Trabajo Práctico Final
Machine Learning

Bianchi, Gabina Luz

17 de julio de 2017

Ejercico A

La implementación utilizada del método *support vector machine* es *SVM Light* [1].
Primero se procedió a particionar el conjunto de datos *heladas* (el único conjunto de datos que se utiliza en este trabajo) en 10, manteniendo la proporción de puntos de cada clase[2], y armar los datos necesarios para poder aplicar validación cruzada. Luego se aplicaron los algoritmos pedidos, ajustando los parámetros correspondientes.

C4.5 y Naive Bayes

Los primeros algoritmos aplicados fueron C4.5 y Naive Bayes con normales, dado que no necesitaban ajustar parámetros. Los resultados se encuentran en la tabla de la Figura 1. En naranja se marca la partición que obtuvo resultados más similares a las medias. El valor que se considera para definir cuál algoritmo presenta, aprentemente, un resultado mejor, es la media. Por lo tanto, en este caso, se concluye que Naive Bayes predice mejor que C4.5.

[1] Se puede descargar de aquí: `http://svmlight.joachims.org/`

[2] En realidad, no fue posible mantener exacta la proporción de puntos de cada clase, dado que existen en el conjunto de datos original, 184 punto de clase 1 y 316 pertenecientes a la clase 0. Se optó por hacer algunos conjuntos con 19 puntos de clase 1 y 31 de clase 0, y otors con 18 puntos de clase 1 y 32 de clase 0.

Partición	C4.5	Naive Bayes
1	80	78
2	82	82
3	74	78
4	80	82
5	74	84
6	76	74
7	76	78
8	68	78
9	76	72
10	82	80
Media	76.8	78.6
DE	4.34	3.66

Figura 1: Porcentaje de aciertos en cada una de las particiones utilizando C4.5 y Naive Bayes con normales.

C	Acierto (%)
0.00001	64
0.0001	64
0.001	64
0.01	64
0.1	74
1	76
10	76
100	76
300	76
500	76
800	76
1000	76
2000	76
3000	76
10000	76
100000	66

Figura 2: Porcentaje de aciertos en la partición 7 con SVM lineal, variando el valor de C.

Support Vector Machine

Kernel lineal

Con la información obtenida a partir de los resultados anteriores se procedió a ajustar el parámetro C para SVM con kernel lineal, de modo de superar los niveles de aciertos. Se utilizó la partición 7, por ser ésta la que obtuvo resultados más parecidos a las medias obtenidas para los algoritmos mencionados anteriormente.
Para comenzar la búsqueda del valor del parámetro C, primero se realizaron las pruebas que se muestran en la tabla de la Figura 2.
Allí se hace un barrido para posibles valores de C, y se observa que utilizando valores que se encuentran entre 1 y 10000 se obtiene el mismo resultado (para una partición particular). Por lo tanto, para seguir con la búsqueda se optó

C	1	15	30	100	200	500	2000
	76	84	86	86	86	86	86
	78	80	82	80	80	80	80
	72	76	74	74	74	74	74
	82	82	82	82	80	80	78
	88	86	86	86	86	86	86
	80	80	80	80	80	80	80
	76	76	76	76	76	76	76
	84	84	84	84	84	84	84
	74	74	74	74	74	74	74
	78	80	80	80	80	80	80
Media	**78.8**	**80.2**	**80.4**	**80.2**	**80**	**80**	**79.8**
DE	**4.83**	**3.94**	**4.5**	**4.47**	**4.42**	**4.42**	**4.47**

Figura 3: Porcentaje de aciertos en cada partición con SVM lineal, variando el valor de C.

por elegir algunos posibles valores de C en el rango mencionado (evitando los valores muy altos), y aplicar el algoritmo a cada una de las particiones generadas, calculando la media y la desviación estándar. Los resultados se presentan en la tabla de la Figura 3. Allí se encontró que el mejor resultado se da para $C = 30$, con una media de 80.4 % de aciertos y una desviación estándar de 4.5. Se podría haber optado por elegir el resultado logrado con $C = 15$, ya que si bien presenta una media menor, la desviación estándar también lo es.

Kernel polinomeal

El kernel no lineal elegido es el polinomeal. Para utilizarlo, además de ajustar el parámetro C, se debe ajustar el grado D del polinomeo utilizado. Siendo que 2 es el valor más popular para el grado, primero se hizo un barrido para los posibles valores de C (análogo al hecho anteriormente), fijando $D = 2$. En la tabla de la Figura 4 se presentan los resultados obtenidos para la partición 7.
Allí se observa que los mejores resultados se obtienen para $C = 0.1$, $C = 1$, $C = 3000$ y $C = 10000$. Luego se procedió a fijar esos valores de C y variar el parámetro D en un rango de 2 a 300. Los resultados se presentan en la tabla de la Figura 5.
En dicha tabla se puede observar que los mejores resultados se obtienen para grados pequeños, entre 2 y 10. Posiblemente para polinomeos con grados muy altos se haga sobreajuste.
Finalmente, con los C en los cuales se obtuvo mejor resultado (0.1 y 1), se corrió el algoritmo para cada partición variando el D entre 2,3,4 y 5. En la tabla de la Figura 6 se presentan los resultados obtenidos. El mejor caso es el logrado con $C = 1$ y $D = 5$, con un 81 % de aciertos y una desviación

C	Aciertos (%)
0.00001	64
0.0001	64
0.001	64
0.01	64
0.1	76
1	76
10	72
100	72
300	72
500	72
800	72
1000	72
2000	72
3000	74
10000	74
100000	48

Figura 4: Porcentaje de aciertos en la partición 7 con SVM polinomeal con grado 2.

	C=0.1	C=1	C=3000	C=10000
D	Aciertos (%)	Aciertos (%)	Aciertos (%)	Aciertos (%)
2	76	76	74	74
3	76	74	74	70
5	76	78	54	54
10	78	72	58	58
15	60	58	58	58
30	66	66	66	66
60	48	48	48	48
100	64	64	64	64
300	64	64	64	64

Figura 5: Porcentaje de aciertos en la partición 7 con SVM polinomeal.

	C=0.1				C=1			
	D=2	D=3	D=4	D=5	D=2	D=3	D=4	D=5
	76	76	76	76	78	82	82	82
	78	78	78	78	76	84	84	78
	72	72	72	72	74	76	76	78
	82	82	82	82	82	82	84	86
	88	88	88	88	86	86	86	86
	80	80	80	80	78	78	78	80
	76	76	76	76	76	74	72	78
	84	84	84	84	84	84	82	82
	74	74	74	74	74	74	76	76
	78	78	78	78	78	78	80	84
Promedio	**78.8**	**78.8**	**78.8**	**78.8**	**78.6**	**79.8**	**80**	**81**
DE	**4.83**	**4.83**	**4.83**	**4.83**	**4.12**	**4.37**	**4.42**	**3.56**

Figura 6: Porcentaje de aciertos obtenidos con SVM polinomeal para todas las particiones.

	Kernel Polinomeal	Kernel Lineal	Naive Bayes	C4.5
Media	81	80.4	78.6	76.8
DE	3.56	4.5	3.66	4.34

Figura 7: Media y desviación estándar del porcentaje de aciertos logrados con cada algoritmo utilizado.

estándar de 3.56.

Resultados Finales

En la tabla de la Figura 7 se presentan las medias y desviaciones estándares de los mejores resultados de los 4 algoritmos utilizados.

Ejercicio B

En este ejercicio se pide realizar dos *t-test* con 95 % de confidencia entre algunos de los resultados obtenidos en la parte A.
En la tablas de la Figura 8 y la Figura 9 se encuentran los valores que se utilizaron para dichos tests. Los dos mejores resultados se obtuvieron utilizando SVM, con kernel polinomeal y con kernel lineal, respectivamente, mientras que el peor fue el logrado con C4.5.

A continuación se presentan algunas consideraciones respecto a los tests realizados.
El parámetro $\bar{\delta}$ representa la media de las diferencias de aciertos observadas entre ambos algoritmos (ver Figura 8 y Figura 9). Dado que el conjunto total de datos se particionó en 10, se tiene $k = 10$, siendo $\bar{\delta}_i$ la diferencia

Partición	Kernel Polinomeal	C4.5	Diferencia
1	82	80	2
2	78	82	-4
3	78	74	4
4	86	80	6
5	86	74	12
6	80	76	4
7	78	76	2
8	82	68	14
9	76	76	0
10	84	82	2

Figura 8: Valores utilizados para el *t-test* entre el mejor y el peor resultado.

Partición	Kernel Polinomeal	Kernel Lineal	Diferencia
1	82	86	-4
2	78	82	-4
3	78	74	4
4	86	82	4
5	86	86	0
6	80	80	0
7	78	76	2
8	82	84	-2
9	76	74	2
10	84	80	4

Figura 9: Valores utilizados para el *t-test* entre el mejor y el segundo mejor resultado.

entre los aciertos para la partición i, variando i=1,...,k. El parámatro $S_{\bar{\delta}}$, el cual representa una estimación de la desviación estándar, está definido como $\sqrt{\frac{\sum_{i=1}^{k}(\bar{\delta}_i-\bar{\delta})^2}{k\ (k-1)}}$. Siendo que el test realizado es de 95 % de confidencia con 9 grados de libertad, se tiene $t_{N,k-1} = 2{,}26$.

Al comparar SVM con kernel polinomeal y C4.5 se tienen los siguientes resultados.

$$\bar{\delta} = 4{,}2$$

$$S_{\bar{\delta}} \approx 1{,}69$$

$$\bar{\delta} \pm S_{\bar{\delta}}\ t_{N,k-1} \approx 4{,}2 \pm 1{,}69\ 2{,}26 \approx [0{,}36;\ 8{,}03]$$

Esto se interpreta de la siguiente manera: hay un 95 % de posibilidad de que la media real de las diferencias de desempeño entre los dos algoritmos estudiados en este caso pertenezca al rango [0.36,8.03]. Por lo tanto, considerando que todos los valores en dicho intervalo son mayores a 0, se puede decir que hay un 0.95 de probabilidad de que el algoritmo SVM con kernel polinomeal

(y los parámatros discutidos en el ejercicio A) tenga, en promedio, mayor cantidad de aciertos que el C4.5.

Por el contrario, al comparar ambos desempeños de SVM, se obtiene el resultado que se presenta a continuación.

$$\bar{\delta} = 0{,}6$$

$$S_{\bar{\delta}} \approx 0{,}99$$

$$\bar{\delta} \pm S_{\bar{\delta}}\ t_{N,k-1} \approx 0{,}6 \pm 0{,}99\ 2{,}26 \approx [-1{,}63;\ 2{,}83]$$

Esto significa que hay 0.95 de probabilidad de que la media real de la diferencia entre ambos algoritmos de SVM pertenezca al intervalo [-1.63, 2.83]. Por lo tanto, siendo que existen valores negativos, no se puede concluír que hay 0.95 de probabilidad de que el kernel polinomeal tenga, en promedio, mayor cantidad de aciertos que el kernel lineal.