Escola Tècnica Superior d'Enginyeria **Informàtica**

Práctica 2. Clasificación de textos

Percepción (PER)

Curso 2018/2019

Departamento de Sistemas Informáticos y Computación

Índice

Objetivos

- Desarrollar un clasificador de textos, más concretamente un filtro *anti-spam*, basado en el clasificador multinomial estudiado en clase de teoría.
- Resolver los problemas prácticos que nos podemos encontrar durante el desarrollo de un clasificador de textos (representación en memoria, suavizado de parámetros, etc.).
- Diseñar un conjunto de experimentos para evaluar el clasificador de textos implementado.
- Analizar el comportamiento (evolución del error) del clasificador implementado en función de sus parámetros.

Corpus TREC 2006

- Tarea: Clasificación de correos electrónicos en dos clases (ham y $spam$)
- 37.822 correos electrónicos **reales** etiquetados por expertos
- Amplio vocabulario = gran número de tokens diferentes
- Echad un vistazo a los datos desde una máquina Linux
- Mas información en: `http://trec.nist.gov`

Representación *bag-of-words*

- Revisad tema 2 de teoría
- Representación como una matriz de correos (filas) por tokens (columnas)
- Cada fila contiene un vector de ocurrencias de los tokens que contiene ese correo
- Última columna es la etiqueta de clase: ham (0) y $spam$ (1)
- La matriz resultante contiene ceros en la mayoría de celdas $\rightarrow$ matriz dispersa
- Utilizaremos representación de matriz dispersa de Octave

Implementación clasificador multinomial (I)

- Revisad tema 5 de teoría
- Lectura de datos
- Diseño de experimento: 90 % entrenamiento y 10 % test de correos barajados
- Entrenamiento: Estimación de parámetros del clasificador multinomial
- Clasificación:
  1. Función discriminante para la clase ham (g_h) y $spam$ (g_s)
  2. Utilizamos estas funciones para clasificar correos de test (y entrenamiento)

$$c^*(\mathbf{x}) = \text{argmax}\,\{g_h(\mathbf{x}),\, g_s(\mathbf{x})\}$$

  3. Comparamos clase obtenida en paso 2 con clase real (última columna de matriz)
  4. Calculamos el error de clasificación en test (y entrenamiento)

Implementación clasificador multinomial (II)

- Suavizado de Laplace

$$\tilde{p}_{cd} = \frac{\hat{p}_{cd} + \epsilon}{\sum_d (\hat{p}_{cd} + \epsilon)}$$

- Estudiamos variación del error en test (y entrenamiento) en función de $\epsilon = 10^{-1}, 10^{-2}, \ldots, 10^{-20}$
- Error depende de la partición entrenamiento-test tras el barajado de correos
- Es necesario calcular el error para 30 bajarados
- Resultados de error promedio e intervalos de confianza al 95 % para cada valor de ϵ

Evaluación

- Esta práctica supone un 15 % del total de la nota de la asignatura (1.5 puntos)

- Para su evaluación se tendrán en cuenta dos actividades:

  1. Memoria del trabajo realizado:
     - Código comentado (0.5 puntos)
     - Gráfica de resultados comparativos (0.25 puntos)
     - Comentarios de los resultados (0.25 puntos)
     - Entrega hasta el **27 de mayo** por tarea en PoliformaT

  2. Competición (0.5 puntos máximo, dependiendo de la tasa de error obtenida)
     - Se proporciona un nuevo conjunto de datos en el mismo formato
     - Objetivo: Buscar la combinación de clasificador (multinomial o Bernoulli), tipo de suavizado y $cantidad$ de suavizado
     - Se realizará durante la última sesión de prácticas