Universidad Católica de Temuco
Escuela de Ingeniería Informática
INFO-1165
Prueba Práctica 2

Pauta de Evaluación

En este proyecto utilizarán el conjunto de datos de Diabetes de los Indios Pima, que se encuentra en el paquete `mlbench` de GNU R. Cada columna $j = 1, \ldots m$, representa una variable médica y cada fila $i = 1, \ldots, n$, corresponde a un individuo. Dividan la base de datos, usando una distribución uniforme, en dos subconjuntos etiquetados como "entrenamiento" y "prueba" (80 % en la primera, 20 % en la segunda). Los subconjuntos deben tener una proporción de "positivos" similar en ambas bases de datos.

Una vez que los datos estén listos, deben implementarán un modelo bayesiano de clasificación. Este modelo utilizará el Teorema de Bayes para calcular la probabilidad posterior de que un paciente tenga diabetes, basándose en las probabilidades previas y en la evidencia proporcionada por las variables del conjunto de entrenamiento. Dado que las variables son continuas, como la glucosa y la presión arterial, es necesario aplicar el supuesto de normalidad para modelar la verosimilitud de las variables condicionadas por cada clase. Deben calcular la media y la desviación estándar de cada variable para cada clase y usarán estas estadísticas para estimar la probabilidad de las variables dada la condición de diabetes, aplicando la función de densidad de la distribución normal, representada por la fórmula:

$$P(x_{ij}|\text{Positivo}) = \frac{1}{\sqrt{2\pi\sigma^2_{\text{Positivo},j}}} e^{-\frac{(x_{ij}-\mu_{\text{Positivo},j})^2}{2\sigma^2_{\text{Positivo},j}}}$$

donde x_{ij} es el valor de la variable j para el individuo i, $\mu_{\text{Positivo},j}$ y $\sigma^2_{\text{Positivo},j}$ son el valor medio y la varianza de la variable j para los individuos de la clase "positivo", respectivamente. Para implementar esto en R, deberán usar la función `dnorm`, que es la función de densidad de la distribución normal. Este cálculo se realizará igualmente para los individuos de la clase "negativo".

El teorema de Bayes en su forma tradicional, al asumir que las variables son independientes una de otra (enfoque *naive*), sería:

$$P(\text{Positivo} \mid x_{i1,\ldots,}x_{im}) = \frac{P(\text{Positivo}) \cdot \prod_{j=1}^{m} P(x_{ij} \mid \text{Positivo})}{\prod_{j=1}^{m} P(x_{ij})}$$

El denominador puede ser complejo de calcular, pero es constante para cada clase, por lo que se puede asumir que:

$$P(\text{Positivo} \mid x_{i1,\ldots,}x_{im}) \propto P(\text{Positivo}) \cdot \prod_{j=1}^{m} P(x_{ij} \mid \text{Positivo})$$

Los mismos cálculos se deben realizar para la clase "negativa". La $P(\text{Positivo})$ se obtiene como la proporción de los individuos que padecen la enfermedad y la $P(\text{Negativo})$ como la de los que no la tienen.

Como trabajarán con probabilidades, estas productorias pueden dar valores muy pequeños y generar inestabilidades numéricas en conjunto de múltiples variables (recordar lo visto en Algoritmos de Optimización), por lo que se aplican logaritmos para calcular las log-verosimilitudes:

$$\log(P(\text{Positivo} \mid x_{i1,\ldots,}x_{im})) = \log\left(P(\text{Positivo})\right) + \sum_{j=1}^{m} \log\left(P(x_{ij} \mid \text{Positivo})\right).$$

Igualmente deben hacer este cálculo para la clase negativo.

Con la base de datos de prueba determinen la clase a la que pertenece cada individuo al comparar las log-verosimilitudes de ambas clases, mediante la siguiente función:

$$\text{Clase}_i = \begin{cases} \text{Positivo}, & \text{si } \log(P(\text{Positivo} \mid x_{i1,\ldots,}x_{im})) - \log(P(\text{Negativo} \mid x_{i1,\ldots,}x_{im})) > \epsilon \\ \text{Negativo}, & \text{en caso contrario} \end{cases}$$

donde ϵ es un umbral de decisión que ajusta la decisión de la asignación de la clase.

Dada la tabla de confusión y los cuatro indicadores de calidad de la clasificación[1], realice una gráfica con `ggplot2` usando un valores equiespaciados de $\epsilon \in [-3, 3]$ (al menos 100 valores).

Se pide que:

[1] Ver https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

1. Expliquen cómo el desbalance de clases en los datos de entrenamiento y prueba puede influir en las métricas de rendimiento del modelo.

2. Evalúen cómo diferentes umbrales de decisión pueden ayudar a mitigar los efectos del desbalance de clases sobre las predicciones del modelo.

3. Seleccionen un umbral de decisión que consideren óptimo para manejar el desbalance de clases y justifiquen su elección con base en los resultados observados en las gráficas.

El trabajo, realizado en GNU R, deberá ser presentado en grupos de tres integrantes el jueves 3 de Octubre de 2024 a partir de las 9:30h. Un integrante será seleccionado al azar para realizar la presentación y responder a las preguntas del docente. Recuerden registrar su hora de presentación en la hoja de cálculo que se proveerá en la plataforma moodle. Si quedan ventanas entre presentaciones o desde la hora de inicio, el docente tiene la libertad de ajustarlas para optimizar el horario.

Cuadro 1: Rúbrica de Evaluación de Presentación Oral

Criterio	3	2	1	0
Planteamiento del problema (5 %)	Problema perfectamente definido.	Problema bien definido, aunque existen deficiencias y/o lagunas de información subsanables.	Problema insuficientemente definido, existiendo deficiencias y/o lagunas de información difícilmente subsanables.	Problema indefinido, existen deficiencias y/o lagunas de información insubsanables.
Cálculos (25 %)	Cálculos perfectamente realizados y presentados.	Cálculos bien realizados, aunque con imprecisiones menores en la presentación.	Cálculos bien realizados, aunque con imprecisiones mayores en la presentación.	Se presenta una de las siguientes situaciones: 1) Cálculos no realizados o incorrectos. 2) El criterio anterior ha sido evaluado con 0 puntos.
Discusión acerca de los resultados obtenidos (25 %)	Discusión acerca de los resultados obtenidos perfectamente realizada.	Discusión acerca de los resultados obtenidos bien realizada, aunque con imprecisiones menores.	Discusión acerca de los resultados obtenidos realizada, aunque con imprecisiones mayores.	Se presenta una de las siguientes situaciones: 1) Discusión acerca de los resultados obtenidos no realizada. 2) El criterio anterior ha sido evaluado con 0 puntos.
Código (10 %)	Código coherente con todo el trabajo realizado.	Código coherente con más del 75 % el trabajo realizado.	Código coherente con más del 50 % del trabajo realizado.	Se presenta una de las siguientes situaciones: 1) Código coherente con menos del 50 % del trabajo realizado. 2) El criterio anterior ha sido evaluado con 0 puntos.

Cuadro 1: Rúbrica de Evaluación de Presentación Oral

Criterio	3	2	1	0
Respuesta a preguntas (25 %)	Se demuestra comprensión y análisis profundo de las preguntas planteadas por el docente.	Se observan ligeras dudas sobre la comprensión y análisis profundo de las preguntas planteadas por el docente.	Se observan grandes dudas sobre la comprensión y análisis profundo de las preguntas planteadas por el docente.	Se presenta una de las siguientes situaciones: 1) Se observa una completa falta de comprensión y análisis profundo de las preguntas planteadas por el docente. 2) El criterio anterior ha sido evaluado con 0 puntos.
Expresión oral (5 %)	La presentación es realizada en un lenguaje técnico adecuado, demostrando respeto por el profesor y los compañeros.	La presentación es realizada en un lenguaje técnico adecuado, pero con algunos problemas menores, aunque demostrando respeto por el profesor y los compañeros.	La presentación es realizada en un lenguaje técnico con problemas graves, aunque demostrando respeto por el profesor y los compañeros.	Se presenta una de las siguientes situaciones: 1) La presentación carece de un lenguaje técnico adecuado. 2) Se falta el respeto por el profesor y los compañeros. 3) El criterio anterior ha sido evaluado con 0 puntos.
Manejo del tiempo (5 %)	Se respeta el intervalo de entre 10 y 15 minutos de presentación oral.	El intervalo es irrespetado en menos de 2 minuto.	El intervalo es irrespetado en menos de 5 minutos.	Se presenta una de las siguientes situaciones: 1) El intervalo es irrespetado en más de 5 minutos. 2) El criterio anterior ha sido evaluado con 0 puntos