Práctica 6: Clasificación de patrones

Semana 6

En esta práctica se utilizarán como datos los archivos a.txt, o.txt y u.txt. Los mismos corresponden a los tres primeros formates F_1, F_2 y F_3 de varias emisiones de las vocales /a/, /o/ y /u/ respectivamente. Cada archivo deberá separse en dos tercios para entrenamiento y un tercio para evaluación.

1. *Clasificación supervisada. Discriminantes lineales:*

   a) Dibuje en un plano F_1 F_2 los dos primeros formantes de las tres vocales para los datos de entrenamiento.

   b) Asumiendo que las muestras se originan a partir de una distribución gausiana, determine la media y la matriz de covarianza de cada una de las vocales usando los datos de entrenamiento.

   c) Dibuje superpuesta al primer gráfico contornos de la elipse que corresponden a las gausianas de cada distribución.

   d) Determine en forma teórica la expresión del discriminante entre clases, y grafique los límites de dichas zonas sobre los puntos de entrenamiento. Explique el criterio para generar ese discriminante.

   e) Grafique los datos de evaluación y corrobore el resultado de clasificación. Determine los porcentajes de error.

2. *Clasificación no supervisada. Algoritmo K-means.*

   a) Con los datos de entrenamiento de las tres vocales (y los dos primeros formates) construya un solo vector para cada formante y ordénelos en forma aleatoria.

   b) Implemente el algoritmo k-means y aplíquelo para los datos del item anterior.

   c) Dibuje los clusters de datos y compárelos con el punto anterior.

   d) Clasificar los datos de test de acuerdo al criterio de máxima probabilidad a posteriori y determine el porcentaje de error cometido.

   e) Repita el procedimiento para tres inicializaciones aleatorias de los centroides y también tomando como inicialización un pequeño subconjunto de muestras de las de prueba (bootstrap). Establezca conclusiones.

   f) Compare los porcentajes de error entre este método, comparando las distintas inicializaciones. Compare también con el método de discriminantes lineales.

3. *Clasificación no supervisada. Algoritmo EM*

   a) Con los datos de entrenamiento de las tres vocales (y los dos primeros formates) construya un solo vector para cada formante y ordénelos en forma aleatoria.

b) Implemente el algoritmo EM y aplíquelo para los datos del item anterior.

c) Grafique en un plano F_1, F_2 los datos y en forma superpuesta los contornos de la elipse que corresponden a las gausianas de cada distribución estimadas con el algoritmo EM.

d) Dibuje los clusters de datos y compárelos con el punto anterior.

e) Determine a que cluster se clasifican los datos de test y determine el porcentaje de error cometido.

f) Repita el procedimiento para tres inicializaciones aleatorias de los centroides, y con un subconjunto de bootstrap. Establezca comparaciones.

g) Compare los porcentajes de error entre este método y los métodos de discriminantes lineales y k-means. Establezca conclusiones.