**Guía de ejercicios Capítulo V**

**Clasificación Bayesiana**

1. **Objetivos de la Unidad**

| **#** | **Descripción** | **Preguntas** |
| --- | --- | --- |
| 1. | Cuantificar probabilidad a priori. |  |
| 2. | Comprender el costo del error de clasificar basado en probabilidad a priori. |  |
| 3. | Cuantificar el riesgo condicional. |  |
| 4. | Usar los conceptos anteriores para obtener un método de clasificación a mínimo riesgo condicional para un problema multivariado. |  |
| 5. | Analizar clasificador Bayesiano simple. |  |
| 6. | Obtener un clasificador mediante criterio distribuido, usando el concepto de redes de clasificación. |  |

1. **Preguntas**

P1. Accidentabilidad conductores

Diseñe un clasificador Bayesiano ingenuo para el siguiente problema. Una compañía de seguros, especializadas en pólizas de automóviles, se requiere tener un modelo que determina si el postulante a un seguro tendrá o no accidente, basado en las variables sexo (M=*masculino*; F=*femenino*) y el tipo de auto que conduce (P=*particular*; T=*trabajo*).

Para diseñar el clasificador se cuenta con la siguiente BD, donde se indica si el cliente ha tenido accidentes previamente:

| **Sexo** | **Tipo** | **Accidente** |
| --- | --- | --- |
| M | P | Sí |
| F | P | No |
| M | P | Sí |
| F | T | No |
| M | T | No |
| F | P | No |
| M | T | Sí |
| F | P | Sí |

P2. Clasificación genes

Para saber un determinado tipo de gen, puede ser “expresado” (Si / No) a partir de dos características principales: tipo de *nucleótido* (A=*Amina*; T=*Tiamina*; C=*Citosina*; G=*Guanina*), o base que se encuentra al inicio, y tipo de *aminoácido* que se encuentra al final, y puede ser de tres tipos (H=*Histidina*; M=*Metionina*; L=*Lisina*).

Aplicando un clasificador Bayesiano Ingenuo en la BD incluida en la siguiente tabla, a) obtenga la “expresión” de genes para todos los casos que comienzan con el nucleótido *Amina*; b) identifique las combinaciones que no poseen ejemplos; y c) mencione una alternativa para obtener las probabilidades de este tipo de casos.

| **Nucleótido** | **Aminoácido** | **Expresión** |
| --- | --- | --- |
| A | L | Sí |
| C | H | No |
| T | H | No |
| qA | M | Sí |
| T | H | No |
| A | L | Sí |
| T | H | No |
| A | H | No |
| C | L | Sí |
| T | M | No |
| T | M | No |
| C | L | Sí |
| A | H | No |
| G | H | Sí |
| A | H | No |
| G | H | No |
| A | H | No |
| C | M | No |
| A | M | No |
| T | M | No |

P3. Evaluación de riesgo en pacientes

Estime las probabilidades para diseñar un clasificador Bayesiano Ingenuo que clasifique el riesgo de los pacientes (riesgoso / no riesgoso), utilizando los datos de sus *edades* y *tipos de cirugías* (codificados en 1, 2 o 3), incluidos en la siguiente tabla.

Aplicando el clasificador determine si los pacientes de 25 años con cirugía tipo 2 son riesgosos o no.

| **Edad paciente** | **Tipo de cirugía** | **Riesgo** |
| --- | --- | --- |
| 15 | 3 | Riesgoso |
| 55 | 1 | No Riesgoso |
| 36 | 1 | No Riesgoso |
| 20 | 2 | Riesgoso |
| 40 | 1 | No Riesgoso |
| 17 | 3 | Riesgoso |
| 33 | 1 | No Riesgoso |
| 18 | 1 | No Riesgoso |
| 62 | 3 | Riesgoso |
| 26 | 2 | No Riesgoso |
| 38 | 2 | No Riesgoso |
| 57 | 3 | Riesgoso |
| 22 | 3 | No Riesgoso |
| 70 | 1 | Riesgoso |
| 16 | 1 | No Riesgoso |
| 72 | 1 | No Riesgoso |
| 24 | 1 | No Riesgoso |
| 45 | 2 | No Riesgoso |
| 15 | 2 | No Riesgoso |
| 26 | 2 | No Riesgoso |

P4. Expresión génica

La *expresión génica* es el proceso mediante el cual la información de un gen se usa en la síntesis de un *producto génico funcional* que le permite *producir productos finales*, proteínas o ARN no codificante, y finalmente afectar un fenotipo, como efecto final.

Para determinar si un gen se “expresa” *G* se utiliza un índice de expresión (Ie) y la presencia de cajas de regulación (CR). La expresión del gen se ha determinado midiendo la fluorescencia (f) del gen en un microarreglo y se dispone de las verosimilitudes para cada gen.

1. Considerando que para el siguiente conjunto de datos de entrenamiento mostrado el 40% de los genes son expresados, determine basado en las verosimilitudes cuáles casos son los expresados.
2. Posteriormente, diseñe un clasificador Bayesiano ingenuo que determine si un gen es expresado de acuerdo con las variables *Ie* y *CR*, y calcule si un gen es expresado cuando el índice *Ie* es 18 y cuando es 36, y en ambos casos existe caja de regulación (CR=1).
3. Muestre las expresiones literales antes de calcular.

| ***Ie*** | ***CR*** | **Verosimilitud de *f* con *G*** |
| --- | --- | --- |
| 40 | 1 | 0,160 |
| 20 | 1 | 0,004 |
| 28 | 0 | 0,040 |
| 47 | 0 | 0,170 |
| 37 | 1 | 0,100 |
| 32 | 0 | 0,072 |
| 41 | 1 | 0,017 |
| 38 | 0 | 0,105 |
| 15 | 1 | 0,0016 |
| 27 | 1 | 0,040 |
| 38 | 0 | 0,110 |
| 42 | 0 | 0,160 |

P5. Evaluación de riesgo de accidentes de tránsito

La SIAT de Carabineros de Chile posee una reducida base de datos (incluida en la siguiente tabla), con las variables edad del conductor y el estado del camino (Bueno / Malo) para predecir el riesgo de accidentes de tránsito.

1. Estime las funciones o proporciones de probabilidades para diseñar un clasificador Bayesiano ingenuo que permita estimar si un conductor tendrá accidente al circular por un determinado camino.
2. Determine además si un conductor con 25 años y transitando por un camino en buen estado producirá accidente.

| **Edad** | **Estado camino** | **Accidente** |
| --- | --- | --- |
| 40 | Malo | Sí |
| 20 | Malo | No |
| 28 | Bueno | No |
| 47 | Bueno | Sí |
| 37 | Malo | Sí |
| 32 | Bueno | No |
| 41 | Malo | No |
| 38 | Bueno | No |
| 15 | Malo | No |
| 27 | Malo | No |
| 38 | Bueno | Sí |
| 42 | Bueno | Sí |

P6. Diagnóstico de diabetes

Para evaluar la diabetes Mellitus existen dos variables de interés, el nivel sobrepeso y el nivel de glucosa en la sangre del paciente. De un hospital se tiene la siguiente tabla.

| **Nivel de sobrepeso** | **Glucosa en mg/100 ml de sangre** | **Diabético** |
| --- | --- | --- |
| Normal | 70 | No |
| Normal | 80 | No |
| Sobrepeso | 90 | No |
| Normal | 100 | No |
| Obeso | 110 | No |
| Sobrepeso | 120 | Sí |
| Sobrepeso | 130 | Sí |
| Obeso | 140 | Sí |
| Obeso | 150 | Sí |

Utilizando un clasificador Bayesiano ingenuo, determine si un paciente obeso con 90 mg/100ml de glucosa en la sangre, será diabético o no.

P7. Trastornos bipolares en apostadores

El profesor Wael El-Deredy de *electrofisiología cognitiva* de la universidad de Manchester, estudia los trastornos bipolares mediante experimentos de juegos de azar. Para modelar estos trastornos plantea un modelo de toma de decisiones Bayesiano. Aquí las probabilidades a priori corresponden a las expectativas que tienen los sujetos de ganar y las probabilidades a posteriori es lo que determina finalmente si se apostará o no. Para probar su hipótesis genera un experimento con dos grupos de sujetos (bipolares y sanos), los que realizan apuestas en un juego al azar, según el nivel de dinero apostado.

Del experimento se obtienen las siguientes probabilidades a priori en función del dinero apostado para sujetos como: las probabilidades a priori para los bipolares son:

Además se tiene la siguiente tabla de verosimilitudes:

| **Nivel de apuesta (d)** | **Verosimilitud Bipolares** | **Verosimilitud Normales** |
| --- | --- | --- |
| 1000 | 0,90 | 0,94 |
| 900 | 0,93 | 0,85 |
| 800 | 0,87 | 0,89 |
| 700 | 0,86 | 0,86 |
| 500 | 0,95 | 0,95 |
| 400 | 0,91 | 0,88 |
| 300 | 0,89 | 0,87 |
| 200 | 0,87 | 0,89 |
| 100 | 0,82 | 0,91 |

Determine la decisión de apostar para cada nivel de apuesta, tanto para bipolares y como para sanos. Con estos resultados identifique si hay diferencias entre bipolares y sanos. Si estas diferencias existen, indique a su juicio cuáles son las causas de estas diferencias, en términos de las expectativas de los sujetos.

1. **Respuestas**

P1. Accidentabilidad conductores

| **#** | **Sexo** | **Tipo** | **Accidente** |
| --- | --- | --- | --- |
| 1 | M | P | Sí |
| 2 | F | P | No |
| 3 | M | P | Sí |
| 4 | F | T | No |
| 5 | M | T | No |
| 6 | F | P | No |
| 7 | M | T | Sí |
| 8 | F | P | Sí |

Nro observaciones = 8

**Probabilidad a priori** *P(Ci)*

P(Accidente=Sí) = 4/8 = 0,5 = 50%

P(Accidente=No) = 4/8 = 0,5 = 50%

Verosimilitudes *P(x|Ci)*

P(Sexo=M | Accidente=Sí) = 3/4 = 0,75 = 75%

P(Sexo=M | Accidente=No) = 1/4 = 0,25 = 25%

P(Sexo=F | Accidente=Sí) = 1/4 = 0,25 = 25%

P(Sexo=F | Accidente=No) = 3/4 = 0,75 = 75%

P(Tipo=P | Accidente=Sí) = 3/4 = 0,75 = 75%

P(Tipo=P | Accidente=No) = 2/4 = 0,50 = 50%

P(Tipo=T | Accidente=Sí) = 1/4 = 0,25 = 25%

P(Tipo=T | Accidente=No) = 2/4 = 0,50 = 50%

**Ejemplo Clasificar:** <Sexo=F, TIpo=T>

ClaseNB = argmax P(ci) ∏ P(xj | ci)

P(Accidente=Sí) P(Sexo=F|Accidente=Sí) P(Tipo=T|Accidente=Sí) = (0,5) (0,25) (0,25) = 0,03125

P(Accidente=No) P(Sexo=F|Accidente=No)P(Tipo=T|Accidente=No) = (0,5) (0,75) (0,50) = 0,1875

**Solución:** ClaseNB = argmax { 0,03125; 0,1875 } => Accidente=No

P3. Evaluación de riesgo en pacientes

**Probabilidad a priori** *P(Ci)*

P(Riesgo=No Riesgoso) = 14/20 = 0,7 = 70%

P(Riesgo=Riesgoso) = 6/20 = 0,3 = 30%

Verosimilitudes *P(x|Ci)*

Edad Paciente|No Riesgoso = { 55, 36, 40, 33, 18, 26, 38, 22, 16, 72, 24, 45, 15, 26 }

Media = X̄(Edad Paciente|Riesgo=No Riesgoso) = 33,3

Desviación estándar = σ(Edad Paciente|Riesgo=No Riesgoso) = 16,1

Distribución (normal) Edad Paciente = *N*(X̄=33.3, σ=16.1)

Edad Paciente|Riesgoso = { 15, 20, 17, 62, 57, 70 }

Media = X̄(Edad Paciente|Riesgo=Riesgoso) = 40,2

Desviación estándar = σ(Edad Paciente|Riesgo=Riesgoso) = 25,4

Distribución (normal) Edad Paciente = *N*(X̄=33.3, σ=16.1)

**Función de Densidad de Probabilidad (Normal) Gaussiana**

P(Tipo de cirugía=1|No Riesgoso) = 8/14 = 0,571 = 57,1%

P(Tipo de cirugía=1|Riesgoso) = 1/6 = 0,167 = 16,7%

P(Tipo de cirugía=2|No Riesgoso) = 5/14 = 0,357 = 35,7%

P(Tipo de cirugía=2|Riesgoso) = 1/6 = 0,167 = 16,7%

P(Tipo de cirugía=3|No Riesgoso) = 1/14 = 0,071 = 7,1%

P(Tipo de cirugía=3|Riesgoso) = 4/6 = 0,667 = 66,7%

**Clasificar:** <Edad Paciente=25, Cirugia=2>

ClaseNB = argmax P(ci) ∏ P(xj | ci)

P(Riesgo=No Riesgoso) P(Edad Paciente=25 | No Riesgoso) P(Cirugia=2 |No Riesgoso) =

(0,7) (0,021) (0,357) = 0,005

P(Riesgo=Riesgoso) P(Edad Paciente=25 | Riesgoso) P(Cirugia=2 | Riesgoso) =

(0,3) (0,013) (0,167) = 0,0006

**Solución:** ClaseNB = argmax { 0,005; 0,0006 } => Riesgo=No Riesgoso