Joan Puigdomenech
Joel Romia

Algorithms for linear classification

Utilizando la herramienta R y una de sus librerías MASS, podemos acceder a una serie de métodos, algunos de las cuales son de análisis de datos utilizados en la estadística y el aprendizaje automático:

- **LDA (Análisis Discriminante Lineal):** Es un método estadístico que se utiliza para clasificar observaciones en diferentes grupos o categorías en función de sus características o variables predictoras. LDA busca encontrar una combinación lineal de las variables predictoras que maximice la separación entre las clases y minimice la variabilidad dentro de cada clase.

- **QDA (Análisis Discriminante Cuadrático):** Es similar a LDA, pero difiere en que permite que las matrices de covarianza sean diferentes para cada grupo. QDA es más flexible que LDA en términos de la forma de las distribuciones de los datos.

- **RDA (Análisis Discriminante Regularizado):** Es una extensión de LDA y QDA que incorpora regularización en la estimación de los coeficientes discriminantes. RDA busca encontrar una combinación lineal de las variables predictoras que maximice la separación entre las clases, pero con la restricción de que la magnitud de los coeficientes esté acotada.

- **Regresión Logística:** Es un método de aprendizaje supervisado utilizado para la clasificación de observaciones en dos o más categorías. La regresión logística utiliza una función logística para modelar la probabilidad de que una observación pertenezca a una categoría específica en función de sus variables predictoras. Se estima un modelo de regresión logística utilizando técnicas de optimización para encontrar los coeficientes que mejor se ajustan a los datos de entrenamiento. Una vez que se estima el modelo, se puede utilizar para predecir la probabilidad de pertenencia a una categoría específica para nuevas observaciones y realizar clasificación.

En esta práctica se nos pide implementar algunos de estos métodos sobre un data set “Musk2.data”, este conjunto de datos describe un conjunto de 102 moléculas, de las cuales 39 son consideradas por expertos como musks y las 63 restantes son consideradas no musks.

El objetivo es aprender a predecir si nuevas moléculas serán musks o no musks. Sin embargo, las 166 características que describen estas moléculas dependen de la forma exacta o conformación de la molécula. Debido a que los enlaces pueden rotar, una sola molécula puede adoptar muchas formas diferentes. Para generar este conjunto de datos, se generaron todas las conformaciones de baja energía de las moléculas para producir 6,598 conformaciones. Luego, se extrajo un vector de características que describe cada conformación.

Esta relación muchos a uno entre vectores de características y moléculas se llama "problema de múltiples instancias". Al aprender un clasificador para estos datos, el clasificador debe clasificar una molécula como "musk" si CUALQUIERA de sus conformaciones se clasifica como musk. Una molécula debe ser clasificada como "no musk" si NINGUNA de sus conformaciones se clasifica como musk.

El conjunto de datos consta de 6,598 instancias y 168 atributos, incluida la clase. Los atributos que incluyen son:

- **molecule_name**: Nombre simbólico de cada molécula. Los almizcles tienen nombres como MUSK-188. Los no almizcles tienen nombres como NO-MUSK-jp13.

- **conformation_name**: Nombre simbólico de cada conformación. Estos tienen el formato MOL_ISO+CONF, donde MOL es el número de moléculas, ISO es el estereoisómero número (generalmente 1), y CONF es el número de conformación.

- **f1 a f162:** Estas son "características de distancia" a lo largo de los rayos, las distancias son medida en centésimas de Angstroms. Las distancias pueden ser negativas o positivas, ya que en realidad se miden en relación con un origen colocado a lo largo de cada rayo, el origen fue definido por una superficie "musk de consenso".

- **f163**: Esta es la distancia del átomo de oxígeno en el molécula a un punto designado en el espacio tridimensional. Esto también se llama OXY-DIS.

- **f164**: OXY-X: X-desplazamiento del punto designado.

- **f165**: OXY-Y: Y-desplazamiento del punto designado.

- **f166**: OXY-Z: Z-desplazamiento del punto designado.

- **clase**: 0 => non-musk, 1 => musk

Primero, se carga el conjunto de datos musk2 en un objeto de marco de datos de R. A continuación, se renombran las columnas para que sean más fáciles de entender y se convierte la variable de clase en un factor. Luego, se realiza un resumen estadístico básico de los datos.

A continuación, se dividen los datos en conjuntos de entrenamiento y prueba mediante muestreo aleatorio sin reemplazo. Dos tercios de los datos se utilizan para entrenar los modelos y un tercio se utiliza para probarlos.

El primer modelo es un modelo de análisis discriminante lineal (LDA). El LDA se utiliza para encontrar una combinación lineal de características que permita separar las dos clases. Se ajusta un modelo LDA a los datos de entrenamiento y se utiliza para predecir la clase de los datos de prueba. La función predict se utiliza para generar las predicciones de clase en los datos de prueba.

```
      Pred
Truth    0    1
    0 1808   33
    1   91  246
```

La tabla que se presenta muestra la matriz de confusión para un modelo de clasificación binario que ha sido aplicado sobre un conjunto de datos de prueba. Esta matriz se utiliza para evaluar la calidad del modelo y su capacidad para predecir correctamente las etiquetas de clase de los datos de prueba.

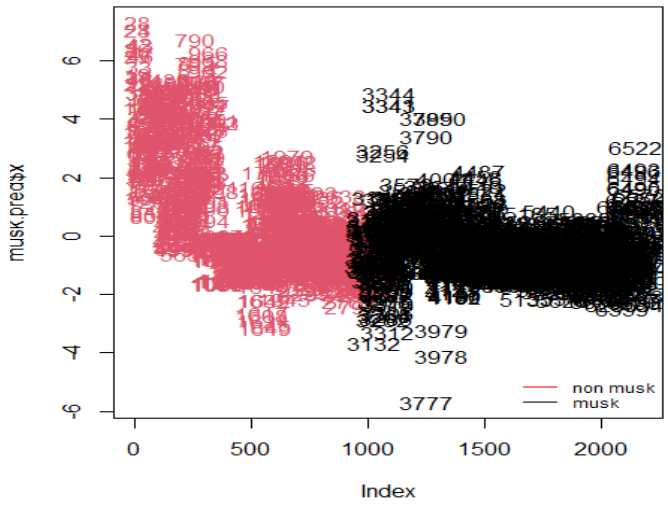
En este caso, la matriz de confusión tiene dos etiquetas de clase, 0 y 1. La etiqueta "Truth" representa la verdadera clase de los datos de prueba, mientras que la etiqueta "Pred" representa las predicciones realizadas por el modelo.

Los números que se presentan en la tabla indican la cantidad de datos que han sido clasificados en cada una de las cuatro posibles combinaciones de predicción y verdadera clase. Específicamente:

- 1808 datos han sido clasificados correctamente como clase 0 (verdadero negativo)
- 33 datos han sido clasificados incorrectamente como clase 1 cuando su verdadera clase es 0 (falso positivo)
- 91 datos han sido clasificados incorrectamente como clase 0 cuando su verdadera clase es 1 (falso negativo)
- 246 datos han sido clasificados correctamente como clase 1 (verdadero positivo)

Esta tabla permite calcular diversas métricas de evaluación del modelo, tales como la precisión, la sensibilidad o la especificidad. Por ejemplo, la precisión se define como la proporción de predicciones positivas que son verdaderamente positivas, es decir, 246/(246+33)=0.881.

Luego se representa gráficamente la salida del modelo.



El segundo modelo es un modelo de regresión logística. La regresión logística se utiliza para modelar la probabilidad de que una observación pertenezca a una clase determinada. Se ajusta un modelo de regresión logística a los datos de entrenamiento y se utiliza para predecir la clase de los datos de prueba. Se utiliza la función step para ajustar el modelo logístico y simplificarlo basándose en el criterio de información de Akaike (AIC). A continuación, se define una función de precisión para evaluar la precisión del modelo. La precisión del modelo se evalúa en los datos de entrenamiento y de prueba, y se informa en términos de porcentaje de aciertos y errores.

La regresión logística se utiliza comúnmente en problemas de clasificación binaria, donde se

busca predecir la pertenencia a una de dos clases posibles en función de un conjunto de variables predictoras. A diferencia de los métodos de LDA (análisis discriminante lineal), QDA (análisis discriminante cuadrático) o RDA (análisis discriminante regularizado), la regresión logística no hace suposiciones sobre la distribución de las variables predictoras, lo que la hace más flexible y adecuada para casos en los que las suposiciones de normalidad o igualdad de covarianzas pueden no ser apropiadas.

Además, la regresión logística es especialmente útil cuando la relación entre las variables predictoras y la variable de respuesta no es lineal, ya que la regresión logística modela la probabilidad de pertenencia a una clase a través de una función logística, que puede ser curvilínea. En contraste, los métodos de LDA, QDA y RDA suponen que las diferencias entre las clases se explican mediante combinaciones lineales de las variables predictoras, lo que puede limitar su capacidad para modelar relaciones no lineales.

En resumen, la regresión logística es una herramienta útil y flexible para la clasificación binaria que se adapta bien a una amplia variedad de situaciones, especialmente cuando las suposiciones de normalidad o igualdad de covarianzas no son apropiadas o cuando se espera que la relación entre las variables predictoras y la variable de respuesta sea no lineal.