Aprendizaje automático

Departamento de Ingeniería en Informática
ITBA

Trabajo Práctico 2

Objetivo: Aprender a clasificar datos utilizando clasificadores bayesianos

Aprendizaje bayesiano

1. Una estación de radio tiene dos grupos de oyentes: los "*jóvenes*" y los "*viejos*". Los dueños de la estación suponen que el saber si un oyente es *joven* o *viejo* es suficiente para terminar si le va a gustar un programa, independientemente de los gustos o preferencias para cualquier otro programa.
   Se sabe que si el oyente es *joven*, hay una probabilidad del 95% de que le guste el programa 1, una probabilidad del 5% de que le guste el programa 2, una probabilidad del 2% de que le guste el programa 3 y una probabilidad del 20% de que le guste el programa 4. Por otra parte, si el oyente es *viejo*, hay una probabilidad del 3% de que le guste el programa 1, una probabilidad del 82% de que le guste el programa 2, una probabilidad del 34% de que le guste el programa 3 y una probabilidad del 92% de que le guste el programa 4. Se sabe además que el 90% de los oyentes son viejos.
   Dado este modelo, y el hecho de que a un nuevo oyente le gustan los programas 1 y 3, pero no le gustan los programas 2 y 4, ¿cuál es la probabilidad de que el nuevo oyente sea joven?

2. Consideremos el siguiente vector de atributos binarios:
   (scones, cerveza, whiskey, avena, fútbol)
   El vector $x = (1,0,1,1,0)$ describiría a una persona a la que a le gustan los scones, pero no la cerveza, bebe whiskey, come avena y no mira partidos de fútbol.
   En las siguientes tablas, por columna, se dan los vectores de preferencias de un grupo de 6 ingleses y de 7 escoceses.

 Ingleses

0	1	1	1	0	0
0	0	1	1	1	0
1	1	0	0	0	0
1	1	0	0	0	1
1	0	1	0	1	0

 Escoceses

1	1	1	1	1	1	1
0	1	1	1	1	0	0
0	0	1	0	0	1	1
1	0	1	1	1	1	0
1	1	0	0	1	0	0

Dado el vector $x = (1,0,1,1,0)$, determinar si esa persona es inglesa o escocesa considerando la hipótesis más probable h_{MAP}.

3. Un supermercado especializado en cereales para el desayuno decide analizar las preferencias de sus clientes. Hacen una pequeña encuesta preguntándoles, a seis clientes elegidos al azar, su edad (mayores o menores de 60 años) y qué cereales prefiere para el desayuno (maíz, granola, azucarados, avena).
   Las respuestas de cada cliente se registran en un vector con entradas 1ó 0 de acuerdo a si le gusta o no cada uno de los cereales. Por ej., el vector (1101) indica que a la persona le gustan todos los cereales excepto los azucarados.
   Los clientes mayores de 60 años dan como respuestas (1000), (1001), (1111) y (0001) y los menores de 60 años (0110), (1110).
   Un nuevo cliente entra en el supermercado y dice que sólo le gustan la granola y los azucarados. Usando el clasificador naïve de Bayes determinar si esa persona es mayor de 60 años o no.

4. Un psicólogo hace una pequeña encuesta sobre la "*felicidad*". Cada participante da como respuestas un vector con entradas 1 ó 0 correspondiendo a las respuestas "si" o "no" a las preguntas. El vector de preguntas tiene atributos:
   $$x = (\text{rico,casado,saludable})$$
   Por ej., una respuesta (101) indica que la persona es rica, soltera y saludable.
   Además, cada participante da como respuesta un valor c tal que
   $$c = \begin{cases} 1 & \text{si la persona está contenta con su vida} \\ 0 & \text{sino} \end{cases}$$
   Se obtuvieron las siguientes respuestas de las personas "contentas": (1,1,1), (0,0,1), (1,1,0), (1,0,1) y de las personas "no contentas": (0,0,0), (1,0,0), (0,0,1), (0,1,0), (0,0,0).
   (a) Calcular la probabilidad de que una persona "pobre", "casada" y "saludable" esté "contenta".
   (b) Calcular la probabilidad de que una persona "pobre" y "casada" esté "contenta".
   (c) Aplicando el clasificador naïve de Bayes, determinar si una persona "pobre", "casada" y "saludable" estaría "contenta".

5. En la década de 1920 un grupo de botánicos recolectó datos de 150 lirios. Estas medidas se conocen como los "datos de los lirios de Fisher" y están disponibles en:
   http://www.stat.ncsu.edu/working_groups/sas/sicl/data/irises.dat
   Los datos consisten en 50 muestras de cada una de 3 especies de lirios: setosa, virginica y versicolor.

setosa

versicolor

virginica

Para cada muestra, se tomaron 4 medidas: largo y ancho del sépalo y del pétalo en centímetros.
Como cada observación corresponde a una especie conocida, se pueden comparar los grupos obtenidos con un clasificador con las especies reales.

(a) Estudiar cómo las medidas de los pétalos y de los sépalos varían entre las especies.
Calcular la media, el desvío estándar, el mínimo, el máximo y el rango de ambas variables para cada especie.

(b) Estudiar gráficamente si algunas de las medidas servirían para clasificar a las especies.
Graficar, por ej., largo vs. ancho del pétalo para los 150 lirios.

(c) Suponiendo que las medidas tomadas sobre los lirios siguen una distribución gaussiana:
   i. Aplicar el clasificador naïve de Bayes para determinar la especie de cada lirio en la muestra dada.
   ii. Calcular el porcentaje de datos mal clasificados con el método anterior.
   iii. Encontrar la matriz de confusión.