Enunciado Práctica 6 para entregar

Para esta entrega se utilizarán como datos los mismos archivos `a.txt`, `o.txt` y `u.txt` utilizados en la práctica. 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 40 muestras para entrenamiento y 10 para evaluación. Para esta práctica se descartará la información del tercer formante.

LDA (Linear Discriminant Analysis), entrenamiento supervisado

1. En el plano de los formantes F_1,F_2 dibuje las muestras de entrenamiento de cada clase con distintos colores. Este gráfico base debe aparecer como fondo del que se pide en el siguiente item.

2. Diuje las medias estimadas de cada clase y las elipses correspondientes a las curvas de nivel de probabilidad $P(\mathbf{x}|\text{clase}_k)$. Para cada clase debe dibujarse la elipse correspondiente a la curva de nivel de probabilidad resultante del entrenamiento con sus puntos solamente, y las correspondientes a la clasificación LDA. En el mismo gráfico dibujar tambien las rectas de separacion entre cada par de clases.

3. Dibuje en otro gráfico la distribución de puntos de testeo con diferentes colores (con puntos) y el resultado de la clasificación con los mismos colores pero con círculos vacíos (de modo que se pueda ver en el gráfico la cantidad de puntos erróneamente clasificados). Dibuje también las medias y las rectas del gráfico anterior. Calcule el porcentaje de error total cometido.

Clasificación no supervisada. Algoritmo K-means y EM (expectation maximization)

En estos ejercicios la separación en dos conjuntos de entrenamiento y testeo será la misma que se utilizó para el ejercicio anterior. Pero en este caso los puntos de entrenamiento se mezclan de modo aleatorio en un solo conjunto del cual se desconoce el etiquetado de clase.

Ambos algoritmos de entrenamiento son iterativos, por lo que se requiere un prototipo inicial de cada clase. Para ambos **se realizarán los ejercicios dos veces, con distintas prototipos iniciales**, una vez con prototipos tomados de pocos ejemplos de entrenamiento y otra con inicialización aleatoria. En la primera inicialización se separa un conjunto de muy pocos datos (no más de 5) de las muestras de entrenamiento de cada clase y con ellas se inicializa un prototipo para cada clase. Luego esas muestras se descartan, y no se usan más para entrenar. En el caso de inicialización aleatoria, se dividirá el espacio en 3 partes, trazando 3 rectas a 120 grados cada una, que confluyan en la media global de los puntos de entrenamiento (la primera recta trazada debe tener una dirección aleatoria en el espacio), y dentro de cada zona se realizarán las medias de los puntos de entrenamiento que se tomarán como prototipos iniciales.

1. Implementar el algoritmo de K-means. Para cada iteración, grafique los prototipos (medias) y los conjuntos de datos “captados” por cada media. En cada gráfico dibuje todas las medias de las iteraciones anteriores.

Muestre en forma de texto la distorsión para cada iteración en el mismo gráfico.

2. Dibuje en otro gráfico la distribución de puntos de testeo con diferentes colores (con puntos) y el resultado de la clasificación con los mismos colores pero con círculos vacíos (de modo que se pueda ver en el gráfico la cantidad de puntos erróneamente clasificados). Dibuje también las medias obtenidas en el entrenamiento y las elipses correspondientes a la curva de nivel de la probabilidad de cada clase. Calcule el porcentaje de error total cometido.

3. Implementar el algoritmo EM. Para cada iteración, grafique los prototipos (medias) y el conjuntos de puntos de entrenamiento, utilizando los "gamma" de cada clase como color en formato `rgb`. En cada gráfico dibuje todas las medias de las iteraciones anteriores. Muestre en forma de texto el likelihood total de la distribución para cada iteración en el mismo gráfico. Utilice como matriz de covarianza inicial la estimación de la covarianza general de todo el conjunto de entrenamiento. Qué pasaría si se inicilizara la matriz de covarianza de modo arbitrario, por ejemplo diagonal de unos?

4. Dibuje en otro gráfico la distribución de puntos de testeo con diferentes colores (con puntos) y el resultado de la clasificación con los mismos colores pero con círculos vacíos (de modo que se pueda ver en el gráfico la cantidad de puntos erróneamente clasificados). Dibuje también las medias obtenidas en el entrenamiento y las elipses correspondientes a la curva de nivel de la probabilidad de cada clase. Calcule los porcentajes de error total cometido.