Taller 1: Preprocesamiento de Datos

Entrega: Miércoles 8 de Octubre de 2008 en clase, impreso

Minería de Datos - 2008-II

---

1. Preprocesamiento de datos

   a) Obtenga el conjunto *Adult* del repositorio **UCI Machine Learning Repository**

   b) Establezca si hay valores faltantes y aplique al menos dos estrategias diferentes para manejarlos. Documente de manera detallada el proceso y los resultados.

   c) Convierta todas los atributos numéricos a categóricos utilizando dos estrategias diferentes. Documente de manera detallada el proceso y los resultados.

   d) Transforme el conjunto de datos de manera que todos los atributos sean numéricos. Documente de manera detallada el proceso y los resultados.

   e) Escoja una técnica para la detección de datos atípico y aplíquela sobre el conjunto de datos. Documente de manera detallada el proceso y los resultados.

2. Reducción de la dimensionalidad:

   a) Sobre el mismo conjunto de datos del punto 1, aplique *PCA* para reducir la dimensionalidad del conjunto a 2 dimensiones.

   b) Grafique el conjunto resultante.

   c) Analice los resultados.

   d) ¿puede encontrar un sector del gráfico en el cual todos los puntos pertenezcan a una única clase?

   e) ¿qué interpretación puede darle a esta región del espacio?

3. Escoja un conjunto de datos diferente del mismo repositorio y repita los pasos de los puntos 1 y 2.

4. Eigenfaces:

   a) Obtenga el conjunto de datos y el programa en Python de la página del curso

   b) Aplique PCA al conjunto de datos con diferentes número de vectores principales (5,10,20,30). Presente los resultados y analícelos de manera detallada.

   c) Dibuje un mapa del conjunto de datos con las coordenadas correspondientes a los dos primeros componentes principales. Analice el resultado:

      1) ¿Qué significan los puntos extremos?
      2) ¿Puede identificar grupos? ¿Qué interpretación les puede dar?
      3) ¿Qué está representando el primer eigenvector principal?
      4) ¿Qué está representando el segundo eigenvector principal?

   d) Basado en la representación generada por el PCA, proponga un método para calcular una *interpolación* entre dos caras. Aplíquela a las caras de los miembros del grupo.