Juan Pablo Naranjo
201730006
Gabriel Mateo Mejía
201718064
Maria Fernanda Martinez
201716576

CASO DE ESTUDIO 8

Procesos Estocásticos - IELE 4010

06-11-2021

Punto 1

Las mediciones de temperatura T (unidades en °C), humedad relativa H (unidades en %), y CO_2 C (partes por millón) en un invernadero real de un cultivo de rosas 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, se tienen en el archivo *greenhouse3d.txt* 300 observaciones conjuntas de las variables de temperatura T (primera columna), humedad H (segunda columna), y CO_2 C (tercera columna).

a)

En la figura 1 se pueden ver las observaciones en espacio tridimensional.

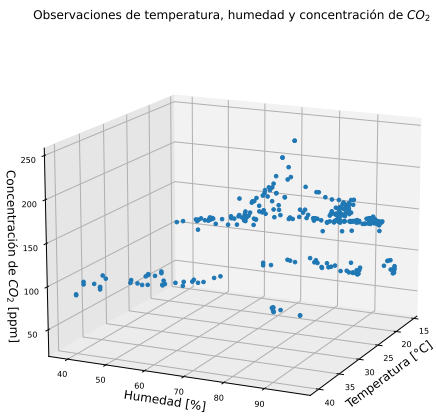

Figura 1: Datos de temperatura, humedad y concentración de CO_2 en 3 dimensiones

Se puede notar una alta concentración de datos alrededor de los 20 °C y 25 °C de temperatura, el 70% y 90% de humedad y entre las 100 y 150 ppm de concentración de dióxido de carbono. También hay otra zona de alta concentración de datos para temperaturas entre los 20 °C y los 36 °C, para concentraciones

de dióxido de carbono inferiores a 100 ppm y humedad relativa entre el 40% y 60%.

b)

Se decidió asumir que el vector Z es un vector aleatorio Gaussiano. Teniendo esto en cuenta, se procedió a usar la estimación MLE para determinar los parámetros de este vector aleatorio, que corresponden a su vector de medias μ_Z y matriz de covarianza Q_Z. Usando la estimación MLE, se sabe que estos dos parámetros van a estar dados por

$$\hat{\mu}_Z = \frac{1}{M}\sum_{i=1}^{M} z_i \tag{1}$$

$$\hat{Q}_Z = \frac{1}{M-1}\sum_{i=1}^{M}(z_i - \mu)(z_i - \mu)^T \tag{2}$$

en donde M es el número total de observaciones que se tiene de cada variable aleatoria y z_i es una observación cualquiera de la forma $z_i = [t_i, h_i, c_i]^T$. Esto quiere decir que la PDF de este vector aleatorio Gaussiano está dada por

$$f_Z(z) = N(z; \hat{\mu}_Z, \hat{Q_Z}) = \frac{1}{2\pi\sqrt{\det \hat{Q_Z}}} \exp\left(-\frac{1}{2}(z-\hat{\mu}_Z)^T \hat{Q_Z}^{-1}(z-\hat{\mu}_Z)\right) \tag{3}$$

donde, realizando los cálculos respectivos:

$$\begin{aligned} \hat{\mu}_Z &= \begin{bmatrix} 24.01 \\ 79.00 \\ 137.50 \end{bmatrix} \\ \hat{Q}_Z &= \begin{bmatrix} 24.085 & -63.57 & 22.11 \\ -63.57 & 224.22 & 167.39 \\ 22.11 & 167.39 & 1753.67 \end{bmatrix} \end{aligned} \tag{4}$$

c)

Ahora se procede a diseñar el detector de fallas en el sistema. Para esto, se definen reglas de decisión para saber si un dato es una observación típica o una observación anómala.

Definimos regla de decisión para observaciones típicas:

$$f_Z(z) \geq \gamma \tag{5}$$

Definimos regla de decisión para observaciones anómalas

$$f_Z(z) < \gamma \tag{6}$$

Para determinar un valor apropiado para el umbral de decisión γ, se procedió a evaluar la función $f_Z(z)$ en cada una de las observaciones originales entregadas en el archivo *greenhouse3d.txt*, para así ver qué valor mínimo toma la función para todos los datos. Conociendo este valor mínimo, se puede establecer el γ en

un valor cercano a este, sabiendo que para cualquier valor de la función $f_Z(z)$ mayor o igual a este, se espera tener una observación típica.

Llevando a cabo este procedimiento, se encontró que, para todos los datos originales, el valor mínimo que tomó la función $f_Z(z)$ fue de aproximadamente 4.74×10^{-9}. Por lo tanto, se decidió elegir un $\gamma = 4.7 \times 10^{-9}$ para la regla de decisión. Siendo así, se determina que una observación será típica si

$$f_Z(z) \geq 4.7 \times 10^{-9} \tag{7}$$

y será anómala si

$$f_Z(z) < 4.7 \times 10^{-9} \tag{8}$$

d)

Ahora se pide reescribir la regla de decisión para observaciones anómalas por medio de una forma cuadrática de la forma

$$(z - \hat{\mu}_Z)^T M (z - \hat{\mu}_Z) > \alpha \tag{9}$$

Es decir, se pide encontrar el valor de la matriz M y la constante α.

En primera instancia, es claro que una forma cuadrática como la mostrada en la desigualdad (9) es la misma encontrada en el término de la exponencial en la definición de la función $f_Z(z)$. Por lo tanto, una elección natural para la matriz M es la inversa de la matriz $\hat{Q}_Z$, obteniendo así $M = \hat{Q}_Z^{-1}$.

Ahora, para encontrar el valor de la constante α se procedió a encontrar para qué valores de z dentro de los datos originales se cumple que $f_Z(z) \geq \gamma$. Una vez teniendo dichos valores de z, se determinó cuánto retornaba la operación $(z-\hat{\mu}_Z)^T \hat{Q}_Z^{-1} (z-\hat{\mu}_Z)$, y dentro de un vector con todos estos valores, se determinó el **mayor** de todos estos, pues este valor mayor indicaría el umbral a partir del cual empiezan a haber mediciones anómalas, lo que significa que este valor máximo es el mismo valor de α. Llevando a cabo este proceso, se encontró que el valor de α es 18.885, por lo que la regla de decisión para observaciones anómalas escrita como una forma cuadrática se ve como

$$(z - \hat{\mu}_Z)^T \hat{Q}_Z^{-1} (z - \hat{\mu}_Z) > 18.885 \tag{10}$$

lo que significa que $M = \hat{Q}_Z^{-1}$ y que $\alpha = 18.885$.

e)

Ahora se tienen unas nuevas observaciones del vector Z dentro del archivo *datosNuevos.txt*. Se procede a usar la regla de decisión definida en el literal c) para determinar el porcentaje de mediciones anómalas que se encontraron dentro de estas mediciones nuevas. Este porcentaje se calculó como

$$\%_{anómalas} = \frac{\#\text{ de anómalas}}{\#\text{ total de observaciones}} \tag{11}$$

y su valor numérico se encontró como 65%.

f)

Ahora se redefinió la función $f_Z(z)$ como una combinación de tres distribuciones Gaussianas, resultando en

$$f_Z(z) = \sum_{k=1}^{3} \alpha_k N(z; \mu_k, Q_k). \tag{12}$$

Para encontrar los parámetros de cada una de estas distribuciones Gaussianas se usó el algoritmo EM. La inicialización de los vectores de medias fue en valores en un rango cercano a los datos (como los descritos en la inspección visual del literal a)), matrices de covarianza como matrices identidad multiplicadas por la constante 100 y α_k como 1/3 cada uno. Al correr el algoritmo, se encontraron los siguientes resultados para cada una de las distribuciones, en donde el subíndice 1 hace referencia a la primera distribución Gaussiana, el subíndice 2 hace referencia a la segunda y el subíndice 3 hace referencia a la tercera:

$$\begin{aligned}
\alpha_1 &= 0.35 \\
\alpha_2 &= 0.36 \\
\alpha_3 &= 0.29 \\
\mu_1 &= \begin{bmatrix} 26.89 \\ 75.62 \\ 174.15 \end{bmatrix} \\
\mu_2 &= \begin{bmatrix} 20.96 \\ 89.70 \\ 149.45 \end{bmatrix} \\
\mu_3 &= \begin{bmatrix} 24.29 \\ 69.81 \\ 78.42 \end{bmatrix} \\
Q_1 &= \begin{bmatrix} 12.59 & -25.27 & -5.98 \\ -25.27 & 87.21 & 45.95 \\ -5.98 & 45.95 & 360.02 \end{bmatrix} \\
Q_2 &= \begin{bmatrix} 1.12 & -4.79 & 2.14 \\ -4.78 & 24.47 & -12.29 \\ 2.14 & -12.29 & 23.98 \end{bmatrix} \\
Q_3 &= \begin{bmatrix} 44.51 & -127.24 & 14.35 \\ -127.24 & 394.81 & -16.42 \\ 14.35 & -16.42 & 271.11 \end{bmatrix}
\end{aligned} \tag{13}$$

Usando esta combinación de distribuciones Gaussianas para modelar el comportamiento de los datos originales, se volvió a determinar el umbral de decisión γ para el cual se formula la regla de decisión para observaciones típicas

$$f_Z(z) \geq \gamma \tag{14}$$

y para observaciones anómalas

$$f_Z(z) < \gamma \tag{15}$$

Siguiendo el mismo procedimiento que se usó para una sola Gaussiana, se evaluó la nueva función $f_Z(z)$ en cada una de las observaciones originales para así ver qué valor mínimo toma la función para todos los datos. De esta forma, conociendo el valor mínimo, se establece el γ en un valor cercano a este. En este caso, el valor mínimo que retornó la función $f_Z(z)$ fue de 1.045×10^{-8}, y por lo tanto el umbral γ que se decidió adoptar en este caso fue de $\gamma = 1.04 \times 10^{-8}$. Siendo así, se determina que una observación será típica si

$$f_Z(z) \geq 1.04 \times 10^{-8} \tag{16}$$

y será anómala si

$$f_Z(z) < 1.04 \times 10^{-8} \tag{17}$$

Aplicando esta regla de decisión para los datos nuevos y usando la combinación de tres Gaussianas $f_Z(z)$, se obtuvo un porcentaje de observaciones anómalas igual al 67% usando el mismo procedimiento que para una sola Gaussiana. En primer lugar, es claro que el porcentaje de observaciones anómalas para una sola Gaussiana fue levemente menor (65%), aunque esto no necesariamente significa que se obtuvo un mejor resultado pues no se sabe con certeza cuáles mediciones fueron anómalas y cuáles no. Sin embargo, sí se sabe que en teoría, la combinación de tres Gaussianas es un mejor estimativo de la función de densidad de probabilidad que gobierna la distribución de los datos originales, pues tiene en cuenta las zonas en las que hay mayor concentración de datos, mientras que una sola Gaussiana no tiene esto en cuenta. Por esta razón, se puede darle más veracidad al resultado obtenido para la combinación de Gaussianas.

En la figura 2 se pueden observar los datos nuevos graficados en $\mathbb{R}^3$, con las observaciones anómalas y normales en diferentes colores para la detección de fallas usando una sola Gaussiana.

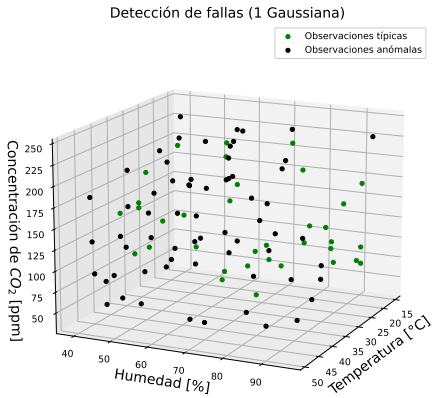

Figura 2: Detección de fallas: 1 Gaussiana

Mientras tanto, en la figura 3 se pueden observar los mismos datos coloreados de acuerdo a su estatus típico o anómalo, pero esta vez para la detección de fallas usando la combinación de tres Gaussianas.

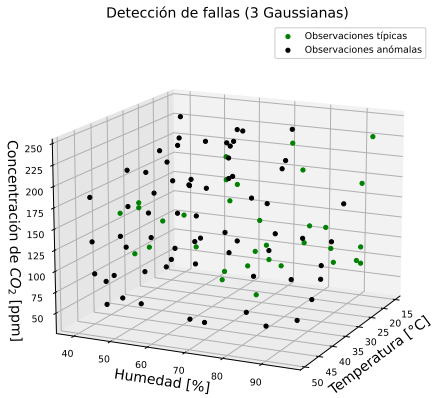

Figura 3: Detección de fallas: combinación de 3 Gaussianas

Comparando las dos anteriores figuras, es claro que tanto el detector de fallas para 1 sola Gaussiana como para la combinación de 3 Gaussianas es muy similar y detecta una cantidad muy similar de fallas en los datos nuevos. Sin embargo, como en teoría la combinación de 3 Gaussianas es una mejor estimación de la PDF que gobierna la distribución de las observaciones reales, se le da más veracidad al detector de fallas diseñado con esta PDF en mente.

Punto 2

Los archivos XtrainVinos.txt y YtrainVinos.txt contienen la información de 1200 muestras de vinos, estas mediciones permiten determinar si el vino es de buena calidad (clase 1) o mala calidad (clase 0).

a)

Utilizando los datos en XtrainVinos.txt y YtrainVinos.txt , se estimaron $f_X(x|clase0)$ y $f_X(x|clase1)$, que son distribuciones Gaussianas, cada una con su vector de medias y matriz de covarianza:

$$\mu_1 = \begin{bmatrix} 0.3039 \\ -0.3549 \\ 0.2757 \\ 0.0214 \\ -0.0588 \\ -0.1258 \\ -0.2383 \\ 0.0203 \\ -0.1199 \\ 0.2260 \\ 0.3750 \end{bmatrix} \tag{18}$$

Cuadro 1: Matriz de covarianza de la clase 1 Q_1

1.1489	-0.2616	0.7864	0.1228	0.0554	-0.1980	-0.1663	0.8202	-0.7670	0.2452	-0.1411
-0.2616	0.8046	-0.5136	0.0208	0.0523	0.0166	0.0479	0.0759	0.2429	-0.1849	-0.1519
0.7864	-0.5136	1.0452	0.1423	0.0873	-0.1431	-0.0486	0.3738	-0.6126	0.2448	0.1676
0.1228	0.0208	0.1423	0.8108	0.0196	0.0184	0.1181	0.2919	-0.0860	-0.0236	0.1029
0.0554	0.0523	0.0873	0.0196	0.6546	-0.0868	-0.0540	0.1953	-0.1409	0.1633	-0.1895
-0.1980	0.0166	-0.1431	0.0184	-0.0868	0.8930	0.5340	-0.1559	0.1700	0.0008	0.0311
-0.1663	0.0479	-0.0486	0.1181	-0.0540	0.5340	0.7242	-0.0820	0.0255	0.0228	-0.0386
0.8202	0.0759	0.3738	0.2919	0.1953	-0.1559	-0.0820	1.2268	-0.3776	0.2123	-0.6487
-0.7670	0.2429	-0.6126	-0.0860	-0.1409	0.1700	0.0255	-0.3776	0.9890	-0.1579	0.1555
0.2452	-0.1849	0.2448	-0.0236	0.1633	0.0008	0.0228	0.2123	-0.1579	0.9087	0.0081
-0.1411	-0.1519	0.1676	0.1029	-0.1895	0.0311	-0.0386	-0.6487	0.1555	0.0081	1.1600

$$\mu_0 = \begin{bmatrix} 0.0247 \\ 0.3106 \\ -0.0739 \\ 0.0142 \\ 0.1513 \\ 0.0173 \\ 0.3089 \\ 0.3347 \\ -0.0368 \\ -0.1682 \\ -0.5215 \end{bmatrix} \tag{19}$$

Cuadro 2: Matriz de covarianza de la clase 0 Q_0

0.8866	-0.1807	0.5566	0.1083	0.0568	-0.0721	-0.0960	0.5318	-0.6193	0.0864	0.0310
-0.1807	0.9932	-0.4860	0.0074	-0.0608	-0.0899	-0.0372	-0.0817	0.2352	-0.1879	0.0336
0.5566	-0.4860	0.9249	0.0736	0.3323	0.0580	0.1598	0.3286	-0.5042	0.2975	-0.0011
0.1083	0.0074	0.0736	0.7967	-0.0428	0.1848	0.1926	0.3016	-0.0333	0.0172	0.1008
0.0568	-0.0608	0.3323	-0.0428	1.4969	0.0453	0.0496	0.0859	-0.3701	0.7532	-0.1219
-0.0721	-0.0899	0.0580	0.1848	0.0453	1.0128	0.8234	0.0675	0.0128	0.1470	-0.0695
-0.0960	-0.0372	0.1598	0.1926	0.0496	0.8234	1.3061	0.0603	-0.1054	0.1890	-0.1389
0.5318	-0.0817	0.3286	0.3016	0.0859	0.0675	0.0603	0.6599	-0.2509	0.1198	-0.1197
-0.6193	0.2352	-0.5042	-0.0333	-0.3701	0.0128	-0.1054	-0.2509	1.0599	-0.3226	0.1658
0.0864	-0.1879	0.2975	0.0172	0.7532	0.1470	0.1890	0.1198	-0.3226	1.1907	-0.0190
0.0310	0.0336	-0.0011	0.1008	-0.1219	-0.0695	-0.1389	-0.1197	0.1658	-0.0190	0.4881

Las dos ecuaciones resultantes con estos parámetros se exponen a continuación:

$$f_X(x) = \frac{1}{\sqrt{det(Q_x)}} \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(x-\mu)^T Q_X^{-1}(x-\mu)}$$
$$f_{clase1}(x) = \frac{1}{\sqrt{det(Q_1)}} \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(x-\mu_1)^T Q_1^{-1}(x-\mu_1)}$$
$$f_{clase0}(x) = \frac{1}{\sqrt{det(Q_0)}} \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(x-\mu_0)^T Q_0^{-1}(x-\mu_0)}$$

b)

Al analizar los resultados se tiene en cuenta que se normalizaron los datos por lo tanto se esperaban tener valores esperados cercanos a 0 y varianzas cercanas a 1 .

Se observa en los vectores correspondientes a los valores esperados seis valores positivos para μ_1 (clase 1) y siete valores positivos para μ_0 (clase 0) de un total de 11 variables, donde se ve que los valores encontrados para cada vector se compensan entre si.

c)

Asumiendo que las probabilidades a priori de cada una de las clases son iguales, la regla MAP de decisión puede escribirse como:

$$\hat{c} = \underset{c \in \{0,1\}}{\operatorname{argmax}} f_X(x|c) \tag{20}$$

Donde las funciones condicionales fueron definidas previamente. Nuestro algoritmo clasificador ingenuo de Bayes predice que de las 399 muestras de prueba 229 son de buena calidad (clase 1).
Se observa que la estimación MAP funciona efectivamente, debido a que los datos que están más cercanos a la media de buena calidad o de mala calidad respectivamente, y debido a que la probabilidad del conocimiento a priori es el mismo en ambas clases, la estimación es equitativa entre los datos de prueba.

d)

Teniendo en cuenta los resultados del anterior inciso, se evidencia que el clasificador cometió 119 fallas. Calculando el error de clasificación como $Error = \frac{Fallas}{Totalmuestras}$ se obtiene un porcentaje de error del 29.82%.

Punto 3 (Bono)

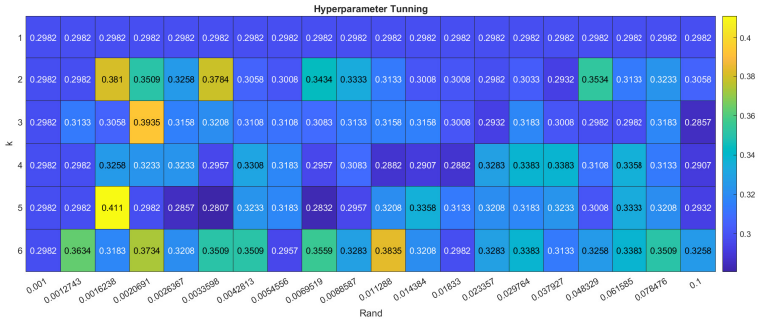

Figura 4: Experimentación con hiperparametros.

Realizando una búsqueda extensiva de hiperparametros, ajustando la distribución inicial, el mejor desempeño global obtenido fue un porcentaje de error del 28.07% mejorando casi 2 puntos porcentuales el desempeño al usar una sola gaussiana. Esto significa que el clasificador cometió un total de 95 fallos reduciendo el valor previo en 24 errores.