INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN5526 - Web Intelligence
Primavera 2016

Profesores: Juan D. Velásquez
Gaspar Pizarro V.
v Prof. auxiliar: Tomás Valdivia H.

Solución control 1

Pregunta 1

(a) ¿Cuáles son las tres divisiones clásicas del *Web Mining*? De una definición corta de cada una. *1 punto*

Web content mining: Minería de texto en las páginas web. Web structure mining: Minería del grafo de la web, de los enlaces, etc. Web usage mining: Minería de los logs de los servidores.

(b) Explique los cinco pasos del proceso KDD. *1 punto*

Selección: Seleccionar los datos con qué trabajar. Preprocesamiento y limpieza: Eliminación de ruido, valores perdidos, filtrado de valores irrelevantes. Transformación: Transformación en vectores. Minería de datos: Aplicación de modelos. Evaluación: Evaluación por parte de experto e interpretación cuando es posible.

(c) Considere las medidas de desempeño o fuerza de una regla de asociación: soporte y confianza. ¿Cuál es el problema cuando se encuentran reglas con mucha confianza pero muy poco soporte? *0,5 puntos*

Cuando una regla tiene mucha confianza pero poco soporte significa que, a pesar de que sea una regla "segura", ocurre muy pocas veces como para tomar decisiones de negocio útiles con ésta.

(d) Describa la estrategia de *k-fold cross validation* y su utilidad en el proceso de construcción de un modelo de clasificación. *1 punto*

Se divide el juego de datos de entrenamiento en k folds, para cada fold se hace lo siguiente: Se entrena el modelo con los otros k-1 folds y se evalua con el fold elegido. El desempeño del modelo es el promedio del desempeño en todos los folds. Sirve para hacer utilizar mejor los datos cuando el juego de entrenamiento es muy pequeño, al contrario de lo que sería si se dividieran los datos en entrenamiento, validación y test.

(e) Describa un criterio para la elección de variables (discretas) para dividir en la construcción de un árbol de decisión. *1* punto

Criterio de ganancia de información: Se decide poner un nodo en la variable cuyo corte genera la mayor ganancia de información, esto es, la que causa el mayor descenso de la entropía de las etiquetas, comparando la entropía de las etiquetas de todos los datos con la entropía de los conjuntos generados por dividir los datos con la variable escogida.

(f) En una máquina de vectores de soporte, ¿cuál es el rol de los vectores de soporte en la definición del hiperplano separador? *0,5 puntos*

Los vectores de soporte son los que definen el hiperplano separador, al ser los que definen la zona de márgen de este. Si estos vectores se mueven, el hiperplano se mueve, al contrario de los otros vectores.

(g) ¿Dónde está la mayor parte del tiempo de computo en la construcción y uso de un clasificador de vecinos más cercanos (K-NN)? Compare con una máquina de vectores de soporte *1 punto*

Está en el uso del clasificador. Al contrario de una máquina de vectores de soporte, el entrenamiento de un knn no hace nada, mientras que el uso de un clasificador knn implica comparar con todos los datos cada vez que se quiere hacer una clasificación.