Bases de Datos Masivas (11088)
Departamento de Ciencias Básicas

TRABAJO PRÁCTICO III: Minería de datos

PARTE 05: Razonamiento probabilístico

Introducción:

En este trabajo se abordará uno de los algoritmos de razonamiento probabilístico basado en aprendizaje bayesiano: Naive Bayes. Este algoritmo de aprendizaje supervisado se utiliza para clasificar y predecir instancias utilizando probabilidad a posteriori.

En primer lugar, se presentan ejercicios orientados a incorporar los fundamentos de este tipo de técnicas, como la estimación de parámetros a través del cálculo de probabilidades.

Luego, se utilizará el lenguaje de Programación **Python** con la librería **Scikit-Learn** con el objetivo de resolver problemas de la disciplina, los cuales son una combinación ejercicios clásicos de minería de datos complementados con ejercicios propuestos por el equipo docente.

Consignas:

1. **Estimación de parámetros en Naive Bayes.** A partir de los conceptos incorporados en relación al Teorema de Bayes e hipótesis MAP, genere el clasificador Naive Bayes del siguiente dataset utilizando el estimador basado en la Ley de la sucesión de Laplace para los atributos discretos:

PRONÓSTICO	TEMPERATURA	HUMEDAD	VIENTO	ASADO
soleado	36	alta	leve	no
soleado	28	alta	fuerte	no
nublado	30	alta	leve	si
lluvioso	20	alta	leve	si
lluvioso	2	normal	leve	si

lluvioso	5	normal	fuerte	no
nublado	11	normal	fuerte	si
soleado	22	alta	leve	no
soleado	9	normal	leve	si
lluvioso	17	normal	leve	si
soleado	19	normal	fuerte	si
nublado	22	alta	fuerte	si
nublado	27	normal	leve	si
lluvioso	21	alta	fuerte	no

a. Una vez generado el clasificador, realice la clasificación de las siguientes instancias:

PRONÓSTICO	TEMPERATURA	HUMEDAD	VIENTO
soleado	19	normal	leve
lluvioso	34	alta	leve
nublado	14	normal	fuerte

b. ¿Qué ventajas observa en los resultados por sobre los métodos de *data mining* vistos antes?
c. El árbol de decisión generado en el TP0501, ¿Hubiera clasificado estas tres instancias de la misma manera? Argumente su respuesta.
d. ¿Qué problemas observa de aplicar *Naive Bayes* sobre este dataset puntualmente?
e. ¿Justifica el empleo del estimador de Laplace en reemplazo del estimador por máxima verosimilitud? ¿Por qué?

2. **Naive Bayes.** Cargue el dataset *zoo* utilizada en el TP0501 en Python y responda:
   a. Estime los parámetros del clasificador Naive Bayes.
   b. Analice las probabilidades calculadas y documente las conclusiones.
   c. ¿Encuentra relación entre las probabilidades y los resultados obtenidos mediante árboles de decisión?

d. Genere al azar 5 instancias y clasifíquelas mediante el clasificador Naive Bayes. Luego, clasifique esas instancias mediante el árbol de decisión, ¿Encuentra diferencias?

3. Ahora, genere el clasificador con el *dataset* spam-data.csv para determinar si un correo corresponde a spam o no. Documente los resultados encontrados en términos de las probabilidades que arroja Naive Bayes.

Referencias sugeridas:

Data Mining: Practical Machine Learning Tools and Techniques http://www.cs.waikato.ac.nz/ml/weka/book.html

Machine Learning, Chapter 6. Tom M. Mitchell, McGraw Hill, 1997.

Data Mining: Concepts and Techniques. Jiawei Han & Micheline Kamber. Morgan Kaufmann. Second Edition. 2006. Chapter 6.4.