Examen - Sistemas Inteligentes

Jueves 30 de junio de 2015

Profesor: Alejandro Figueroa **Ayudante:** Alexander Espina

- 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.

Sección 1 (20 puntos)

Indique si cada una de las siguientes aseveraciones es verdadera(V) o falsa(F). No es necesario justificar las falsas. Cada respuesta correcta vale 2 puntos, mientras que las incorrectas descuentan 1 punto.

1. ..V.. Una forma de hacer más robusto a K-Means es ejecutarlo un número grande de veces.
2. ..V.. Si se tiene los siguientes tres puntos (1,2) (2,3) y (7,2), Single-link uniría (1,2) y (2,3) en la siguiente iteración.
3. ..V.. Cuando hacemos 3-opt tenemos dos nuevas formas de re-conectar los arcos.
4. ..F.. De acuerdo a lo visto en clases, ACS tiende a favorecer la exploración por sobre la explotación.
5. ..F.. K-Means puede utilizar sólo un subconjunto de los datos para lidiar con los outliers.
6. ..V.. La Varianza Interna calcula la suma de las desviaciones cuadradas entre todos los ítems de datos y su centro asociado.
7. ..F.. Una de las desventajas de k-opt (i.e., k=2,3,…) es que hay que calcular la función objetivo para cada una de las recombinaciones.
8. ..V.. En Simulated Annealing es posible emporar soluciones en cualquier momento del proceso iterativo.
9. ..V.. K-Means es una estrategia de clustering particional.
10. ..F.. Una forma de hacer más robusto a K-Means es ejecutarlo un número grande de iteraciones.

Sección 2 (30 puntos)

1. Un investigador etiqueta un conjunto de datos de manera binaria. Al finalizar el proceso de etiquetado, calcula la entropía y obtiene uno. Al aplicar K-Means, el investigador obtiene dos clústers de igual tamaño y misma entropía (0,721928095). Argumente si este es un buen o mal resultado utilizando la métrica Purity. (10 puntos)

Si los datos tienen una entropía de uno, quiere decir que tenemos la misma cantidad de cada una de las etiquetas. Supongamos que tenemos 100 datos, entonces tendremos 50 y 50.

Por otro lado, el algoritmo K-Means generó dos clústers de igual tamaño, es decir con 50 y 50 vectores. Si la entropía es 0,72, entonces en un clúster tenemos 40 ejemplos positivos y 10 negativos; y en el otro tendremos 40 negativos y 10 positivos.

Dado que el Purity de un clúster es la proporción de su clase mayoritaria, tendremos 0.8 en ambos casos, lo que arrojará un purity de 0.8.

2. Calcule el breakeven point para la siguiente salida (en orden de más a menos relevante): ++--+-+---. Asuma que no existen documentos relevantes adicionales. (10 puntos)

pos	Relevancia	Recall	Precision
1	+	1/4	1/1
2	+	2/4	2/2
3	-	2/4	2/3
4	**-**	**2/4**	**2/4**
5	+	3/4	3/5
6	-	3/4	3/6
7	+	4/4	4/7
8	-	4/4	4/8
9	-	4/4	4/9
10	-	4/4	4/10

3. Si el desempeño de un clasificador está gobernado por la ecuación TPR = (e^{FPR}-1)/(e-1), enuncie su curva AUC y comente su desempeño. (10 puntos)

La curva AUC se obtiene integrando la curva de desempeño, por lo que está gobernada por la relación: TPR=(e^{FPR}-X+C)/(e-1). La constante de integración "C" se determina con la condición que en (0,0) el valor debe ser cero, por ende C=-1. Para diferentes valores obtenemos:

0	0	0
0,1	0,0612070246	0,0030093539
0,2	0,1288512481	0,0124559067
0,3	0,2036096767	0,0290166646
0,4	0,2862305179	0,0534398351
0,5	0,3775406688	0,0865523154
0,6	0,4784539921	0,129267968
0,7	0,5899804623	0,1825967675
0,8	0,7132362737	0,2476549082
0,9	0,849455012	0,3256759758
1	1	0,4180232931*

Es decir, es un clasificador que tiene un desempeño menor a un clasificador aleatorio (AUC=0,5), por lo que sería conveniente invertir las etiquetas de salida.

*sólo es necesario calcular este punto para comentar el desempeño, el resto de los número se presentan con un fin educativo.

Sección 3 (50 puntos)

1. Compare SVM y K-Means. De cinco aspectos. (15 puntos)

SVM	K-Means
Supervisado o semi-supervisado	No-supervisado
Parte el espacio en dos secciones	Parte el espacio en almenos dos secciones.
Intenta encontrar la partición que maximiza la separación de las clases.	Intenta encontrar el centroide que representa a cada uno de los grupos de la mejor forma.

Determinista. Cada vez que lo ejecutas con los mismos vectores produce el mismo modelo.	Cada vez que se ejecuta, se obtiene un modelo diferente.
No necesita una instanciación inicial	Necesita una instanciación inicial.

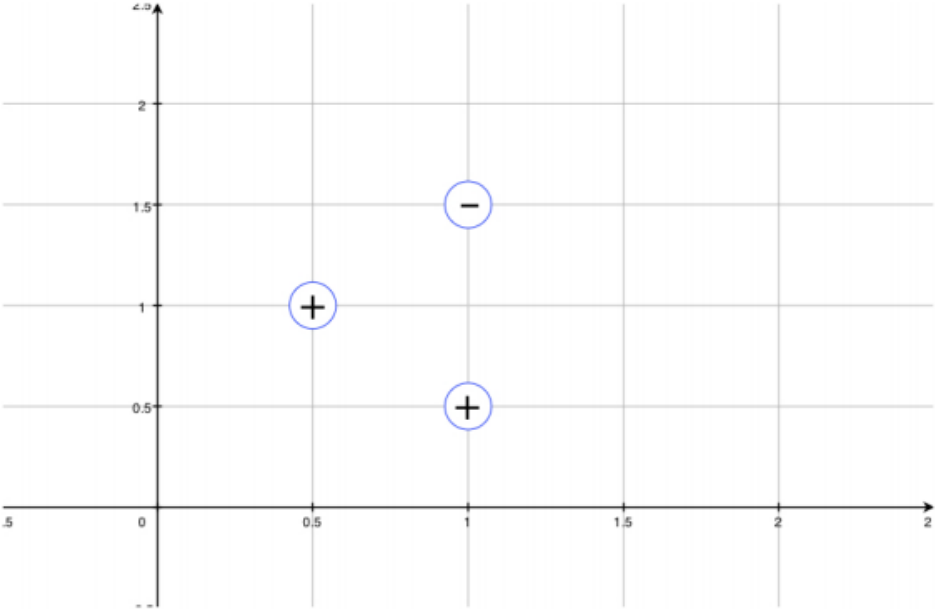
2. ¿Cree ud. que se podría implementar el "kernel trick" en K-means? De poderse, en qué casos teoriza que podría ser útil. (15 puntos)

K-Means no es útil cuando los datos no son particionables, como por ejemplo las distribuciones de hiper-esfera. Algo similar ocurre con SVM. Para esas situaciones, uno podria utilizar el "kernel trick" o algo similar, para calcular las distancias al centroide en un espacio alternativo, donde los puntos si sean particionable.

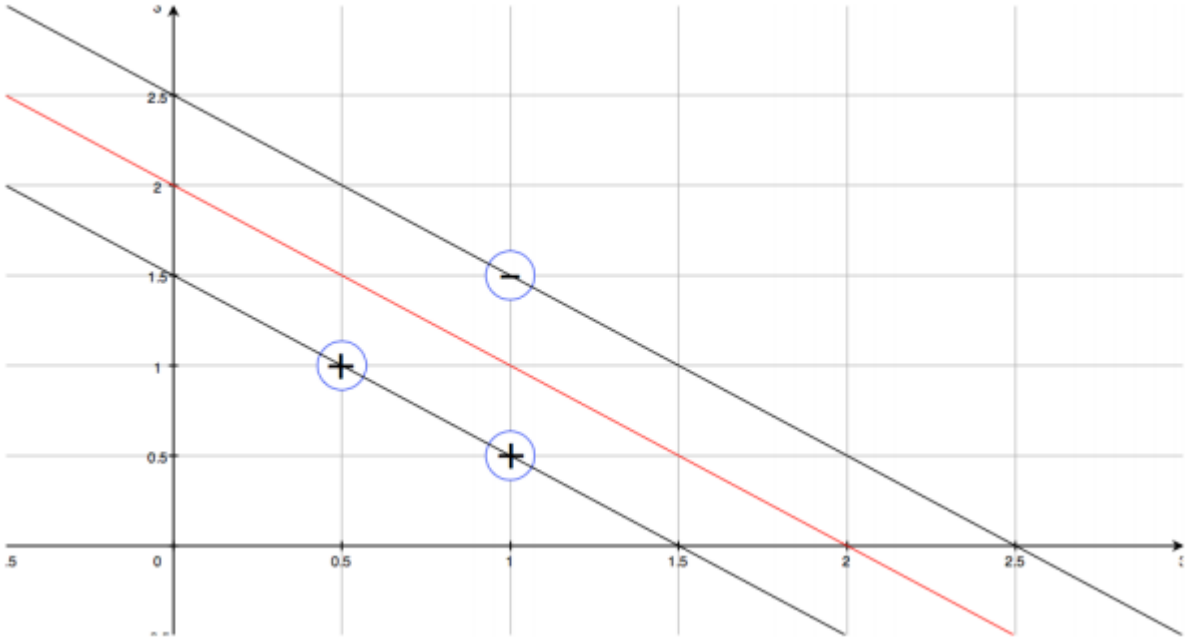
3. Considere un conjunto de datos con entropía 1 y con etiquetas binarias. Analice la accuracy de un clasificador de acuerdo a las diferentes entropías de sus salidas. (10 puntos)

Entropia	Accuracy	proporción
0	0,5	100/0
0	0,5	0/100
0,46899	0,4	90/10
0,46899	0,4	10/90
0,721928	0,3	20/80
0,721928	0,3	80/20
0,88129	0,2	70/30
0,88129	0,2	30/70
0,97095	0,1	40/60
0,97095	0,1	60/40
1	0	50/50

4. Considere los tres vectores de entrenamiento mostrados en la siguiente figura. Nótese que son linealmente separables. Encuentre el SVM lineal que separa las clases de manera óptima mediante la maximización del margen. (10 puntos)



Los tres vectores son vectores de soporte. El plano H+ pasa por los dos puntos positivos. El margen H- pasa por el punto negativo. H- debe ser paralelo a H+. El límite debe pasar por el medio de H+ y H-, lo que nos da la ecuación -x+2=0. La ecuación de H+ es y=-x+1.5 y la de H- es y=-x+2.5.


2.5
2
1.5
1
0.5
0
0.5
1
1.5
2
2.5