MINERIA DE DATOS UTILIZANDO SISTEMAS INTELIGENTES

PRACTICA 3 - NAIVE BAYES

Material de Lectura: Capítulo 10 del Libro Introducción a la Minería de Datos de Hernández Orallo

Ejercicios a entregar: 4 y 5 (todos los incisos).

1. **Repaso de distribuciones normal y categórica**

a) Calcular el valor de la función de densidad de probabilidad (*fdp*) correspondiente a una distribución Normal con parámetros $\mu = 10$ y $\sigma = 2$ para los siguientes valores de X

X	$P(X)$
0	
5	
10	
15	
20	

Nota: Para calcular la fdp (*probability density function (pdf)* en inglés), puede utilizar:

- Python: script para calcular la fpd
- DanielsOper: calculadora online (requiere σ como parámetro)
- PlanetCalc: calculadora online (requiere σ^2 como parámetro)
- Python con google colab: script en Python utilizando Numpy/SciPy y Matplotlib para dibujar.

b) Se obtiene una muestra de una variable aleatoria X con los valores $1, 5, 5, 9$. Asumiendo que $XNormal(\mu, \sigma)$, estime los valores de μ y σ en base a la muestra. Recuerde que los estimadores insesgados de μ y σ son:

$$\mu_X = \frac{x_1+x_2+\ldots+x_n}{n} \qquad var_X = \frac{(x_1-\mu_x)^2+(x_2-\mu_x)^2+\ldots+(x_1-\mu_x)^2}{n-1} \qquad \sigma_X = \sqrt{(var_X)}$$

c) La variable aleatoria X tiene una distribución categórica con 4 valores posibles: Bajo, Medio, Alto y Extremo. Se obtiene una muestra de la variable aleatoria X con los valores Alto, Bajo, Extremo, Alto, Bajo, Bajo. Estime la probabilidad de cada valor de X en base a la muestra. Para estimar, utilice la frecuencia relativa de cada valor, es decir, $\frac{cant.muestrasconelvalor}{cant.totaldemuestras}$.

d) Repetir c), pero ahora utilizando la corrección de Laplace con α=1 para estimar las probabilidades.

e) Repetir c), pero ahora utilizando la corrección de Laplace con α=100 para estimar las probabilidades.

X	c) $P(X)$ (α=0)	d) $P(X)$ (α=1)	e) $P(X)$ (α=100)
Bajo			
Medio			
Alto			
Extremo			

2. Cálculo de probabilidades con atributos nominales

Dado el siguiente conjunto de datos:

Calor	**Clase**
Alto	A
Alto	B
Bajo	C
Bajo	B
Bajo	A
Alto	A
Alto	A
Bajo	C

a) Calcular la probabilidad de cada clase, sin modelar el valor de sus atributos

	Clase **A**	Clase **B**	Clase **C**
P(Clase)			

b) La tabla anterior determina un modelo de NB trivial, es decir, que no considera ningún atributo. Utilizando dicho modelo, ¿qué clase se le asigna a cualquier ejemplo?

c) Calcular la probabilidad de cada clase para cada valor del atributo Calor. Utilizar la siguiente tabla para ordenar los resultados

P(Atributo = Valor |Clase)	Clase **A**	Clase **B**	Clase **C**
Calor = Bajo			
Calor = Alto			

d) Observe la tabla de c) ¿los valores de cada columna suman 1? ¿y los de las filas? ¿y la suma de los valores de toda la tabla? ¿por qué?

e) Si se cambian los valores de los ejemplos de la clase A, ¿cambiarán las probabilidades para las clases B y C?

f) La tabla generada en c) es la que utiliza NB para almacenar el modelo. Está compuesta por valores correspondientes a P(Atributo = Valor | Clase). Para comprender mejor esta cantidad, genera ahora dos tablas parecidas pero cuyos valores ahora indiquen P(Clase | Atributo = Valor) y P(Atributo = Valor y Clase) y compará los valores:

P(Clase|Atributo = Valor)	Clase **A**	Clase **B**	Clase **C**
Calor = Bajo			
Calor = Alto			

P(Atributo = Valor y Clase)	Clase **A**	Clase **B**	Clase **C**
Calor = Bajo			
Calor = Alto			

Recordá que en el caso de **P(Atributo = Valor y Clase)**, el espacio muestral es el total de los ejemplos.

g) Para cada una de las tablas generadas en **f)** ¿los valores de cada columna suman 1? ¿y los de las filas? ¿y la suma de los valores de toda la tabla? ¿Por qué? Comparar con la tabla de **c)**.

3. Cálculo de probabilidades con atributos continuos y la distribución normal

Para un atributo continuo es común en NB, aunque no obligatorio, que asumamos que sus valores siguen una distribución normal o gaussiana. Dado el siguiente conjunto de datos:

Calor	Clase
7	A
8	B
3	C
4	B
6	A
9	A

10	A
1	C

a) Calcule los parámetros de la distribución normal (μ y σ) para cada una de las clases, o sea, para los ejemplos que pertenecen a cada una de ellas:

Atributo / Valor	Clase **A**	Clase **B**	Clase **C**
Calor (μ)			
Calor (σ)			

b) Dados los parámetros calculados anteriormente, calcule la probabilidad de que cada valor pertenezca a una clase, utilizando sólo la distribución P(Calor | Clase) (y no P(Clase)).

Valor de Calor	Clase **A**	Clase **B**	Clase **C**
0			
2.5			
5			
7.5			
10			

c) Hacer un gráfico 2D, donde el eje X representa a los valores de cada una de las distribuci las tres distribuciones para ver como separan las clases. ¿En qué puntos dos clases tienen la misma probabilidad?

d) Si se cambian los valores de los ejemplos de la clase A, ¿cambiarán las probabilidades para las clases B o C?

4. Aplicación de un modelo Naive Bayes

Dado el siguiente modelo NB para clasificar las frutas en **Pera** o **Manzana** en base a los atributos **Color** y **Esfericidad:**

Atributo / Valor	Clase **Manzana**	Clase **Pera**
Color = Amarillo	0.8	0.1

Color = Mezcla	0.0	0.6
Color = Rojo	0.2	0.3
Esfericidad (μ)	0.5	0.8
Esfericidad (σ)	0.3	0.2

a) Si ambas clases son equiprobables (las probabilidades de clase a priori son P(**Pera**)=0.5 y P(**Manzana**)=0.5), y no se utiliza corrección de Laplace, indicar cómo clasificaría los siguientes 2 ejemplos, incluyendo los cálculos realizados, en la siguiente tabla:

Color	Esfericidad	P(x |Pera)	P(x| Manzana)	P(x | Pera) * P(Pera)	P(x| Manzana) * P(Manzana)	Predicción
Amarillo	0.6					
Mezcla	0.8					

b) Si las probabilidades de clase a priori son P(**Manzana**)=0.01 y P(**Pera**)=0.99, y no se utiliza corrección de Laplace, indicar cómo clasificaría los siguientes 2 ejemplos, incluyendo los cálculos realizados en la siguiente tabla:

Color	Esfericidad	P(x |Pera)	P(x| Manzana)	P(x | Pera) * P(Pera)	P(x| Manzana) * P(Manzana)	Predicción
Amarillo	0.6					
Mezcla	0.8					

5. Generación de un modelo NB

a) En base a los datos del archivo **estrellas.xslx**, generar un modelo de NB para clasificar su tipo espectral (F o K), sin utilizar corrección de Laplace. No usar Rapidminer o herramienta similar. Incluir sus cálculos.

Atributo / Valor	Clase **F**	Clase **K**
Temperatura μ		
Temperatura σ		
Habitable = Si		

Habitable = No		
Luminosidad μ		
Luminosidadσ		

P(Clase F)	
P(Clase K)	

b) Utilizando el modelo anterior, clasifique los **3 primeros ejemplos** del conjunto de datos **estrellas.xslx**. Utilice una tabla como la siguiente para realizar los cálculos. Utilizar notación científica con 2 decimales para escribir las probabilidades para números menores a 0.01. Por ejemplo, 0.0000436 se escribe como 4.36e-5. y 7.3214123e-12 se redondea a 7.32e-12.

Temperatura	Habitable	Luminosidad	P(x| F)	P(x| K)	P(x| F) * P(F)	P(x| K) * P(K)	Predicción
...	...	...					

c) Las predicciones del modelo ¿coinciden con las etiquetas del dataset? Calcule las predicciones para el resto de los ejemplos (sin llenar la tabla, utilizando RapidMiner).

Luego, calcule el **accuracy** (porcentaje de ejemplos clasificados correctamente) del modelo para ese conjunto de datos. Por ejemplo, dados 10 ejemplos, si el modelo acierta la clase de 7 de ellos, entonces el accuracy es 7/10=0.7 o 70%.

d) Se agrega un ejemplo al conjunto de datos:

Temperatura	Habitable	Luminosidad	Clase
20000	No	35	K

Vuelva a generar el modelo, ahora incluyendo este ejemplo (no es necesario incluir los cálculos). ¿Cómo afecta al modelo de cada atributo o clase?

e) Vuelva a calcular el accuracy para los datos y el modelo del punto **d)** ¿Cambió? Si es así, ¿Qué ejemplo cambió su predicción? ¿Por qué?

6. Atributos útiles para el modelo NB

a) Dado un modelo NB con siguientes distribuciones para los atributos continuos Ancho y Alto ¿qué atributos, individualmente, son más informativos para el modelo (permiten clasificar mejor utilizando solo ese atributo)? Ordenarlos y dibujar las distribuciones para corroborar el resultado.

Atributo / Valor	Clase **A**	Clase **B**
Ancho (μ)	10	20
Ancho (σ)	1	1
Alto (μ)	10	20
Alto (σ)	2	2
Profundo (μ)	5	25
Profundo (σ)	7	7

b) Dado un modelo NB con siguientes distribuciones de clase para los atributos discretos Color y Forma ¿qué atributo es más informativo para cada clase? ¿y para el modelo en general?

Atributo / Valor	Clase **A**	Clase **B**	Clase **C**
Color = Rojo	0.70	0.10	0.50
Color = Verde	0.10	0.75	0.45
Color = Azul	0.20	0.15	0.05
Forma = Redonda	0.05	0.50	0.33’
Forma = Cuadrada	0.05	0.25	0.33’
Forma = Triangular	0.90	0.25	0.33’

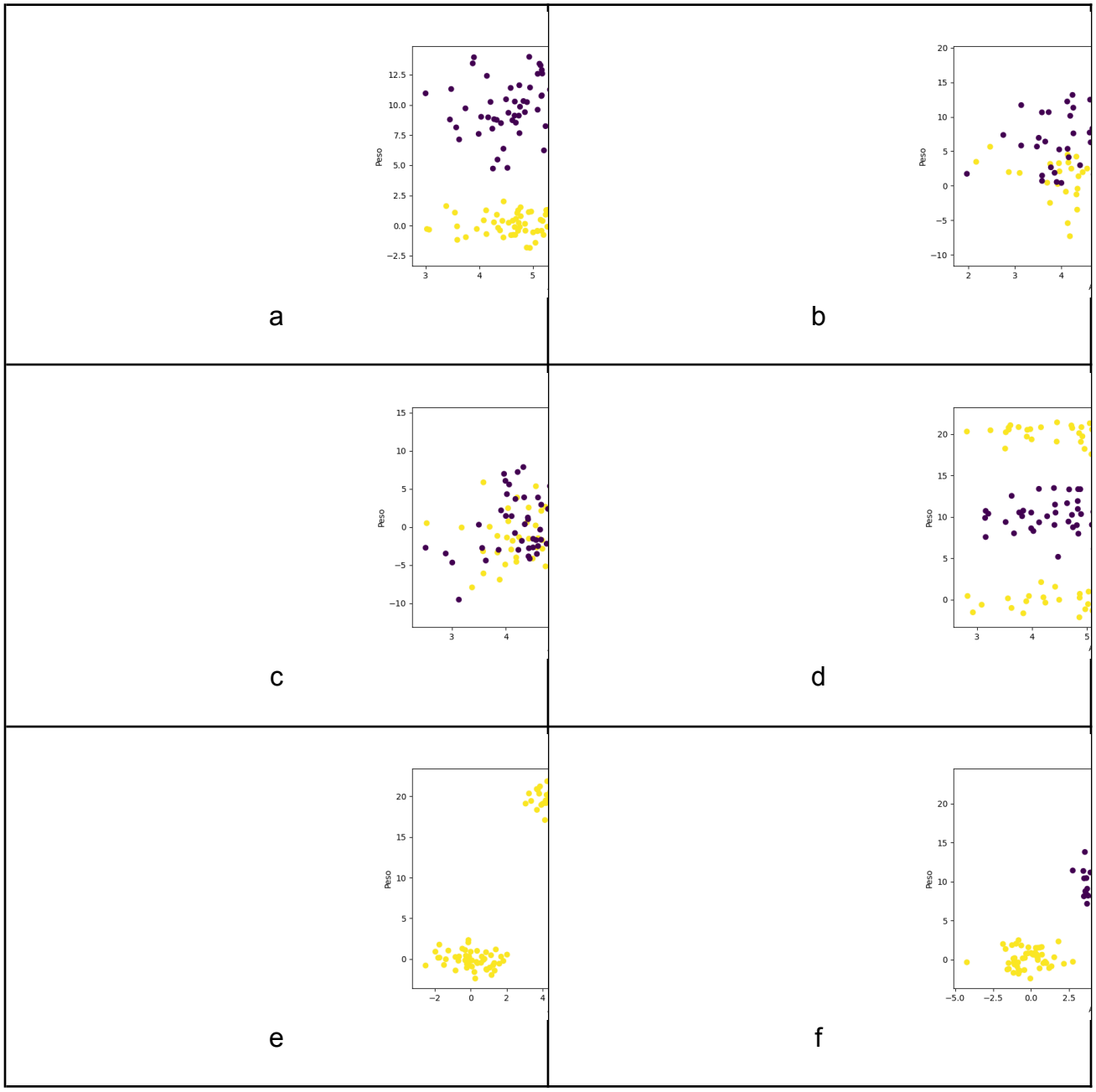
c) Dado un modelo de NB, con N atributos. Supongamos que agregamos un atributo adicional **A**, tal que ahora hay N+1 atributos. Supongamos también que **A** es un atributo nominal, que para todas las clases tiene una distribución uniforme (P(A=Valor| c) = 1/M, donde M es la cantidad de valores de A). ¿Puede este atributo empeorar el desempeño del modelo NB? Pista: Estudiar la fórmula de cálculo de probabilidades para un modelo con dos clases y 3 atributos:

$P(a_1, a_2, a_3 | c_1) = P(c_1) * P(a_1 | c_1) * P(a_2 | c_1) * P(a_3 | c_1)$

$P(a_1, a_2, a_3 | c_2) = P(c_2) * P(a_1 | c_2) * P(a_2 | c_2) * P(a_3 | c_2)$

7. Limitaciones de un modelo Naive Bayes

a) Dados los siguientes conjuntos de datos, con dos variables continuas **Peso** y **Altura**, y dos clases **Violeta** y **Amarillo** ¿en qué casos podrá entrenar un modelo NB para clasificar correctamente la gran mayoría de los ejemplos? ¿qué atributos permiten clasificar en cada caso?



b) En los casos en que no puede clasificar utilizando una distribución normal, ¿qué sucede si discretiza un atributo, con intervalos adecuados, para generar un atributo nominal? Por ejemplo, discretizando en tres valores [Bajo, Medio, Alto]. ¿Podría utilizar una distribución categórica para clasificar en ese caso?

8) Clasificación de datos sintéticos con NB

Entrene un modelo NB para los siguientes conjuntos de datos. Para cada conjunto, i) calcule el accuracy del modelo ii) analice los parámetros del modelo, para determinar qué atributos dan más información al modelo.

a) El archivo ***Iris.xls*** contiene información referida a la longitud y al ancho de sépalos y pétalos de tres especies de flores, *iris setosa*, *iris versicolor* e *iris virginica*. Este es un problema famoso en Minería de Datos porque ha sido utilizado en varios libros como conjunto de datos para ejemplificar problemas de clasificación.

b) El SENASA quiere proponer una nueva clasificación de los vinos para que su venta que la elección de vinos en la compra sea más simple. El archivo ***Vinos.xls*** tiene información referida a 13 características químicas y/o visuales de varias muestras de vinos pertenecientes a las 3 clases.

9) Clasificación de Jugadores de Fútbol

El archivo **jugadores_2023.xlsx** contiene datos de atributos físicas y habilidades de jugadores de fútbol en varias posiciones. Este conjunto de datos es complejo, y por lo tanto no es trivial entender sus atributos, valores y las relaciones entre los mismos. El atributo “Best position” indica la posición en que el jugador se desempeña usualmente. Sus valores son:

7. **LB -** Defensor Izquierdo
8. **CB**– Defensor Central
9. **RB** – Defensor Derecho
10. **CDM**– Mediocampista Defensivo Central
11. **LM -** Mediocampista Izquierdo
12. **CM** –Mediocampista Central
13. **RM** – Mediocampista Derecho
14. **CAM –**Mediocampista Atacante Central
15. **LW** – Atacante Izquierdo
16. **CF** –Atacante Central
17. **RW** – Atacante Derecho
18. **ST** –Atacante Goleador

*Figura SEQ Figure * ARABIC 1: Posiciones de futbol típicas.*

a) Entrene un modelo NB para predecir la mejor posición de un jugador, en base a los atributos que indican como juega el mismo y sus características físicas. Para ello, divida el conjunto de datos, dejando 80% de los datos para el conjunto de entrenamiento y un 20% para el de prueba, y evalúe el accuracy del modelo con cada conjunto. ¿Cuántas clases tiene el modelo? ¿Puede interpretar los parámetros del mismo?

b) El atributo de mejor posición tiene muchos valores distintos, lo que dificulta al modelo para obtener buenos resultados. Recodifique este atributo para dejar solo 4 valores distintos: arquero, defensor, mediocampista, y atacante y vuelva a evaluar el accuracy con la misma modalidad que en a). Visualice e interprete los parámetros del modelo para cada clase.

c) Entrene un modelo para predecir el valor de venta de un jugador en base a otros de sus atributos. Para ello, discretizar este atributo en tres valores distintos: alto, medio y bajo. Para encontrar intervalos adecuados para la discretización, visualice el atributo y elija los límites de forma manual. Evalúe también el accuracy del modelos como en a). Visualice la matriz de confusión para ambos conjuntos de entrenamiento y prueba. ¿Son igual de importantes todos los errores del modelo?

10) Predicción de batallas Pokemon

El archivo **pokemon_combats.csv** indica el ganador de 50000 batallas entre dos Pokemon distintos, cada uno identificado por su código.

El archivo **pokemon_attributes.csv** contiene información de distintos Pokemon del juego del mismo nombre. Cada Pokemon tiene varios atributos básicos: puntos de vida (HP), habilidad de ataque, defensa, ataque especial y defensa especial, y velocidad. También se conoce si el Pokemon es legendario, en qué generación fue introducido (8 distintas), su nombre y su identificador. Por último, también tiene hasta 2 tipos característicos, como Agua, Tierra, Aire, etc. El orden de los tipos no es lo importante, y la mayoría de los Pokemon tiene un solo tipo.

a) En base a esta información, entrene un modelo capaz de predecir, en base a las características de dos Pokemon, el ganador de una batalla Pokemon.

Note que para ello deberá unir los datos de ambos archivos, y recodificar varios de ellos, de modo de tener, para cada combate, las características de cada uno de los dos pokemon y un atributo que indique con 1 si ganó el primero de los Pokemon o 0 si ganó el 2do.

b) Visualice e interprete los parámetros del modelo. ¿Qué limitaciones tiene un modelo de NB para codificar este tipo de problemas?