Trabajo Práctico 2
Selección de variables

AGUSTÍN MISTA
Universidad Nacional de Rosario
Tópicos de Minería de Datos
Rosario, 6 de Noviembre de 2017

Apartado 2.

Aplique los 4 métodos (los 3 desarrollados más el forward) a los dos datasets de ejemplo (datosA y datosB).

Para este apartado se utilizaron dos datasets sintetizados a partir de variables independientes con ruido uniforme. La clase correspondiente a cada muestra esta definida por un criterio distinto en ambos casos. A continuación se presentan ambos datasets junto con los resultados obtenidos al seleccionar las variables más importantes usando los distintos métodos estudiados.

Dataset A

Este dataset posee 2000 muestras con 10 features cada una. Cada muestra es generada usando ruido uniforme en el conjunto $[-1, 1]$, las cuales son clasificadas usando el siguiente criterio:

- 50% de los datos $\mapsto$ signo de la variable 8.
- 20% de los datos $\mapsto$ signo de la variable 6.
- 10% de los datos $\mapsto$ signo de la variable 4.
- 5% de los datos $\mapsto$ signo de la variable 2.

Dado este criterio de clasificación, podemos observar que la feature más importante es la número 8, ya que logra clasificar al 50% de los datos por si misma, seguida de las features 6, 4 y 2 respectivamente. El resto de las features no aporta datos, por lo cual sería deseable identificarlas y eliminarlas de nuestro dataset.

A continuación se muestran los resultados obtenidos al seleccionar las features más importantes para este dataset usando los métodos vistos en el curso.

Forward Selection	RF	8	6	2	9	1	4	10	3	7	5
	LDA	8	9	1	5	10	7	3	2	4	6
	SVM	8	9	10	3	1	5	2	7	4	6
Backward Elimination	RF	8	9	10	7	3	5	2	4	1	6
	LDA	8	9	1	10	3	7	5	2	4	6
	SVM	8	1	5	2	4	10	3	9	7	6
Recursive Feature Elimination	RF	8	6	4	3	5	1	2	7	10	9
	SVM	8	6	4	2	7	9	10	5	3	1
Kruskal-Wallis		8	6	4	2	7	9	5	10	3	1

A partir de la tabla anterior podemos destacar algunos aspectos interesantes. En principio, todos los métodos logran identificar a la feature número 8 como la de mayor importancia.

Los métodos basados en Forward Selection y Backward Elimination obtuvieron los peores resultados, ya que buscan encontrar relación entre las distintas features, y en este caso todas son independientes. Junto a esto, estos métodos además presentan una performance temporal muy pobre.

En el otro extremo, el filtro basado en Kruskal-Wallis obtuvo el mejor resultado, ordenando correctamente las cuatro features con mayor aporte de información, debido a que éste analiza cada una de ellas de manera independiente y no considera el caso de que algún par de features se encuentre correlacionada. Todo esto junto a una performance muy buena.

Finalmente, los métodos basados en Recursive Feature Elimination también obtuvieron muy buenos resultados, ordenando de manera ideal las features de nuestro dataset para el caso de la estimación basada en Support Vector Machines—y de manera extremadamente similar al método de Kruskal-Wallis—, mientras que usando una estimación basada en Random Forest se obtiene un resultado casi ideal, con solo una feature mal ordenada.

Dataset B

Este dataset posee 2000 muestras con 8 features cada una y, al igual que el dataset anterior, cada muestra es generada usando ruido uniforme en el conjunto $[-1, 1]$. En este caso, la clasificación de cada muestra está dada por el *xor* del signo de las primeras dos features. Además se hace que las features número 3 y 4 tengan un 50% de correlación con la clase de cada muestra. De esta manera, las features 1 y 2 por separado no son suficientes para predecir la clase de cada muestra—se necesita considerarlas en conjunto. Por otro lado, las features 3 y 4 sólo son capaces de predecir el 50% de las muestras por lo que no resultan buenos predictores para este dataset.

A continuación se muestran los resultados obtenidos al seleccionar las features más importantes para este dataset usando los métodos vistos en el curso.

Forward Selection	RF	3	4	8	2	1	5	7	6
	LDA	4	7	6	2	1	5	8	3
	SVM	4	6	7	8	2	3	5	1
Backward Elimination	RF	1	2	6	8	5	7	4	3
	LDA	3	2	8	6	7	1	5	4
	SVM	4	6	2	7	8	1	3	5
Recursive Feature Elimination	RF	2	1	3	4	8	7	5	6
	SVM	3	4	6	1	5	2	7	8
Kruskal-Wallis		3	4	1	7	2	8	5	6

Si analizamos la tabla anterior, podemos observar como este dataset requiere del uso de métodos que consideren las posibles correlaciones entre distintas features.

El resultado obtenido usando el filtro basado en Kruskal-Wallis es incapaz de reconocer a las features 1 y 2 como importantes, ya que las analiza por separado, y de esta manera ninguna de ellas aporta información útil para clasificar nuestro dataset. Luego, el mismo selecciona como importantes las features 3 y 4, las cuales por separado son las únicas que poseen información útil.

Por otro lado, los únicos métodos capaces de rankear correctamente a las features 1 y 2 fueron el de Backward Elimination y el de Recursive Feature Elimination, ambos usando Random Forests.

Ĺos métodos basados en Forward Selection no obtuvieron resultados muy favorables, debido a que en cada paso seleccionan de manera greeedy la variable que aporta mayor ganancia de información. De esta manera, en los primeros pasos tienden a seleccionar a las variables que estan levemente correlacionadas con la clase de cada muestra.

Como conclusión de este apartado, podemos notar que el problema de asignar el grado de importancia a cada feature presente en un dataset no es sencillo, y no existe un método mágico que funcione lo suficientemente bien en todos los casos, por lo cual resultaría conveniente aplicar distintos métodos y/o conocer la naturaleza del dataset a la hora de seleccionar cuales features serán utilizadas y cuales descartadas.

Apartado 3.

Prepare un dataset del problema diagonal del práctico 1 con 10 variables, 50 puntos por clase y sigma igual a dos. Agregue 90 variables de ruido uniforme. Aplique los 4 métodos. Repita el experimento 30 veces. Calcule para cada método el porcentaje de aciertos en el ranking (cuántas de las 10 variables "originales" están en los primeros 10 lugares).

Para este apartado generamos datasets basados en el generador de datos *diagonal* con 100 muestras de 10 features cada una, a los cuales se le agregaron 90 features de ruido uniforme en cada caso.

Cada dataset fue evaluado 30 veces con cada uno de los métodos de ranking de features vistos anteriormente. Luego se calculó para cada método el porcentaje medio de aciertos, es decir, la cantidad de veces que cáda método logró seleccionar a las diez variables relevantes del dataset en los diez primeros lugares del ranking. Los resultados de este experimento se muestran a continuación.

Forward Selection	RF	42%
	LDA	52%
	SVM	43%
Backward Elimination	RF	39%
	LDA	39%
	SVM	48%
Recursive Feature Elimination	RF	91%
	SVM	67%
Kruskal-Wallis		99%

Como se puede observar en la tabla anterior, el mejor resultado fue el obtenido mediante el filtro basado en Kruskal-Wallis, ya que todas las features relevantes de este dataset son independientes entre sí.

Por otro lado, los wrappers basados en Forward Selection y Backward Elimination obtuvieron resultados relativamente malos, dado que los mismos intentan encontrar una correlación entre las distintas features, la cual no existe en nuestro dataset, por lo que son propensos a guiarse por la gran cantidad de ruido que introducimos en el mismo.

Finalmente, los resultados obtenidos para el caso de los métodos basados en Recursive Feature Elimination dependen en gran medida del modelo

subyacente utilizado. En el caso de Random Forests, el ranking obtenido es casi ideal, puesto que este modelo se ajusta muy bien a los problemas con features independientes. En cambio, aquel basado en Support Vector Machines con kernel lineal obtuvo un resultado bastante peor, dado que el kernel utilizado se ve altamente afectado por la gran cantidad de ruido que fue introducido al dataset.