INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

fcfm

Tarea 3
CSI: Santiago
Otoño 2010

IN643/IN4521 - Introducción a la Minería de Datos
Prof. Richard Weber
Aux. Gastón L'Huillier

INDICACIONES:

- Fecha entrega: 19/07/2010 antes de las 23:59 hrs.
- Entrega: Vía u-cursos
  1. Un informe técnico específicando todos los detalles del modelo propuesto. Incluir todas las referencias.
  2. El modelo y su esquema de evaluación para verificar el poder predictivo de su modelo.
- En grupos de máximo 4 personas.
- **Tarea coeficiente 2**.

Como experto analísta en minería de datos, el laboratorio *CSI: Santiago* de la PDI lo ha contactado para desarrollar un modelo predictivo que permitirá resolver crímenes cuya evidencia sean muestras de vidrios. La idea es que dado ciertas propiedades fisicoquímicas, se pueda determinar el origen del tipo de vidrio presente en la escena del crimen. Muchas veces este tipo de evidencia es clave para descartar sospechosos, y dirigir correctamente la investigación.

Se tomaron muestras de vidrios de varios crímenes a lo largo de todo Chile, y sus propiedades se tabularon en una base de datos con 214 muestras (**glass.csv**) cuyas características están descritas en el archivo adjunto (**glass.names**). Dadas las características de los productores de vidrio, en Chile se tienenn 7 tipos de vidrios, de los cuales sólo 6 fueron registrados en las distintas escenas policiales.

En esta competencia están concursando distintas empresas y grupos de investigación expertos en minería de datos. Las bases de la competencia son simples y están basados en la siguiente pauta de evaluación:

1. Debe considerar **obligatoriamente** un modelo de múltiples clasificadores.
2. Debe evaluar su modelo utilizando *Cross Validation* de 10 *folds*.
3. Puede modificar a su gusto la base de datos entregada, es decir, puede crear nuevos atributos, agregar observaciones, etc. (**glass.csv**)
4. Los equipos serán evaluados en base a los siguientes parámetros:
   (a) Modelos cuyo Accuracy sea menor a un 60% serán descartados (y tendrán nota final de predicción 4.0)
   (b) Modelos cuyo Accuracy tenga al menos un `Class Recall` o `Class Precission` de 0.0% serán aceptados con observaciones de tipo 1. (optan a una nota de predicción máxima 5.3)
   (c) Modelos cuya desviación estándar del Accuracy (en 10-folds Cross Validation) sea mayor a un 15%, serán aceptados con observaciones de tipo 2. (optan a una nota de predicción máxima 5.8)

(d) Modelos cuyo Accuracy sea mayor a 60% y menor o igual que 65% serán descartados de la fase I (nota de predicción 4.5)

(e) Modelos cuyo Accuracy sea mayor a 65% y menor o igual que 70% serán descartados de la fase II (nota de predicción 5.2)

(f) Modelos cuyo Accuracy sea mayor a 70% y menor o igual que 75% serán descartados de la fase III (nota de predicción 5.6)

(g) Modelos cuyo Accuracy sea mayor a 75% y menor o igual que 80% serán descartados de la fase final (nota de predicción 6.3)

(h) Modelos cuyo Accuracy sea mayor a 80% serán elegidos en la fase final (nota de predicción máxima 7.0)

(i) La nota final será

$$\text{Nota Final} = \frac{[\text{Nota Informe} + \text{Nota Predicción}]}{2}$$

.

(j) Por día de atrazo i, se hará un descuento de

$$\text{Nota Final}^{\text{Atrazo}} = \text{Nota Final} - \frac{1}{2} exp\left(\frac{1}{3} \cdot i\right)$$

(k) El equipo con mayor nota general ganará el concurso y tendrá **un punto extra en el Examen**.

5. El informe técnico debe contener los siguientes puntos:

(a) Estrategia de múltiples clasificadores utilizada y justificación. [1 punto]

(b) Explicación de los modelos de clasificación utilizados. [1 punto]

(c) Estrategia de selección de parámetros por cada modelo. [1 punto]

(d) Evaluación del modelo. [1 punto]

(e) Conclusiones. [1 punto]

(f) Referencias. [1 punto]