.

SVM

2. (2 puntos) Para la base de datos *iris* y utilizando las variables *Sepal Width* y *Sepal Length* ajustar:

   (a) Un SVM lineal.

   (b) Un SVM usando un kernel polinomial de grado 3.

   (c) Un SVM usando el kernel gaussiano.

   Para cada modelo el objetivo es dibujar las regiones generadas, obtener la matriz de confusión y argumentar cuál es el mejor modelo.

Árboles de Decisión

3. (2 puntos) Para la base de datos del *Titanic* ajustar:

   (a) Un árbol de decisión.

   (b) Un bosque aleatorio.

   (c) Los tres métodos *boosted* vistos en clase, i.e. *AdaBoost*, *GBoost* y *XGBoost*.

   Esto con el fin de predecir que pasajeros sobreviven dados sus atributos. Reportar la metodología completa incluyendo la elección de híper-parámetros, las matrices de confusión y la discusión sobre el mejor modelo.

DBSCAN

4. (2 puntos) En este ejercicio se estudiará el algoritmo DBSCAN por lo cual es necesario

   (a) Simular 500 observaciones de 5 distribuciones normales bivariadas (100 observaciones por cada distribución) de parámetros

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \mu_2 = \begin{pmatrix} 5 \\ 5 \end{pmatrix} \qquad \mu_3 = \begin{pmatrix} -5 \\ -5 \end{pmatrix} \qquad \mu_4 = \begin{pmatrix} -5 \\ 5 \end{pmatrix} \qquad \mu_5 = \begin{pmatrix} 5 \\ -5 \end{pmatrix}$$

   y

$$\Sigma_1 = \begin{pmatrix} 1.5 & 0 \\ 0 & 1.5 \end{pmatrix} \qquad \Sigma_2 = \begin{pmatrix} 1 & .5 \\ .5 & 1.5 \end{pmatrix} \qquad \Sigma_3 = \begin{pmatrix} 2 & -1 \\ -1 & 2.5 \end{pmatrix}$$

$$\Sigma_4 = \begin{pmatrix} 1 & -1 \\ -1 & 1.5 \end{pmatrix} \qquad \Sigma_5 = \begin{pmatrix} 1 & -1 \\ -1 & 2.5 \end{pmatrix}$$

(b) Graficar los datos y comentar en la estructura de ellos.

(c) Utilizar el algoritmo *DBSCAN* para realizar un aprendizaje no supervisado para diferentes parámetros de *min_puntos* y ϵ. Reportar las métricas de clasificación para los mejores parámetros encontrados.

(d) Finalmente, utilizar el algoritmo de k-means para clasificar los datos. ¿Cómo se compara con *DBSCAN*?

GMM

5. (2 puntos) Con los datos generados en el ejercicio anterior

   (a) Ajusta un modelo de mezclas gaussiano para diferentes valores de k. ¿Cuál es el mejor valor para k?

   (b) Para el valor de k encontrado graficar las curvas de nivel generadas por el GMM. ¿Hay evidencia de posibles anomalías?

   (c) ¿Cómo se compara el GMM contra los métodos del ejercicio anterior, i.e., DBSCAN y k-means?