TEMA 3

MÓDULO:
TÉCNICAS AVANZADAS DE DATA MINING

ANÁLISIS DISCRIMINANTE (LDA) SELECCIÓN DE VARIABLES

FERRAN ARROYO

Licenciado en Empresariales y en Ciencias Actuariales y Financiaras por la UB. Máster Executive en Data Science por la MBIT School. Data Scientist.

Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

ÍNDICE

OBJETIVOS ESPECÍFICOS

- Utilizar el LDA como clasificador.
- Utilizar el LDA como reductor de dimensionalidad.
- Entender el proceso de selección de variables mediante la regresión stepwise.

1. INTRODUCCIÓN AL ANÁLISIS DISCRIMINANTE

El Análisis Discriminante Lineal (ADL o LDA por sus siglas en inglés) es una generalización del discriminante lineal de Fisher, un método utilizado en estadística cuyo objetivo es encontrar una combinación lineal de rasgos que caracterizan o separan dos o más clases de objetos o eventos. La combinación resultante puede ser utilizada como un clasificador lineal o para la reducción de dimensionalidad.

Así pues, un mismo problema, como es el caso de la reducción de dimensionalidad, puede ser enfocado con un algoritmo de aprendizaje no supervisado, como sería el caso del PCA, cuyo objetivo se centra en maximizar la varianza en un conjunto de datos, pero también puede ser enfocado con un algoritmo de aprendizaje supervisado, como sería el caso del LDA, cuyo objetivo es maximizar la separabilidad entre clases.

El LDA está estrechamente relacionado con el análisis de varianza (ANOVA) y el Análisis de Regresión, el cual, también, intenta expresar una variable dependiente como la combinación lineal de otras características o medidas. Sin embargo, el análisis ANOVA usa variables independientes categóricas y una variable dependiente continua, mientras que el Análisis Discriminante tiene variables independientes continuas y una variable dependiente categórica (o sea, la etiqueta de clase).

El LDA está también, estrechamente, relacionado con el análisis de Componentes Principales (PCA) en que ambos buscan combinaciones lineales de variables que sean capaces de explicar mejor los datos.

El LDA, explícitamente, intenta modelar la diferencia entre las clases de datos, mientras que el PCA no toma en cuenta cualquier diferencia entre las clases.

Es importante comentar que el Análisis Discriminante requiere que las variables independientes sean contínuas.

Existe una técnica llamada Análisis Discriminante de Correspondencia que es equivalente al Análisis Discriminante Lineal y que admite variables categóricas.

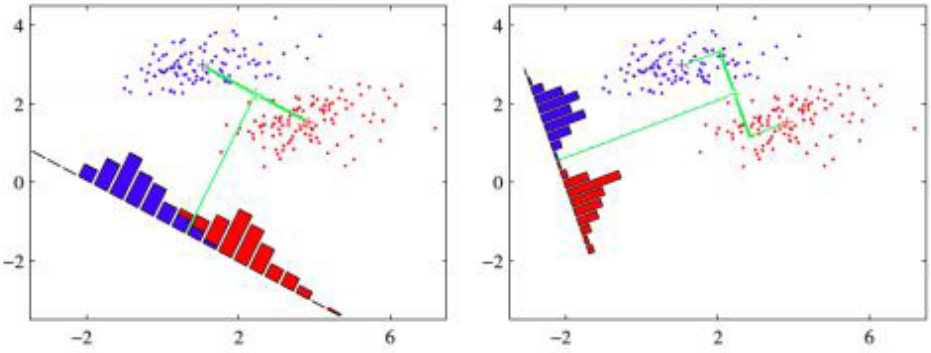
La idea detrás del LDA es simple: matemáticamente hablando, necesitamos encontrar un nuevo espacio para proyectar los datos para maximizar la separabilidad de las clases.

El primer paso consiste en encontrar una forma de medir la separación de cada nuevo espacio de variables. La distancia entre las medias proyectadas de cada clase podría ser una de las medidas, sin embargo, esta no sería una métrica demasiado útil, ya que no tiene en cuenta aspectos importantes como la dispersión.

En 1988, un estadístico llamado Ronald Fisher propuso la siguiente solución: **maximizar la distancia entre la media de cada clase y minimizar la dispersión dentro de la propia clase.** Por lo tanto, llegamos a **dos medidas: la que hace referencia a dentro de la clase y la que hace referencia a las clases.** Sin embargo, esta formulación sólo es posible si asumimos que el conjunto de datos tiene una distribución Normal.

Esta suposición es muy importante y, si la distribución es significativamente no gaussiana, el LDA podría no funcionar demasiado bien.

Como una imagen vale más que mil palabras, a continuación, puedes ver, gráficamente, el input y el output del proceso:



Fuente: https://mc.ai/fischers-linear-discriminant-analysis-in-python-from-scratch/

SABÍAS QUE...

Existe una técnica llamada Análisis Discriminante de Correspondencia que es equivalente al Análisis Discriminante Lineal y que admite variables categóricas.

Los siguientes papers adjuntos pueden serte de utilidad para profundizar más en la técnica del Análisis Discriminante:

- **Análisis_Discriminante_de_Correspondencia.pdf:** donde aprenderás más acerca del Análisis Discriminante de Correspondencia.
- **Análisis_Discriminante_Clasificador.pdf:** Teoría del LDA como clasificador.
- **Análisis_Discriminante_Reductor.pdf:** Teoría del LDA como reductor de dimensionalidad.

2. INTRODUCCIÓN A LA SELECCIÓN DE VARIABLES

La selección de variables es un proceso importante a la hora de realizar cualquier modelo, ya que mejora su interpretabilidad y la complejidad computacional (aunque esto último, actualmente, ya no es tan relevante, ciertamente).

Muchas veces el proceso de selección de variables puede ayudarnos a lidiar con el archiconocido problema del overfitting. Este problema, básicamente, consiste en que nuestro modelo es incapaz de generalizar las predicciones para nuevas instancias y se ajusta excesivamente bien al subconjunto de entreno.

Como una imagen vale más que mil palabras:

Fuente: https://pbs.twimg.com/media/Ef5yD-j2X0AIkHEe.jpg

Al reducir el número de variables explicativas del modelo, lo convertimos en más generalizable ante nuevas observaciones, a costa de reducir la precisión en el conjunto de entreno.

El proceso para seleccionar variables es relativamente sencillo, básicamente, deberías preguntarte lo siguiente:

- ¿La adición de nuevas características aumenta necesariamente el rendimiento del modelo de manera significativa?
- Si no es así, ¿por qué añadir esas nuevas características que sólo van a aumentar la complejidad del modelo?

Existen distintos métodos de selección de variables con distintos enfoques. La **Regresión Stepwise** es uno de ellos.

2.1 LA REGRESIÓN STEPWISE

En estadística, la Regresión Stepwise es un método de ajuste de modelos de regresión en el que la elección de las variables predictivas se lleva a cabo mediante un procedimiento automático.

En cada iteración del proceso, se considera añadir o quitar una variable del conjunto de variables explicativas en base a algún criterio preestablecido.

En el material complementario, encontrarás los principales criterios, que suelen ser:

- F-test.
- T-test.
- R2 ajustado.
- Indice de Akaike.

2.1.1 PRINCIPALES ENFOQUES

El modo de seleccionar variables se puede enfocar de 3 modos distintos:

- **Forward Selection:** consiste en empezar el modelo con 1 variable, probar la adición de cada variable utilizando un criterio de ajuste del modelo elegido, añadir la variable (si la hay) cuya inclusión da la mejora más significativa estadísticamente del ajuste, y repetir este proceso hasta que ninguna mejore el modelo en un grado estadísticamente significativo.
- **Backward Elimination:** consiste en empezar con todas las variables posibles, probar la eliminación de cada variable utilizando un criterio de ajuste del modelo elegido, eliminar la variable (si la hay) cuya pérdida da el mínimo deterioro, estadísticamente, significativo del ajuste del modelo, y repetir este proceso hasta que ninguna otra variable pueda eliminarse sin una pérdida de ajuste estadísticamente significativa.
- **Stepwise o Eliminación Bidireccional:** una combinación de lo anterior, probando en cada paso las variables que deben incluirse o excluirse.

2.2 CRITERIOS DE SELECCIÓN

Un algoritmo ampliamente utilizado en la actualidad fue propuesto, por primera vez, por *Efroymson* en los años 60. Se trata de un procedimiento automático para la selección de modelos estadísticos en los casos en que hay un gran número de posibles variables explicativas y no existe una teoría subyacente en la que basar la selección del modelo.

El procedimiento se utiliza, principalmente, en el análisis de regresión, aunque el enfoque básico es aplicable en muchas formas de selección de modelos. Se trata de una variación de la selección con el enfoque *Forward Selection.* En cada etapa del proceso, después de añadir una nueva variable, se hace una prueba para comprobar si algunas variables pueden suprimirse sin aumentar, apreciablemente, la *suma residual de los cuadrados* (RSS).

El procedimiento termina cuando la medida se maximiza (localmente), o cuando la mejora disponible cae por debajo de algún valor crítico.

2.3 PRECISIÓN DEL MODELO

Una forma de comprobar los errores en los modelos creados por Regresión Stepwise consiste en no depender del estadístico F, sino en evaluar el modelo frente a un conjunto de datos no utilizados para crear el modelo. Esto se hace, habitualmente, construyendo un modelo basado en el 70% de las observaciones (conjunto de entreno) y se utiliza el 30% de las observaciones restantes para testar el modelo (conjunto de testeo), **aunque estas proporciones pueden variar en función de distintos criterios** tales como el imbalanceo de clases o el número de observaciones total.

IMPORTANTE

La **precisión del modelo se suele medir mediante distintos métodos de ajuste del modelo** (accuracy, standard error o MAPE).

3. ACTIVIDAD GUIADA

El ejercicio propuesto consiste en una tarea de clasificación sobre un conjunto de datos públicos. El conjunto de datos contiene 846 observaciones y 18 variables de 4 tipos diferentes de vehículos.
Cada uno de los atributos se corresponde con medidas físicas relativas a la forma de los vehículos. Para nuestro caso concreto, únicamente tendremos en cuenta 3 de las 4 clases, agrupando la clase OPEL y SAAB como car.

Se aplicará el LDA sobre el mismo conjunto de datos dos veces con enfoques diferentes:

- En el **primer enfoque,** el LDA actuará como un clasificador **(aproximación supervisada).**

- En **el segundo enfoque,** el LDA actuará como un **reductor de dimensionalidad** y utilizaremos el algoritmo del Random Forest para realizar la tarea de clasificación **(aproximación no supervisada).**

3.1 EL LDA COMO CLASIFICADOR

En primer lugar, cargaremos el conjunto de datos. Esta versión del conjunto de datos ya viene con la nueva clase "car" y no será necesario realizar ningún filtrado ni agrupación.

Posteriormente, eliminaremos las observaciones con valores nulos (verás que hay pocas) y normalizaremos las variables en rango 0-1.

Para terminar, crearemos un subconjunto de entreno y otro de testeo para aplicar el algoritmo de clasificación, a posteriori.

```
```{r,warning=FALSE,message=FALSE}
# Carga paquetes necesarios
require(dplyr)
require(MASS)
require(caret)
require(randomForest)
```
```

```
# Carga del conjunto de datos
data_raw <- read.csv( "https://raw.githubusercontent.com/LaxmiChaudhary/Classifying-Silhouettes-of-Vehicles/master/vehicle.csv",
stringsAsFactor = FALSE )

# Elimino los NA's
data<-na.omit(data_raw)

# Normalizo las variables del datset en rango 0-1
maxs <- apply( data[,1:18], 2, max )
mins <- apply( data[,1:18], 2, min )
dataset <- as.data.frame( scale( data[,1:18], center = mins, scale = maxs - mins ) )
dataset <- cbind( dataset, "class" = data$class )

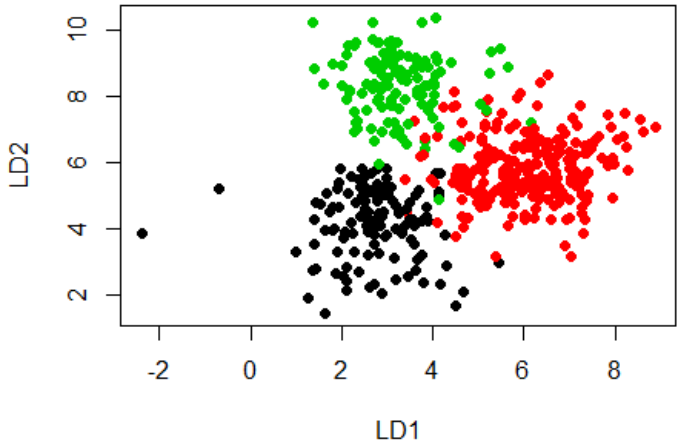
# Split dataset (60% conjunto de entreno - 40% conjunto de testing)
index <- sample( 1:nrow( dataset ), round( nrow( dataset )*0.6 ), replace = FALSE )
X_train <- dataset[ index, ]
test <- dataset[ -index, ]
```

La implementación que utilizaremos del algoritmo del LDA la encontrarás en la librería MASS mediante la función lda. Esta función trae un par de parámetros del modelo como las probabilidades previas de los grupos, las medias de los grupos y los coeficientes de discriminante lineal.

El resultado más importante, aquí, son los coeficientes, son valores que describen el nuevo espacio de variables en el que se proyectarán los datos.

```
# Creo el objeto con el modelo LDA llamado model
set.seed(12345)
model <- lda( class ~ ., data = X_train )

# Grafico las dos nuevas dimensiones creadas por el modelo LDA
projected_data <- as.matrix( X_train[, 1:18] ) %*% model$scaling
plot( projected_data, col = X_train[,19], pch = 19 )
```




```
# Cálculo de las predicciones del modelo
X_test <  test[, !( names( test ) %in% c( "class" ) ) ]
model.results <- predict( model, X_test )

# Matriz de confusión
t – table( model.results$class, test$class )
print(confusionMatrix(t))
```

```
Confusion Matrix and Statistics


          bus car van
  bus  81   4   0
  car   2 159   2
  van   1   7  69

Overall Statistics

               Accuracy : 0.9508
                 95% CI : (0.9213, 0.9716)
    No Information Rate : 0.5231
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9204

 Mcnemar's Test P-Value : 0.2173

Statistics by Class:

                     Class: bus Class: car Class: van
Sensitivity              0.9643     0.9353     0.9718
Specificity              0.9834     0.9742     0.9685
Pos Pred Value           0.9529     0.9755     0.8961
Neg Pred Value           0.9875     0.9321     0.9919
Prevalence               0.2585     0.5231     0.2185
Detection Rate           0.2492     0.4892     0.2123
Detection Prevalence     0.2615     0.5015     0.2369
Balanced Accuracy        0.9738     0.9547     0.9702
```

El LDA reduce la dimensionalidad del número original de variables a C - 1, donde C es el número de clases. En este caso, tenemos 3 clases, por lo tanto, el nuevo espacio tendrá sólo 2 variables (LD1 y LD2).

En la imagen de arriba, puedes ver el nuevo plano con las dos únicas variables. Tal y como podrás ver, hay algunos puntos que se superponen entre las tres clases, pero en general, el conjunto de datos es bastante separable.

Después de la fase de entrenamiento, necesitamos medir la precisión del modelo obtenido para saber cómo está clasificando las clases. Para ello, imprimiremos la matriz de confusión.

Tal y como puedes ver en la imagen de arriba, el LDA alcanza una precisión por encima del 94% ¿Es buen resultado? Bueno, según como se mire y a quién se lo presentes:

Fuente: https://www.memecenter.com/fun/1430563/rmx-villain-039-s-accuracy

Ahora que hemos visto cómo aplicar el LDA como clasificador mediante una implementación del algoritmo en R, pasamos a ver cómo aplicarlo como reductor de dimensionalidad.

3.2 EL LDA COMO REDUCTOR DE DIMENSIONALIDAD

Tal y como has visto en el ejemplo anterior, el Análisis Discriminante reduce la dimensionalidad del número original de variables a C - 1, donde C es el número de clases. En este ejercicio, tenemos 3 clases, por lo tanto, el nuevo espacio hemos visto que sólo tenía 2 variables (LD1 y LD2).

En el siguiente ejemplo, tenemos 3 clases y 18 variables.

El LDA reducirá de 18 variables a solo 2. Después de la reducción de variables, se aplicará el algoritmo del Random Forest para la tarea de clasificación. El procedimiento aquí es casi el mismo que el anterior. Echa un vistazo:

Tal y como puedes ver en el código, utilizaremos los coeficientes ya definidos en la fase de entrenamiento anterior para proyectar el conjunto de datos de entrenamiento en el nuevo espacio de variables, este nuevo espacio será el nuevo conjunto de datos de entrenamiento. Este será el dataset de entreno para el modelo.

```
# Creación del nuevo dataset de entreno
new_X_train <- as.matrix( X_train[,1:18] ) %*% model$scaling
new_X_train <- as.data.frame( new_X_train )
new_X_train$class <- X_train$class
head(new_X_train)
```

	LD1 <dbl>	LD2 <dbl>	class <fctr>
263	1.7942267	15.61282	van
622	0.3341986	13.81608	van
661	1.1311754	13.99743	van
657	0.3379571	11.27818	bus
579	3.3074998	11.36947	car
61	3.5428872	12.61724	car

6 rows

```r
# Creación del nuevo dataset de testing
new_X_test <- as.matrix( X_test[,1:18] ) %*% model$scaling
new_X_test <- as.data.frame( new_X_test )
head(new_X_test)
```

	LD1 <dbl>	LD2 <dbl>
9	0.22722577	14.754666
12	3.28615950	13.492740
13	0.47132817	10.710429
18	-0.01463246	8.605316
19	1.83266227	11.977992
22	1.79488508	12.289095

6 rows

Posteriormente, procedemos a entrenar el modelo.

En posteriores temas, aprenderás más acerca de los parámetros del algoritmo del Random Forest, por lo que no nos pararemos en detalle a ver cómo se construye el clasificador ni a su interpretación.

Para este caso concreto, únicamente necesitaremos ver las predicciones para poderlo comparar con el LDA actuando como clasificador:

```r
# Entreno el modelo con random forest
set.seed(12345)
modfit.rf <- randomForest(class ~. , data=new_X_train)

# Predicciones con random forest
predictions.rf <- predict(modfit.rf, as.data.frame(new_X_test), type = "class")

# Matriz de confusión
t = table( predictions.rf, test$class)
print(confusionMatrix(t))
```

```
Confusion Matrix and Statistics


predictions.rf bus car van
           bus  81   4   0
           car   3 159   4
           van   0   7  67

Overall Statistics

               Accuracy : 0.9446
                 95% CI : (0.9139, 0.9668)
    No Information Rate : 0.5231
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.91

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: bus Class: car Class: van
Sensitivity              0.9643     0.9353     0.9437
Specificity              0.9834     0.9548     0.9724
Pos Pred Value           0.9529     0.9578     0.9054
Neg Pred Value           0.9875     0.9308     0.9841
Prevalence               0.2585     0.5231     0.2185
Detection Rate           0.2492     0.4892     0.2062
Detection Prevalence     0.2615     0.5108     0.2277
Balanced Accuracy        0.9738     0.9451     0.9581
```

Tal y como podrás comprobar en los resultados arriba expuestos, el Random Forest alcanza una precisión por encima del 95%, quedando por debajo en poco más de medio punto porcentual del LDA actuando como clasificador.

3.3 CONCLUSIÓN

Tal y como habrás podido comprobar, los resultados son ligeramente mejores para el primer enfoque del problema. Es decir, **se mejora la precisión, ligeramente, si el LDA actúa directamente como clasificador en lugar de utilizarlo como reductor de dimensionalidad para, posteriormente, utilizar un clasificador distinto.**

Así pues, **¿a qué enfoque concederías la victoria?**

Fuente: http://memegenerator.net/instance/46083336/michael-scott-with-win-win-win-we-all-win

En realidad no hay una respuesta universal. Pese a que el primer enfoque nos da un resultado, ligeramente, superior al segundo y es más directo, en este caso, sí parece ser la solución más directa y efectiva al problema. Sin embargo, esto no debes tomarlo como norma y puede haber determinados escenarios en los que te interese utilizar el LDA como reductor de dimensionalidad para, posteriormente, aplicar el nuevo set de variables dentro de un modelo mucho más grande.

IDEAS CLAVE

- El LDA está estrechamente relacionado con el análisis de varianza (ANOVA) y el Análisis de Regresión.
- El LDA se puede utilizar como reductor de dimensionalidad.
- La selección de variables es nuestra primera línea de defensa contra el overfitting.
- Se puede enfocar la selección de variables mediante la regresión stepwise de tres modos distintos, empezando con una sola variable e ir añadiendo el resto (forward selection), empezando con todas las variables posibles e ir eliminando una a una (backward elimination) o mediante la eliminación bidireccional (stepwise) en el que cada paso es una mezcla de los dos anteriores.
- Es importante definir el método de ajuste para determinar si un modelo es efectivo o no, con métricas como la precisión o el MAPE.