Aprendizaje automático

Grupo Sociofísica
Sebastián Pinto, Guillermo Pasqualetti, Gustavo Landfried

24 de septiembre de 2016

1. Introducción

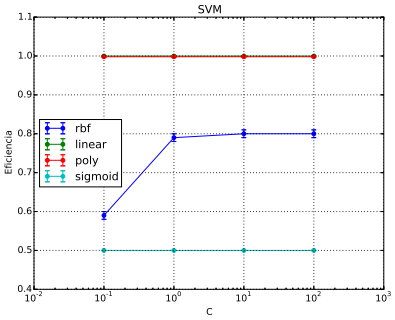
Usamos el lenguaje de progrmación python.
Usamos conceptos del libro "An Introduction to Statistical Learning" [**?**].

2. Extracción de atributos

3. Modelos

3.1. SVM

Decidimos realizar el estudio en un principio para SVM con un subconjunto del dataset original. Del tutorial de *scikit-learn* (CITAR!!!) vimos que el tiempo de ejecución es cuadrático con el número de muestras originales, por lo cual vuelve a este método muy lento de entrenar y validar respecto de los observados anteriormente. De un dataset acotado a 10000 mails (50 % ham - 50 % spam) obtuvimos las eficacias de la figura **??** para distintos kernels y valores del parámetro C.



(a)

(SI TERMINA;, VOY A PONER SVM CON C=1 Y KERNEL = POLY CON TODA LA DATA)

4. Reducción de dimensionalidad

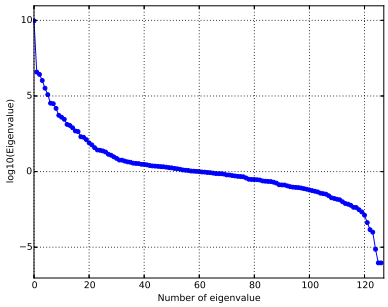
4.1. PCA

De los atributos seleccionados, realizamos un reducción de la dimensionalidad mediante la técnica de PCA. Dada la matriz de *mails* x *atributos*, calculamos la matriz de covarianza de los atributos, y realizamos una descomposición en valores singulares (SVD). La matriz de covarianza

Cuadro 1

Autovector	Componentes principales
1	Largo del documento.
2	Cantidad de espacios en blanco.
3	Términos: germ, hi, how, think, valuable, enron, republic, content-class, thread-index.
4	Términos: x-origin, x-filename, x-cc, binary
5	Términos: receive, email, upgrade, fast, spam

es una matriz simétrica, por lo tanto sus valores singulares coinciden con sus autovalores, que además son reales y no negativos. En la figura **??**, observamos el valor de los mismos, ordenados de mayor a menor.



(a)

Los autovectores obtenidos de la factorización son una combinación lineal de los atributos elegidos originalmente. Dichos autovectores pueden ser tomados como nuevos atributos, los cuales, a partir de sus autovalores asociados, sabemos en cuáles hay una mayor variabilidad de los datos. Para dar una interpretación a los nuevos atributos, observamos la representación de los autovectores en el espacio de atributos originales. Como criterio, estudiamos qué componentes tienen un valor absoluto mayor a 0,1 en el espacio de atributos originales. En la tabla 1 mostramos el resultado para las 5 direcciones principales. De la tabla vemos que los dos principales atributos coindicen con dos aitrutos originales, mientras los otros tres tienen un grupo de términos, de los cuales el 4 y 5 autovector muestran una correlación más visible entre los miembros del grupo.

5. Resultados

6. Discusión