Práctica 2: Probabilidad y Variable Aleatoria (6 horas)

1. Ejercicio

1.1. Clasificador Naive Bayes

El objetivo central de esta práctica será calcular la probabilidad de los sucesos aprobar/suspender la asignatura de Estadística en Junio, en base a sucesos relacionados con aprobar/suspender las prácticas y/o evaluación parcial. De una forma más formal, definiremos los siguientes sucesos:

- AJ: Suceso aprobar en Junio.
- GA: Suceso pertenecer al grupo A.
- AP1: Suceso aprobar Práctica 1.
- AP2: Suceso aprobar Práctica 2.
- AP3: Suceso aprobar Práctica 3.
- AP: Suceso aprobar Total prácticas.
- AEP: Suceso aprobar Evaluación parcial

Nuestro objetivo es calcular probabilidades condicionadas de este tipo:

$$P(AJ|GA \cap AP1 \cap AP2 \cap AP3 \cap AP \cap AEP),$$

es decir, la probabilidad de que apruebe dado que pertenezco al grupo A, he aprobado la práctica 1, etc. A la hora de calcular esta probabilidad podemos aplicar la regla de Bayes:

$$P(AJ|GA \cap AP1 \cap AP2 \cap AP3 \cap AP \cap AEP) = \frac{P(GA \cap AP1 \cap AP2 \cap AP3 \cap AP \cap AEP|AJ)P(AJ)}{P(GA \cap AP1 \cap AP2 \cap AP3 \cap AP \cap AEP)}$$

El clasificador Naive Bayes asume que $P(GA \cap AP1 \cap AP2 \cap AP3 \cap AP \cap AEP|AJ)$ se puede calcular de la siguiente manera:

$$P(GA \cap AP1 \cap AP2 \cap AP3 \cap AP \cap AEP|AJ) = P(GA|AJ)P(AP1|AJ) \cdots P(AP|AJ)P(AEP|AJ),$$

donde la probabilidad de un suceso dado otro suceso pueden a su vez ser calculadas siguiendo la regla de Laplace. Por ejemplo,

$$P(AP1|AJ) = \frac{n(AP1 \cap AJ)}{n(AJ)}$$

se expresa como el cociente entre el número de alumnos que aprobaron la práctica 1 y aprobaron en Junio dividido entre el número de alumnos que aprobaron en Junio.

Al final si la probabilidad de $P(AJ|GA \cap AP1 \cap AP2 \cap AP3 \cap AP \cap AEP)$ es mayor o igual que 0.5, se predice que voy a aprobar en Junio y si no, se predice que voy a suspender.

En el paquete de R "e1071" podéis encontrar una implementación del clasificador Naive Bayes, a través del comando "naiveBayes".

1.2. Conjunto de Datos

Para esta práctica vamos a partir del conjunto de datos "`notasP2.csv`", que es el mismo de la práctica anterior con unos leves cambios en los nombres de las columnas.

1.3. Objetivo 1 (Sesión 1, Tema 2)

El primero objetivo es crear un clasificador Naive Bayes para predecir la Nota Final de Junio (aprobado/suspenso) partiendo de las variables que dan información acerca del grupo, de las prácticas y de la evaluación parcial. Para ello realizan los siguientes pasos:

1. Transformar las variables cuantitativas en variables cualitativas con valores aprobado/suspenso. Asume que un alumno no presentado a una práctica o a la evaluación parcial ha suspendido.

2. Aplicar un clasificador Naive Bayes y evaluar su capacidad de predicción calculando las siguientes medidas:

   a) La probabilidad de que acierte en la predicción de aprobado dado que el/la alumno/a ha aprobado.

   b) La probabilidad de que acierte en la predicción de suspenso dado que el/la alumno/a ha suspendido.

   c) La probabilidad de que el/la alumno/a aprueba dado que el clasificador predice que va a aprobar.

   d) La probabilidad de que el/la alumno/a suspenda dado que el clasificador predice que va a suspender.

   e) La probabilidad de que el clasificador acierte en su predicción.

3. Evalúa cómo van cambiando las medidas anteriores cuando el clasificador Naive Bayes usa los siguientes conjuntos de variables:

   a) Utilizando solo información de grupo.

b) Utilizando información de grupo, de Práctica 1 y de Práctica 2.

c) Utilizando información de grupo, de Práctica 1, de Práctica 2 y de Evaluación parcial.

d) Utilizando toda la información disponible de grupo, prácticas y evaluación parcial.

1.4. Objetivo 2 (Sesión 2, Tema 3)

Realiza el mismo análisis que en el objetivo anterior (predecir aprobado/supenso en Junio) con los siguientes nuevos elementos:

1. Considera la información de que un alumno puede aparecer como "No-Presentado" en alguna práctica/cuestionario de forma explícita.
2. Transforma las variables cuantitativas en variables cualitativas con valores suspenso, aprobado, notable y sobresaliente y evalúa si esta transformación es mejor que la transformación en suspenso/aprobado.
3. Obtén las tablas de probabilidad condicionada estimadas por el clasificador Naive Bayes y comenta las más relevantes.
4. Analiza si es posible hacer una predicción más ajustada, donde en vez de predecir suspenso/aprobado en Junio, podamos predecir también su nota en base a suspenso, aprobado, notable y sobresaliente.

1.5. Objetivo 3 (Sesión 3, Tema 4 y Tema 5)

Realiza el mismo análisis que en el Objetivo 1 con los siguientes nuevos elementos:

1. Utiliza las variables cuantitativas sin transformar asumiendo que se distribuyen de forma Normal dada la variable a predecir y evalúa si esto da lugar a mejores predicciones.
2. Analiza si es posible mejorar la capacidad de predicción del clasificador transformando algunas variables en aprobado/suspenso y dejando otras sin transformar.

Calcula la predicción de Aprobar en Junio para cada uno de los miembros del grupo que ofrece alguno de los clasificadores Naive Bayes que habéis construido en base a las notas que ya tenéis disponibles de esta asignatura.

2. Evaluación

La práctica se realizará en **grupos de 5**, pero será evaluada de forma individual en clase (durante la primera sesión de la tercera práctica). Para evaluar la práctica será necesario adjuntar, antes de la **fecha límite (12 de mayo)**, **un único fichero zip** comprimido como respuesta a la tarea iniciada en el aula virtual (1 por grupo), con los siguientes archivos:

- Memoria en formato `.pdf` que recoja los análsis descritos en cada uno de los tres objetivos.
- 3 archivos con el código fuente de R usado para resolver los tres objetivos de la práctica.

¡OBLIGATORIO! *** Dentro del informe (`.pdf`) deben aparecer los nombres de **todos los integrantes del grupo**, en caso contrario sólo se evaluará a las personas cuyos nombres ahí figuren.***

Puntuación

- La práctica se evaluará sobre 10 puntos.
- La no presentación de la práctica en el formato establecido supondrá una penalización de 5 puntos.
- El contenido de los ficheros entregados puntuará hasta 6 puntos.
- La defensa de la práctica (en clase) puntuará hasta 4 puntos.

Consideraciones

- Si falta alguno de los dos archivos la evaluación de la práctica será automáticamente de 0 puntos.
- La ausencia durante la sesión de evaluación de alguno de los miembros del grupo supondrá la no calificación de la defensa de la práctica (4 puntos) solamente para ese miembro del grupo.
- No se aceptarán entregas de prácticas fuera de plazo.
- Se utilizará el programa antiplagio Turnitin de la biblioteca. Detectado el plagio entre dos prácticas, ninguna de las dos puntuará.