Facultad de Ciencias Exactas, Ingeniería y Agrimensura

Tópicos de Minería de Datos

Trabajo Práctico 2 Selección de Variables

Alumno Rodríguez Jeremías

3 de octubre de 2017

1. Ejercicio 1

Las implementaciones se encuentran en codigo_practico_2.R, al comienzo del archivo.

2. Ejercicio 2

2.1. Aplicando métodos al dataset DatosA

Recordemos que el dataset DatosA tiene dimensiones n=1000 y p=10, donde en principio todo es ruido uniforme. Sobre ese dataset, se aplican las siguientes modificaciones

- Al 50 % de los datos al azar se les asigne el signo de la variable 8 como clase
- Al 20 % de los datos al azar se les asigne el signo de la variable 6 como clase
- Al 10 % de los datos al azar se les asigne el signo de la variable 4 como clase
- Al 5 % de los datos al azar se les asigne el signo de la variable 2 como clase

Observemos que los features son independientes entre si, y que (obviamente) el ranking de *importancia* de variables comienza con 8-6-4-2 seguido por las otras variables (cualquier orden entre las últimas será por chance). Las restantes variables tienen la misma importancia y no aportan información últil.

A continuación se encuentran los rankings arrojados por los métodos estudiados en este trabajo

```
Forward rf      8 1 3 7 5 2 9 6 4 10
Forward lda     8 7 1 10 3 5 9 2 4 6
Forward svm     8 10 3 9 7 1 5 2 6 4
Backward rf     8 4 5 7 1 6 9 3 10 2
Backward lda    8 3 9 4 6 10 7 1 5 2
Backward svm    8 9 3 7 2 1 10 6 5 4
Filtro Kruskal  8 6 4 9 2 3 5 1 7 10
RFE rf          8 6 4 5 3 2 10 1 9 7
RFE svm         8 6 4 9 3 2 5 1 7 10
```

- Todos lograron identificar que la variable 8 es la más importante.
- El filtro con Kruskal dio el mejor resultado, retornando casi el ranking ideal, excepto por un 9 mezclado (por ruido). Es el mejor para este problema porque los features son independientes, y al analizar las variables individualmente obtiene el mejor resultado. Otro pro a mencionar es que es muy rápido.
- Los métodos de RFE funcionan también muy bien para este problema, y también son rápidos. No aprenden ruido porque son aproximados.
- Los métodos backward/forward no son muy buenos para este dataset, no solo tardan mucho en dar una respuesta sino que realizan overfitting al intentar descubrir relaciones entre las variables, que sólo estan presentes en el dataset por chance.

2.2. Aplicando métodos al dataset DatosB

El dataset DatosB tiene las mismas dimensiones que DatosA y en principio también es ruido uniforme donde se aplican las siguientes modificaciones

- A todas las muestras se les asigna la clase correspondiente al XOR de los signos de las variables 1 y 2
- A la mitad de las muestras, se le cambia el signo (de ser necesario) a la variable 3 para que coincida con el de la clase.
- A la mitad de las muestras, se le cambia el signo (de ser necesario) a la variable 4 para que coincida con el de la clase.

Por lo tanto, a priori sabemos que

- Las únicas variables correlacionadas son 1 y 2
- Las únicas variables que aportan información útil para clasificar son 1,2,3 y 4
- La variable 1 (2) por si sola no aporta ninguna información útil
- Las variables 1 y 2 juntas son capaces de predecir la clase de todas las muestras por si solas
- La variable 3 (4) es capaz de predecir solo algunas de las clases de las muestras correctamente.

Los rankings arrojados por los métodos estudiados son

```
Forward rf      3 4 2 1 7 8 5 6
Forward lda     3 8 1 2 5 7 6 4
Forward svm     3 1 8 5 2 7 4 6
Backward rf     2 1 6 7 8 5 4 3
Backward lda    3 1 2 5 7 6 8 4
Backward svm    3 1 2 6 7 5 4 8
Filtro Kruskal  3 4 5 1 7 6 8 2
RFE rf          2 1 3 4 8 7 6 5
RFE svm         3 4 7 6 1 2 5 8
```

- Como era de esperar, kruskal sólo reconoce como importantes las variables 3 y 4 que, de forma independiente, proveen información. Cómo 1 y 2 precisan combinarse, kruskal es incapaz de colocarlas al principio del ranking.
- Por otro lado, métodos backward y RFErf logran efectivamente darse cuenta que la clasificación depende de 1 y 2 combinadas; y posicionan a ambas juntas en el comienzo del ranking.
- Los métodos backward funcionan bien pues detectan la conveniencia de mantener 1 y 2 en el conjunto. En cambio los métodos forward comienzan agregando al feature 3 porque ven una primera ganancia (individual) allí. Recién cuando agregan a la variable 1 (2), los métodos forward lda y rf se dan cuenta de que conviene luego incorporar a 2 (1). Usando svm, al ser un kernel no conveniente, no se logra detectar esta relación.

3. Ejercicio 3

Consideremos datasets formados por n=100 puntos con p=100 features. Las primeras 10 features y la clase responden al dataset diagonal del trabajo práctico 1; las otras 90 variables son ruido uniforme.

Corrí los 9 métodos en 30 de estos datasets; y calculé la media de la cantidad de veces que las 10 variables originales aparecen en el top-10 del ranking.

Los resultados fueron los siguientes

```
Porcentaje de aciertos
 FORWARD RF     0.39
 FORWARD LDA    0.55
 FORWARD SVM    0.47
 BACKWARD RF    0.4
 BACKWARD LDA   0.38
 BACKWARD SVM   0.45
 FILTER KRK     0.98
 RFE RF         0.91
 RFE SVM        0.65
```

- Era de esperarse que el filtro de el mejor resultado, pues las 10 features significativas son independientes entre sí. No llega a 100 % por ruido ocasional.
- Los wrappers dan malos resultados, pues se guian por el ruido e intentan encontrar relaciones entre variables, sobreajustando. Además tardan mucho.
- Los métodos RFE dan un mejor resultado pues no sobreajustan tanto. En particular, RFE con random forest da un resultado muy bueno; pues los árboles de decisión son buenos con features independientes.

4. Opcional

Elegí el dataset **Student Performance Data Set**[1]. El objetivo es predecir la performance en la educación secundaria de estudiantes. En concreto, me limité a analizar sólo la asignatura matemáticas. Este dataset tiene 649 instancias y 33 features.

4.1. Información sobre el Dataset

Se recolectó información sobre las calificaciones académicas de dos escuelas secundarias portuguesas. Cada instancia corresponde a un alumno y los atributos incluyen las calificaciones obtenidas (en este trabajo me limito sólo a la calificación en la asignatura matemática), información demográfica, social y escolar relacionada al alumno. Fue recolectada usando reportes escolares y cuestionarios.

[1] http://archive.ics.uci.edu/ml/datasets/student+performance

4.2. Atributos

En principio, el dataset consta de 33 features:

```
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade),
        2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade),
        2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services',
         'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services'
        , 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation',
        'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1: <15m, 2: <30m, 3 - <1h, or 4>1 h)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours,
           or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)
```

4.3. Preprocesamiento del dataset

Usando las técnicas que aprendimos las primeras clases, tomé las siguientes decisiones:

- Eliminé las variables categóricas Mjob,Fjob y reason pues tienen muchos valores posibles (categorías) y para convertirlas a variables numéricas debería introducir demasiadas features nuevas.

- Convertí todas las variables binarias a variables numéricas 1-0.
- Convertí la variable nominal guardian, que asumía los valores 'mother', 'father' o 'other' en tres variables numéricas 0-1 llamadas guardian_father, guardian_mother, guardian_other.
- Verifiqué que (casi seguro) no haya valores faltantes (NA/0/<empty space>)
- Eliminé las variables G1 y G2 pues son los resultados del primer y segundo semestre; y por lo tanto serían falsos predictores.
- A efectos de convertir este problema en un problema de clasificación sencillo, transformé la variable target G3 (calificación final numerica entre 0 y 20) en una nueva variable 0-1 (llamada smart) tal que un estudiante está etiquetado como inteligente si su nota final es mayor a 15.
- Finalmente (no estoy muy seguro de que se deba hacer) normalicé todos los datos para que estén en el intervalo [0-1].

De este modo, el dataset final consta de todas variables numéricas, y deseamos predecir si el estudiante está entre los más destacados, o no.

4.4. Selección de variables

El objetivo es ver cuales de las aproximadamente 30 variables consideradas son las más importantes a la hora de predecir si un estudiante será destacado en matemáticas o no. En principio, pienso que claramente muchas de estas variables estan correlacionadas; a diferencia de varios de los ejemplos que vimos anteriormente en este trabajo.

El resultado fue el siguiente:

```
Forward rf       1 2 4 5 6 21 24 9 12 11 7 17 23 10 16 14 13 18 26 15 25 3 19 22 20 27 28 29 8
Forward lda      1 2 3 4 5 6 7 8 9 10 11 12 13 17 18 20 15 16 21 22 26 19 27 28 29 14 23 24 25
Forward svm      27 4 13 14 17 18 8 2 6 15 21 22 24 20 11 3 25 16 10 23 1 19 5 12 7 28 9 29 26
Backward rf      29 22 21 7 14 18 25 27 9 17 15 12 13 11 6 5 16 2 28 8 26 3 4 1 20 19 24 23 10
Backward lda     23 22 21 19 18 14 12 7 5 2 17 25 10 16 8 15 29 9 28 27 26 24 13 11 4 3 6 1 20
Backward svm     23 21 8 16 28 13 9 25 27 14 22 20 12 10 6 29 2 26 18 17 1 3 7 19 5 11 15 24 4
Filtro Kruskal   7 11 24 12 8 14 16 2 18 3 19 17 9 23 20 25 22 10 6 26 5 1 4 29 21 27 15 13 28
RFE rf           7 8 14 6 11 12 13 24 2 26 21 15 10 20 23 18 9 17 22 3 28 16 29 5 1 25 4 19 27
RFE svm          6 11 29 18 24 22 16 19 25 1 21 9 4 5 20 26 10 13 23 27 3 28 14 7 15 17 2 12 8
```

Me resulta difícil interpretar estos resultados, en primer lugar porque es probable que al no entender mucho del tema haya cometido algún error en el pre-procesamiento y estos números no tengan sentido.

Suponiendo que esos resultados están bien, tomaría más en cuenta los resultados de RFE y de los wrappers backward porque considero que en este problema los datos deben estar correlacionados. Por ejemplo, podría decirse que las variables 21-22-23 son importantes pues aparecen en los top-5 de los tres métodos backward.

Por otro lado, de los métodos de kruskal y forward deduzco que de forma individual las variables que más aportan son 1 2 y 4; o bien la 7ma.