CP 7 Inteligencia Artificial

Tema: Clasificación Supervisada e introducción a la herramienta sklearn

Objetivo

Resolver un problema de clasificación pasando por todo el proceso.

Que los estudiantes sepan resolver un problema aplicando las técnicas de aprendizaje supervisado, clasificación. Deben aprender a entender el problema y representar los datos en una matriz de características y un vector de clases. Además Utilizar la herramienta sklearn para crear un clasificador y evaluarlo.

Introducción

¿Qué es un problema difícil? Se han preguntado alguna vez qué es un problema difícil. ¿Cuándo sienten que están resolviendo un problema difícil?

Aclarar que el concepto de difícil en la computación no es basado en lo que resulta difícil para ellos. Es en lo que resulta difícil para una computadora. Un ejemplo de problemas difíciles son los NP-duros.

Problema

Hoy vamos a estar resolviendo un problema difícil.

Se tienen críticas de cine positivas y negativas. 1000 de cada una. Cada grupo (positivo, negativo) en una carpeta y cada archivo es una crítica diferente.

Descripción de la clase

en este momento ya están trabajando por equipos.

Ejercicio 1

¿Qué hacemos? ¿Por dónde empezamos? ¿Cuál es nuestro problema?

Primero deben reponderse estas preguntas hasta que lleguen a una idea de lo que tienen que hacer:

- Convertir un archivo en una cadena de texto.
- Hacer lo mismo para cada archivo.
- Tener 2 listas con el texto de las críticas positivas y negativas.

Esto más bien se les escribe, no hay que dejar que lo hagan solos. Pero si se quiere que describan el proceso a grandes rasgos, para que sientan que lo hicieron solos.

Para comprobar el resultado de este proceso, comprobar la longitud de las listas.

Ejercicio 2

¿Cómo creamos la matriz de características?

Después que describan el proceso a grandes rasgos se les propone utilizar sklearn, que ya tiene resuelto este problema. Vamos a utilizar sklearn, CountVectorizer

¿Cuántas características hay? ¿Qué por ciento de los valores es distinto de cero?

¿Cómo creamos el vector de clases?

Ejercicio 3

Ahora ya se tienen la Matriz de Características y el vector de clases.

¿Qué hacemos?

¿Qué clasificador utilizamos? ¿Por qué?

Que reflexionen que clasificadores teóricamente se deberían comportar mejor dadas las características del problema. Deberían descartar KNN, por la cantidad de dimensiones y Árboles de decisión porque hay muchos rasgos igualmente importantes. Los prometedores parecen ser Naive Bayes y SVM.

Vamos a probar con Naive Bayes utiliznado sklearn.

Llegado este punto es necesario explicar por qué el Naive Bayes de sklearn se llama Gaussian Naive Bayes. Porque está diseñado (a diferencia del visto en conferencia) para lidiar con características que sean valores continuos. En vez de tener probabilidad del feature dada la clase que es $P(w|c)$ ahora lo que se tiene es la probabilidad de observar un valor tan alto como x_i, esto es $P(x \leq x_i|c)$

Ejercicio 4

Probar Naive Bayes con todo el conjunto para entrenar y todo el conjunto para testear.

¿Por qué da un valor tan alto?

Estamos haciendo trampa, ya que el clasificador ya vio estos ejemplos.

Probar con un conjunto de entrenamiento y un conjunto de prueba. Fijarse en como bajan los resultados.

Ejercicio 5

Probar con todos los clasificadores vistos KNN, Árboles de Decisión y SVM.

Ejercicio 6

Hablarles de cómo se deben hacer los experimentos.

Tarea

Hacer un Naive Bayes que supere al de sklearn utilizado en clase para el corpus utilizado (rotten tomatoes). Debe dar un 80%.