Tarea Breve: Modelos predictivos, obtener la solución de un examen tipo test

Elias Barba Moral
Minería de Datos

April 17, 2017

Modelos predictivos Un fichero recoge las contestaciones de un examen tipo test de 25 preguntas y 5 alternativas por pregunta(a,b,c,d y e). La última columna recoge la puntuación total de respuestas acertadas, sin restar las preguntas erróneas o no contestadas. El numero de alumnos ha sido 23. Y queremos intentar usar models predictivos para descubrir el patrón correcto del examen. Todas las preguntas han sido correctamente respondidas por al menos un alumno. Hay una sóla pregunta que ha sido contestada correctamente por un único alumno y el resto de preguntas han sido acertadas por varios alumnos. Ninguno ha contestado a todas las preguntas correctamente, y algunos alumnos han dejado respuestas en blanco.

1 Algoritmo genético

En primer lugar intenté un algoritmo genético que obtuviera la solución, y a partir de ella tener una guía al aplicar el modelo predictivo. Se podría pensar que un posible patrón inicial sobre el que trabajar podría ser el de la respuesta mas contestada. Esto se probará cuando empecemos a trabajar el modelo predictivo, pero otro posible acercamiento inicial es el tomar la respuesta mas contestada, pero asignando un peso a cada individuo en función de su puntuación final. Así las respuestas de alumnos que hayan obtenido una puntuación mayor será mas relevante que las respuestas de aquellos con una puntuación menor.
Para evaluar los diferentes patrones que vayamos obteniendo usaremos dos medidores: la desviación de la puntuación de cada individuo si es corregido por el patrón que estamos probando con respecto a la puntuación final que obtuvieron, y la media de la desviación de la puntuación asignada por el patrón que estamos probando con respecto a la puntuación real, que cuando sea 0 significará que hemos encontrado la solución.
Para el patrón que acabamos de explicar los resultados son los siguientes:

```
> sum(abs(Q1251617[,26]-calification))/23
[1] 2.304348
```

```
Q1251617[,26]-calification
[1]  0 -4 -3 -4 -3 -1 -2 -1  2 -1  1 -3 -1 -3 -2 -1 -2  1 -5 -2 -2 -4 -5
```

Vemos que el patrón sobreestima la mayoría de las puntuaciones, lo cual es esperable dado que la respuesta de la mayoría otorgará mas puntos que el patrón (sabiendo además que una respuesta solo ha sido contestada correctamente por un alumno).
A continuación aplicaremos un algoritmo genético para encontrar el patrón correcto. Para ello crearemos una población inicial, tomando solo las opciones que aparecen contestadas por al menos un alumno, creando 200 individuos que iremos mejorando hasta encontrar la solución. Estos individuos son creados tomando un patrón inicial, en este caso el patrón anteriormente obtenido. Es posible acelerar el algoritmo tomando como patrón inicial una solución previamente obtenida que sabemos se acerca a la solución que queremos.
Para mejorar a la población usaremos dos técnicas; el cruce de dos padres para producir dos

hijos, y la mutación de un individuo.
Para el cruce de padres usaremos una selección de padres por torneo. De cinco posibles padres seleccionados aleatoriamente de la población, seleccionaremos los dos con menor desviación media de la puntución de los individuos. Estos dos padres se cruzaran en una posición aleatoria, tomando como hijos la adición de la primera parte del primer padre hasta dicha posición aleatoria y el resto del segundo padre, y viceversa.
Para la mutación tomaremos una posición aleatoria dentro del patrón de solución y la cambiaremos por cualquier otra que haya sido elegida al menos por un estudiante. Incluye además una condición para aumentar la cantidad de individuos mutados en caso que todas las soluciones empiecen a parecerse mucho.
Finalmente añadimos un filtro para ir seleccionando las mejores soluciones. Descartamos las 4 peores soluciones e incluimos los nuevos patrones creados en el cruce y la mutación.
El algoritmo tiene definidas 100000 iteraciones o encontrar una solución como mecanismo de parada. También hay que mencionar que en caso de estancamiento el algoritmo es capaz de crear una nueva población y seguir buscando soluciones.
La evolución del algoritmo se puede ver en el siguiente dibujo:

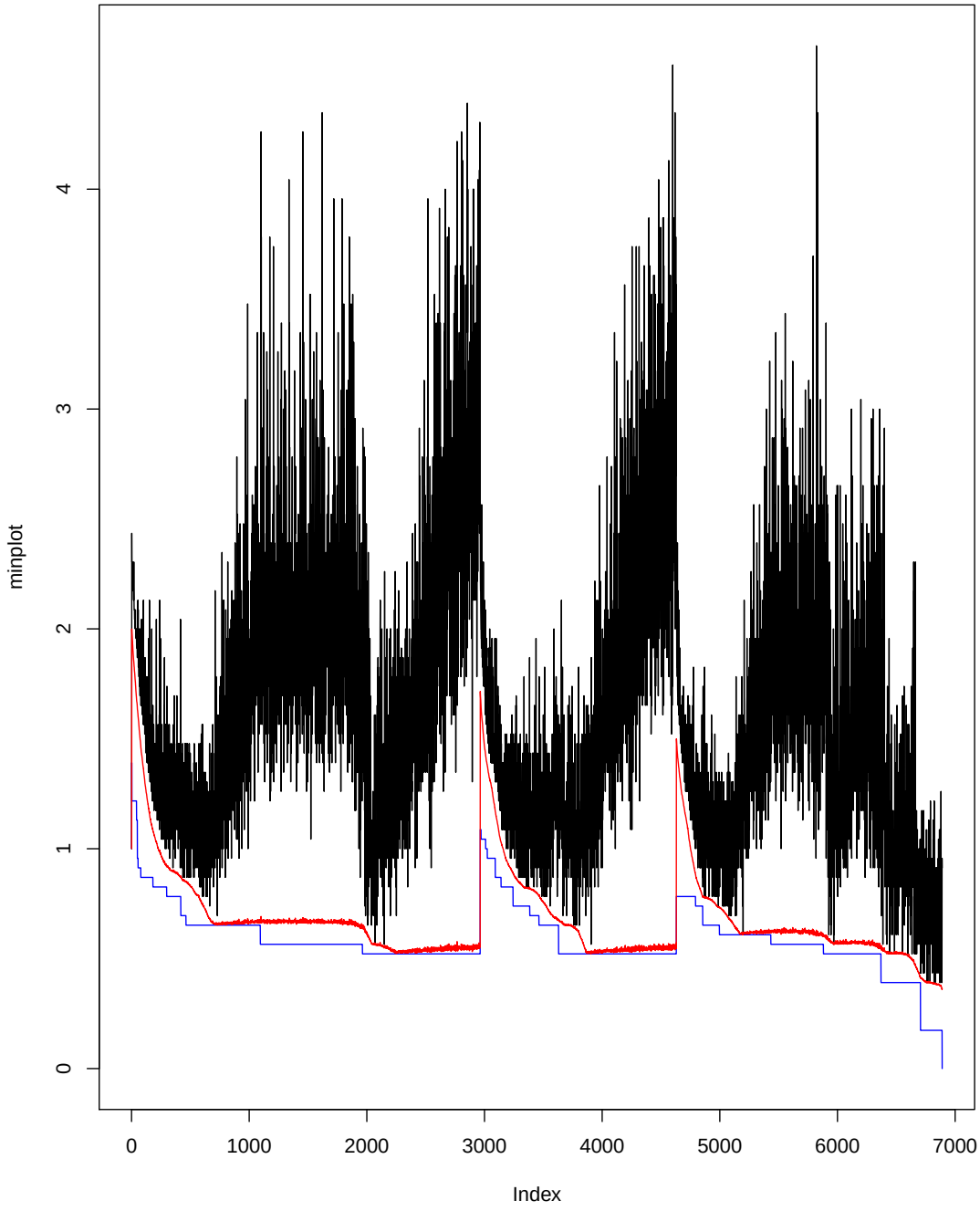

Figure 1: La linea negra es el individuo de la población con la máxima desviación, el rojo es la desviación media de la población y el azul es la desviación mínima.

Con resultado final:

```
6889 : media= 0.3613043  min= 0  max= 0.9130435  contador para nueva poblacion= 183
solucion:  C B C D B D E C D E A A B A B A D B C B C E B C B[1] 6889
```

Los picos observados en las lineas azules y rojas son los momentos en los que, devido a estancamiento del mínimo, una nueva población es creada basada en el mínimo anterior. El algoritmo como se puede ver, si se está estancando tiende a buscar en zonas mas diversas cada vez, como se puede apreciar en los aumentos en las lineas negras.

2 Modelo Predictivo

Usando lo aprendido en clase, podemos intentar aplicar un modelo predictivo para obtener un resultado, en un tiempo computacional mucho mas corto. Para ello aplicaremos un modelo de tipo lasso.
Primero crearemos una matriz binaria con los datos, tomando el valor uno si una opción de respuesta ha sido elegida por un alumno en cierta pregunta o cero en caso contrario. El siguiente paso es crear el modelo, y para ello nos ayudaremos de las librerias caret y dummy. Una de las opciones es el comando train, que ajusta una serie de modelos, devolviendo que tan complejo es el mejor modelo. Esto se usa posteriormente cuando usando el comando lars (least angle regression), obtenemos una cantidad de modelos, y ayudandonos de lo obtenido en el comando train, podemos seleccionar el mejor modelo.
Con este método obtenemos un patrón de respuesta para 19 preguntas (2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 16, 17, 18, 21, 22, 24 y 25) en el último modelo. Además, haciendo caso a lo obtenido por el comando train, podemos añadir la respuesta a una pregunta (la 19) que no aparece en el último modelo. Esto no es una respuesta completa a nuestro problema, pero es una primera aproximación.
Se puede notar que inicialmente hemos eliminado la contribución del individuo en la fila 3 dado que tiene la respuesta 16 sin contestar. Podemos intentar aplicar también el modelo añadiendo este individuo, asumiendo que contesta la opción mas popular y añadiendo un punto en la puntuación final. Aplicado esto obtenemos una respuestas para otras preguntas: la 23 en el mejor modelo según el comando train, y la 20 en el modelo mas complejo. Hay que mencionar que la pregunta 17 obtiene diferentes resultados, pero trabajaremos con la opción D que aparece en este último patrón.
Con esto somos capaces de montar un patrón de respuestas para 22 preguntas, quedándonos las preguntas 1,10 y 15 sin poder decir nada. Analizando las respuestas a las preguntas elegimos las respuestas mas comunes para las tres preguntas. En el caso de la pregunta 1, la mayoría se decanta por la C; en la pregunta 10, las opciones pueden ser la B o la E; y finalmente en la pregunta 15 las opciones pueden ser la B y la C. Para decidir que combinación usar, evaluaremos las desviaciones y nos quedaremos con la que tenga menor desviación media, en este caso siendo la E para las pregunta 10 y B para la pregunta 15.
Con todo es evidente que el patrón obtenido no es perfecto. Podemos ver que como mínimo existen dos preguntas en el patrón que son erroneas. Sabemos además que esas preguntas están dando puntos de mas, y por tanto la verdadera respuesta tiene que dar una respuesta diferente a la que está en el patrón que actualmente tenemos:

```
> Q1251617[,26]-calification
 [1]  0  0  0 -1  0 -1 -1  1 -1  0  0 -1 -2  0 -1  0 -1  0 -2  0 -2  0 -1
>
> Q1251617[13,]
   q1 q2 q3 q4 q5 q6 q7 q8 q9 q10 q11 q12 q13 q14 q15 q16 q17 q18 q19 q20 q21 q22 q23 q24 q25 total
13  C  B  A  C  C  D  E  E  C   B   D   A   B   B   D   B   A   E   C   B   A   B   B   C   C    10
> Q1251617[19,]
   q1 q2 q3 q4 q5 q6 q7 q8 q9 q10 q11 q12 q13 q14 q15 q16 q17 q18 q19 q20 q21 q22 q23 q24 q25 total
19  A  B  B  E  C  A  E  E  D   A   D   A   B   B   C   B   D   C   C   E   D   C   E   E   D     7
> Q1251617[21,]
   q1 q2 q3 q4 q5 q6 q7 q8 q9 q10 q11 q12 q13 q14 q15 q16 q17 q18 q19 q20 q21 q22 q23 q24 q25 total
21  E  B  C  D  C  A  E  C  B   E   A   A   B   B   B   A   D   B   C   E   D   E   E   C   A    16
```

Podemos ver que las preguntas 2,5,7,12,13,14 y 19 tienen el mismo patrón de respuesta en los tres alumnos y por tanto son las susceptibles de ser cambiadas. El cambio de la pregunta 2 no mejora el patrón, ergo es descartada. Pero el cambio en la pregunta 5 si mejora, luego nos quedamos con ella. Vemos además que hemos reducido de 2 a 1 el numero mínimo de preguntas mal corregidas. Tomando todas las preguntas en las que se ve una mala asignación deducimos que la pregunta restante que está mal en nuestro patrón es la 14, que además cumple con el requisito que solo ha sido contestada correctamente (opción A) por un alumno.

3 Conclusiones

A modo de conclusiones podemos decir que, mientras que el algoritmo genético ofrece una solución segura con suficiente tiempo, el modelo predictivo puede conseguir buenos resultados en un tiempo mucho menor. El modelo predictivo sin embargo, requiere generalmente un trabajo posterior, intentando completar y corregir la solución, y por tanto requiere mas intervención para obtener el resultado final. Es posible buscar otros modelos predictivos que devuelvan una solución mas completa que el implementado en este trabajo.