Tarea 1

Machine Learning
Rodrigo S. Cortez Madrigal

UNAM iimas INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS

Un estudiante de posgrado tiene como tarea clasificar a los pacientes en dos categorías, sano (S) y enfermo (E) basándose en sus características médicas. Las características incluyen la temperatura corporal, la presión arterial y la comorbilidad. Los registros disponibles son los siguientes:

Paciente	Temperatura (°C)	Presión arterial (mmHg)	Comorbilidad	Clasificación
1	36.5	120	Obesidad	S
2	37.2	140	Obesidad	E
3	36.8	130	Ansiedad	S
4	37.5	150	Asma	E
5	36.9	125	Asma	S
6	37.0	145	Arritmia	E
7	35.3	129	Ansiedad	S
8	37.1	141	Obesidad	E
9	34.9	153	Ansiedad	S
10	38.1	151	Obesidad	E
11	36.7	125	Ansiedad	S
12	36.5	158	Asma	E
13	37.22	160	Arritmia	E

Entrena un clasificador bayesiano ingenuo usando estimación por máxima verosimilitud y otro usando estimación por máximo a posteriori. Reporta los parámetros que obtuviste en ambos casos y usa los clasificadores entrenados para predecir la clase de los siguientes vectores:

x1 = (35.9, 143, Obesidad), x2 = (36.0, 140, Asma), x3 = (37.2, 125, Asma), x4 = (36.4, 120, Arritmia) y x5 = (36.8, 162, Obesidad).

Considera un intervalo de ±.1 para los atributos de temperatura y
presión arterial. Describe de forma detallada el procedimiento que seguiste tanto en el entrenamiento como en la predicción y discute los resultados obtenidos.

Considera un intervalo de ±.1 para los atributos de temperatura y
presión arterial. Describe de forma detallada el procedimiento que seguiste tanto en el entrenamiento como en la predicción y discute los resultados obtenidos.

Para el entrenamiento del clasificador por máximo a posteriori considera los siguientes valores
para las distribuciones correspondientes:

Salud	Comorbilidad	Temp μ_0	Temp σ_0^2	Temp σ^2	Presión arterial μ_0	Presión arterial σ_0^2	Presión arterial σ^2
S	2, ∀k	36.8	0.5	0.7	130	17.0	15.6
E	2, ∀k	38.1	0.6	0.9	150	35.0	71.0

Procedimiento

Preprocesar datos

- Aplicar One-hot encoding

Cálcular probabilidades

- Usar distribuciones normales para modelar temperatura y presión arterial.
- Calcular probabilidades condicionales para la variable categórica *Comorbilidad*.
- Usar la estimación de máxima verosimilitud y máxima a posteriori.

Hacer Inferencia

- Para cada nuevo paciente, calcular la probabilidad de pertenecer a cada clase (*S* o *E*) y asignar la clase con mayor probabilidad.

Preprocesamiento

Dado que estamos usando *Naïve Bayes*, tratamos cada atributo de forma independiente, por lo que la comorbilidad debe manejarse como una probabilidad discreta en función de la clase.

Paciente	Temperatura (°C)	Presión arterial (mmHg)	Obesidad	Ansiedad	Asma	Arritmia	Clasificación
1	36.5	120	1	0	0	0	S
2	37.2	140	1	0	0	0	E
3	36.8	130	0	1	0	0	S
4	37.5	150	0	0	1	0	E
5	36.9	125	0	0	1	0	S
6	37.0	145	0	0	0	1	E
7	35.3	129	0	1	0	0	S
8	37.1	141	1	0	0	0	E
9	34.9	153	0	1	0	0	S
10	38.1	151	1	0	0	0	E
11	36.7	125	0	1	0	0	S
12	36.5	158	0	0	1	0	E
13	37.22	160	0	0	0	1	E

Cálculo de parámetros por Máxima Verosimilitud (MLE)

Ahora necesitamos calcular los parámetros del MLE, es decir μ y σ^2 para cada variable.

$$\mu_C = \frac{1}{N_C} \sum_{i=1}^{N_C} x_i$$

$$\sigma_C^2 = \frac{1}{N_C} \sum_{i=1}^{N_C} (x_i - \mu_C)^2$$

Para cada clase (*S* y *E*), debemos estimar la media (μ) y la varianza (σ^2) de los atributos.

Temperaturas

Pacientes sanos: 1, 3, 5, 7, 9, 11
Temperaturas: 35.6, 36.8, 36.9, 35.3, 34.9, 36.7

$$\mu_S = \frac{36.5 + 36.8, 36.9 + 35.3 + 34.9 + 36.7}{6} = 36.183$$

$$\sigma_S^2 = \frac{(36.5 - \mu_S)^2 + (36.8 - \mu_S)^2, (36.9 - \mu_S)^2 + (35.3 - \mu_S)^2 + (34.9 - \mu_S)^2 + (36.7 - \mu_S)^2}{6} = 0.61472$$

Pacientes enfermos: 2, 4, 6, 8, 10, 12, 13
Temperaturas: 37.2, 37.5, 37.0, 37.1, 38.1, 36.5, 37.22

$$\mu_E = \frac{37.2 + 37.5 + 37.0 + 37.1 + 38.1 + 36.5 + 37.22}{7} = 37.23$$

$$\sigma_E^2 = \frac{(37.2 - \mu_E)^2 + (37.5 - \mu_E)^2, (37.0 - \mu_E)^2 + (37.1 - \mu_E)^2 + (38.1 - \mu_E)^2 + (36.5 - \mu_E)^2 + (37.22 - \mu_E)^2}{7} = 0.2048$$

Presión arterial

Pacientes sanos: 1, 3, 5, 7, 9, 11
PA: 120, 130, 125, 129, 153, 125

$$\mu_S = \frac{120 + 130 + 125 + 129 + 153 + 125}{6} = 130.33$$

$$\sigma_S^2 = \frac{(120 - \mu_S)^2 + (130 - \mu_S)^2, (125 - \mu_S)^2 + (129 - \mu_S)^2 + (153 - \mu_S)^2 + (125 - \mu_S)^2}{6} = 113.22$$

Pacientes enfermos: 2, 4, 6, 8, 10, 12, 13
PA: 140, 150, 145, 141, 151, 158, 160

$$\mu_E = \frac{140 + 150 + 145 + 141 + 151 + 158 + 160}{7} = 149.29$$

$$\sigma_E^2 = \frac{(140 - \mu_S)^2 + (150 - \mu_S)^2, (145 - \mu_S)^2 + (141 - \mu_S)^2 + (151 - \mu_S)^2 + (158 - \mu_S)^2 + (160 - \mu_S)^2}{7} = 52.49$$

Comorbilidades

Después de haber preprocesado los datos, podemos utilizar una dist categórica. Para esto calculamos las q_k con $k = 4$

Recordemos que esto viene de

$$P(\text{Comorbilidad}|C) = \frac{\text{Número de veces que aparece la comorbilidad en la clase } C}{\text{Número total de muestras en la clase } C}$$

que en clase vimos como $q_k = \frac{1}{n} C_k$

Para los Sanos

$q_1 = P(\text{Obesidad}|S) = \frac{1}{6}$
$q_2 = P(\text{Ansiedad}|S) = \frac{2}{6}$
$q_3 = P(\text{Asma}|S) = \frac{1}{6}$
$q_4 = P(\text{Arritmia}|S) = \frac{0}{6} = 0$

Ahora, pensando en S y E como una variable Bernoulli entonces tenemos que

$$q_S = \frac{6}{13}$$

Para los Enfermos

$q_1 = P(\text{Obesidad}|E) = \frac{3}{7}$
$q_2 = P(\text{Ansiedad}|E) = 0$
$q_3 = P(\text{Asma}|E) = \frac{2}{7}$
$q_4 = P(\text{Arritmia}|E) = \frac{2}{7}$

$$q_E = \frac{7}{13}$$

Cálculo de parámetros por Máximo a posteriori (MAP)

$$\mu_C^{MAP} = \frac{\sigma^2 \mu_0 + N_C \sigma_0^2 \bar{x}}{\sigma^2 + N_C \sigma_0^2}$$

$$\sigma_C^{2\,MAP} = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + N_C \sigma_0^2}$$

Recordemos que:

Salud	Comorbilidad	Temp μ_0	Temp σ_0^2	Temp σ^2	Presión arterial μ_0	Presión arterial σ_0^2	Presión arterial σ^2
S	2, ∀k	36.8	0.5	0.7	130	17.0	15.6
E	2, ∀k	38.1	0.6	0.9	150	35.0	71.0

Temperaturas

Pacientes sanos: 1, 3, 5, 7, 9, 11
Temperaturas: 35.6, 36.8, 36.9, 35.3, 34.9, 36.7

Salud	Temp μ_0	Temp σ_0^2	Temp σ^2
S	36.8	0.5	0.7

$$\mu_S = \frac{(0.5)(36.5 + 36.8, 36.9 + 35.3 + 34.9 + 36.7) + ((0.7)(36.8))}{(6)(0.5) + (0.7)} = 36.3$$

$$\sigma_S^2 = 0.7$$

Pacientes enfermos: 2, 4, 6, 8, 10, 12, 13
Temperaturas: 37.2, 37.5, 37.0, 37.1, 38.1, 36.5, 37.22

Salud	Temp μ_0	Temp σ_0^2	Temp σ^2
E	38.1	0.6	0.9

$$\mu_E = \frac{(0.6)(37.2 + 37 + 37.0 + 37.1 + 38.1 + 36.5 + 37.22) + ((0.9)(38.1))}{(0.6)(7) + (0.9)} = 37.384$$

$$\sigma_E^2 = 0.9$$

Presión arterial

Pacientes sanos: 1, 3, 5, 7, 9, 11
PA: 120, 130, 125, 129, 153, 125

Salud	Presión arterial μ_0	Presión arterial σ_0^2	Presión arterial σ^2
S	130	17.0	15.6

$$\mu_S = \frac{(17)(120+130+125+129+153+125) + ((15.6)(130))}{(6)(17)+(15.6)} = 130.28$$

$$\sigma_S^2 = 15.6$$

Pacientes enfermos: 2, 4, 6, 8, 10, 12, 13
PA: 140, 150, 145, 141, 151, 158, 160

Salud	Presión arterial μ_0	Presión arterial σ_0^2	Presión arterial σ^2
E	150	35.0	71.0

$$\mu_E = \frac{(0.5)(140+150+145+141+151+158+160) + ((71)(150))}{(7)(35)+(71)} = 149.44$$

$$\sigma_E^2 = 71$$

Comorbilidades

Recordemos que

$$q_k = \frac{C_k + \alpha_k - 1}{n + \sum_{k=1}^{k} \alpha_k - K}$$

y

$$q_C = \frac{\sum_{i=1}^{n} X^i + \alpha_k - 1}{n + \beta + \alpha_k - 2}$$

Sanos

Consideremos que $\alpha_k = 2, \forall_k \implies \forall_k - 1$, $n = 6$, $K = 4$
entonces el denominador es $6 + 4(2) - 4 = 10$ y para el nominador $C_k + (2) - 1$ = $C_k + 1$

$q_1 = P(\text{Obesidad}|S) = \frac{1+1}{10}$
$q_2 = P(\text{Ansiedad}|S) = \frac{4+1}{10}$
$q_3 = P(\text{Asma}|S) = \frac{2}{10}$
$q_4 = P(\text{Arritmia}|S) = \frac{1}{10}$

Ahora hay que usar Bernoulli

$q_S = \frac{6+2-1}{13+2+2-2} = \frac{7}{15}$

Enfermos

Consideremos que $\alpha_k = 2, \forall_k \implies \forall_k - 1$, $n = 6$, $K = 4$
entonces el denominador es $7 + 4(2) - 4 = 11$ y para el nominador $C_k + (2) - 1$ = $C_k + 1$

$q_1 = P(\text{Obesidad}|E) = \frac{4}{11}$
$q_2 = P(\text{Ansiedad}|E) = \frac{1}{11}$
$q_3 = P(\text{Asma}|E) = \frac{3}{11}$
$q_4 = P(\text{Arritmia}|E) = \frac{3}{11}$

$q_E = \frac{7+2-1}{13+2+2-2} = \frac{8}{15}$

Inferencia

Recordemos que

$P(C|x) \propto P(x|C)P(C)$

que dado el vector y los calculos anteriores en realidad es

$P(C|x) \propto P(x[1]|C)P(x[2]|C)P(x[3]|C)P(C)$

y que

$$P(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Con MLE

1. $x_1 = [35.9, 143, [1, 0, 0, 0]]$ # Obesidad

Temperatura

$P(x_1[1]|S)$ = $P(x_1[1]|36.1, 0.61)$ = P(35.9 | 36.1, 0.61) = $\frac{1}{\sqrt{2\pi(0.61)}} e^{-\frac{(35.9-36.1)^2}{2(0.61)}}$

No obstante el problema requiere que calculemos la probabilidad de que una observación x_1 caiga dentro del intervalo $[x[i] - 0.1, x[i] + 0.1]$.

Para este caso en el intervalo $[36.0, 36.2]$

$z_{\text{inferior}} = \frac{36.0-36.1}{\sqrt{0.61}} = \frac{-0.1}{\sqrt{0.61}}$

$z_{\text{superior}} = \frac{36.2-36.1}{\sqrt{0.61}} = \frac{0.1}{\sqrt{0.61}}$

Entonces c

$P(36.0 \leq x_1[1] \leq 36.2|C) = \Phi(z_{\text{superior}}) - \Phi(z_{\text{inferior}})$

que en python podemos calcular así

```python
import scipy.stats as stats
import math

x = float(input("Introduce el valor de x: "))
mu = float(input("Introduce la media: "))
sigma = math.sqrt(float(input("Introduce la varianza: ")))
intervalo = float(input("Introce el intervalo de confianza:"))

x_inferior = x - intervalo
x_superior = x + intervalo

z_inferior = (x_inferior - mu) / sigma
z_superior = (x_superior - mu) / sigma

P_inferior = stats.norm.cdf(z_inferior)
P_superior = stats.norm.cdf(z_superior)

P = P_superior - P_inferior

print("P(%.2f < X < %.2f) = %.4f" % (x_inferior, x_superior, P))
```

Entonces

$P(x_1[1]|S)$ = P(35.9 | 36.1, 0.61) = 0.0994
$P(x_1[2]|S)$ = P(143 | 130.33, 113.22) = 0.0037
$P(x_1[3]|S)$ = 1/6
P(S) = 6/13

$P(S|x) \propto (0.0994) * (0.0037) * (1/6) * (6/13) = 2.829076923076923e-05$

Enfermo

$P(x_1[1]|E)$ = P(35.9 | 37.23, 0.2048) = 0.0025
$P(x_1[2]|E)$ = P(143 | 149.29, 52.49) = 0.0076
$P(x_1[3]|E)$ = 3/7
P(E) = 7/13

$P(E|x) \propto (0.0025) * (0.0076) * (3/7) * (7/13) = 4.384615384615385e-06$

Entonces $P(E|x) < P(S|x)$ por lo tanto lo clasificamos como Saludable.

2. $x_2 = [36.0, 140, [0, 0, 1, 0]]$ # Asma

$P(x_1[1]|S)$ = P(36.0 | 36.1, 0.61) = 0.1011
$P(x_1[2]|S)$ = P(140 | 130.33, 113.22) = 0.0050
$P(x_1[3]|S)$ = 1/6
P(S) = 6/13

$P(S|x) \propto (0.1011) * (0.0050) * (1/6) * (6/13) = 3.888461538461538e-05$

$P(x_1[1]|E)$ = P(36.0 | 37.23, 0.2048) = 0.0046
$P(x_1[2]|E)$ = P(140 | 149.29, 52.49) = 0.0048
$P(x_1[3]|E)$ = 2/7
P(E) = 7/13

$P(E|x) \propto (0.0046) * (0.0048) * (2/7) * (7/13) = 3.3969230769230765e-06$

Entonces $P(E|x) < P(S|x)$ por lo tanto lo clasificamos como Saludable.

3. $x_3 = [37.2, 125, [0, 0, 1, 0]]$ # Asma

$P(x_1[1]|S)$ = P(37.2 | 36.1, 0.61) = 0.0380
$P(x_1[2]|S)$ = P(125 | 130.33, 113.22) = 0.0066
$P(x_1[3]|S)$ = 1/6
P(S) = 6/13

$P(S|x) \propto (0.1011) * (0.0050) * (1/6) * (6/13) = 3.888461538461538e-05$

$P(x_1[1]|E)$ = P(37.2 | 37.23, 0.2048) = 0.1745
$P(x_1[2]|E)$ = P(125 | 149.29, 52.49) = 0.0001
$P(x_1[3]|E)$ = 2/7
P(E) = 7/13

$P(E|x) \propto (0.1745) * (0.0001) * (2/7) * (7/13) = 2.6846153846153843e-06$

Entonces $P(E|x) < P(S|x)$ por lo tanto lo clasificamos como Saludable.

4. $x_4 = [36.4, 120, [0, 0, 0, 1]]$ # Arritmia

$P(x_1[1]|S)$ = P(36.4 | 36.1, 0.61) = 0.09467306
$P(x_1[2]|S)$ = P(120 | 130.33, 113.22) = 0.00468078
$P(x_1[3]|S)$ = 0
P(S) = 6/13

$P(S|x) \propto (0.09467306) * (0.00468078) * (0) * (6/13) = 0$

$P(x_1[1]|E)$ = P(36.4 | 37.23, 0.2048) = 0.03342467
$P(x_1[2]|E)$ = P(120 | 149.29, 52.49) = 0.00000311
$P(x_1[3]|E)$ = 2/7
P(E) = 7/13

$P(E|x) \propto (0.03342467) * (0.00000311) * (2/7) * (7/13)$

Entonces $P(E|x) > P(S|x)$ por lo tanto lo clasificamos como Enfermo.

5. $x_5 = [36.8, 162, [1, 0, 0, 0]]$ **# Obesidad**

$P(x_1[1]|S)$ = P(36.8 | 36.1, 0.61) = 0.06832990
$P(x_1[2]|S)$ = P(162 | 130.33, 113.22) = 0.00008941
$P(x_1[3]|S)$ = 1/6
P(S) = 6/13

$P(S|x) \propto (0.06832990) * (0.00008941) * (1/6) * (6/13) = 4.6995202761538457e-07$

$P(x_1[1]|E)$ = P(36.8 | 37.23, 0.2048) = 0.11216836
$P(x_1[2]|E)$ = P(162 | 149.29, 52.49) = 0.00236393
$P(x_1[3]|E)$ = 3/7
P(E) = 7/13

$P(E|x) \propto (0.11216836) * (0.00236393) * (3/7) * (7/13) = 6.119034259726153e-05$

Entonces $P(E|x) > P(S|x)$ por lo tanto lo clasificamos como Enfermo.

Con MAP

1. $x_1 = [35.9, 143, [1, 0, 0, 0]]$ **# Obesidad**

$P(x_1[1]|S)$ = P(35.9 | 36.3, 0.7) = 0.08491027
$P(x_1[2]|S)$ = P(143 | 130.28, 15.6) = 0.00011314
$P(x_1[3]|S)$ = 1/5
P(S) = 7/15

$P(S|x) \propto (0.08491027) * (0.00011314) * (1/5) * (7/15) = 8.966298084613334e-07$

$P(x_1[1]|E)$ = P(35.9 | 37.384, 0.9) = 0.02481017
$P(x_1[2]|E)$ = P(143 | 149.446, 71) = 0.00706687
$P(x_1[3]|E)$ = 3/7
P(E) = 8/15

$P(E|x) \propto (0.02481017) * (0.00706687) * (3/7) * (8/15) = 4.007548481552e-05$

Entonces $P(E|x) > P(S|x)$ por lo tanto lo clasificamos como Enfermo.

2. $x_2 = [36.0, 140, [0, 0, 1, 0]]$ **# Asma**

$P(x_1[1]|S)$ = P(36.0 | 36.3, 0.7) = 0.08924250
$P(x_1[2]|S)$ = P(140 | 130.28, 15.6) = 0.00097837
$P(x_1[3]|S)$ = 1/5
P(S) = 7/15

$P(S|x) \propto (0.08924250) * (0.00097837) * (1/5) * (7/15) = 8.149137241e-06$

$P(x_1[1]|E)$ = P(36.0 | 37.384, 0.9) = 0.02907849
$P(x_1[2]|E)$ = P(140 | 149.446, 71) = 0.00505151
$P(x_1[3]|E)$ = 3/11
P(E) = 8/15

$P(E|x) \propto (0.02907849) * (0.00505151) * (3/11) * (8/15) = 2.136585934834909e-05$

Entonces $P(E|x) > P(S|x)$ por lo tanto lo clasificamos como Enfermo.

3. $x_3 = [37.2, 125, [0, 0, 1, 0]]$ # Asma

$P(x_1[1]|S)$ = P(37.2 | 36.3, 0.7) = 0.05349106
$P(x_1[2]|S)$ = P(125 | 130.28, 15.6) = 0.00826714
$P(x_1[3]|S)$ = 1/5
P(S) = 7/15

$P(S|x) \propto (0.05349106) * (0.00826714) * (1/5) * (7/15) = 4.127368763171734e-05$

$P(x_1[1]|E)$ = P(37.2 | 37.384, 0.9) = 0.08239043
$P(x_1[2]|E)$ = P(125 | 149.446, 71) = 0.00014082
$P(x_1[3]|E)$ = 3/11
P(E) = 8/15

$P(E|x) \propto (0.08239043) * (0.00014082) * (3/11) * (8/15) = 1.6875956876509091e-06$

Entonces $P(E|x) < P(S|x)$ por lo tanto lo clasificamos como Saludable.

4. $x_4 = [36.4, 120, [0, 0, 0, 1]]$ # Arritmia

$P(x_1[1]|S)$ = P(36.4 | 36.3, 0.7) = 0.09446494
$P(x_1[2]|S)$ = P(120 | 130.28, 15.6) = 0.00068334
$P(x_1[3]|S)$ = 1/10
P(S) = 7/15

$P(S|x) \propto (0.09446494) * (0.00068334) * (1/10) * (7/15) = 3.0124113646480004e-06$

$P(x_1[1]|E)$ = P(36.4 | 37.384, 0.9) = 0.04912053
$P(x_1[2]|E)$ = P(120 | 149.446, 71) = 0.00002111
$P(x_1[3]|E)$ = 3/11
P(E) = 8/15

$P(E|x) \propto (0.04912053) * (0.00002111) * (3/11) * (8/15) = 1.5082682011636363e-07$

Entonces $P(E|x) < P(S|x)$ por lo tanto lo clasificamos como Saludable.

5. $x_5 = [36.8, 162, [1, 0, 0, 0]]$ # Obesidad

$P(x_1[1]|S)$ = P(36.8 | 36.3, 0.7) = 0.07964783
$P(x_1[2]|S)$ = P(162 | 130.28, 15.6) = 0.000000001
$P(x_1[3]|S)$ = 1/5
P(S) = 7/15

$P(S|x) \propto (0.07964783) * (0.000000001) * (1/5) * (7/15) = 7.433797466666667e-12$

$P(x_1[1]|E)$ = P(36.8 | 37.384, 0.9) = 0.06950743
$P(x_1[2]|E)$ = P(162 | 149.446, 71) = 0.00312111
$P(x_1[3]|E)$ = 4/11
P(E) = 8/15

$P(E|x) \propto (0.06950743) * (0.00312111) * (4/11) * (8/15) = 4.2073277061294536e-05$

Entonces $P(E|x) > P(S|x)$ por lo tanto lo clasificamos como Enfermo.