Cuestiones

Responda breve y **razonadamente** a las siguientes cuestiones:

1. **(2)** Considere el gráfico de dispersión de la figura 1 que representa las cuatro variables del problema iris. ¿Qué información se puede extraer acerca del problema iris y de cada una de las variables de la gráfica?

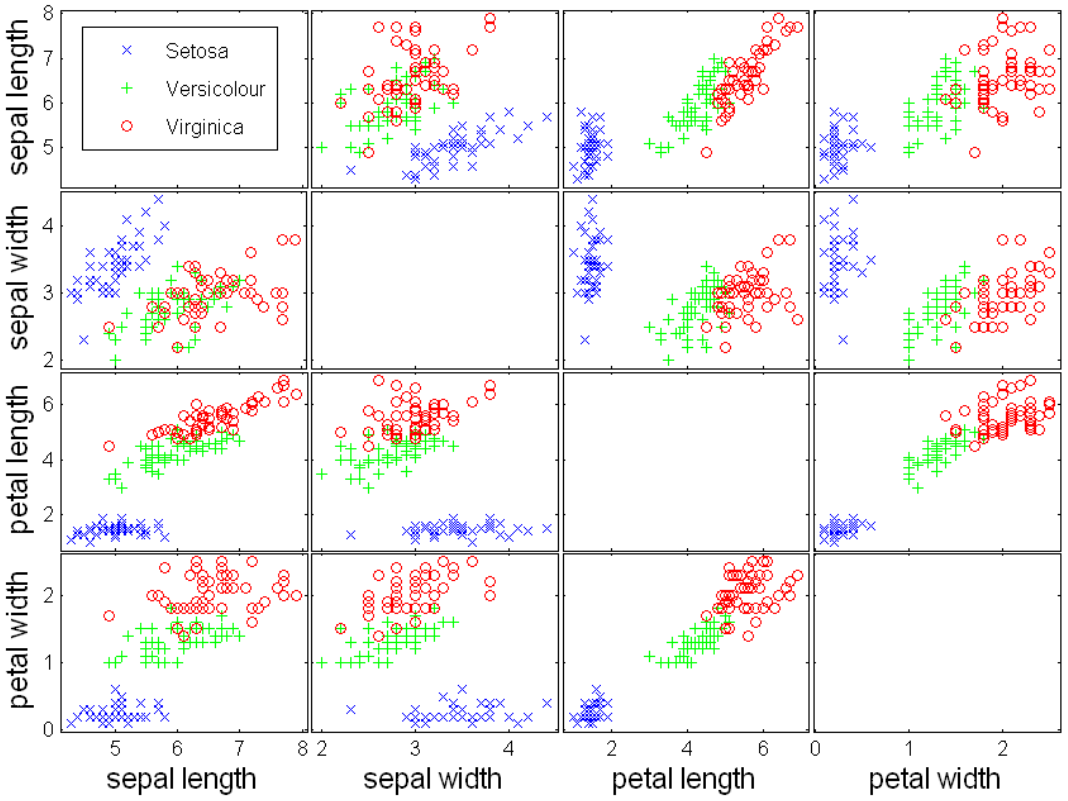


Figure 1: Gráfico de dispersión del problema iris.

RESPUESTA: Se pueden obtener diferentes conclusiones. Como idea general podemos afirmar que el problema será fácil de resolver porque hay varios pares de variables que muestran una clara separación entre las clases. Cualquier algoritmo de clasificación será capaz por lo tanto de aprender una clasificación correcta. Por otro lado, respecto a las variables, vemos que *sepal-width* muestra en general una peor separación entre clases mientras que las otras tres variables son mucho más discriminantes.

2. **(2)** Indique qué test de hipótesis sería el más adecuado en cada una de las siguientes situaciones:

   (a) Deseamos comparar dos métodos de clasificación aplicados a una serie de problemas de prueba para establecer si uno de los dos es mejor que el otro.

   (b) Deseamos comprobar si un método que hemos desarrollado es superior a 4 métodos previos. Hemos probado los cinco métodos en un conjunto de problemas de prueba.

(c) Deseamos comprobar si existen diferencias en un conjunto de problemas de prueba entre 6 métodos de clasificación estándar.

RESPUESTA: Para el primer caso podemos aplicar el test de Wilcoxon ya que es un test diseñado para comparar dos métodos sobre un conjunto de problemas. En el segundo caso las comparaciones en parejas usando el test de Wilcoxon no serían apropiadas porque acumularíamos los errores. El procedimiento de Holm sería lo recomendable. Finalmente cuando comparamos un grupo de métodos entre sí el test de Nememyi sería el apropiado siempre y cuando hayamos hecho previamente un test global de diferencias como Friedman o Iman-Davenport.

3. **(2)** ¿Qué función cumple el parámetro C en la construcción de una máquina de vectores soporte (SVM)?

   RESPUESTA: Los SVM tratan de resolver el problema de clasificación mediante un modelo lineal. En la mayoría de los casos no es posible separar todas las instancias usando un clasificador lineal, en esos casos C corresponde a la penalización por clasificar incorrectamente un patrón.

4. **(2)** ¿Qué pros y contras encuentra en el algoritmo bisecting k-medias respecto al algoritmo k-medias estándar?

   RESPUESTA: Como ventajas tiene la no necesidad de fijar los centroides iniciales y que provee como solución final un clustering jerárquico además del particional. Como inconveniente su mayor coste computacional.

5. **(2)** Considere un problema de clasificación en dos clases, Valor = {Bajo, Alto}, con los siguientes atributos:

   - Aire acondicionado = {Funcionando, Roto}
   - Motor = {Bueno, Malo}
   - Kilometraje = {Alto, Medio, Bajo}
   - Corrosión = {Sí, No}

   Considere un clasificador basado en reglas con el siguiente conjunto de reglas:

   Kilometraje = Alto → Valor = Bajo
   Kilometraje = Bajo → Valor = Alto
   Aire acondicionado = Funcionando, Motor = Bueno → Valor = Alto
   Aire acondicionado = Funcionando, Motor = Malo → Valor = Bajo
   Aire acondicionado = Roto → Valor = Bajo

   Responda a las siguientes preguntas:

   (a) ¿Son las reglas mutuamente excluyentes?
   RESPUESTA: No.

   (b) ¿Es el conjunto de reglas exhaustivo?
   RESPUESTA: Sí.

   (c) ¿Es necesario ordenar las reglas?
   RESPUESTA: Sí, porque hay instancias que disparan más de una regla.

   (d) ¿Es necesario definir una clase por defecto para el conjunto de reglas?
   RESPUESTA: No, porque el conjunto es exhaustivo.

Problemas

Resuelva los siguientes problemas.

1. **(3)** La siguiente tabla representa la matriz de distancias entre los elementos de un cierto conjunto de datos compuesto por 6 puntos:

	p1	**p2**	**p3**	**p4**	**p5**	**p6**
p1	0.00	—	—	—	—	—
p2	0.72	0.00	—	—	—	—
p3	0.41	0.73	0.00	—	—	—
p4	0.18	0.59	0.65	0.00	—	—
p5	0.97	0.21	0.38	0.35	0.00	—
p6	0.76	0.11	0.87	0.49	0.33	0.00

Realice el dendograma correspondiente al clustering jerárquico mediante los métodos de *single link* y *complete link* y comente los resultados.

RESPUESTA: ***Single link.***

Tablas de distancias:

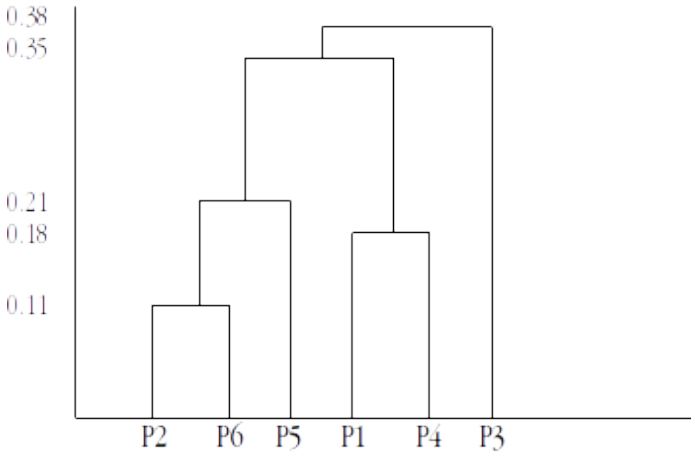
	p1	**p2,6**	**p3**	**p4**	**p5**
p1	0.00	—	—	—	—
p2,6	0.72	0.00	—	—	—
p3	0.41	0.73	0.00	—	—
p4	**0.18**	0.49	0.65	0.00	—
p5	0.97	0.21	0.38	0.35	0.00

	p1,4	**p2,6**	**p3**	**p5**
p1,4	0.00	—	—	—
p2,6	0.49	0.00	—	—
p3	0.41	0.73	0.00	—
p5	0.35	**0.21**	0.38	0.00

	p1,4	**p2,5,6**	**p3**
p1,4	0.00	—	—
p2,5,6	**0.35**	0.00	—
p3	0.41	0.38	0.00

	p1,2,4,5,6	**p3**
p1,2,4,5,6	0.00	—
p3	0.38	0.00

Dendrograma:



Complete link.

Tablas de distancias:

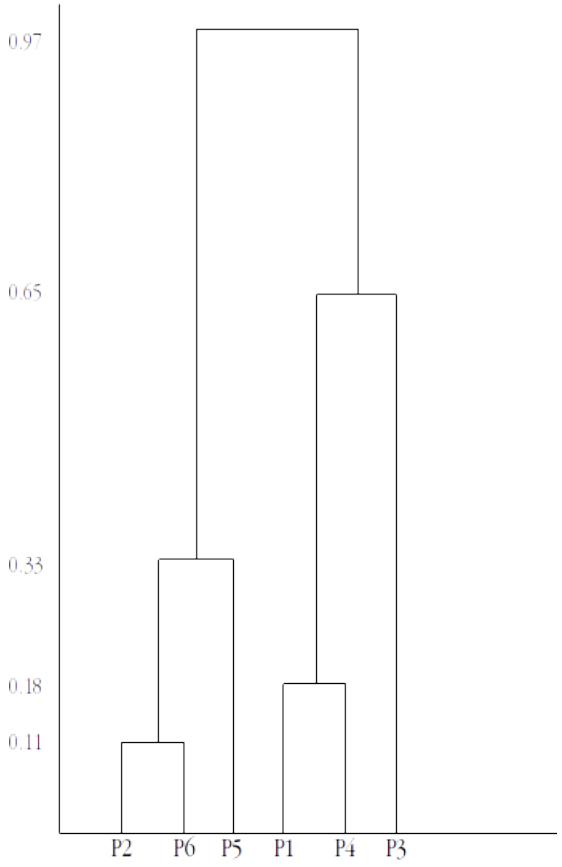
	p1	**p2,6**	**p3**	**p4**	**p5**
p1	0.00	—	—	—	—
p2,6	0.76	0.00	—	—	—
p3	0.41	0.87	0.00	—	—
p4	**0.18**	0.59	0.65	0.00	—
p5	0.97	0.33	0.38	0.35	0.00

	p1,4	**p2,6**	**p3**	**p5**
p1,4	0.00	—	—	—
p2,6	0.76	0.00	—	—
p3	0.65	0.87	0.00	—
p5	0.97	**0.33**	0.38	0.00

	p1,4	**p2,5,6**	**p3**
p1,4	0.00	—	—
p2,5,6	0.97	0.00	—
p3	**0.65**	0.87	0.00

	p1,3,4	**p2,5,6**
p1,3,4	0.00	—
p2,5,6	0.97	0.00

Dendrograma:



Comentarios: Como suele ser frecuente, *single link* tiende a crear clústers por agregación mientras que *complete link* tiende a crear clúster más pequeños que se van agrupando conforme avanza el método.

2. **(4)** Considere un conjunto de datos dónde cada instancia tiene 3 atributos, dos de ellos de tipo lógico y un tercero nominal y tres clases, c_1, c_2 y c_3. La siguiente tabla indica el número de instancias para cada una de las tres clases en función de los valores posibles de los atributos:

Atributos			**Número de instancias**		
x_1	x_2	x_3	c_1	c_2	c_3
V	V	a	36	7	23
V	V	b	10	44	8
V	V	c	7	1	50
V	F	a	7	16	23
V	F	b	6	8	4
V	F	c	6	34	9
F	V	a	28	10	22
F	V	b	11	8	45
F	V	c	5	4	32
F	F	a	3	18	0
F	F	b	34	3	0
F	F	c	15	0	35

Construya un árbol de decisión **binario** utilizando una estrategia voraz y cómo criterio de división de cada nodo el error de clasificación. Para crear una nueva división es necesario que el nuevo subárbol mejore al nodo padre.

RESPUESTA: En fichero adjunto "Problema 2.pdf".

3. **(3)** Considere dos clasificadores C_1 y C_2 cuyas salidas para el conjunto de patrones de test se muestran en la siguiente tabla:

Instancia	Clase	C_1	C_2
1	+	0.74	0.61
2	+	0.69	0.13
3	-	0.44	0.68
4	-	0.55	0.41
5	+	0.87	0.85
6	+	0.47	0.09
7	-	0.08	0.38
8	-	0.15	0.05
9	+	0.25	0.80
10	-	0.35	0.25

(a) Obtenga las curvas ROC para cada uno de los dos clasificadores y muéstrelas gráficamente.

(b) Obtenga el área bajo la curva ROC de cada uno de los clasificadores y compárelos con dicho valor.

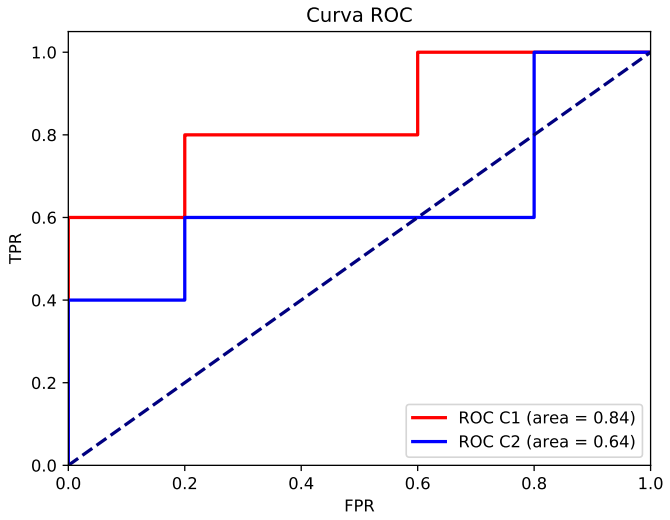
RESPUESTA: Tabla de la curva ROC para C1:

Clase	-	-	+	-	-	+	-	+	+	+	
Umbral $\geq$	0.08	0.15	0.25	0.35	0.44	0.47	0.55	0.69	0.74	0.87	1.00
TP	5	5	5	4	4	4	3	3	2	1	0
FP	5	4	3	3	2	1	1	0	0	0	0
TN	0	1	2	2	3	4	4	5	5	5	5
FN	0	0	0	1	1	1	2	2	3	4	5
TPR	1.0	1.0	1.0	0.8	0.8	0.8	0.6	0.6	0.4	0.2	0.0
FPR	1.0	0.8	0.6	0.6	0.4	0.2	0.2	0.0	0.0	0.0	0.0

Tabla de la curva ROC para C2:

Clase	-	+	+	-	-	-	+	-	+	+	
Umbral $\geq$	0.05	0.09	0.13	0.25	0.38	0.41	0.61	0.68	0.8	0.85	1.00
TP	5	5	4	3	3	3	3	2	2	1	0
FP	5	4	4	4	3	2	1	1	0	0	0
TN	0	1	1	1	2	3	4	4	5	5	5
FN	0	0	1	2	2	2	2	3	3	4	5
TPR	1.0	1.0	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0.0
FPR	1.0	0.8	0.8	0.8	0.6	0.4	0.2	0.2	0.0	0.0	0.0

Curvas ROC y área bajo la curva para ambos clasificadores:



Tiempo de realización: **6** horas. Calificación de cada ejercicio entre paréntesis.