Facultad de Ciencias Exactas, Ingeniería y Agrimensura

Introducción al Aprendizaje Automatizado

TRABAJO FINAL

Alumno: Rodríguez Jeremías

20 de julio de 2017

1. Ejercicio 1

En este ejercicio trabajé con el datasets Heladas, utilizando varias de las herramientas que aprendimos en el curso: árboles de decisión (c4.5), aprendizaje Bayesiano (naive Bayes con normales) y support vector machines.

Se evaluaron los métodos con una estimación en 10-Folds, particionando el conjunto de datos en 10 subconjuntos con la misma proporción de puntos de cada clase. Luego usé alternativamente uno de esos conjuntos para test y el resto para training.

Respecto al método SVM, utilicé la implementación SVM^{light}[1] con dos kernels: linear y radial basis function. Elegí este último porque, en ausencia de conocimiento experto, ha demostrado ser un buen kernel default.

Para estos dos kernels tuve que ajustar los parámetros c y (c, γ) respectivamente:

- Con el kernel lineal optimicé el parámetro respecto al primer fold, variando c desde 2^{-10} hasta 2^{20} subiendo de a una unidad el exponente. El valor obtenido fue $c = 2^6$, y sucesivos intentos de mejorar la precisión (probando intervalos más ajustados) mantuvieron el error uniforme.
- Con el kernel RBF realicé un procedimiento similar variando c de 2^{-10} hasta 2^{20} subiendo de a una unidad el exponente, y γ de 2^{-10} hasta 2^{10} subiendo de a una unidad el exponente. Un par de valores óptimos (elegido al azar entre todos) resultó ser $\gamma = 1$ y $c = 2^5$, aunque hubo combinaciones produciendo el mismo valor. Intentos de afinar los valores resultaron en obtener siempre la el mismo error.

La siguiente tabla reúne los errores porcentuales de las 10 ejecuciones para cada método, en las últimas dos filas se pueden ver las medias y desviaciones estandar:

n°	C4.5	N. Bayes	SVM Lineal	SVM RBF
1	19.6	16.32	15.69	13.73
2	21.6	14.28	17.65	19.61
3	15.7	20.40	17.65	19.61
4	25.5	16.32	19.61	19.61
5	28.0	30.61	28.00	28.00
6	18.0	16.32	16.00	14.00
7	30.6	22.44	20.41	22.45
8	30.6	22.44	30.61	30.61
9	22.4	22.44	18.37	20.41
10	22.4	20.40	14.29	12.24
μ	23.440	20.197	19.828	20.023
σ	5.135	4.757	5.348	5.956

Como vemos, en promedio los dos métodos usando SVM fueron mejores que los otros métodos estudiados, aunque Naive Bayes obtuvo resultados muy similares. Respecto a SVM, parece ser que los datos son linealmente separables, o al menos el kernel RBF no mejoró el resultado mucho para los datos particulares que analizamos. Además, SVM es un método adecuado para datasets no muy extensos (como Heladas).

El hecho de que naive Bayes con normales funcione relativamente bien podría también sugerir que los datos siguen una distribución similar a normal en cada uno de sus features, aunque este método está asumiendo que los features son independientes cuando claramente tiene que haber relacion entre ellos.

[1] `svmlight.joachims.org`

c4.5 dio el peor resultado, pero junto con NB tienen la ventaja de que no hay que optimizar parámetros, ni elegir kernels, ni preocuparse por si los datos son o no linealmente separables, etc.

2. Ejercicio 2

Realicé un t-test con 95 % de confianza entre el método que dio mejor resultado (SVM lineal) y el que dio peor resultado (c4.5). El resultado es el intervalo $3{,}612 \pm 2{,}667$.

Por otro lado, realicé un t-test con 95 % de confianza entre el método que dio el mejor resultado (SVM lineal) y el segundo mejor (SVM RBF). El resultado es el intervalo $0{,}195 \pm 1{,}240$.

Por lo tanto, el primer t-test es positivo; dado que con un 95 % de probabilidad SVM lineal será mejor que c4.5 para este problema. El segundo t-test es negativo, pues no podemos asegurar con ese porcentaje de certeza que SVM lineal sea mejor que RBF (en el intervalo hay valores negativos!).

Como conclusión, si tuviera que elegir entre SVM lineal y C4.5 elegiría claramente SVM lineal para este problema; pero si tuviera que elegir entre SVM lineal y RBF necesitaría realizar otros análisis (o el mismo sobre más datos) para definir si uno es realmente mejor que el otro con tanta confianza.