Caso de Estudio 11
IELE-4010 Procesos Estocásticos 2021-II
Profesor: Luis Felipe Giraldo

Detección de fallas en datos reales

1. (40 puntos) Se tiene que intentar resolver todos los enunciados planteados en este ejercicio para obtener puntos en el mismo. La falta de solución de alguno de los enunciados implicará automáticamente una calificación de 0 puntos en todo el ejercicio.

   Las mediciones de temperatura T (unidades en $^\circ C$), humedad relativa H (unidades en $\%$), y CO2 C (partes por millón) en un invernadero real de un cultivo de rosas [1] se modelan como variables aleatorias continuas a través del vector aleatorio $Z = [T, H, C]^T$. Se desea diseñar un sistema de detección automático basado en MLE que genere una alarma cuando se presente una medición anómala. Para esto, nos dan en el archivo `greenhouse3d.txt` 300 observaciones conjuntas de las variables de temperatura T (primera columna), humedad H (segunda columna), y CO2 C (tercera columna).

   a) Realice una gráfica 3-dimensional de las observaciones de temperatura vs humedad vs CO2, y realice una inspección visual de los datos analizando su comportamiento.

   b) Estime la mejor distribución Gaussiana de Z utilizando MLE, e indique la función resultante (es decir, una sola Gaussiana). Recuerde que para estimar la distribución Gaussiana basta con estimar el vector de valores esperados y la matriz de covarianza. **No puede utilizar funciones predefinidas para estimar el vector de medias y la matriz de covarianza. Implemente su propia rutina para hacer este cálculo.**

   c) Determine el umbral de decisión γ que define la detección. Es decir, dada una nueva observación $z \in \mathbb{R}^3$, se considera como dato anómalo si $f_Z(z) < \gamma$. Escriba la ecuación completa de la regla de decisión. Explique y justifique la elección de γ. Para esta explicación, proponga alguna forma de realizar una inspección visual de la regla de decisión. Por ejemplo, una forma de hacerlo es generar datos aleatorios uniformes dentro de una caja que encierre las observaciones de la base de datos del invernadero, y pintar de un color diferente aquellos que son considerados como anómalos de acuerdo con la regla de decisión.

   d) Reescriba la regla de decisión en el enunciado b) a través de una forma cuadrática. Es decir, encuentre la constante α y la matriz M, tal que, dada una observacion $z \in \mathbb{R}^3$, ésta es considerada como anómala si $(z-\mu)^T M(z-\mu) > \alpha$.

   e) Considere las nuevas mediciones dentro del invernadero en el archivo `datosNuevos.txt`. Dada la regla de decisión encontrada en c) o d) (las reglas en estos enunciados son exactamente la misma pero escrita de diferente forma), determine el porcentaje de mediciones que son anómalas.

   f) Repita los enunciados b), c), y e) ahora con una distribución de la forma $f_Z(z) = \sum_{k=1}^{3} \alpha_k N(z; \mu_k, Q_k)$, encontrada con el algoritmo EM. **No puede utilizar funciones predefinidas para encontrar la distribución combindada. Implemente su propio algoritmo EM.** Compare los resultados de realizar detección de anomalías usando una sola Gaussiana en el enunciado e) con los obtenidos al utilizar una combinación de Gaussianas (por ejemplo, el procentaje de muestras detectadas como anómalas, una gráfica comparando las muestras del dataset inicial utilizando para ejecutar el algoritmo EM con las muestras consideradas como anómalas, etc.).

[1] Tomado de IEEE Data Port

Clasificación en datos reales

2. (60 puntos) Se tiene que intentar resolver todos los enunciados planteados en este ejercicio para obtener puntos en el mismo. La falta de solución de alguno de los enunciados implicará automáticamente una calificación de 0 puntos en todo el ejercicio.

   Una empresa desea determinar la calidad de vino rojo en un proceso automatizado. Para eso, se miden las siguientes 11 variables en el líquido que se consideran dan información sobre la calidad de un vino [2]: acidez fija, acidez volatil, acides cítrica, azúcar residual, nivel de cloruro, dióxido de sulfuro libro, dióxido de sulfuro total, densidad, pH, sulfatos, y alcohol. Se midieron 1200 muestras de vinos y catadores determinaron si el vino es de buena calidad (clase 1) o mala calidad (clase 0). Esta información se encuentra en los archivos `XtrainVinos.txt` (mediciones de las 11 variables por cada muestra) y `YtrainVinos.txt` (variable de calidad). Las mediciones de las 11 variables fueron estandarizadas, es decir, se les restó la media de los datos y se dividió por la desviación estándar. El orden de las variables en el archivo es el mismo en el que se mencionaron anteriormente.
   Para resolver este problema, modele cada variable medida en el vino como una variable aleatoria en un vector aleatorio $X = [X_1, X_2, \ldots, X_{11}]^T$, y encuentre un clasificador ingenuo de Bayes Gaussiano como modelo para clasificar muestras desconocidas.

   a) Asuma que $f_X(x|clase\ 0)$ y $f_X(x|clase\ 1)$ (donde $x \in \mathbb{R}^{11}$) son distribuciones Gaussianas, cada una con su vector de medias y matriz de covarianza. Utilizando los datos en `XtrainVinos.txt` y `YtrainVinos.txt`, estime $f_X(x|clase\ 0)$ y $f_X(x|clase\ 1)$. Escriba las dos ecuaciones resultantes. **No puede utilizar funciones predefinidas para estimar el vector de medias y la matriz de covarianza. Implemente su propia rutina para hacer este cálculo.**

   b) Analice y compare los valores esperados, varianzas, covarianzas entre cada par de variables para cada una de las clases (buena calidad y mala calidad) encontradas en a).

   c) Se tomaron 399 nuevas muestras de vino rojo, guardados en el archivo `XtestVinos.txt`. La calidad del vino es desconocida. Asumiendo que la probabilidad a priori de las clases es la misma, utilizando los modelos estimados en a) y la regla MAP, determine la calidad de cada uno de las 399 muestras de vino. Analice los resultados. **No puede utilizar funciones predefinidas para realizar este procedimiento de clasificación. Se recomienda realizar la clasificación utilizando el logaritmo natural de las distribuciones evaluadas en cada una de las observaciones para evitar problemas numéricos.**

   d) Después de tener catadores que evaluaron estas 399 muestras, se obtienen los resultados de calidad, es decir, los valores considerados como reales de las clases. Estos se guardaron en el archivo `YtestVinos.txt`. Cuente cuántas fallas tuvo el clasificador en el enunciado c) con respecto a estos valores reales de las clases. Calcule el error de clasificación como el número de fallas dividido por el total de muestras que es 399.

Bono

3. (10 puntos) Opcional: Intente reducir este error de clasificación del punto 2c al utilizar el algoritmo EM para tener un mejor estimado de las distribuciones $f_X(x|clase\ 0)$ y $f_X(x|clase\ 1)$. **No puede utilizar funciones predefinidas para encontrar la distribución combindada. Implemente su propio algoritmo EM.**

[2] Tomado de UCI Machine Learning Repository del Center for Machine Learning and Intelligent Systems