UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE INGENIERÍA

Departamento de Informática

Laboratorio 4 - Análisis Estadístico

Clasificador Bayesiano

Bryan Santelices

Matias Coronado

Profesor: Max Chacón Pacheco

Ayudante: Javier Arredondo Contreras

Santiago – Chile

TABLA DE CONTENIDO

Índice de ilustraciones **iii**

1 Introducción **1**

2 Marco Teórico **2**

3 Obtención del clasificador **3**

4 Análisis de resultados y comparación **4**

5 Conclusión **6**

ÍNDICE DE ILUSTRACIONES

CAPÍTULO 1. INTRODUCCIÓN

Entre los cáncer que mas afectan a la población mundial, y acotando a las mujeres, es el cáncer de mama es unos de los canceres mas frecuentes entre las mujeres, estudios indican que aproximadamente 1 de cada 8 mujeres desarrollaran esta enfermedad a lo largo de su vida. Es una enfermedad que tiene como origen el proceso en que las células mamarias comienzan una metástasis desenfrenada, es decir que no paran de reproducirse, lo cual genera un tumor dentro del seno. Una vez se detecta esta enfermedad, es de vital importancia identificar el tumor como benigno o maligno, debido a que este ultimo representa un peligro mortal para paciente, ya que tiene la capacidad de extender su área de afección a zonas externas al seno, logrando generar metástasis en tejidos mas delicados como pueden ser órganos vitales.

En este contexto, se solicita realizar una estudio relacionado con la base de datos de Breast Cancer Wisconsin, con tal de luego de haber analizado las características de la información que se dispone en las experiencias previas, encontrar un **clasificador bayesiano** en base a la contribución que entrega a la probabilidad bayesiana para determinar el nivel que toma en la clase de estudio, en este caso, determinar si tiene tumores benignos o malignos. Todo esto para finalmente extraer conocimiento de los datos recopilados.

Así, en el presente de este reporte se presenta el detalle de el proceso para obtener el clasificador bayesiano en si, para posteriormente realizar el proceso de análisis de los resultados obtenidos.

Los objetivos que se abordados a lo largo de la experiencia son los siguientes:

1. Extraer nuevo conocimiento del problema por medio de un clasificador bayesiano.
2. Comparar los resultados obtenidos con un set de datos de prueba respecto a la clasificación real, para así evaluar la calidad del predictor probabilístico.

CAPÍTULO 2. MARCO TEÓRICO

1. **Clasificador bayesiano ingenuo**: Se trata de un clasificador simple basado en la aplicación del teorema de Bayes con supuestos de independencia fuertes llamados $ingenuos$. Se encuentra entre los modelos de redes bayesianas más simples, pero cuenta con la característica de que es capaz de lograr niveles de precisión más altos. La idea principal es que cada variable contribuya de manera individual con la probabilidad de la clasificación. Así, la formula correspondiente es la 2.1:

$$argmax_{i=1}^{n} P(c)_i * \prod_{j=0}^{m} P(a_j|c_j) \tag{2.1}$$

2. **Probabilidad posteriori**: Es un tipo de probabilidad condicional, la cual se obtiene después de que se tome en cuenta la probabilidad a priori.

3. **Probabilidad a priori**: Este tipo de probabilidad con la que se empieza el experimento, esta se utiliza para generar nuevo conocimiento, en base al razonamiento que se le puede dar a la información de entrada.

CAPÍTULO 3. OBTENCIÓN DEL CLASIFICADOR

Antes que nada hay que definir el atributo a predecir, el cual para este caso es la variable **class**, la que indica si paciente presenta un cáncer benigno o maligno. Por otro lado, se utilizaran el resto de los parámetros para predecir el tipo de clase.

Para el proceso de obtener el clasificador bayesiano para este conjunto de datos, resulta necesario separar las observaciones en 2 conjuntos, uno que sera utilizado para entrenar al clasificador (70 $\%$ de los datos) y el otro para con este ya creado (restante 30 $\%$ de los datos), probar tanto la eficacia como la calidad del mismo.

Dentro del set de datos existe una distribución de los casos, en donde el $64{,}91\,\%$ corresponde a Benigno, mientras que el $35{,}07\,\%$ al Maligno.

En base a lo anterior, se conformo el clasificador bayesiano correspondiente a los datos de entrenamiento. Se presenta a continuación una matriz de confusión que distribución de casos de cáncer clasificados como cancerígenos por parte del clasificador en contraparte de la clasificación real proporcionada por el dataset.

		Realidad	
		Benigno	Maligno
Clasf.	Benigno	125	7
	Maligno	0	71

CAPÍTULO 4. ANÁLISIS DE RESULTADOS Y COMPARACIÓN

En base a la tabla de contingencia creada, es posible evaluar que tan preciso es el modelo bayesiano generado. A continuación, se aprecian las correspondientes formulas para determinar el nivel de Accuracy para el caso positivo (Benigno) y caso negativo (Maligno).

$$Accuracy_{Positivo} = \frac{VP}{VP + FP} \tag{4.1}$$

$$Accuracy_{Negativo} = \frac{VN}{FN + VN} \tag{4.2}$$

En donde:

1. **VP**: Verdadero positivo
2. **FP**: Falso positivo
3. **VN**: Verdadero negativo
4. **FN**: Falso negativo

Al reemplazar los valores en ambas formulas se obtiene:

$$Accuracy_{Positivo} = \frac{127}{127 + 7} = 94{,}77\,\% \tag{4.3}$$

$$Accuracy_{Negativo} = \frac{71}{0 + 71} = 100\,\% \tag{4.4}$$

A razón de las matricas calculadas anteriormente, se puede observar que tanto para los casos el clasificador bayesiano señala como malignos o benignos, lo logra con una exactitud casi perfecta, es decir, el clasificador es capas de predecir correctamente en casi todos los casos en base a las variables del dataset si un paciente debería ser diagnosticado con tumores cancerígenos o no.

Cabe destacar que como se puede ver en la matriz de confusión, las observaciones que se utilizaron para el set de datos para el entrenamiento (con el que se construyo el clasificador

bayesiano), contaba con más observaciones (pacientes) diagnosticados con tumores benignos que malignos, por ende es entendible que como se tienes mas variedad de entradas para el modelo del clasificador de estos casos, se logra distinguir con mayor precisión la implicación de las variables en la detección del cáncer, lo que explica que a la hora de utilizar el set de datos de prueba para evaluar la calidad del modelo, la detección de estos casos no es tan precisa como para los casos negativos, lo que en realidad es bueno ya que indica que el clasificador esta utilizando criterios mas específicos para distinguir si un paciente cuenta con tumores cangerigenos o no.

CAPÍTULO 5. CONCLUSIÓN

El laboratorio nos ayudo en el entendimiento de los fundamentos de este tipo de clasificador, a partir de la aplicación practica realizada en R, el cual nos permitió generar exitosamente un clasificador bayesiano ingenuo.Esto ultimo nos ayudo a entender a fondo las métricas que se utilizan en este tipo de estudios, y en como estas afectan a los porcentajes de precisión que entrego el modelo.

En base a este ultimo, se estima que el modelo predicativo es altamente confiable, esto debido a los altos porcentajes de preciso que obtuvo para ambos tipos de clasificación, siendo $93{,}77\,\%$ y $100\,\%$ respectivamente para los casos Benigno y Maligno. Como se menciono anteriormente, dentro del segmento del análisis comparativo, se cree que que la baja cantidad de casos Malignos beneficio el valor del porcentaje, debido a que estos representaban aproximadamente el $30\,\%$ de los datos de entrenamiento.

Finalmente, para futuras experiencias se espera profundizar en la utilización de este tipo de clasificador, además de los factores que influyen en las métricas del modelo bayesiano.

BIBLIOGRAFÍA

[1] Wikipedia. (No indica). A priori y a posteriori. Recuperado de: https://es.wikipedia.org/wiki/A_priori_y_a_posteriori

[2] Victor Roman. (2019, Abril 25). Algoritmos Naive Bayes: Fundamentos e Implementación . Recuperado de: https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementaci%C3%B3n-4bcb24b307f

[3] F. Villalba (2018, Obctubre 10). Naive Bayes- clasificación bayesiano ingenuo . Recuperado de: https://fervilber.github.io/Aprendizaje-supervisado-en-R/ingenuo.html