DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Redes Neuronales

Trabajo práctico 1

Parser

Resumen

Entrenamiento de Redes Neuronales

Integrante	LU	Correo electrónico
Negri, Franco	893/13	franconegri2004@hotmail.com
Rey, Martin Gaston Podavin	483/12	marto.rey2006@gmail.com

Palabras claves:

TP

Índice

1. Introduccion

En este trabajo practico nos propondremos clasificar mails en spam como en no spam (conocido tambien como ham).

Para esto analizaremos los datos en busca de atributos utiles y luego experimentaremos con diversos clasificadores con la intención de encontrar los que mejor clasifiquen.

Ademas utilizaremos tecnicas de reduccion de dimencionalidad con la intención de mejorar aun mas los modelos de los puntos anteriores.

Para tener una buena metrica de los resultados obtenidos, apartamos parte del set de datos que nos fue entregado para ser utilzado al final de la experimentación y asi tener una visión realista de que tan buenos son los clasificadores elegidos. Con el resto de los datos, realizamos kfolds con k igual a diez con la intención de minimizar en cierta medida el overfiting de datos.

2. Desarrollo

Como ya adelantamos en la seccion anterior el objetivo de este tp será clasificar mails en spam y no spam, para ellos, comenzaremos el trabajo seleccionando atributos adecuados que concideramos dividen bien el problema.

2.1. Selección De atributos

Para la selección de atributos observamos parte del json entregado en busca de palabras claves que pudieran distinguir entre spam y no spam (desde ahora, ham).

De este analisis se encontró que palabras tales como *viagra* o *nigeria* son muy frecuentes en los mensajes de spam, asi tambien aquellas que hagan referencia a negocios o a dinero, por lo que incluimos atributos que cuenten las veces que son mencionados estas palabras en el cuerpo y el encabezado del mensaje. Así tambien observamos que los mails de spam tienden a tener enlaces hacia sitios web o contenido html en el cuerpo del mensaje por lo que tambien contavilizamos la cantidad de ocurrencias de palabras tales como html o http y simbolos especiales como la barra invertida, o el hashtag.

De esta menera reunimos al rededor de 100 atributos que utilizaremos a continuación para la experimentación.

2.2. Experimentación

Los clasificadores que utilizaremos para experimentar en este trabajo serán:

- Arboles de Deciciones
- Naive bayes Multinomial
- Vecinos mas Cercanos
- SVC
- Random Forest

3. Concluciones