# Práctico Aprendizaje Automático - v0.2

Aprendizaje Automático sobre Grandes Volúmenes de Datos

September 17, 2014

http://aprendizajengrande.net/practico1.pdf

## 1 Ejercicio Naïve Bayes

Tenemos un problema de clasificación binaria con dos *features* numéricas $a$ y $b$ que no son mutualmente independientes[1] dadas la clase objetivo. Los datos pueden resumirse como

$$y = \begin{cases} 2a + b > 5 & \to 1 \\ 2a + b \leq 5 & \to 0 \end{cases}$$

Con los siguientes datos (en formato $(a, b, y)$ ), calcular[2] los estimadores ML y MAP para $(a, b) = (1, 3)$ usando (1) el Teorema de Bayes[3] (2) la simplificación Naïve Bayes[4]. (En ambos casos estimar los priors a partir de los datos dados.)

{ (5,0,1); (9,1,1); (4,0,1); (6,0,1); (4,1,1); (1,2,0); (8,2,1); (7,3,1); (0,2,0); (2,3,1); (4,1,1); (7,4,1); (8,1,1); (0,0,0); (9,0,1); (2,3,1); (4,4,1); (9,2,1); (4,1,1); (7,2,1) }

Dividir los datos de entrenamiento en dos partes de forma aleatoria y calcular precision/recall de Naïve Bayes usando cros-validación.

Para Naïve Bayes, usar dos técnicas de smoothing: considerar que los eventos inexistentes aparecen 0.5 veces y considerar que aparecen 1e-4 veces.

## 2 Ejercicio Árboles de Decisión

Armar un ejemplo de datos (con al menos tres *features* y 24 instancias) de forma tal que el algoritmo de árboles de decisión ID3 usando *Information Gain* produzca resultados subóptimos.

[1] Sí, eso contradice la hipótesis de Naïve Bayes, sin embargo lo mismo sucede en muchas aplicaciones del algoritmo, en este ejercicio veremos un poco el impacto de que no ocurra.

[2] Son 4 números distintos, el ML usando el Teorema de Bayes, el MAP usando Teorema de Bayes, el ML usando NB y el MAP usando NB.

[3] Usar la forma cerrada de $P(a, b|y)$.

[4] Calcular $P(a|y), P(b|y)$ a partir de los datos

Para mostrar que los resultados son subóptimos, armar un árbol de decisión a mano y mostrar que tiene mejores resultados comparados con una evaluación cros-validación en tres partes.

## 3 Ejercicio Regresión Logística

Explique en media página porqué necesitamos una restricción de regularización en la búsqueda de coeficientes de regresión logística.

## 4 Ejercicio Clustering

Dados un conjunto de datos, un número $k$ de clústers y una matriz de proximidad completa, es posible obtener un conjunto de $k$ clústers de dos maneras:

1. Utilizando un criterio de corte en el dendograma inducido por la matriz de proximidad.
2. Utilizando k-means empezando con $k$ elementos al azar.

¿Bajo qué circunstancias de inicialización de $k$-Means y condiciones de corte en el dendograma se obtendrán los mismos $k$ clusters en (1) y (2)?

## 5 Ejercicio Recomendación

Explique cómo se podría usar un algoritmo de clasificación para hacer recomendación. ¿Cuál es el espacio de *features*? ¿Es el algoritmo de recomendación por ítems más eficiente en su uso de los datos? ¿Es posible agregar más información sobre los ítems en esta otra formalización? Máximo una página.