

Tarea 1 de Inteligencia Artificial

Prof. Carlos Hurtado

4 de septiembre de 2008

El objetivo de la tarea es familiarizar al alumno con el proceso de inducción de un modelo a partir de datos reales. También se espera que alumno aprenda a manejar el software Weka¹ e investigue algunas de sus funcionalidades.

En la tarea usaremos un conjunto de datos de 1000 registros con información de crédito de un banco real en Alemania (que serán entregados en un archivo denominado *credit-g.arff*). Cada registro representa un crédito, con información referente al deudor, plazos, monto, condiciones, etc. La variable de la clase, “Class”, toma el valor “good” o “bad” dependiendo de si el deudor cumplió o no sus obligaciones en el crédito.

Como resultado del trabajo, se debe entregar un informe explicando los resultados obtenidos para cada una de las secciones que se explican a continuación.

Su informe debe ser claro y bien escrito, mostrar análisis y presentar resultados con gráficos, cuando sea necesario.

Para la inducción y prueba de los modelos se deben considerar la siguiente partición de los datos:

- Datos de Prueba: corresponden a todos los registros cuyo id sea múltiplo de 3.
- Datos de Entrenamiento: registros restantes.

Para evaluar los modelos a construir se considera la tasa de error sobre los datos de prueba.

El día Lunes 22 de Septiembre deben entregar el mejor modelo obtenido hasta el momento, como archivo “model object file”. Solo deben entregar el modelo, sin informe.

El día Viernes 26 de Septiembre se publicará en ucursos un ranking de los modelos entregados con sus tasas de error. La idea de esta etapa es que el grupo pueda evaluar cómo se encuentra su modelo con respecto a los modelos de otros grupos del curso.

El Lunes 06 de Octubre deberán entregar el informe final de la tarea con los archivos de los mejores modelos generados en cada categoría.

¹ www.cs.waikato.ac.nz/ml/weka/

1. Inducción de Modelos

1. Construir el mejor árbol de Decisión usando el método J48 de Weka. Lo más importante en esta etapa es determinar el mejor Árbol usando los mecanismos de poda de J48, es decir, evitando sobreajuste.
2. Construir el mejor clasificador Bayesiano Naive de Decisión usando alguno de los métodos de Weka.
3. Encontrar la mejor red Bayesiana usando el método *BayesNet*.

Investigue los distintos algoritmos de búsqueda de estructura que provee Weka, con distintos parámetros, para ver cuál encuentra la mejor estructura: considere métodos basados en verificación de independencia (bajo categoría “ci”) y métodos basados en puntajes locales (bajo categoría “local”).

2. Comparación de Modelos

Use el módulo de experimentación de Weka para producir un reporte que compare los tres modelos obtenidos en los pasos anteriores. Escriba un breve análisis de los resultados.

3. Entrenamiento Sensible al Costo

Para cada uno de los tres modelos seleccionados, Weka debe entregar un reporte sensible al costo, considerando la siguiente función de costo:

$$\text{Costo} = 5 \times FP + 1 \times FN,$$

donde FP son los falsos positivos y FN son los falsos negativos contados sobre los 300 registros de prueba.

Después de esto, deberá analizar cada modelo obtenido como una función discriminante y encontrar el umbral de la clase *good* que minimice el costo.

Deberá determinar el modelo y umbral con el menor costo.