Tarea 7
IELE-4017 Análisis Inteligente de Señales y Sistemas
Profesor: Luis Felipe Giraldo Trujillo
2021-I

En esta tarea no pueden utilizar funciones predefinidas para evaluar funciones Gaussianas, estimar vectores de medias y matrices de covarianza, combinar Gaussianas, hacer automáticamente particiones de los conjuntos de entrenamiento, validación y prueba, realizar la clasificación, calcular errores de clasificación, calcular matrices de confusión. Alguna falta de este tipo en cualquiera de los enunciados de esta tarea implicará que la tarea completa tenga una calificación de 0.0.

1. (40 puntos) **Combinación de Gaussianas.** La falta de la solución en alguno de los enunciados de esta tarea implicará una nota de 0.0 en este ejercicio.
   El archivo `datosIris.txt` contiene 150 observaciones del largo y ancho de los sépalos de flores tipo Iris. Estos datos son reales obtenidos en el repositorio UCI (`http://archive.ics.uci.edu/ml/datasets/iris`).

   a) Grafique los 150 puntos y comente la distribución de los datos.

   b) Implemente el algoritmo EM, y estime la función de densidad que genera estos datos 2-dimensionales como una combinación de **tres** funciones de densidad Gaussianas. Es decir, $p(x) = \sum_{k=1}^{2} \alpha_k N(x; \mu_k, Q_k)$. Grafique los puntos en el archivo, y traslape el contorno de la función de densidad combinada $p(x)$ (NO el contorno de cada función de densidad base por separado) para varias de las iteraciones del algoritmo. Comente los resultados obtenidos.**No puede utilizar funciones preestablecidas de Gaussianas, ni funciones que combinan distribuciones.**

   c) Repita el procedimiento del punto b) para una combinación de **cuatro** Gaussianas.

2. (60 puntos) **Clasificación de expresiones faciales.** La falta de la solución en alguno de los enunciados de esta tarea implicará una nota de 0.0 en este ejercicio.
   La carpeta `faces` contiene 974 imágenes de personas posando expresiones faciales y los respectivos landmarks (n landmarks por cada expresión facial). Recuerde que cada landmark es una coordenada en dos dimensiones. La penúltima letra del nombre de cada archivo indica qué tipo de expresión facial la persona está posando. La notación es la siguiente: `n` para neutral, `h` para feliz, `s` para triste, `a` para furioso, `d` para disgusto, y `f` para miedo. Sólo vamos a utilizar los landmarks en este ejercicio. Las imágenes son para referenciar los landmarks.

   a) Divida el conjunto total de imágenes en tres particiones: conjunto de entrenamiento con 70 % de los datos, conjunto de validación con 15 %, y conjunto de prueba con 15 %. Estos conjuntos se deben definir de forma aleatoria, pero asegurando que los datos de las clases son balanceados en todos los conjuntos (es decir, no pueden haber más datos de una clase que de otras).

   b) **Sólo con los datos de entrenamiento**, calcule la media de Procrustes. Utilizando esta media de Procrustes, alinee los landmarks en las particiones de entrenamiento, validación, y prueba utilizando las rutinas de la tarea pasada (tenga en cuenta que en esta tarea hay más expresiones faciales). Este paso sería el de preprocesamiento (es decir, no hay necesidad de estandarizar los datos).

   c) Con los landmarks alineados, cree un vector de características de tamaño $2n$ por cada expresión facial en todos los conjuntos. Las primeras n características corresponden a las componentes horizontales de los landmarks de la expresión facial, y las n características siguientes corresponden a las componentes verticales de los landmarks de la expresión facial.

d) **Gaussian Naive Bayes Classifier:** Con los datos de entrenamiento, estime la media y la matriz de covarianza para cada clase (es decir, debe haber un vector de medias y una matriz de covarianza por clase). **No pueden utilizar funciones preestablecidas para estimar la media y matriz de covarianza.** Estime el likelihood de los datos (es decir, la función de distribución) por cada clase asumiendo que son Gaussianos con las medias y covarianzas estimadas. Asuma que las clases tienen igual probabilidad a priori.

   A tener en cuenta: en este caso, es probable que las matrices de covarianza sean cercanas a singulares, y existan problemas a la hora de invertirlas. Por lo tanto, a cada matriz de covarianza estimada súmele un término λI, donde I es la matriz identidad del mismo tamaño de la matriz de covarianza, y λ es un número mayor que cero. En otras palabras, la matriz de covarianza de la clase i es $Q_i = C_i + \lambda I$, donde C_i es la matriz de covarianza estimada con los datos de la clase i.

e) Construya un clasificador con la regla MAP y las distribuciones en el enunciado d) (es decir, construya el Gaussian Naive Bayes Classifier). Calcule el error de clasificación sobre el conjunto de validación para al menos 20 valores de $\lambda > 0$. Grafique una curva con los valores de error versus λ. Analice los resultados obtenidos, e identifique el valor de λ que produce menor error de clasificación en el conjunto de validación.

f) Con el valor de λ seleccionado en el enunciado e), en los datos de prueba calcule el porcentaje de error de clasificación y la matriz de confusión. Comente los resultados obtenidos. Analice en detalle la confusión del clasificador entre clases.