Análisis Exploratorio de Datos (Data Mining)
Máster en Técnicas Estadísticas
Examen de Febrero - Curso 2007-2008

Nombre y Apellidos: _____

A) El fichero *jardin.xls* en su Hoja1 contiene los datos de una muestra de 825 matrimonios de una provincia que residen en casa unifamiliar, indicando: salario de él y de ella en euros, régimen de utilización de la vivienda, e indicación de si mantienen o no jardín con carácter permanente. Utilizando el programa EXCEL, respóndase a las siguientes cuestiones:

1. Categorizando previamente la nueva variable "Renta Salarial Familiar" (suma de los salarios del matrimonio) en una nueva variable "Nivel de Renta" con cinco categorías (1: <1200€, 2: [1200,2400), 3: [2400,3600), 4: [3600,4800) y 5: ≥4800€), obténganse:

   a. La distribución de frecuencias relativas:

Concepto	Nivel de Renta				
	1	2	3	4	5
Frecuencia relativa (%)					

   b. Media: _____; Moda: _____; Cuasivarianza: _____;

   c. 1º Cuartil: _____; 2º Cuartil: _____; 3º Cuartil: _____;

2. Indicar las siguientes medidas de las variables condicionadas: Salario_Ella | Nivel_de_Renta = 2, 3, 4:

Concepto	Nivel de Renta		
	2	3	4
Media			
Desviación típica			
Coeficiente de Variación de Pearson			

   ¿Qué nivel de Renta es más heterogéneo en términos del Salario de Ella?. El nivel ___ porque _____ _____.

3. Proporcionar los siguientes valores:

   a. Frecuencia de: Jardín = Si | (Nivel_de_Renta = 2) _____.

   b. Frecuencia relativa de: Jardín = No | (Nivel_de_Renta = 4) = _____.

   c. Moda de la variable Nivel_de_Renta | (Jardín = No)_____.

4. Proporcionar los siguientes valores:

   a. Frecuencia relativa de (Nivel_de_Renta, Vivienda) = (3, Propiedad): _____.

   b. Media de Salario_Ella | (Vivienda = Alquiler, Jardín = Si) = _____.

5. Estudiar la dependencia entre las variables Vivienda y Jardín:

   a. Estadístico: _____ = _____; p-value = _____.

b. Por tanto, las variables Vivienda y Jardín son _____.

c. Medida de asociación: _____ = _____.

d. Es decir, el grado de asociación entre ambas puede considerarse _____.

6. Modelizar linealmente la posible relación entre las variables Salario_El y Salario_Ella:

   a. SI/NO existe cierto grado de relación de tipo lineal ya que _____ = _____.

   b. Esa relación indicaría que:

      i. Por cada euro adicional en la variable Salario_Él la variable Salario_Ella se incrementaría en _____ €.

      ii. Para un salario nulo de él, ella poseería un salario medio de: _____ €.

   c. En todo caso, la relación entre ambas variables posee un DÉBIL/FUERTE carácter lineal ya que _____ = _____, es decir, la variable Salario_El explica linealmente un _____ de la variabilidad de Salario_Ella.

B). Cargar en R el fichero **jardin.csv** con las variables continuas normalizadas en [0,1] y discretas en {0,1}, utilizando la instrucción *read.table("jardin.csv", header=TRUE)*.

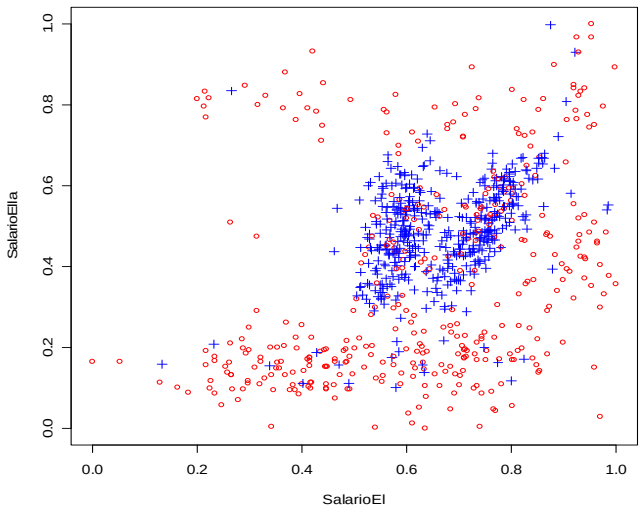
Se trata de determinar las condiciones salariales del matrimonio que favorecen el mantenimiento del jardín, es decir, predecir el valor de la variable Jardin a partir de los valores de las variables Salario_El y Salario_Ella.

Para ello, constrúyase un modelo logístico (**MASS**) y un modelo CART (**rpart**), respondiendo a las siguientes cuestiones:

1. Cumplimentar la siguiente tabla:

Modelo	Tasa de Error en Test (200 puntos)	Probabilidad a Posteriori: Jardin = 1| x = (0.5, 0.5)	Covariable(s) Relevante(s) para la Clasificación
Logístico			
CART			

2. Dibujar en el gráfico de la siguiente página las fronteras entre las clases estimadas por cada una de las dos técnicas.



C) Respóndase a las siguientes preguntas:

1. Formular el modelo de la Support Vector Machine (SVM) para regresión interpretando sus elementos desde un punto de vista conexionista y representando gráficamente su estructura.

2. Describir la secuencia de técnicas de regresión (modelos y problemas de optimización/aprendizaje), desde el modelo lineal de regresión hasta el modelo SVM para regresión, explicando las modificaciones (y su finalidad) que se introducen en cada técnica (modelo/problema de aprendizaje) para dar lugar a la siguiente.