

Solemne 1 - Sistemas Inteligentes

martes 22 de Septiembre 2015

Profesor: Alejandro Figueroa

Ayudantes: Alexander Espina - Daniel Palomera

- Está prohibido el uso de teléfonos celulares durante el desarrollo de la prueba.
- La prueba debe responderse con un lápiz de tinta indeleble, de lo contrario no hay opción a correcciones.
- Cualquier alumno que sea sorprendido intentando copiar será calificado con una nota 1.
- Está prohibido conversar durante la prueba. Recuerde que su compañero puede estar concentrado y el ruido puede perturbarlo en el desarrollo de su prueba.
- Utilice sólo las hojas entregadas para escribir sus respuestas.

Pregunta 1

Considere la siguiente salida de un clasificador. Calcule la matriz de confusión, la entropía del conjunto de datos de acuerdo a las etiquetas manuales y de la distribución asignada por el clasificador agregado el feature <pr>.

```
Feature: <pr>
Reading training examples...done
Training set properties: 12 features, 6228857 examples (1999586 pos / 4229271 neg)
Iter 1: *(NumConst=1, SV=1, CEps=100.0000, QPEps=0.0000)
Iter 2: *(NumConst=2, SV=2, CEps=84.6617, QPEps=0.0000)
Iter 3: *(NumConst=3, SV=2, CEps=31.9446, QPEps=12.9094)
Iter 4: *(NumConst=4, SV=3, CEps=36.1103, QPEps=8.5406)
Iter 5: *(NumConst=5, SV=4, CEps=8.7654, QPEps=3.7114)
Iter 6: *(NumConst=6, SV=4, CEps=4.7753, QPEps=2.2752)
Iter 7: *(NumConst=7, SV=4, CEps=1.6511, QPEps=0.7965)
Iter 8: *(NumConst=8, SV=4, CEps=0.4580, QPEps=0.1789)
Iter 9: (NumConst=8, SV=4, CEps=0.0000, QPEps=0.1789)
Iter 10: (NumConst=8, SV=4, CEps=0.0000, QPEps=0.1789)
Final epsilon on KKT-Conditions: 0.17894
Upper bound on duality gap: 1.07269
Dual objective value: dval=536.79317
Primal objective value: pval=537.86586
Total number of constraints in final working set: 8 (of 8)
Number of iterations: 10
Number of calls to 'find_most_violated_constraint': 57380149
Number of SV: 4
Norm of weight vector: |w|=1.07979
Value of slack variable (on working set): xi=26.86414
Norm of longest difference vector: ||Psi(x,y)-Psi(x,ybar)||=178.95240
Runtime in cpu-seconds: 6.75
Compacting linear model...done
Writing learned model...done
Reading model...done.
Reading test examples...done.
Classifying test examples...done
```

```
Runtime (without IO) in cpu-seconds: 0.17
Average loss on test set: 7.9771
Zero/one-error on test set: 100.00% (0 correct, 1 incorrect, 1 total)
NOTE: The loss reported above is the percentage of errors. The zero/one-error
      is the multivariate zero/one-error regarding the whole prediction vector!
Accuracy : 92.02
Precision: 86.14
Recall   : 90.19
F1       : 88.12
PRBEP    : 85.49
ROCArea  : 86.69
AvgPrec   : 79.25
```

Desarrollo:

Total=6228857 y accuracy 92,02 \Rightarrow TP+TN=5731794 y FP+FN=497063

$0,9019 \cdot (TP+FN) = 0,8614 \cdot (TP+FP) \Rightarrow 1.4792709969 \text{ FN} = \text{FP}$

$2.4792709969 \text{ FN} = 497063 \Rightarrow \text{FN} = 200488 \text{ y } \text{FP} = 296575$

$0,9019 \cdot (TP+FN) = \text{TP} \Rightarrow \text{TP} = 1843222 \text{ y } \text{TN} = 3888572$

Prob. Positivos = $1999586 / 6228857 = 0.32$

Prob. Negativos = $4229271 / 6228857 = 0.68$

Entropía = $-(0.32 \cdot \log_2(0.32) + 0.68 \cdot \log_2(0.68)) = 0.905486$

Prob. Positivos = $(1843222 + 296575) / 6228857 = 0.3435$

Prob. Negativos = $(3888572 + 200488) / 6228857 = 0.6564$

Entropía = $-(0.3435 \cdot \log_2(0.3435) + 0.6564 \cdot \log_2(0.6564)) = 0.928156$

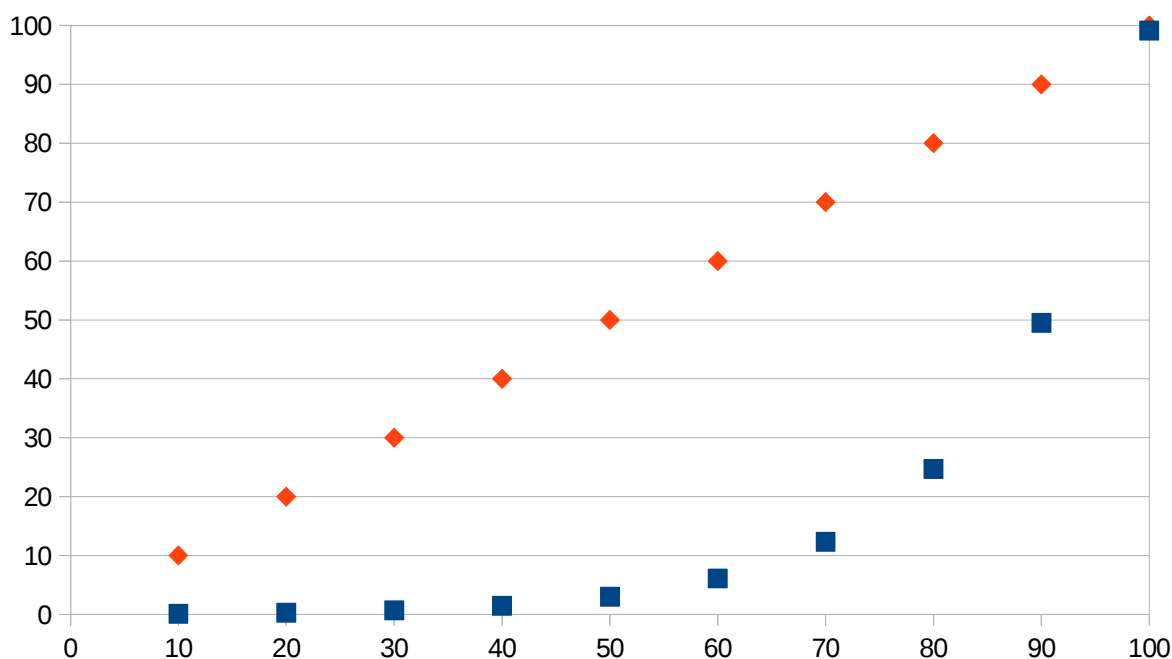
Pregunta 2

Calcule la curva lift para un conjunto de 10,230 datos positivos. Utilice 10 paquetes, además considere que el primero tiene 10 y el ultimo 5120 elementos positivos, respectivamente. La distribución de los ejemplos positivos sigue una progresión geométrica.

Desarrollo:

Si el primer término es 10, y el ultimo 5120, podemos deducir que la razón de la progresión geométrica es 2.

1	10	10	0.0968992248	10
2	20	30	0.2906976744	20
3	40	70	0.6782945736	30
4	80	150	1.4534883721	40
5	160	310	3.003875969	50
6	320	630	6.1046511628	60
7	640	1270	12.30620155	70
8	1280	2550	24.709302326	80
9	2560	5110	49.515503876	90
10	5120	10230	99.127906977	100
		10230		



Pregunta 3

De 3 similitudes y 3 diferencias entre boosting y bagging.

Boosting	Bagging
Método de combinación por votación de clasificadores.	Método de combinación por votación de clasificadores.
Utilizan el mismo clasificador base.	Utilizan el mismo clasificador base.
Normalmente, ambos utilizan un conjunto de entrenamiento de igual tamaño al original.	Normalmente, ambos utilizan un conjunto de entrenamiento de igual tamaño al original.
Tienden a mejorar el desempeño cuando el algoritmo es inestable.	Tienden a mejorar el desempeño cuando el algoritmo es inestable.
Asigna pesos a los distintos ejemplos en el conjunto de datos, de acuerdo a si están o no bien clasificados.	Hace un muestro del conjunto de datos donde pueden haber repetidos en el nuevo set.
Genera una secuencia de clasificadores.	Genera un conjunto de clasificadores.
Clasificadores con gravitancia de acuerdo a su error.	Clasificadores de igual importancia, o bien aprendida con algún método como regresión lineal.

Pregunta 4

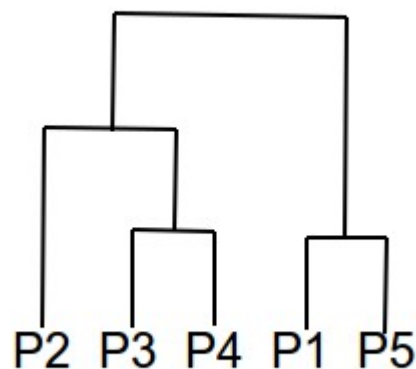
Para los siguientes cinco puntos, desarrolle un clustering aglomerativo utilizando single-link junto con Manhattan y Canberra como métricas de distancia. Grafique los dendrogramas.

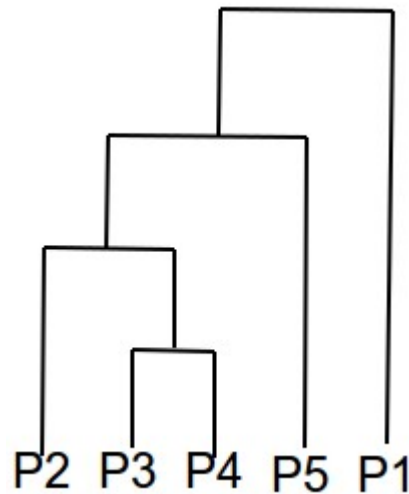
P1:<1,5,6,7>; P2:<3,2,3,1>; P3:<7,7,3,1>; P4:<8,5,3,1>; P5:<5,3,2,7>

Desarrollo:

Distancia Manhattan (1,2)=14.0
 Distancia Manhattan (1,3)=17.0
 Distancia Manhattan (1,4)=16.0
 Distancia Manhattan (1,5)=10.0
 Distancia Manhattan (2,3)=9.0
 Distancia Manhattan (2,4)=8.0
 Distancia Manhattan (2,5)=10.0
 Distancia Manhattan (3,4)=3.0
 Distancia Manhattan (3,5)=13.0
 Distancia Manhattan (4,5)=12.0















Distancia Canberra (1,2)=2.011904761904762
 Distancia Canberra (1,3)=2.0
 Distancia Canberra (1,4)=1.8611111111111112
 Distancia Canberra (1,5)=1.4166666666666665
 Distancia Canberra (2,3)=0.9555555555555556
 Distancia Canberra (2,4)=0.8831168831168831
 Distancia Canberra (2,5)=1.4
 Distancia Canberra (3,4)=0.23333333333333334
 Distancia Canberra (3,5)=1.5166666666666666
 Distancia Canberra (4,5)=1.4307692307692308





Pregunta 5

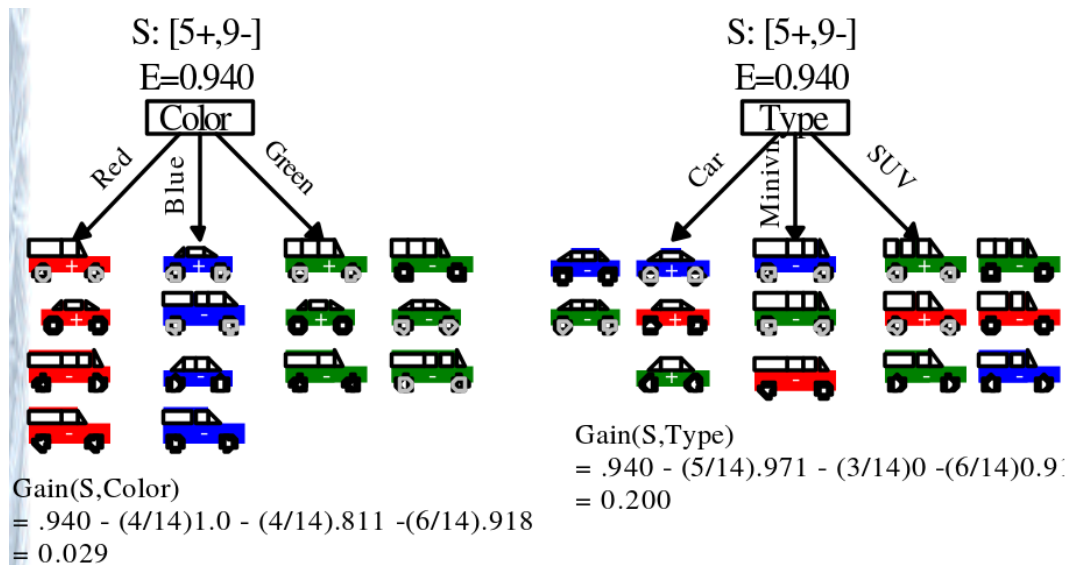
Construya el árbol de decisión para el siguiente conjunto de datos¹:

Color	Type	Doors	Tires	Class	
Red	SUV	2	Whitewall	+	
Blue	Minivan	4	Whitewall	-	
Green	Car	4	Whitewall	-	
Red	Minivan	4	Blackwall	-	
Green	Car	2	Blackwall	+	
Green	SUV	4	Blackwall	-	
Blue	SUV	2	Blackwall	-	
Blue	Car	2	Whitewall	+	
Red	SUV	2	Blackwall	-	
Blue	Car	4	Blackwall	-	
Green	SUV	4	Whitewall	+	
Red	Car	2	Blackwall	+	
Green	SUV	2	Blackwall	-	
Green	Minivan	4	Whitewall	-	

Desarrollo:

Tenemos 5 ejemplos positivos y 9 ejemplos negativos, lo que nos da una entropía para el conjunto de datos de 0.940.

¹Ejercicio obtenido en <http://www.d.umn.edu/~rmaclin/cs5751/notes/Chapter03.PDF>

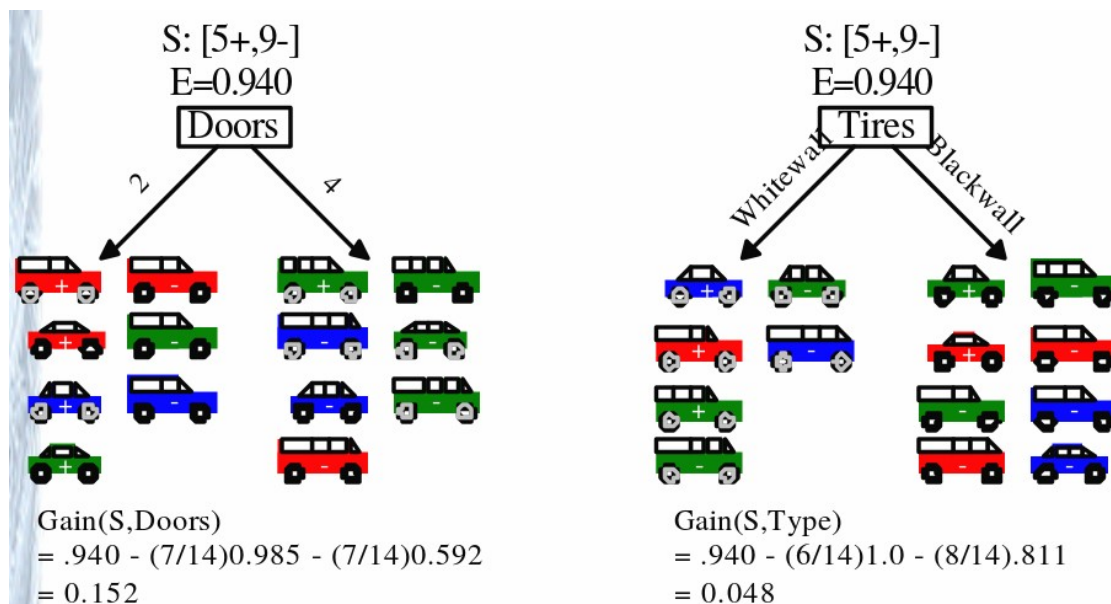


$$\text{Entropy(Blue)} = -\left(\frac{1}{4} \log_2 \frac{1}{4} + \frac{3}{4} \log_2 \frac{3}{4}\right) = 0.811$$

$$\text{Entropy(Green)} = -\left(\frac{2}{6} \log_2 \frac{2}{6} + \frac{4}{6} \log_2 \frac{4}{6}\right) = 0.918$$

$$\text{Entropy(Car)} = -\left(\frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} \log_2 \frac{2}{5}\right) = 0.971$$

$$\text{Entropy(SUV)} = -\left(\frac{2}{6} \log_2 \frac{2}{6} + \frac{4}{6} \log_2 \frac{4}{6}\right) = 0.918$$

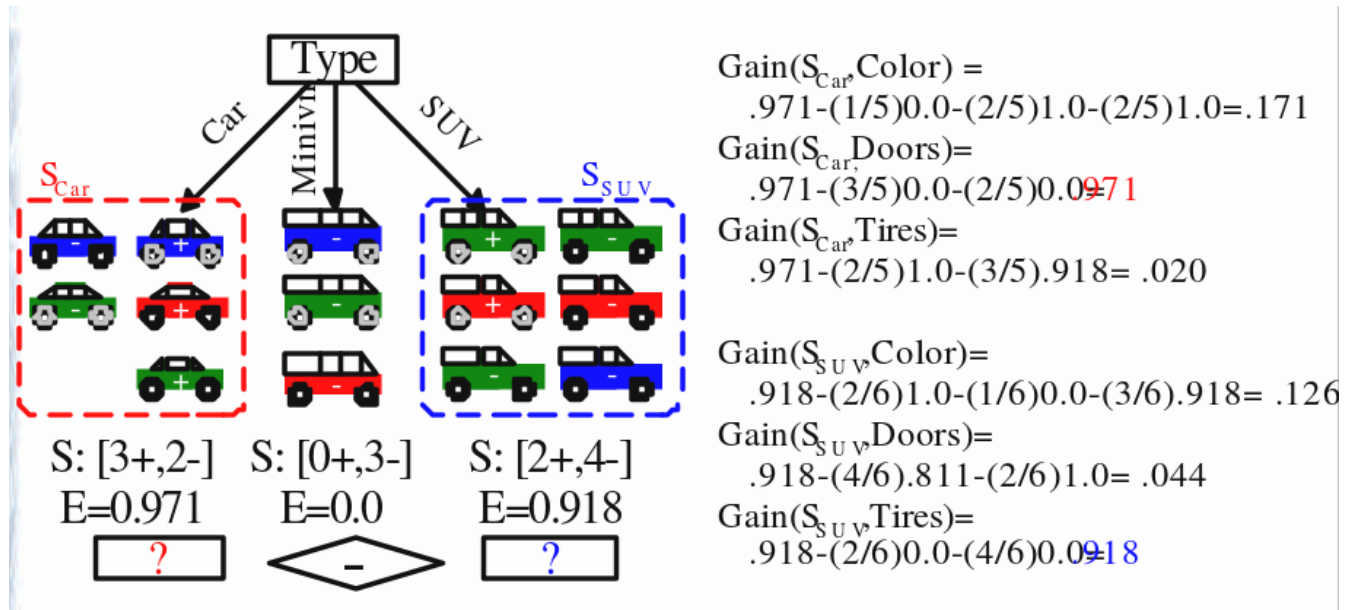


$$\text{Entropy(2)} = -\left(\frac{4}{7} \log_2 \frac{4}{7} + \frac{3}{7} \log_2 \frac{3}{7}\right) = 0.985$$

$$\text{Entropy(4)} = -\left(\frac{1}{7} \log_2 \frac{1}{7} + \frac{6}{7} \log_2 \frac{6}{7}\right) = 0.592$$

$$\text{Entropy}(\text{Blackwall}) = -\left(\frac{2}{8} \log_2 \frac{2}{8} + \frac{6}{8} \log_2 \frac{6}{8} \right) = 0.811$$

El mejor atributo es type, porque nos da un information gain de 0.2. En ese árbol la rama de Minivan no es necesario abrirla porque tenemos sólo ejemplos negativos.



$$\text{Entropy}(S_{\text{Car}}, \text{Tires} = \text{Whitewall}) = -\left(\frac{1}{3} \log_2 \frac{1}{3} + \frac{2}{3} \log_2 \frac{2}{3} \right) = 0.918$$

$$\text{Entropy}(S_{\text{SUV}}, \text{Color} = \text{green}) = -\left(\frac{1}{3} \log_2 \frac{1}{3} + \frac{2}{3} \log_2 \frac{2}{3} \right) = 0.918$$

