# Trabajo Práctico Final: SVM: Máquina de Vectores Soporte

## Introducción al Aprendizaje Automatizado


Alumno: Navall, Nicolás Uriel. N-1159/2.

a) Para optimizar los valores de la SVM lineal lo que hice fue iterar sobre una lista de posibles valores de c que iban desde 10^(-5) hasta 10^5. Luego de entrenar y ver los resultados de errores llegue a que 100 era el mejor valor de c de entre los valores dados, y a partir de esto busque mejores valores de c cercanos al obtenido, por lo cual volví a entrenar la SVM con los valores 40, 70, 100, 400 y 700 y de esta manera llegue a qué 40 es el valor óptimo de c.

Realicé un proceso similar para la SVM gaussiana, solo que en lugar de optimizar solo el valor de c optimice la combinación de valores (c,gamma), iterando con c sobre valores que iban desde 10^(-5) hasta 10^5 y gamma sobre valores de 0.1 a 10, ya que gamma está asociado a la distancia entre los puntos de los datos, y como los datos con los que estoy trabajando tienen 102 componentes y cada componente tiene un valor entre 0 y 1 entonces la máxima distancia entre dos puntos es la distancia entre el punto con todas sus componentes en 0 y el punto con todas sus componentes en 1, la cual es dist(v0,v1)=sqrt((0-1)^2+(0-1)^2+...+(0-1)^2)=sqrt(1+1+..+1)=sqrt(102)≅10.1, por lo cual la distancia media entre los puntos se tiene que encontrar entre 0 y 10.1, cotas que use para iterar el valor de gamma. Además de utilizar los valores descritos para el gamma, la SVM también fue entrenada con valores de gamma llamados 'scale' y 'auto', que se corresponden a los valores de gamma 1/(n_features * X.var()) y 1/n_features. De esta manera obtuve un valor óptimo de c de 100 y de gamma de 0.5, por lo cual pasé a entrenar nuevamente SVM gaussianas con valores alrededor de los obtenidos. Y de esta manera llegué a que los valores óptimos de c y gamma para el método fueron 40 y 0.2 respectivamente.

| | SVM Lineal | SVM Gaussiana | Árboles | Naieve Bayes |
|---|---|---|---|---|
| Media error de test | 0.305 | 0.265 | 0.21 | 0.607499999 |
| Media error de entrenamiento | 0.101333333333 | 0.00906666666667 | 0.0223 | 0.561866668 |
| Desviación estándar | 0.0524404424085 | 0.0459468291736 | 0.0529674952736 | 0.0425734706778 |

Los bajos valores que obtuvimos en las desviaciones estándar con cada método (cerca del 5%) nos indican que estos no varían mucho sobre la media, y esto significa que la media de los errores en el conjunto de datos de test nos da una buena idea del comportamiento de cada método sobre la muestra de datos dada.

El hecho de que hayamos obtenido un error tan grande con naive bayes y tan “chico” con el resto sugiere que el problema no tiene asociado un concepto de probabilidad, sino que utiliza reglas más “duras” (como se modelan con los métodos de árboles o las SVM), por lo cual modelar un clasificador utilizando un método probabilístico generativo como naive-bayes no representa bien el problema.

Al mirar los árboles de decisión generados podemos observar como utiliza muy pocas componentes de las 102 disponibles en el set de datos, lo cual tiene sentido ya que difícilmente se necesiten todos los datos para determinar si un compuesto puede o no atravesar la barrera entre la sangre y el cerebro.

El hecho de que los árboles pueden categorizar cada componente con un valor de ganancia le permite al método discernir cuales son las componentes que son relevantes para la clasificación, a diferencia de los otros métodos. Razón por la cual el error al utilizar árboles es tan bajo (o al menos lo es en comparación con otros métodos que no identifican las componentes más importantes) en un problema con tantos componentes en los datos como este.

b) Al realizar el t-test entre el método de árboles y el método de naive bayes (el mejor y peor método) obtuvimos un valor de t de 4.61348902045, y cómo estamos realizando el test con 95% de confidencia entonces podemos descartar que los método de Árboles y Naive-Bayes son iguales (con hasta más de un 99% incluso).

En el t-test entre el método de árboles y la svm gaussiana (el mejor y segundo mejor método) obtuvimos un valor de t de 0.654782526859, con lo cual no llegamos al 95% necesario para descartar que los dos métodos son iguales, pero eso no implica que los dos métodos sean iguales, solo que no podemos descartar la hipótesis nula.