Plantilla de ficha de formación

Título	ANÁLISIS DISCRIMINANTE LINEAL
Palabras clave (metaetiquetas)	análisis discriminante, clasificación, R, análisis bayesiano
Idioma	Español
Objetivos / Metas / Resultados de aprendizaje	**el objetivo de este módulo es introducir y explicar los conceptos básicos del Análisis Discriminante Lineal (ADL).** **Al final de este módulo podrás:** - **Identificar situaciones en las que ADL puede ser útil** - **Calcular funciones ADL** - **Interpretar los resultados producidos por ADL descriptivo y predictivo**

Curso de entrenamiento:	
Alfabetización en ciencia de datos	
Módulo de visualización de datos y análisis visual	X
Introducción a la ciencia de datos para las ciencias humanas y sociales	
Ciencia de datos para siempre	
Periodismo de datos y storytelling	

Descripción	En este módulo de capacitación, se le presentará el uso del análisis discriminante lineal (ADL). El ADL es un método para encontrar combinaciones lineales de variables que separa mejor las observaciones en grupos o clases, y fue desarrollado originalmente por Fisher (1936). Este método maximiza la relación entre la varianza entre clases y la varianza dentro de la clase en cualquier conjunto de datos en particular. Al hacer esto, se maximiza la variabilidad entre grupos, lo que da como resultado una separabilidad máxima. ADL se puede utilizar con fines puramente de clasificación, pero también con objetivos predictivos.

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

Contenidos dispuestos en 3 niveles	**1. INTRODUCCIÓN: MOTIVACIÓN POR UN EJEMPLO ILUSTRATIVO** Supongamos que tenemos una muestra de individuos y observamos el modo de transporte (automóvil, transporte público o caminando) que suelen tomar para moverse dentro de una ciudad. Sabemos que la elección del modo de transporte está parcialmente influenciada por su situación económica, y observamos datos sobre su edad en años y el ingreso anual del hogar, junto con el medio de transporte elegido:  Queremos saber cómo estas dos variables ayudan a clasificar (es decir, discriminar) a los individuos asignándolos a una categoría específica de modo de transporte. Podemos ver que no existe una clasificación perfecta: las personas con altos ingresos tienden a usar el automóvil con mayor frecuencia, pero existe una gran superposición de las categorías "caminar" y "transporte público" para aquellos con ingresos más bajos. Y hay una mayor superposición entre las categorías con respecto a su distribución por edad: las personas mayores no caminan, pero en valores más jóvenes, la edad no es un buen predictor del modo de transporte. Este es el problema típico que aborda ADL. **2. ADL para clasificación** **2.1. Formulación** Las funciones ADL se pueden recuperar para ayudar con la clasificación de los datos en función de una matriz de variables $\mathbf{X}$. De manera similar al análisis de componentes principales (PCA), las funciones ADL tienen

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

como objetivo encontrar una combinación lineal de los datos originales como:

$$\mathrm{LDA} = \mathbf{u}^{\mathbf{T}}\mathbf{X}$$

donde la varianza entre clases ($\mathbf{B}$) se maximiza en relación con la varianza dentro de la clase ($\mathbf{W}$), que puede abordarse como un problema generalizado de valores propios:

$$\mathrm{u} = \arg\max_{\mathbf{u}} \frac{\mathbf{u}^{\mathbf{T}}\mathbf{B}\mathbf{u}}{\mathbf{u}^{\mathbf{T}}\mathbf{W}\mathbf{u}}$$

Las coordenadas discriminantes se obtienen a partir de los vectores propios de $\mathbf{W}$ ^(− $\mathbf{1}$) $\mathbf{B}$.

2.2. Un ejemplo

Como ejemplo ilustrativo, resolvemos el problema de clasificación del modo de transporte en función de la edad y los ingresos por ADL en R. Esto se puede hacer fácilmente mediante la función "lda" dentro de la biblioteca "mass". Para todo el análisis presentado aquí, necesitaremos instalar y cargar los siguientes paquetes R:

```
# LDA packages
install.packages("MVN")
install.packages("heplots")
install.packages("caret")
install.packages("MASS")
library(MVN)
library(heplots)
library(caret)
library(tidyverse)
library(MASS)
```

Los datos estudiados vienen en un archivo csv (llamado "trasnport_example"), que se puede importar fácilmente a R ejecutando este código:

```
# Get Data
transport <- read.csv(transport_example.csv)
View(transport)
transport <- as.data.frame(transport)
```

Para tener una primera impresión de los datos, podemos representar gráficamente la muestra en forma de diagrama de dispersión como:

Co-funded by the Erasmus+ Programme of the European Union

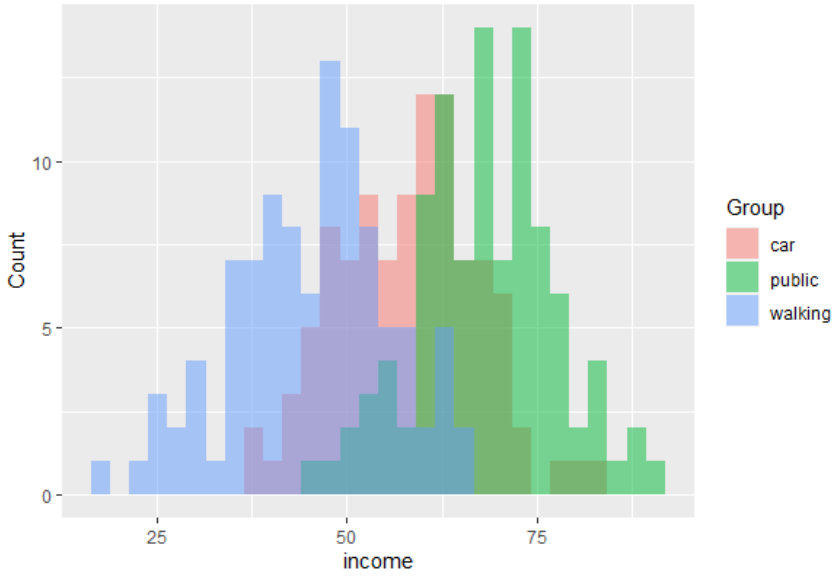
Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

```
#scatterplots
ggplot(transport, aes(age, income)) +
  geom_point(aes(color = group))
```

Las líneas de código anteriores producen el diagrama de dispersión que se muestra en la sección introductoria de este documento. Alternativamente, podríamos trazar los datos como una serie de histogramas como:

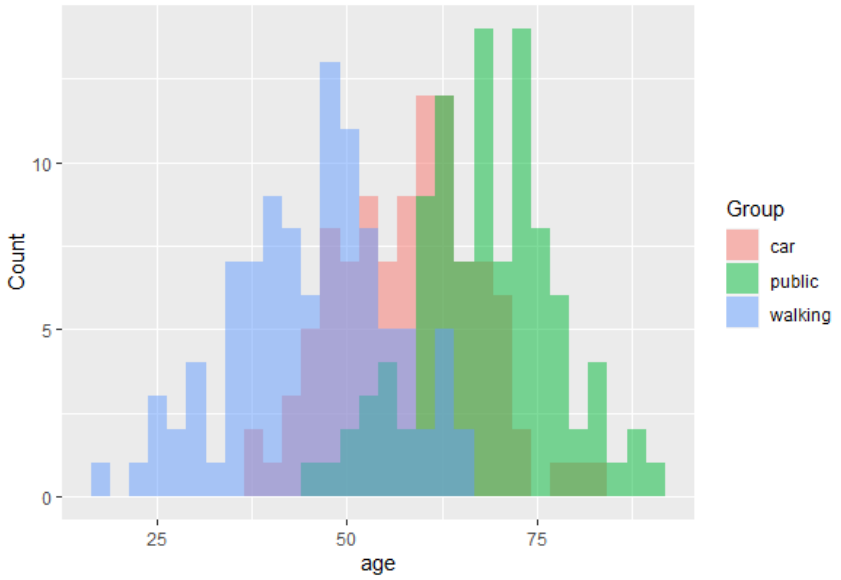
```
#histograms for income
ggplot(transport, aes(x = income, fill = group)) +
  geom_histogram(alpha = 0.5, position = "identity") +
  labs(x = "income", y = "Count", fill = "Group")

#or
ldahist(data = transport$income, g = transport$group)
```

Al ejecutar cualquiera de estas dos líneas, podemos tener una idea de cómo se distribuyen los modos de transporte entre los valores, la edad y los ingresos. Por ejemplo:



O:

Co-funded by the Erasmus+ Programme of the European Union
Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.



ADL se lleva a cabo simplemente ejecutando:

```
#######################################
### Case Classification ###
#######################################
# Run the LDA using the lda function
output <- lda(group ~ ., transport)
output
```

La salida típica muestra las medias iniciales por grupo, los coeficientes en las proyecciones de LD y la proporción de la varianza entre clases (*between* o traza) que explica cada coordenada de LD:

```
Group means:
           age income
car      58.32  89.44
public   68.40  49.82
walking  45.52  52.89

Coefficients of linear discriminants:
               LD1        LD2
age     -0.1177011 0.08844338
income   0.1376988 0.02050334

Proportion of trace:
   LD1    LD2
0.8997 0.1003
```

Co-funded by the Erasmus+ Programme of the European Union

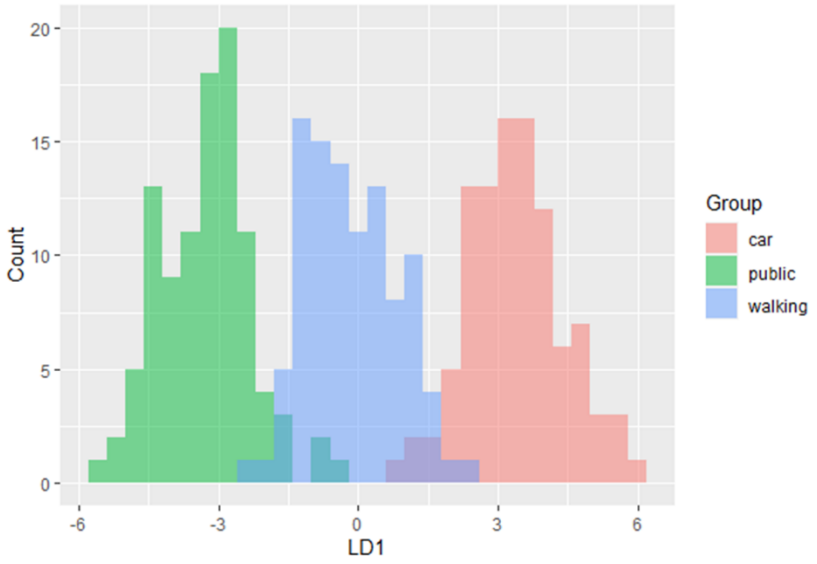
Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

En nuestro ejemplo, la primera coordenada LD está positivamente correlacionada con el ingreso y negativamente con la edad, y contiene casi el 90% de la variabilidad entre clases. La segunda función LD muestra una correlación positiva pero más débil con ambas variables, y solo representa aproximadamente el 10% de esta variabilidad.

Las nuevas coordenadas se producen proyectando los puntos de datos originales con los coeficientes ADL mediante la expresión $\mathbf{u}^T\mathbf{X}$. En estas nuevas coordenadas, las observaciones están más claramente separadas entre grupos. En nuestro ejemplo, tenemos dos coordenadas LD para cada individuo, dadas su edad e ingresos. Las coordenadas correspondientes a la primera función LD tienen el mayor poder discriminante. Podemos ver fácilmente este poder discriminante trazando en R un histograma, poniendo ahora las primeras coordenadas LD en el eje horizontal:

```
#histograms: first LDA
ggplot(lda.data, aes(x = LD1, fill = group)) +
  geom_histogram(alpha = 0.5, position = "identity") +
  labs(x = "LD1", y = "Count", fill = "Group")
```

Obtención:



Este gráfico muestra cómo la cantidad de superposición disminuye considerablemente. En otras palabras, la primera coordenada LD (recuerda que es un "compuesto" que se correlaciona negativamente

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

con la edad y positivamente con el ingreso) discrimina adecuadamente entre las categorías de transporte.

3. ADL predictivo

3.1 El procedimiento

ADL se puede utilizar no solo con fines de clasificación (descriptivos), sino también con el objetivo de predecir la pertenencia a una clase. Por ejemplo, supongamos que tenemos datos de la edad y los ingresos familiares anuales de una persona (dentro o fuera de la muestra) y nos gustaría predecir el modo de transporte que es más probable que utilice. ADL puede ser útil para proporcionarnos una predicción, de manera similar a los modelos probit o logit multinominal.

Para este propósito predictivo, se requieren algunas supuestos:

- los grupos son normales multivariados
- varianzas-covarianzas iguales entre grupos

La formulación del ADL predictivo está relacionada con la formulación del teorema de Bayes para actualizar probabilidades: Sea g el número de grupos y q_i la probabilidad inicial (generalmente frecuencias relativas observadas) para el grupo i. Sea $\mathbf{x}$ un vector de observaciones de variables para un individuo. La probabilidad (a posteriori) de pertenecer al grupo G_i condicionada a $\mathbf{x}$, P(G_i | $\mathbf{x}$), se puede expresar como:

$$\mathrm{P}(G_i|\mathbf{x}) = \frac{q_i \mathrm{P}(\mathbf{x}|G_i)}{\sum_{j=1}^{g} q_j \mathrm{P}(\mathbf{x}|G_j)}$$

Este es un enfoque bayesiano que actualiza las probabilidades previas q_i basándose en las probabilidades condicionales P($\mathbf{x}$ | G_i). Bajo los supuestos de normalidad:

$$\mathrm{P}(\mathbf{x}|G_i) = (2\pi)^{(-p/2)} |\mathbf{W}|^{(-1/2)} e^{(-D_i^2/2)}$$

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.

donde | $\mathbf{W}$ | es el determinante de la matriz de varianza dentro de la clase y $D_{\ i}$ es$D_i \ = (\mathbf{x} - \bar{\mathbf{x}}_\mathbf{i})^T \mathbf{W}^{-1} \ (\mathbf{x} - \bar{\mathbf{x}}_\mathbf{i})$. Reemplazando la expresión de $\mathrm{P}(\mathbf{x}|G_i)$ en la fórmula para $\mathrm{P}(G_i|\mathbf{x})$, tenemos:

$$\mathrm{P}(G_i|\mathbf{x}) = \frac{q_i e^{(-D_i^2/2)}}{\sum_{j=1}^{g} q_j \, e^{\left(-D_j^2/2\right)}}$$

3.2. Un ejemplo con R

La rutina ADL en R puede producir probabilidades posteriores basándose en los supuestos y la formulación detallada anteriormente. Las funciones ADL permiten predecir la pertenencia a la clase más probable para cualquier individuo, dado un vector de variables (edad e ingresos familiares en el ejemplo).

Como ilustración, la tabla que se muestra a continuación contiene las probabilidades pronosticadas para cada grupo para un subconjunto de individuos en la muestra. Se supone que los $q_i$son idénticas para cada uno de los tres modos de transporte ($q_i = 1/3$).

group	income	age	LD1	LD2	predclass	pred_car	pred_public	pred_walk
walking	26	47	1.349620208	-3.127883266	walking	1.983231e-03	2.965401e-07	9.980165e-01
walking	27	51	1.782714373	-2.957426532	walking	1.245493e-02	1.063176e-07	9.875450e-01
walking	28	35	-0.538167997	-3.197036576	walking	2.241897e-06	9.299867e-05	9.999048e-01
walking	29	34	-0.793567966	-3.129096536	walking	1.034985e-06	2.354290e-04	9.997635e-01
walking	30	45	0.603417987	-2.815116429	walking	2.575777e-04	5.608833e-06	9.997368e-01
walking	31	35	-0.891271423	-2.931706440	walking	1.062902e-06	4.699394e-04	9.995290e-01
walking	31	43	0.210319191	-2.767679728	walking	7.042531e-05	2.095366e-05	9.999086e-01
walking	32	42	-0.045080777	-2.699739689	walking	3.251705e-05	5.305263e-05	9.999144e-01
walking	34	45	0.132613419	-2.461342914	walking	9.528279e-05	4.866634e-05	9.998561e-01
walking	37	37	-1.322080621	-2.360039490	walking	6.786660e-07	5.490035e-03	9.945093e-01
walking	56	60	-0.391329301	-0.208038498	walking	1.019914e-03	2.021248e-02	9.787676e-01
walking	54	48	-1.808312938	-0.630965323	walking	1.821514e-06	4.269625e-01	5.730357e-01
walking	63	78	1.263341587	0.780125254	car	6.956557e-01	2.513904e-04	3.040929e-01
public	69	56	-2.4722395	0.859712069	public	4.567507e-08	9.909603e-01	9.039690e-03
public	87	70	-2.6630764	2.738739632	public	1.118083e-08	9.998736e-01	1.264240e-04
public	73	50	-3.7692370	1.090465551	public	8.209481e-12	9.998980e-01	1.019855e-04
public	46	33	-2.9321862	-1.646062437	public	2.043553e-09	7.715710e-01	2.284290e-01
public	62	64	-0.5467408	0.404635130	walking	1.713491e-03	1.000701e-01	8.982164e-01
public	68	42	-4.2823219	0.484221945	public	2.851843e-13	9.999323e-01	6.768416e-05
public	50	31	-3.6783884	-1.333295600	public	1.790602e-11	9.845625e-01	1.543752e-02
public	71	36	-5.4616183	0.626532048	public	1.118031e-16	9.999987e-01	1.298905e-06
public	56	43	-2.7322094	-0.556595260	public	8.609396e-09	9.387842e-01	6.121583e-02
public	60	45	-2.9276163	-0.161815068	public	2.388442e-09	9.839053e-01	1.609473e-02

La clase predicha corresponde a la $\mathrm{P}(G_i|\mathbf{x})$ más alta para cada individuo. Se calculan aplicando la siguiente rutina en Rstudio :

Co-funded by the Erasmus+ Programme of the European Union

	`` ##################################### ### Predicting classifications ### ##################################### # Get the posterior values and predicted classification for each case pred <- predict(output) # Posterior values for each class for each case posteriors <- pred$posterior # Predicted Class predclass <- pred$class # Putting Data (including actual class) next to predicted class and posterior values post_transport <- cbind(lda.data,predclass,posteriors) colnames(post_transport) <- c("group","income","age","LD1","LD2","predclass", "pred_car","pred_public","pred_walk") ``` En la mayoría de los casos, ADL predice correctamente el grupo al que pertenece cada individuo. Hay algunos casos, sin embargo, para los cuales ADL no predice correctamente. Estos casos corresponden a las observaciones superpuestas que aún permanecen en la clasificación ADL
Autoevaluación (preguntas y respuestas de opción múltiple)	El ADL es una técnica estadística que permite: a) Clasificación de datos en grupos b) Predicción de la pertenencia a una clase c) Ambas respuestas son verdaderas Los supuestos necesarios para aplicar ADL con fines predictivos son: a) Normalidad multivariada entre grupos b) Igualdad de varianzas-covarianzas entre grupos c) Ambas respuestas son verdaderas ADL se basa en maximizar la ratio: a) Variabilidad “entre grupos” versus “dentro de la clase” b) Variabilidad “Dentro de los grupos” versus total c) Variabilidad total versus “dentro de los grupos”
Recursos (videos, enlaces a referencias)	
material relacionado	
PPT relacionado	
Bibliografía	Boedeker, P., & Kearns, N. T. (2019). Linear discriminant analysis for prediction of group membership: A user-friendly primer. Advances in Methods and Practices in Psychological Science, 2, 250-263.
Proporcionado por	[Unioví]

Co-funded by the Erasmus+ Programme of the European Union

Con el apoyo del programa Erasmus+ de la Unión Europea. Este documento y su contenido reflejan la puntos de vista únicamente de los autores, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en el mismo.