Solemne 2 – Sistemas Inteligentes (Pauta)

Jueves 15 de junio de 2016

Profesor: Alejandro Figueroa
Ayudante: Jean Contreras

- 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.
- La nota 4.0 se alcanza con 60 de los 100 puntos que tiene la prueba.

Pregunta 1 (8 puntos)

Demuestre que es posible construir un clasificador SVM para el siguiente conjunto de entrenamiento.

Feature X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Clase	P	P	P	N	N	N	N	N	P	P	P

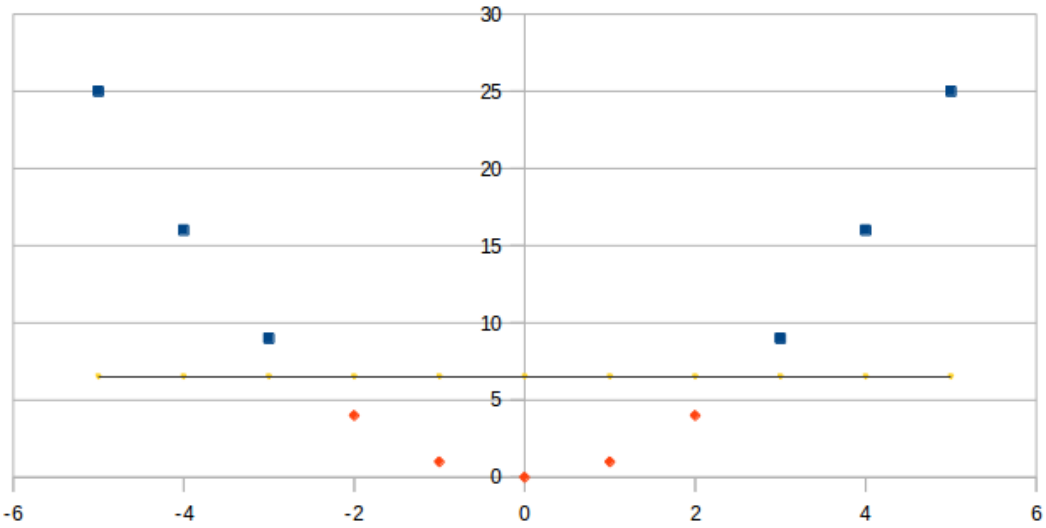
Utilizando el “kernel trick”, podemos crear un feature adicional Y=X*X, donde los elementos “P” estarán con el valor mayor o igual a nueve y los elementos “N” menor e igual a cuatro. Entonces, un plano separador podría pasar por Y= (9+4)/2. En resumen, podemos utilizar un kernel polinomial de grado dos para convertir este data-set en linealmente separable (aumentando la dimensionalidad), ergo poder construir una máquina de soporte vectorial.

Pregunta 2 (12 puntos)

Considere los siguientes datos[1]: ((3,3),r) ((-1,-4),b) ((2,3),r) ((0,-5),b) ¿Qué categoría le entregaría un clasificador kNN al punto (3,4)? Argumente y explicite sus supuestos.

Asumiendo una métrica de distancia euclidiana, el punto más cercano es (3,3) con una distancia de uno. Ergo, si k es igual a uno, entonces la etiqueta asignada sería “r”. El segundo punto más cercano es (2,3) y la distancia respectiva es 1.4. La etiqueta es “r”, ergo la etiqueta asignada al nuevo punto

1 http://www.indiana.edu/~dll/B657/B657_lec_kmeans.pdf



sería “r”. Los otros dos puntos están a una distancia de 8.9 y 9.4, y son “b”. No hay clase mayoritaria, ergo no pueden usarse los cuatro puntos.

Si k es uno, dos y tres la etiqueta asignada es “r”. El caso cuatro no es factible con este dataset, y si se usara, sería aleatorio. ¿Si usáramos pesos? El escenario sería el mismo porque los puntos más cercanos son todos “r” (pensando en estrategias clásicas de asignación de pesos).

Pregunta 3 (20 puntos)

Considere los siguientes datos: (3,3) (-1,-4) (2,3) (0,-5) Muestre el proceso iterativo de K-means. Asuma los centroides iniciales c1=(3,3) y c2=(2,3). Argumente y explicite sus supuestos.

Asumiendo distancia euclidiana: El punto (3,3) está a cero y uno de c1 y c2, respectivamente. El punto (-1,-4) está a 8.1 y 7.6 de c1 y c2, respectivamente. El punto 1 está más cerca del centroide uno y el punto dos del centroide dos. El punto (2,3) está a uno del centroide c1 y a cero del centroide c2, y el punto (0,-5) está a 8.5 y 8.2 de c1 y c2, respectivamente. Ergo, estos dos últimos puntos son asignados a c2.
Ahora es necesario re-calcular los centroides. El primero queda invariante, el segundo es ahora (0.3,-2).
Con este nuevo centroide se calculan las distancias. El punto (3,3) queda a 0 y 5.7 de c1 y c2, respectivamente. El punto (2,3) queda a uno y 5.3 de c1 y c2, respectivamente. Ambos son asignados a c1. Los otros dos puntos están más cerca de c2 (2.4 y 3.0). Las distancias a c1 son 8.1 y 8.5.

Se recalculan los centroides: c1=(2.5,3) y c2=(-0.5,-4.5). Las nuevas distancias a c1 son: 0.5, 7.8, 0.5 y 8.3; en el caso de c2 son: 8.2, 0.7, 7.9 y 0.7. Al recalcular los centroides, quedan igual. No hay re-asignaciones.

Pregunta 4 (30 puntos)

Calcule el plano separador (SVM) para el conjunto de datos: ((3,3),r) ((-1,-4),b) ((2,3),r) ((0,-5),b). Explicite sus supuestos y cálculos.

Asumimos distancia euclidiana, no hay permisividad para ruidos (slack y penalización).
La distancia entre (3,3) y (3,2), (0,-5) y (-1,-4) es uno, 8.5 y 8.1, respectivamente. La distancia entre (3,2) y (0,-5)/(-1,-4) es 7.6 y 7.2 respectivamente. La distancia entre (0,-5) y (-1,-4) es 1.4.

Los dos puntos más cercanos de clases distintas son: (3,2) y (-1,-4). Asumimos estos dos como vectores de soporte, es decir el hiperplano debe pasar exactamente por el medio de esos dos puntos (3.6 de distancia).

La recta que une estos dos puntos está dada por y=mx+b, con m= (-4-2)/(-1-3)=-6/-4=3/2. Reemplazando (4/2=9/2+b) se obtiene b=-5/2.

Sabemos que la distancia entre (3,2)/(-1,-4) al hiperplano es 3.6. Para ser más exactos sqrt(13). Entonces, asumamos que buscamos un punto (x,y) equidistante a ambos puntos:

(3-x)(3-x)+(2-y)(2-y)=13
(-1-x)(-1-x)+(-4-y)(-4-y)=13

Reemplazando la ecuación de la recta en la primera, tenemos: 13x2-78x+65=0. Las raíces son: x1=5 y x2=1 (el primero es infactible para nuestro caso). Con x=1, tenemos y=3/2-5/2=-1

Con la pendiente (perpendicular a la recta que une ambos puntos) y el punto podemos encontrar el plano separador: -1=-3/2*1+b, lo que implica b=0.5.

Pregunta 5 (30 puntos)

Calcule y dibuje el árbol de decisión completo para los datos contenidos en la siguiente tabla[2]:

Ejemplo	Sitio de acceso A_1	1ª cantidad gastada A_2	Vivienda (zona) A_3	Última compra A_4	Clase
1	1	0	2	Libro	Bueno
2	1	0	1	Disco	Malo
3	1	2	0	Libro	Bueno
4	0	2	1	Libro	Bueno
5	1	1	1	Libro	Malo
6	2	2	1	Libro	Malo

Desarrollo:
Hay dos clases equi-probables, ergo la entropía del conjunto de datos es 1.

2 http://banzai-deim.urv.net/~riano/teaching/id3-m5.pdf

$$
\begin{aligned}
I(A_1) = & \sum_{j=1}^{nv(A_1)} \frac{n_{ij}}{n} I_{ij} = \sum_{j=1}^{3} \frac{n_{ij}}{6} I_{ij} = \\
& \frac{n_{10}}{6} I_{10} + \frac{n_{11}}{6} I_{11} + \frac{n_{12}}{6} I_{12} = \frac{1}{6} I_{10} + \frac{4}{6} I_{11} + \frac{1}{6} I_{12} \\
& I_{10} = -\sum_{k=1}^{2} \frac{n_{10k}}{n_{10}} \log_2 \frac{n_{10k}}{n_{10}} = -\frac{1}{1} \log_2 \frac{1}{1} - \frac{0}{1} \log_2 \frac{0}{1} = 0 \\
& I_{11} = -\sum_{k=1}^{2} \frac{n_{11k}}{n_{11}} \log_2 \frac{n_{11k}}{n_{11}} = -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 1 \\
& I_{12} = -\sum_{k=1}^{2} \frac{n_{12k}}{n_{12}} \log_2 \frac{n_{12k}}{n_{12}} = -\frac{0}{1} \log_2 \frac{0}{1} - \frac{1}{1} \log_2 \frac{1}{1} = 0
\end{aligned}
$$

$$I(A_1) = \frac{1}{6}I_{10} + \frac{4}{6}I_{11} + \frac{1}{6}I_{12} = \frac{1}{6}0 + \frac{4}{6}1 + \frac{1}{6}0 = 0,66$$

$$I(A_2) = \frac{2}{6}I_{20} + \frac{1}{6}I_{21} + \frac{3}{6}I_{22} = \frac{2}{6}1 + \frac{1}{6}0 + \frac{3}{6}(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}) = 0,79$$

$$I(A_3) = \frac{1}{6}I_{30} + \frac{4}{6}I_{31} + \frac{1}{6}I_{32} = \frac{1}{6}0 + \frac{4}{6}(-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}) + \frac{1}{6}0 = 0,54$$

$$I(A_4) = \frac{1}{6}I_{4Disco} + \frac{5}{6}I_{4Libro} = \frac{1}{6}0 + \frac{5}{6}(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}) = 0,81$$

El atributo escogido es A3 (vivienda). La próxima iteración es como sigue (la entropía del conjunto de datos es ahora 0.81):

$$I(A_1) = \frac{1}{4}I_{10} + \frac{2}{4}I_{11} + \frac{1}{4}I_{12} = \frac{1}{4}0 + \frac{2}{4}0 + \frac{1}{4}0 = 0$$

$$I(A_2) = \frac{1}{4}I_{20} + \frac{1}{4}I_{21} + \frac{2}{4}I_{22} = \frac{1}{4}0 + \frac{1}{4}0 + \frac{2}{4}1 = 0,5$$

$$I(A_4) = \frac{1}{4}I_{4Disco} + \frac{3}{4}I_{4Libro} = \frac{1}{4}0 + \frac{3}{4}(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}) = 0,23$$