Universidad de Talca
Facultad de Ingeniería
Departamento de Ciencias de la Computación

Tarea Nº 2 - Naïve Bayes

Machine Learning

Gonzalo Mardones Baeza
gmardones10@alumnos.utalca.cl
19 de septiembre de 2017

1. Introducción

El presente informe da a conocer el estudio del algoritmo de clasificación **Naïve Bayes**, que tiene como finalidad predecir la clase de un objeto, por medio de un conjunto de entrenamiento cuyas clases son conocidas *con anterioridad*. Además, de comentar ventajas y desventajas del clasificador, resultados generados tras la ejecución de un programa desarrollado que hace uso de una base de datos públicas y entregue matriz de confusión, gráficos Gaussianos y certeza de la predicción, ente otros.

2. Acerca de Naïve Bayes

Es un método intuitivo que utiliza las probabilidades de un atributo perteneciente a cada clase para realizar una predicción. Es un aprendizaje supervisado que simplifica el cáclulo de las probabilidades suponiendo que la probabilidad de cada atributo que pertenece a un valor de la clase dada es independiente de todos los otros atributos. Para hacer una predicción, podemos calcular las probabilidades de la instancia que pertenece a cada clase y seleccionar el valor de la clase con la más alta probabilidad [1].

El método de Naïve Bayes tiene varias ventajas, como el hacer predicciones a partir de datos parciales, pueden manejar base de datos incompletas y el ser rápido. Su principal desventajas está el no ser apto para el manejo de variables aleatorias continuas [3].

3. Acerca de la Base de Datos

Se utiliza la base de datos que aplica al problema de la *diabetes de los indios Pima* [5]. La cual posee detalles médicos para pacientes mujeres de 21 años o más del pueblo indio Pima - Canadá. La base de datos se compone de 768 observaciones, todos ellos númericos y sus unidades varían de atributo en atributo. Cada instancia tiene un valor de clase que indica si la paciente sufrió la aparición de diabetes dentro de los 5 años del momento en que se tomaron las mediciones, la clase 1

indica que sufrió diabetes y 0 que no. Una buena precisión de la predicción es de 70 % en adelante. Disponible como archivo en formato CSV en *Pima Indians Diabetes Data Set* [4]. La base de datos se compone de la siguiente manera:

1. Número de instancias: 768.
2. Número de atributos por instancia: 8 + clase.
3. Cada atributo es númerico, y esta compuesto en el siguiente orden:
   {Número de veces embarazadas, Concentración de glucosa en plasma a 2 horas en una prueba oral de tolerancia a la glucosa, Presión arterial diastólica (mm Hg): es la presión máxima que se alcanza en el sístole, Espesor del pliegue cutáneo del tríceps (mm), Horas de insulina en suero (mu U/ml), Indice masa corporal, Función pedigree de la diabetes, Edad (años), Clase del atributo (0 ó 1) }.

Para el último atributo cada instancia tiene un valor de clase que indica si el paciente sufrio una aparición de diabetes dentro de los 5 años del momento en que se tomaron las mediciones: 0 : No y 1 : Si.

4. Descripción Gráficas de Gaussianas

Se eligió el atributo que correspondía al **indice masa corporal** (situado en el ese de las ordenadas), correspondiente al elemento de la posición 6 de una instancia. Se aprecia en la Figura [1], que la clase 0 que corresponden a las personas que no presentaron diabetes, en donde la media (mu) fue de $30,448$ es menor que para los casos de la clase 1 que fue de $35,152$. La clase 0 presentó una desviación estandar (desv_est) levemente mayor que fue de 7,73 contra 7,40 de la clase 1, dado que existieron un par de datos extremos que hicieron variar levemente la gráfica, no así como el caso de la clase 1 donde los datos se encontraban más agrupados.

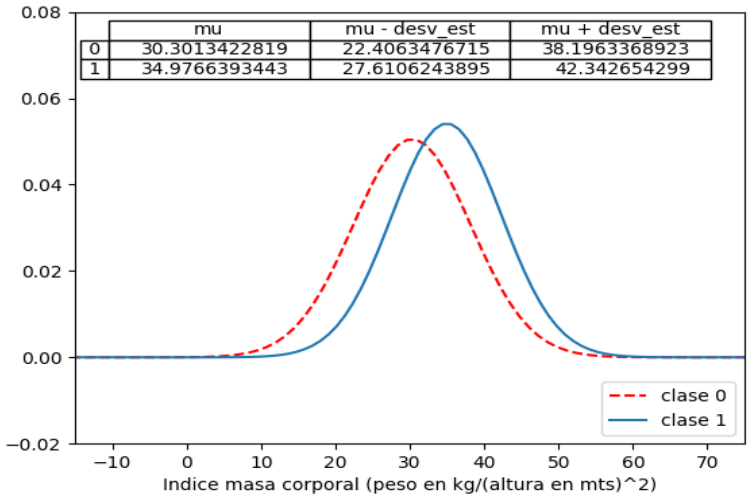

Figura 1: *Gráficas de Distribución Gaussianas de las clases del conjunto de datos.*

5. Matriz de Confusión

La matriz de confusión que permite visualizar el desempeaño de un algoritmo de aprendizaje supervisado, en donde cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias de la clase [2]. Al ejecutar el programa, es posible apreciar la matriz generada, donde:

```
MP-GM:Tarea_NaiveBayes gmardones$ python tarea_naivebayes.py
+-----------------------+
| Matriz de Confusion   |
+-----------------------+
| clase |   0  |   1    |
+-----------------------+
| 0     |   44 |  9     |
| 1     |   6  |  18    |
+-----------------------+
Exactitud (Certeza) Modelo: 80.5194805195%
```

Donde, de 77 instancias en la matriz de confusión, se infiere que de 53 instancias de la clase 0, el error es de $9/53 = 0,16981\,\%$, mientras que la exactitud fue de $44/53 = 0,83018\,\%$. Para la clase 1 la que contó con 24 instancias, el error es de $6/24 = 0{,}25\,\%$, mientras que la exactitud es de $18/24 = 75\,\%$. El desempeño del clasificador para esta ejecución es del $80,51\,\%$.

6. Tecnologías y Requisitos del Sistema

El programa se desarrollo bajo el IDE `Sublime Text - Build 3126`, en el lenguaje de programación `Python` versión `2.7.13`, S.O `macOS Sierra` versión `10.12.6`, es necesario contar con los módulos `numpy, matplotlib, scipy` y `sklearn` de Python que deben ser incluidos externamente, por medio de la sentencia en el terminal:

```
MP-GM:Tarea_NaiveBayes gmardones $ sudo pip install numpy
MP-GM:Tarea_NaiveBayes gmardones $ sudo pip install matplotlib
MP-GM:Tarea_NaiveBayes gmardones $ sudo pip install sklearn
MP-GM:Tarea_NaiveBayes gmardones $ sudo pip install scipy
```

El programa debe contar con los archivos: `PID_dataset.csv` que posee la base de datos y `tarea_naivenayes.py` que es el programa que debe ser ejecutado por medio del terminal con la sentencia:

```
MP-GM:Tarea_NaiveBayes gmardones $ python tarea_naivebayes.py
```

7. Conclusión

El estudio e implementación del clasificador de Naïve Bayes para el conjunto de datos del pueblo indio Pima permitió comprender el algoritmo para valores conocidos con anticipación, comprender sus ventajas y desventajas. Adicionalmente de realizar la comparativa de gráficas Gaussianas para atributos del conjunto de diferentes clases.

Por medio de la matriz de confusión fue posible conocer la certeza de la predicción de sus clases, probabilidades de aciertos y errores de la predicción. Además de conocer la certeza del modelo.

Finalmente, por medio de la implementación fue posible comprender el potencial de este clasificador y de las tecnologías utilizadas para problemas de aprendizaje supervisado.

Referencias

[1] Alpinu. *Naive Bayes*. Wikipedia.org, 2016.

[2] Ignacio Icke. *Matriz de Confusión*. Wikipedia.org, 2008.

[3] Rodolfo García Flores Samuel D. Pacheco Leal, Luis Gerardo Díaz Ortiz. *El clasificador Naïve Bayes en la extracción de conocimiento de bases de datos*. Posgrado en Ingeniería de Sistemas, FIME-UANL, 2005.

[4] Vincent Sigillito. *Pima Indians Diabetes Data Set*. National Institute of Diabetes and Digestive and Kidney Diseases, 1990.

[5] Carl Waldman. *Encyclopedia of Native American Tribes*. Springer US, 1999.