Cátedra Dr. Javier García Fronti

Análisis Numérico

Prof. Mg. Mauro Speranza

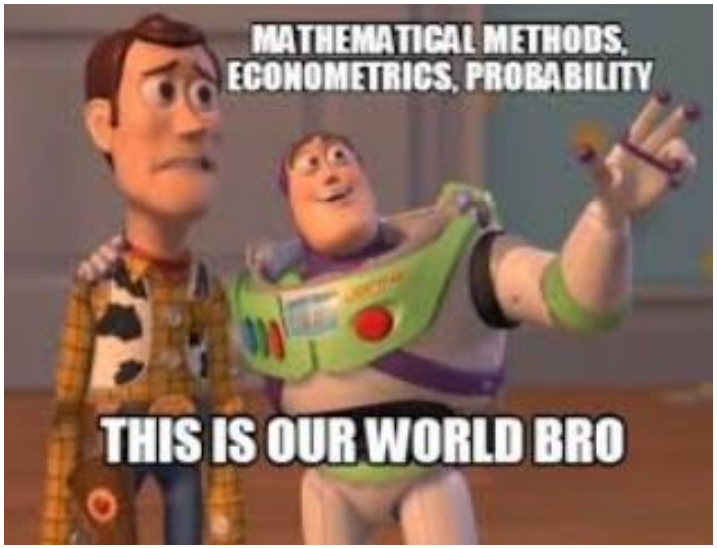
Clase especial

"Introducción a la ciencia de datos desde una perspectiva actuarial"

Dr. Martín E. Masci

Material disponible en: https://github.com/martinmasci/AnalisisNumerico2020

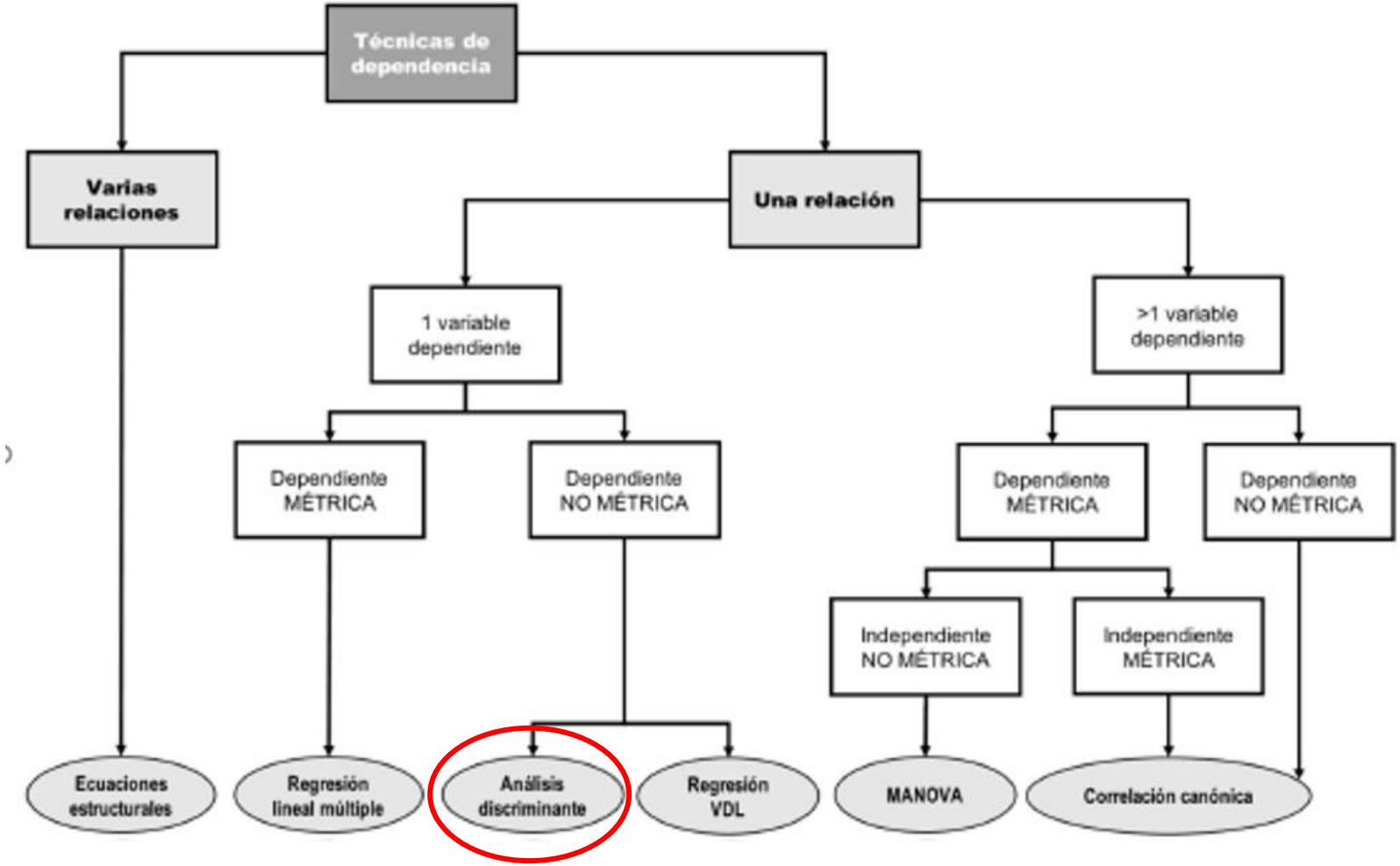
AGENDA *



- Minería de datos / Actualidad Actuarial
- Clustering
- Análisis de Componentes Principales
- Análisis de discriminante
- Aplicación en R

** La presente clase forma parte de contenidos generados por el Prof. Mg. Rodrigo Del Rosso*

Técnicas de Análisis de Dependencias



FUENTE: Manzano, J. A., & Jiménez, E. U. (2017). *Análisis multivariante aplicado con R.* Ediciones Paraninfo, SA.

Técnicas de Análisis de Interdependencias

FUENTE: Manzano, J. A., & Jiménez, E. U. (2017). *Análisis multivariante aplicado con R.* Ediciones Paraninfo, SA.

Syllabus Actuarial
(Internacional)

Syllabus Actuarial Internacional

Society of Actuaries (SOA – *USA*)

https://www.soa.org/research/about-research/

Institute and Faculty of Actuaries (*UK*)

https://www.actuaries.org.uk/system/files/field/document/IFoA%20changes%20to%20the%202019%20syllabus%20for%202020.pdf

Casualty Actuarial Society (CAS - *USA*)

https://www.casact.org/pubs/?fa=cas

Canadian Institute of Actuaries (CIA – *Canadá*)

https://www.cia-ica.ca/publications/standards-of-practice

Definiciones

Institute and Faculty of Actuaries (*UK*) // Cambios en el Syllabus 2019 al 2020

https://www.actuaries.org.uk/system/files/field/document/IFoA%20changes%20to%20the%202019%20syllabus%20for%202020.pdf

Institute and Faculty of Actuaries

Changes to the 2019 syllabus for 2020

There have been very minor changes to the Syllabus. These are detailed below:

***CM1** (2019) Syllabus objective 1.1 now removed for 2020. It now becomes Syllabus objective 2.1 for **CS1**. It is repeated below*

2.1 Data analysis

2.1.1 Describe the possible aims of data analysis (e.g. descriptive, inferential, and predictive).

2.1.2 Describe the stages of conducting a data analysis to solve real-world problems in a scientific manner and describe tools suitable for each stage.

2.1.3 Describe sources of data and explain the characteristics of different data sources, including extremely large data sets.

2.1.4 Explain the meaning and value of reproducible research and describe the elements required to ensure a data analysis is reproducible.

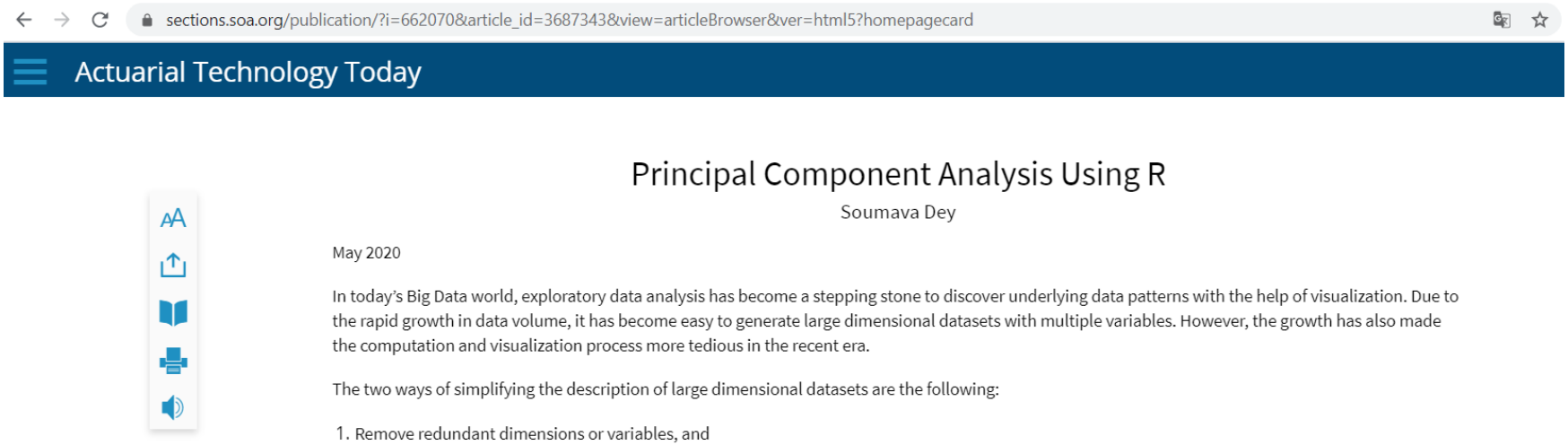
CM: Actuarial Mathematics

CS: Actuarial Statistics

Definiciones

Society of Actuaries (SOA – *USA*)

https://sections.soa.org/publication/?i=662070&article_id=3687343&view=articleBrowser&ver=html5?homepagecard

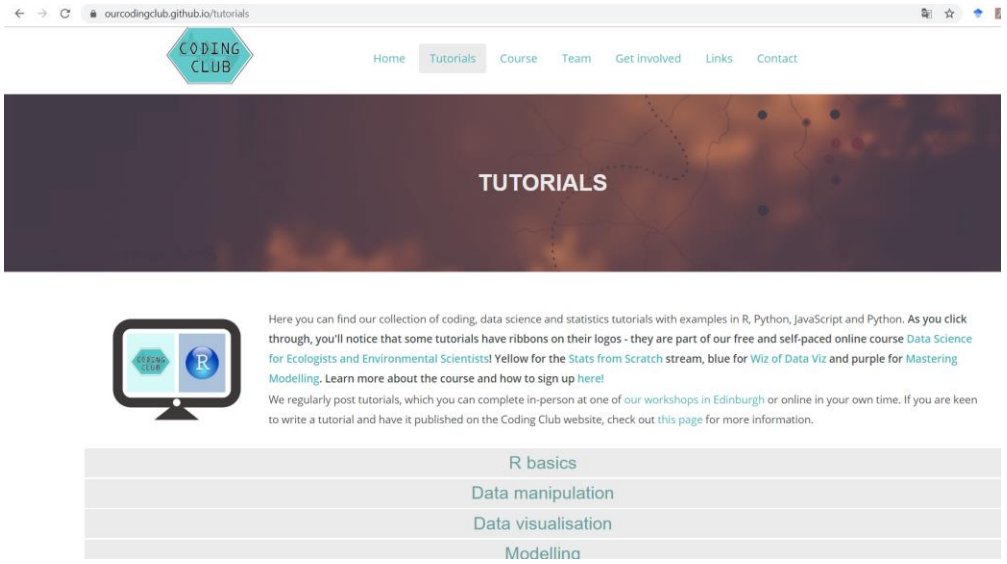


Definiciones

Society of Actuaries (SOA – *USA*)

Métodos de Machine Learning:
https://www.soa.org/resources/research-reports/2019/machine-learning-methods/

Es muy recomendable el repositorio de Our Coding Club:
https://ourcodingclub.github.io/



Definiciones

Canadian Institute of Actuaries (CIA – *Canadá*)

https://www.cia-ica.ca/perfectionnement/meeting-archives/past-meeting-programs/2019/2019-predictive-analytics-seminar/program

La CIA viene haciendo hace mucho eventos científicos que combinan la labor actuarial con la Ciencia de Datos.

https://www.cia-ica.ca/docs/default-source/2018/218081e.pdf

Definiciones

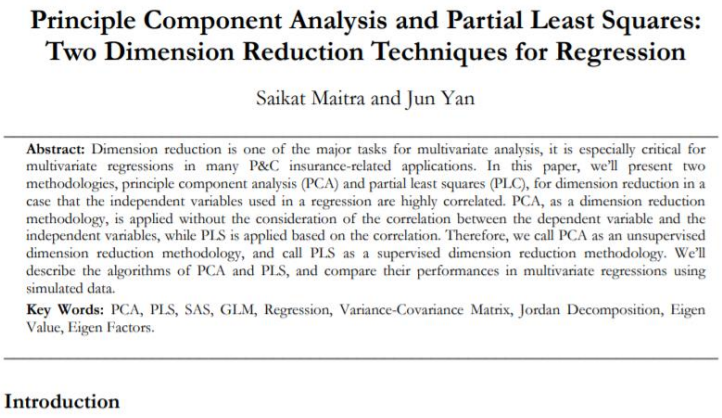
Casualty Actuarial Society (CAS - *USA*)

Artículos del repositorio con Data Analytics:

https://www.casact.org/research/dare/index.cfm?fa=adv_search_rs&keyword=data+analytics&title=&abstract2=&authorFirstName=&authorLastName=&publication=&prizeID=&isCASSyll=&hasAudio=&hasRelated=&category=&search=Search

Paper sobre componentes principales (2008)

https://www.casact.org/pubs/dpp/dpp08/08dpp76.pdf



Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression

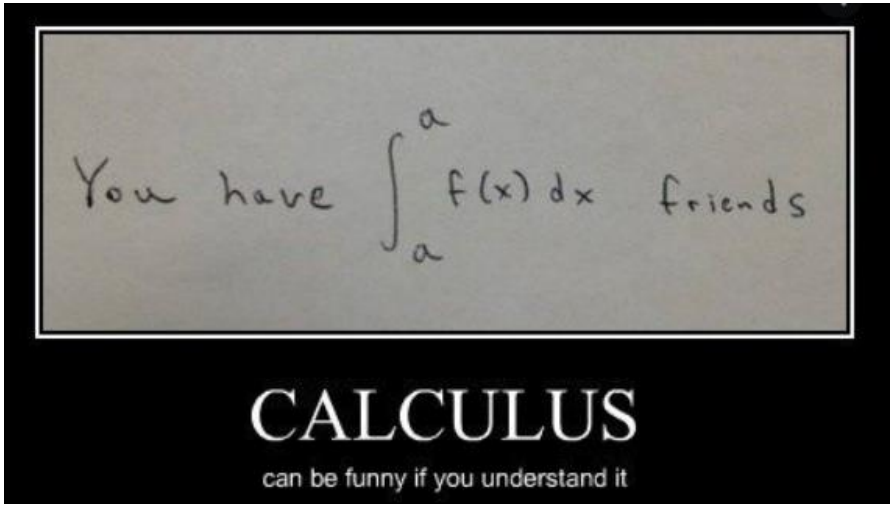
Saikat Maitra and Jun Yan

Abstract: Dimension reduction is one of the major tasks for multivariate analysis, it is especially critical for multivariate regressions in many P&C insurance-related applications. In this paper, we'll present two methodologies, principle component analysis (PCA) and partial least squares (PLC), for dimension reduction in a case that the independent variables used in a regression are highly correlated. PCA, as a dimension reduction methodology, is applied without the consideration of the correlation between the dependent variable and the independent variables, while PLS is applied based on the correlation. Therefore, we call PCA as an unsupervised dimension reduction methodology, and call PLS as a supervised dimension reduction methodology. We'll describe the algorithms of PCA and PLS, and compare their performances in multivariate regressions using simulated data.

Key Words: PCA, PLS, SAS, GLM, Regression, Variance-Covariance Matrix, Jordan Decomposition, Eigen Value, Eigen Factors.

Introduction

Minería de Datos

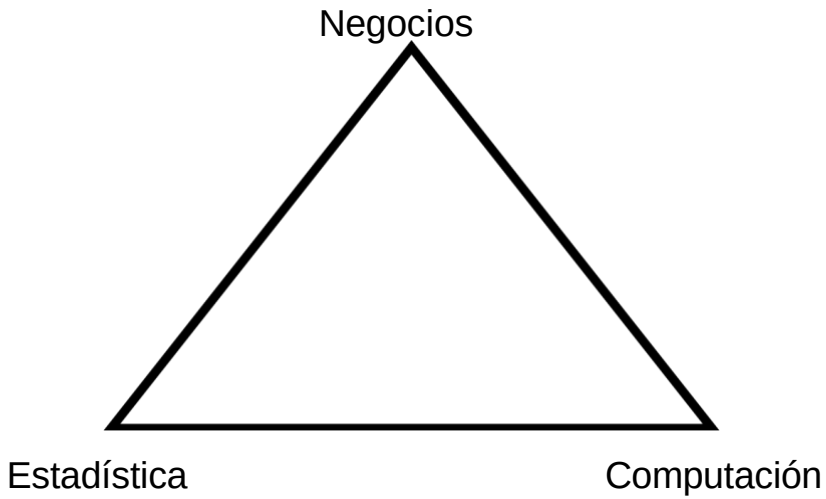


Definiciones

¿Qué es la Minería de datos?

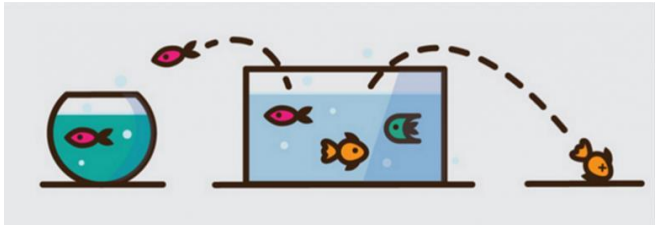
"*At a high level, **data science** is a set of fundamental principles that guide the extraction of knowledge from data. Data mining is the extraction of knowledge from data, via technologies that incorporate these principles.*" (Provost & Fawcett, 2013)

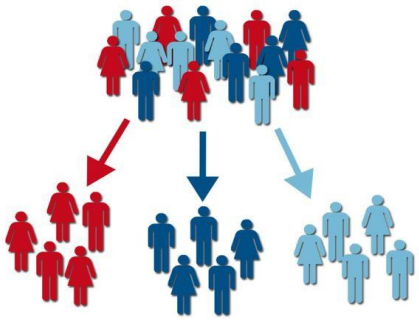
"*Data mining is a business process for exploring large amounts of data to discover meaningful patterns and rules.*" (Gordon & Berry, 2011). En este sentido, se requiere de habilidades en (al menos) tres campos:

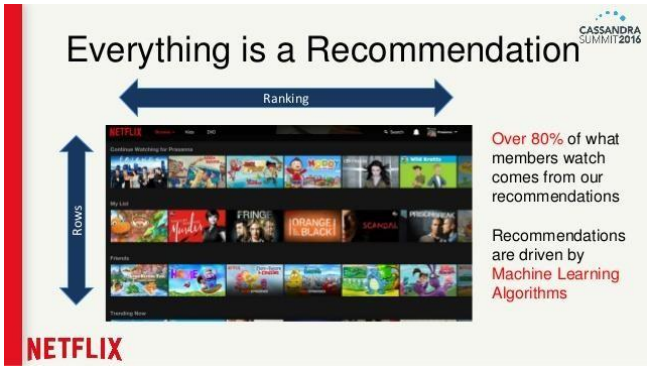


Aplicaciones en Negocios

- Churn (attrition) de clientes.
- Segmentación de clientes.
- Recomendación de productos.
- Publicidades personalizadas.
- Credit scoring.
- Predicción de valor de un cliente.
- Análisis de sentimiento.







Aprendizaje Automático (panorama)

Existen diversas técnicas para descubrir patrones en grandes volúmenes de datos (por ejemplo, a través de la exploración "manual").

Definición “casi canónica”:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." (Mitchell, T., 1997, Machine Learning).

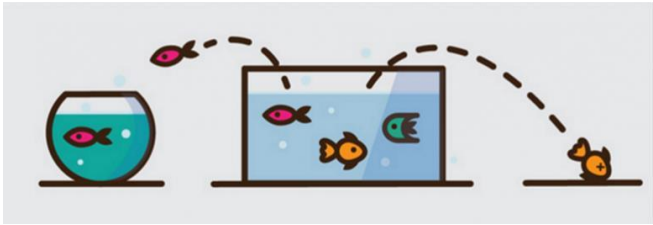
- *Computer program* → modelo estadístico.
- *Experience E* → datos.
- *Task T* → tarea (e.g., predecir el peso de unx chicx de acá a medio año).
- *Performance measure P* → medida de performance (e.g., *logloss*).

Aprendizaje Automático (panorama)

Existen tres grandes familias de algoritmos de Aprendizaje Automático (Supervisado, No Supervisado, Por Refuerzos).

1. Supervisado

"*For each observation of the predictor measurement(s)* x_i, $i = 1, ..., n$ *there is an associated response measurement* y_i".



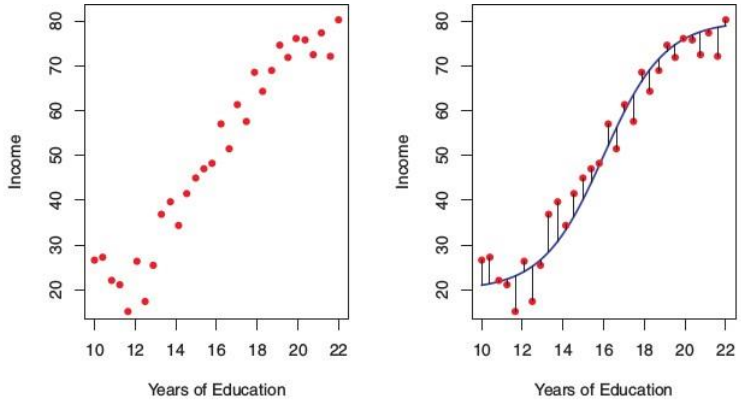
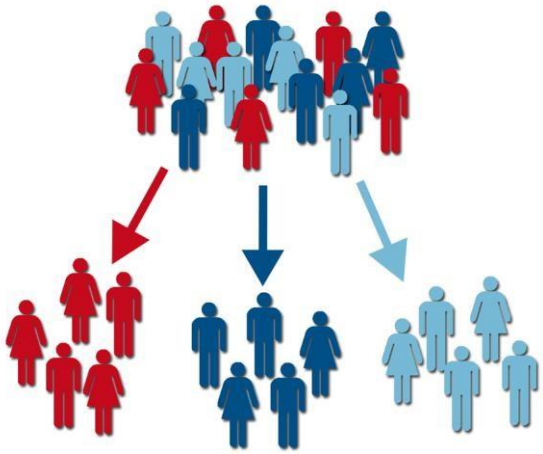


FIGURE 2.2. *The* Income *data set.* Left: *The red dots are the observed values of* income *(in tens of thousands of dollars) and* years of education *for* 30 *individuals.* Right: *The blue curve represents the true underlying relationship between* income *and* years of education*, which is generally unknown (but is known in this case because the data were simulated). The black lines represent the error associated with each observation. Note that some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve). Overall, these errors have approximately mean zero.*

2. No Supervisado

"*Unsupervised learning describes the somewhat more challenging situation in which for every observation i = 1, ..., n, we observe a vector of measurements* x_i *but no associated response* y_i *... We can seek to understand the relationships between the variables or between the observations*".



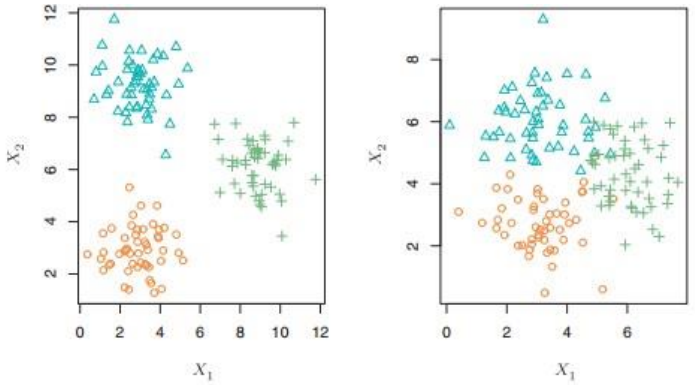
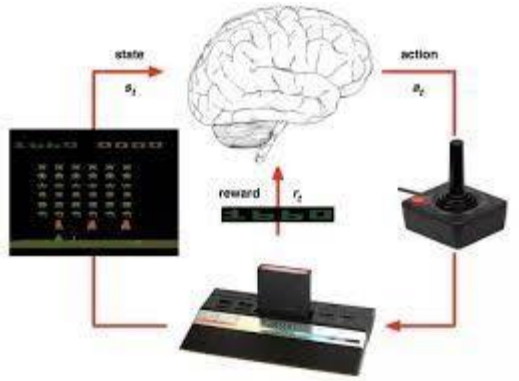


FIGURE 2.8. *A clustering data set involving three groups. Each group is shown using a different colored symbol.* Left: *The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups.* Right: *There is some overlap among the groups. Now the clustering task is more challenging.*

3. Por Refuerzos

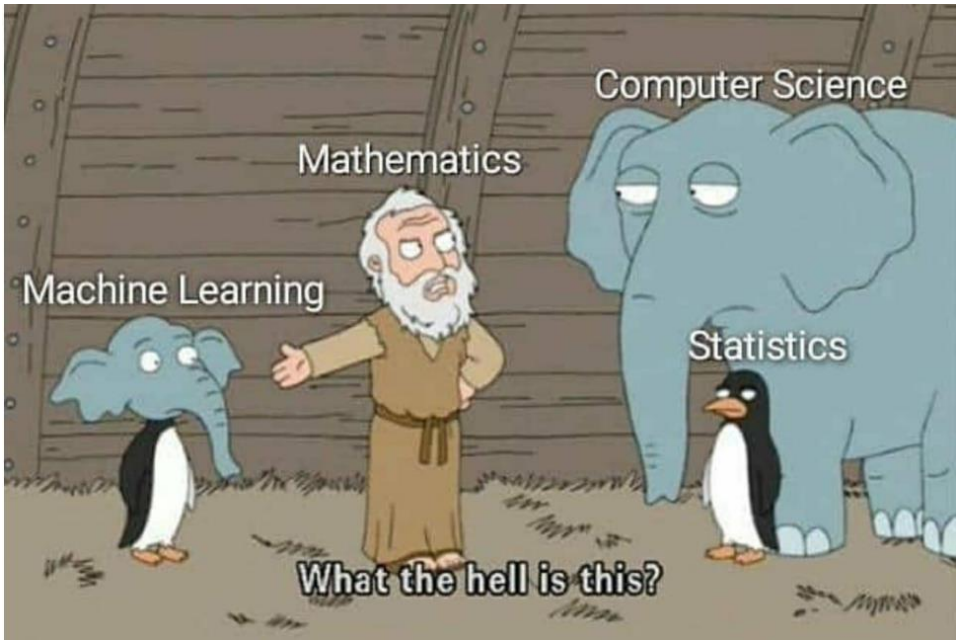
A grandes rasgos, se enfoca en lograr que agentes maximicen un beneficio esperado mediante la interacción con un ambiente (el cual muchas veces no conocen)... Muchos problemas de la realidad se ajustan a este esquema.



ADVERTENCIA

Desde nuestra humilde apreciación, aquí debemos parar un segundo y aclarar algo.

Necesitas tomar riendas de los procesos y GENERAR VALOR ORGANIZACIONAL con los modelos que vamos a utilizar.



También, repensar y construir RESPONSABLEMENTE un concepto de gobernanza de la información y la ética en la utilización de datos "públicos".

Aprendizaje Supervisado

Lo que caracteriza al aprendizaje supervisado es que uno quiere predecir una variable. Puede haber dos tipos de variables a predecir:

Continua → regresión
Categórica → clasificación (binaria, multiclase) $Y = f(X)$

Distintos problemas se atacan con distintos algoritmos (algunos sirven tanto para regresión como clasificación).

¿Qué tipo de problema son los siguientes?

- Predecir la cantidad de ventas de líneas telefónicas en un día dado.
- Predecir si un cliente en particular se dará de alta dada nuestra campaña de marketing.
- Predecir si un día determinado tendremos o no más de 200 líneas nuevas vendidas.

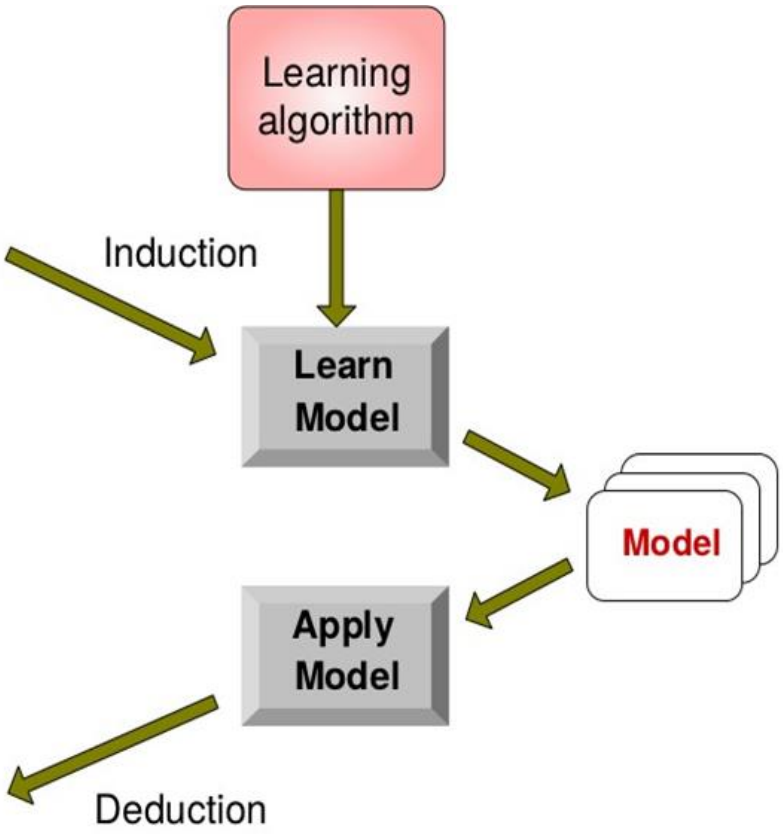
Aprendizaje Supervisado

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

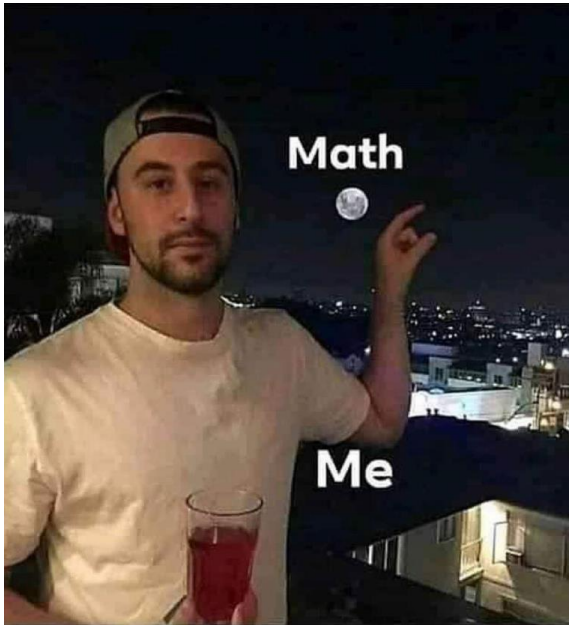
Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



$$Y = f(X)$$



Clustering

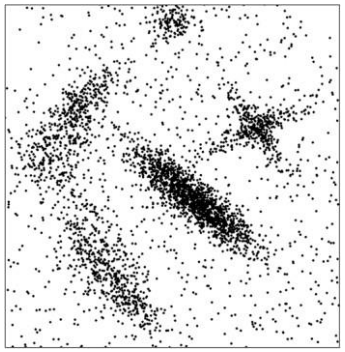
Técnicas

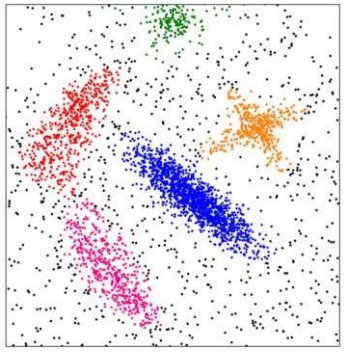
Clustering se refiere a una serie de técnicas cuyo objetivo es encontrar subgrupos o "*clusters*" en un conjunto de datos.

La idea es particionar los datos de tal manera que:

1. Las observaciones que pertenecen a un grupo sean similares entre ellas.
2. Las observaciones de grupos distintos sean distintas entre ellas.

Esto plantea el "problema" de definir cuando dos observaciones son similares o distintas entre sí. (Algo que en gran medida puede depender del dominio en donde se esté trabajando.)





Clustering

Existen múltiples familias de algoritmos de clustering:

1. *Partitioning*
2. *Hierarchical*
3. *Density-based*
4. *Grid-based*
5. *Model-based*

Nosotros veremos los dos algoritmos más tradicionales:

- *K-means clustering* (partitioning)
- *Hierarchical*

Clustering

Divide al conjunto de datos en K subconjuntos distintos sin solapamiento. Uno
debe fijar el valor de K antes de correr el algoritmo de K medias.

Los clusters deben cumplir las siguientes condiciones:

1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$. In other words, each observation belongs to at least one of the K clusters.

2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. In other words, the clusters are non-overlapping: no observation belongs to more than one cluster.

Clustering

K-means asume que una buena asignación es aquella que dado un valor de K minimiza lo más posible la variabilidad intra cluster (*within-cluster variation*).

Si $W(C_j)$ es una medida que indica cuánto las observaciones de un cluster j difieren entre ellas, el problema se puede escribir como:

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

Si uno usa la distancia euclídea como medida de disimilaridad $W(C_j)$ puede escribirse de la siguiente manera:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

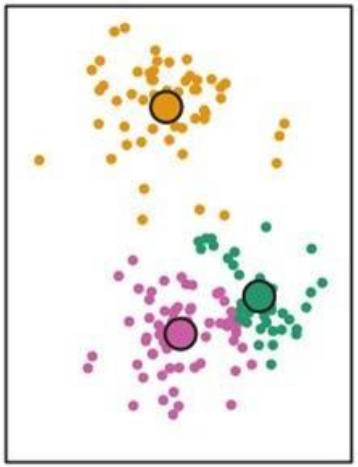
Clustering

De esta forma el problema se puede reescribir como:

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

En donde $W(C_j)$ se puede expresar como la distancia a un centroide:

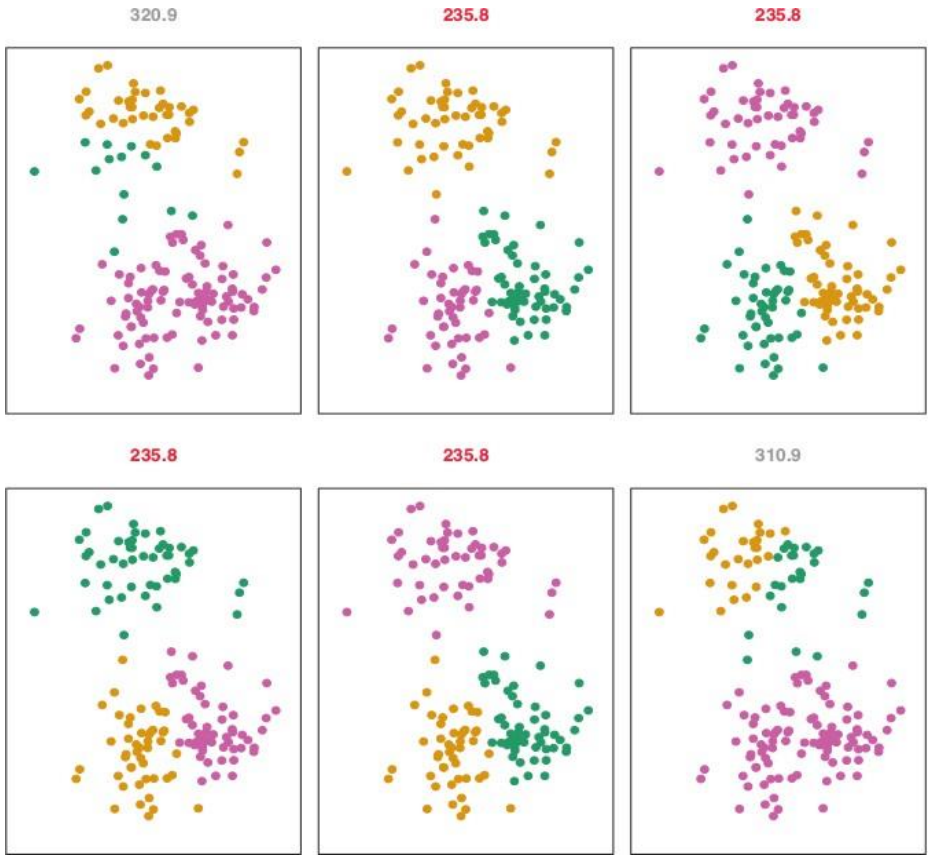
$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$



Clustering

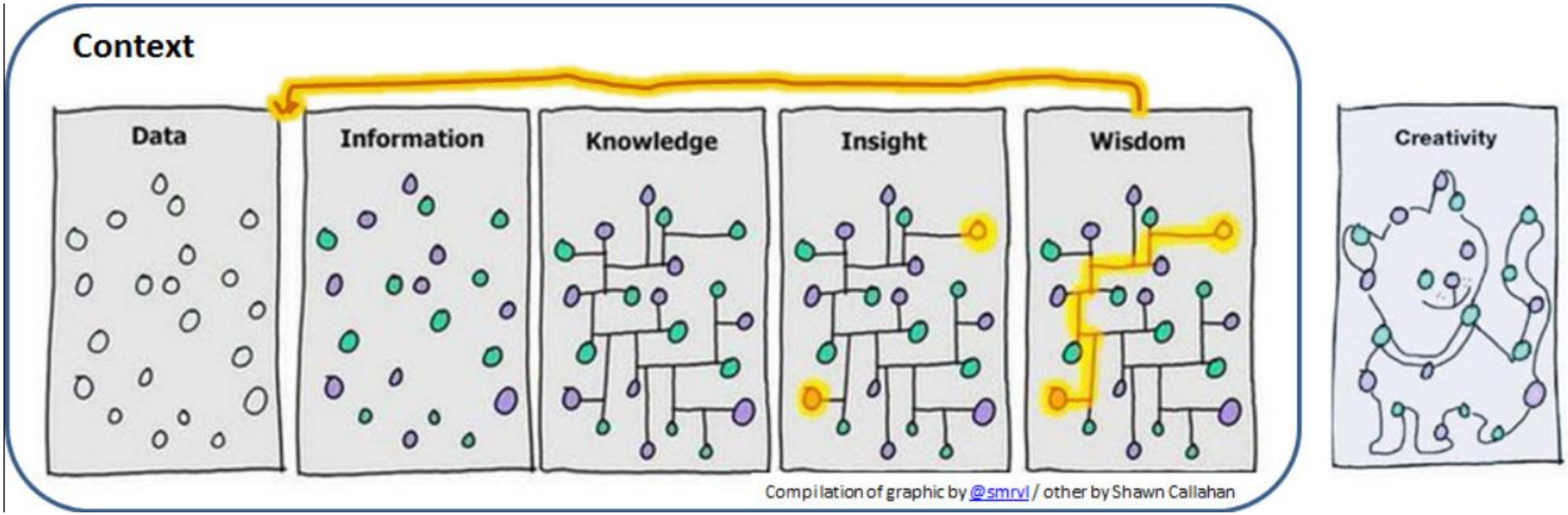
La solución obtenida por K-means depende en gran medida de los valores iniciales de asignación a clusters.

Por este motivo **se suele correr el algoritmo muchas veces y quedarse con la mejor solución.**



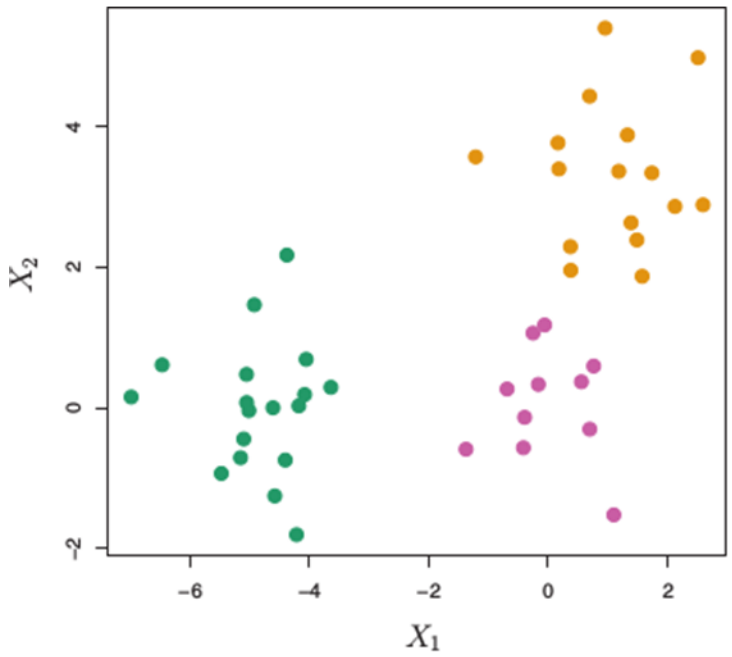
Clustering

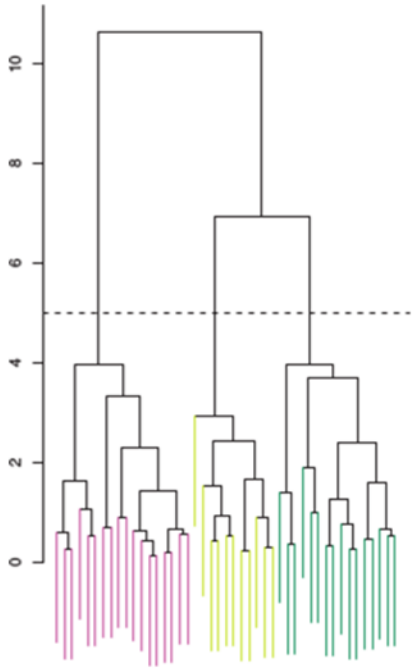
Los algoritmos jerárquicos requieren más tiempo...



Clustering Jerárquico (1)

Es un método bottom up o aglomerativo. A diferencia de K-means, uno no se debe comprometer a un número de clusters antes de ejecutar el algoritmo.
El objetivo va a ser obtener un **dendrograma**.

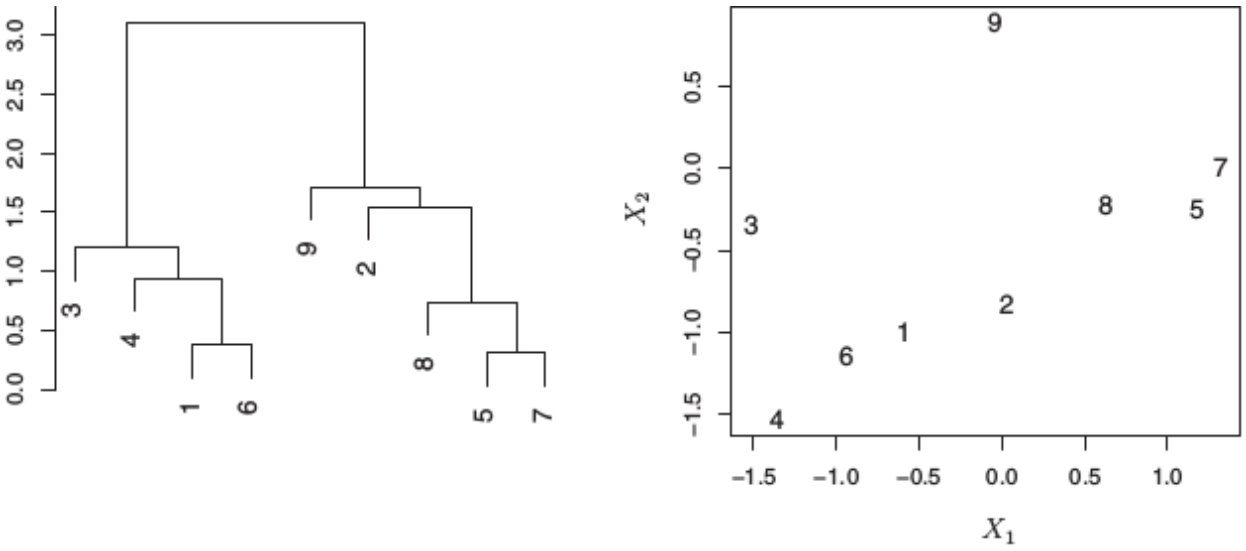




Clustering Jerárquico (2)

Dendrograma:
Cada hoja representa una observación.
Las hojas se unen un grupos similares en entre sí.
A medida que uno sube en el dendograma los grupos se unen entre sí.
Mientras más "bajo" se haga una unión, más cercanos son los elementos.

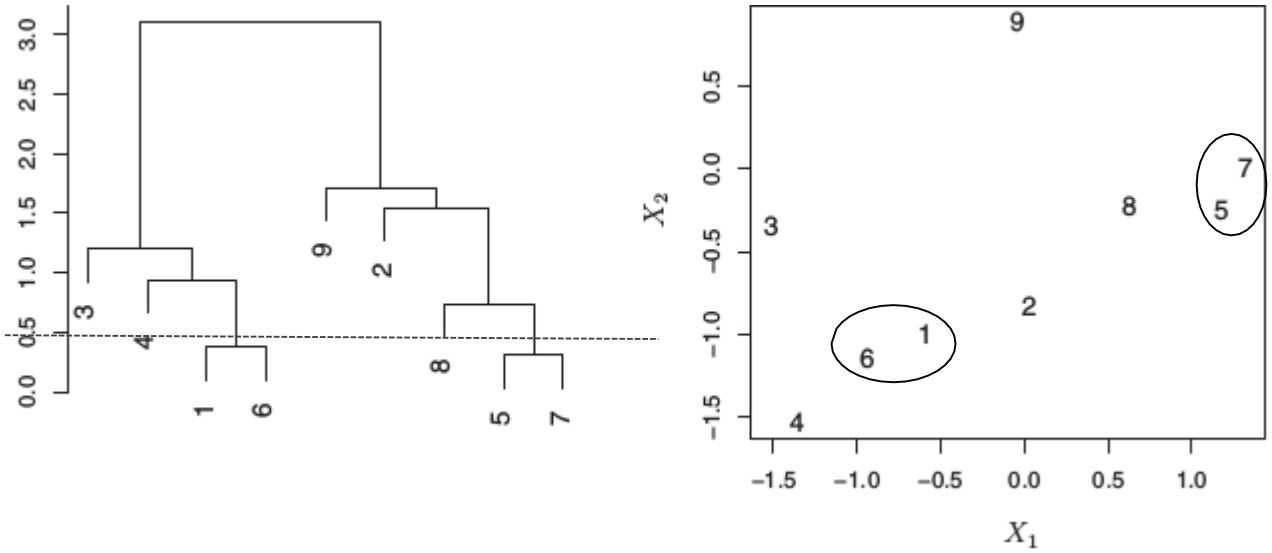
Veamos el siguiente ejemplo...



Clustering Jerárquico (4)

Dendrograma:
Primera aproximación.

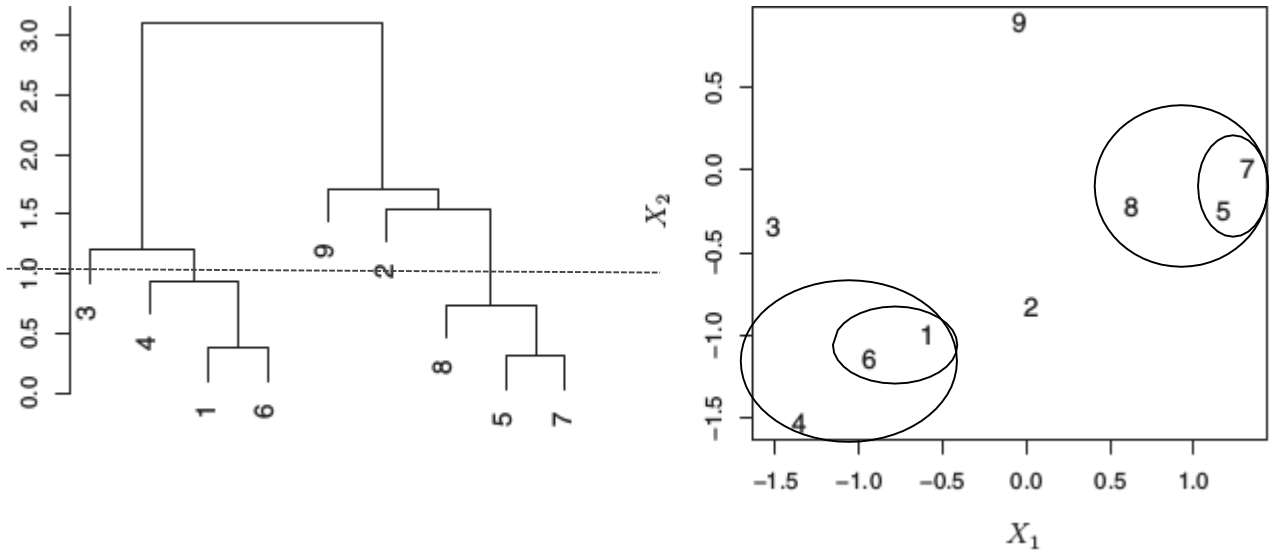
La identificación por cercanía comienza en las observaciones 5 y 7
También en las observaciones 6 y 1 pero claramente entre ambos grupos hay distancia significactiva.



Clustering Jerárquico (6)

Dendrograma:
Segunda aproximación.

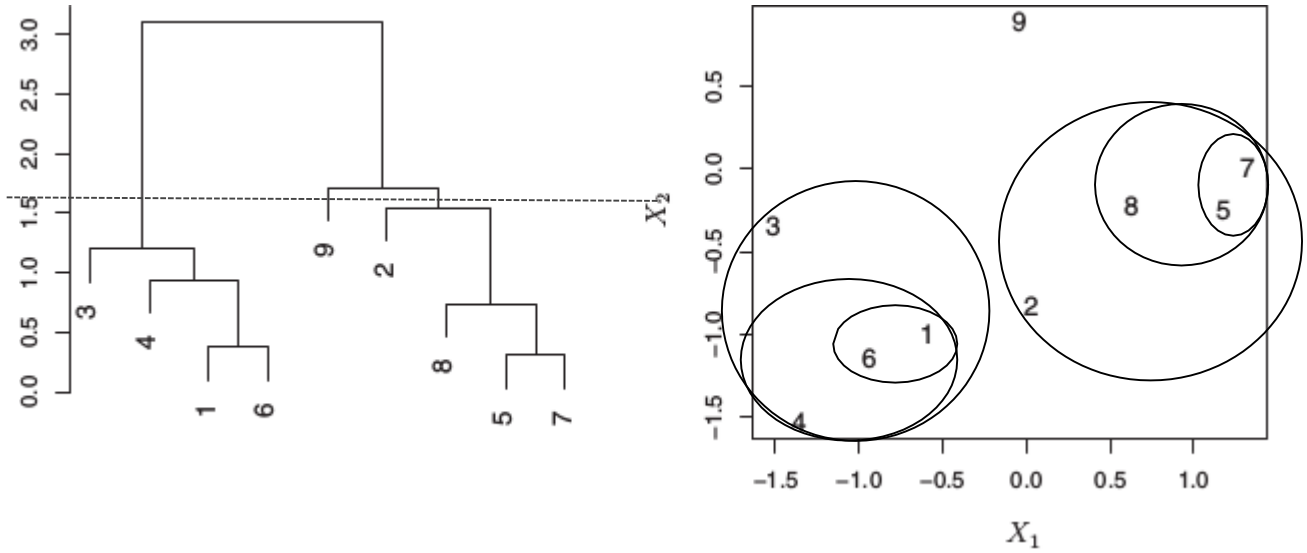
Luego, la cercanía de los grupos a otras observaciones cercanas.



Clustering Jerárquico (8)

Dendrograma:
Tercera aproximación.

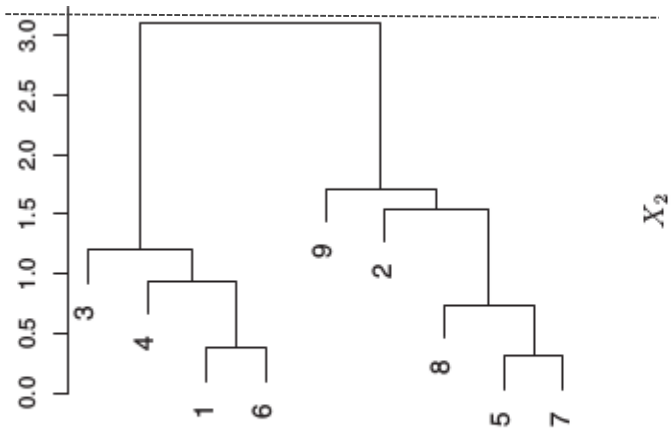
Luego, observen que la 3 cae en el grupo {6,1}. Mientras que la 2 la identificamos con el grupo {7,5}

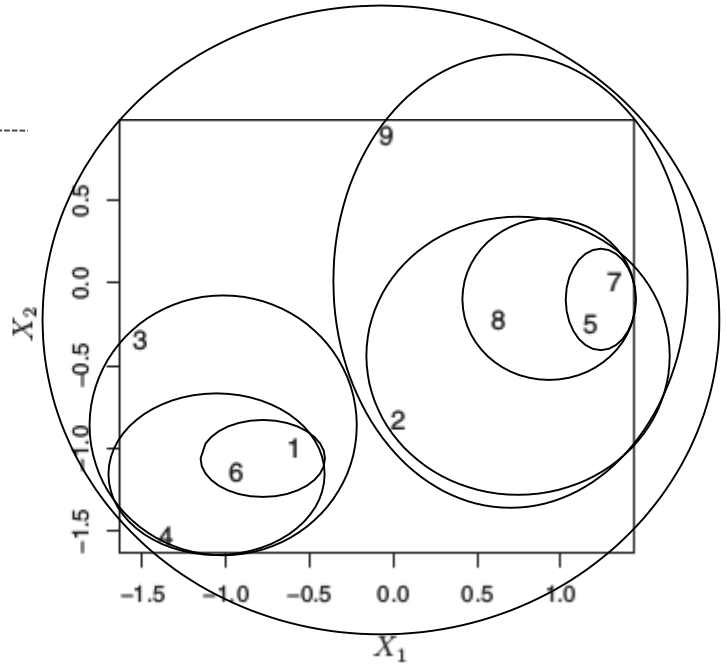


Clustering Jerárquico (10)

Dendrograma:
Cuarta aproximación.

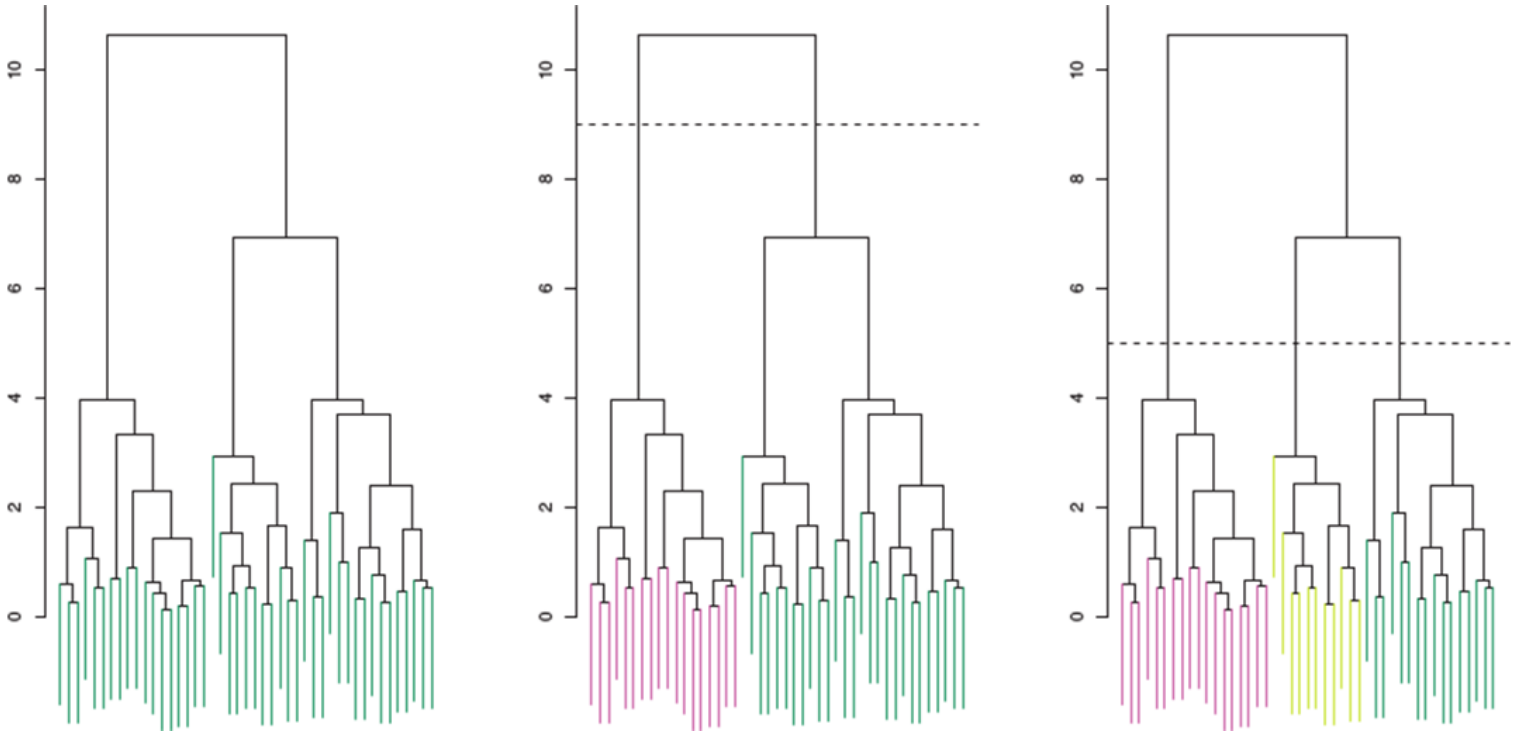
Finalmente, observen que la 9 cae en el grupo {7,5}.





Clustering Jerárquico (11)

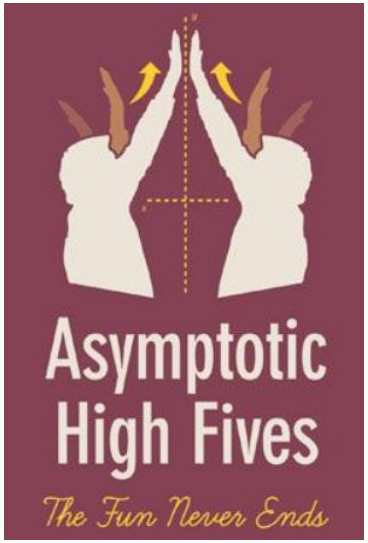
Los clusters se encuentran haciendo cortes a distintos niveles del dendrograma.



Clustering Jerárquico (16)

Consideraciones a tener en cuenta al hacer clustering:

- Pequeñas decisiones pueden tener grandes consecuencias (en la práctica se prueban muchas opciones y se analiza la robustez de los resultados).
- No existe un consenso referido a cómo validar clusters encontrados.
- ¿Necesariamente una observación debe pertenecer 100% a un cluster?
- Las particiones pueden ser poco estables al quitar un pequeño subconjunto de observaciones.



Análisis de Componentes Principales

PCA (1)

Es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información.

Supongase que existe una muestra con n individuos cada uno con p variables $(X_1, X_2, \dots, X_p)$, es decir, el espacio muestral tiene p dimensiones.

PCA permite encontrar un número de factores subyacentes $(z < p)$ que explican aproximadamente lo mismo que las p variables originales.

Donde antes se necesitaban p valores para caracterizar a cada individuo, ahora bastan z valores.

Cada una de estas z nuevas variables recibe el nombre de componente principal.

PCA (2)

Pertenece a la familia de técnicas conocida como *APRENDIZAJE NO SUPERVISADO*. Los métodos supervisados tienen el objetivo de predecir una variable respuesta *Y* a partir de una serie de predictores. Para ello, se dispone de *p* características (*X1, X2,... Xp*) y de la variable respuesta *Y* medidas en *n* observaciones.

En el caso de *los no supervisados*, la variable respuesta *Y* no se tiene en cuenta ya que el objetivo no es predecir *Y* sino extraer información empleando los predictores, por ejemplo, para identificar subgrupos.

El principal problema al que se enfrentan los métodos de *estos métodos* es la dificultad para validar los resultados dado que no se dispone de una variable respuesta que permita contrastarlos.

El método de PCA permite por lo tanto "condensar" la información aportada por múltiples variables en solo unas pocas componentes.

PCA (3)

Cada componente principal Z_i se obtiene por combinación lineal de las variables originales. Se pueden entender como nuevas variables obtenidas al combinar de una determinada forma las variables originales.

La primera componente principal de un grupo de variables $(X_1, X_2, \dots, X_p)$ es la combinación lineal normalizada de dichas variables que tiene mayor varianza,

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

Que la combinación lineal sea normalizada implica que,

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

Los términos $\phi_{11}, \dots, \phi_{p1}$ reciben en el nombre de *loadings* y son los que definen a la componente. ϕ_{11} es el *loading* de la variable X_1 de la primera componente principal.

PCA (4)

Los ***loadings*** pueden interpretarse como el peso/importancia que tiene cada variable en cada componente y, por lo tanto, ayudan a conocer que tipo de información recoge cada una de las componentes.

Dado un set de datos X con n observaciones y p variables, el proceso a seguir para calcular la primera componente principal es:

- **Centralización de las variables**: se resta a cada valor la media de la variable a la que pertenece. Con esto se consigue que todas las variables tengan media cero.
- **Se resuelve un problema de optimización para encontrar el valor de los *loadings* con los que se maximiza la varianza**. Una forma de resolver esta optimización es mediante el cálculo de eigenvector-eigenvalue de la matriz de covarianzas.

PCA (5)

Una vez calculada la primera componente (Z_1) se calcula la segunda $(\boldsymbol{Z_2})$ **repitiendo el mismo proceso, pero añadiendo la condición de que la combinación lineal no pude estar correlacionada con la primera componente.**

Esto equivale a decir que $\boldsymbol{Z_1}$ **y** $\boldsymbol{Z_2}$ **tienen que ser perpendiculares.**

El proceso se repite de forma iterativa hasta calcular todas las posibles componentes $(\min(n-1, p))$ o hasta que se decida detener el proceso.

El orden de importancia de las componentes viene dado por la magnitud del eigenvalue asociado a cada eigenvector.

Escalado de Variables

El proceso identifica aquellas direcciones en las que la varianza es mayor.

Como la varianza de una variable se mide en su misma escala elevada al cuadrado, si antes de calcular las componentes no se estandarizan todas las variables para que tengan media 0 y desviación estándar 1, aquellas variables cuya escala sea mayor dominarán al resto.

De ahí que sea **recomendable estandarizar siempre los datos**.

Reproducibilidad de las componentes

El proceso genera siempre las mismas componentes principales independientemente del software utilizado, es decir, el valor de los *loadings* resultantes es el mismo.

La única diferencia que puede darse es que el signo de todos los *loadings* esté invertido. **Esto es así porque el vector de *loadings* determina la dirección de la componente, y dicha dirección es la misma independientemente del signo** (la componente sigue una línea que se extiende en ambas direcciones).

Del mismo modo, el valor específico de las componentes obtenido para cada observación (*Principal Component Scores*) es siempre el mismo, a excepción del signo.

Influencia de Outliers

Al trabajar con varianzas, el método PCA es altamente sensible a *outliers*, por lo que es altamente recomendable estudiar si los hay.

La detección de valores atípicos con respecto a una determinada dimensión es algo relativamente sencillo de hacer mediante comprobaciones gráficas.

Sin embargo, cuando se trata con múltiples dimensiones el proceso se complica. Por ejemplo, considérese un hombre que mide 2 metros y pesa 50 kg. Ninguno de los dos valores es atípico de forma individual, pero en conjunto se trataría de un caso muy excepcional.

La distancia de Mahalanobis es una medida de distancia entre un punto y la media que se ajusta en función de la correlación entre dimensiones y que permite encontrar potenciales *outliers* en distribuciones multivariante.

Proporción de la Varianza Explicada

¿Cuánta información presente en el set de datos original se pierde al proyectar las observaciones en un espacio de menor dimensión?

¿Cuanta información es capaz de capturar cada una de las componentes principales obtenidas?

Para contestar a estas preguntas se recurre a la proporción de varianza explicada por cada componente principal.
Asumiendo que las variables se han normalizado para tener media cero, la varianza total presente en el set de datos se define como,

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{x_{ij}^2}{n}$$

Proporción de la Varianza Explicada

Asumiendo que las variables se han normalizado para tener media cero, la varianza total presente en el set de datos se define como,

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{x_{ij}^2}{n}$$

y la varianza explicada por la componente m es,

$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2$$

Por lo tanto, **la proporción de varianza explicada por la componente m viene dada por el ratio,**

$$\frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} \frac{x_{ij}^2}{n}}$$

PCA (10)

Tanto la **proporción de varianza explicada** como la **proporción de varianza explicada acumulada** son dos valores de gran utilidad a la hora de decidir el número de componentes principales a utilizar en los análisis posteriores.

El sumatorio de la proporción de varianza explicada acumulada de todas las componentes es siempre 1.

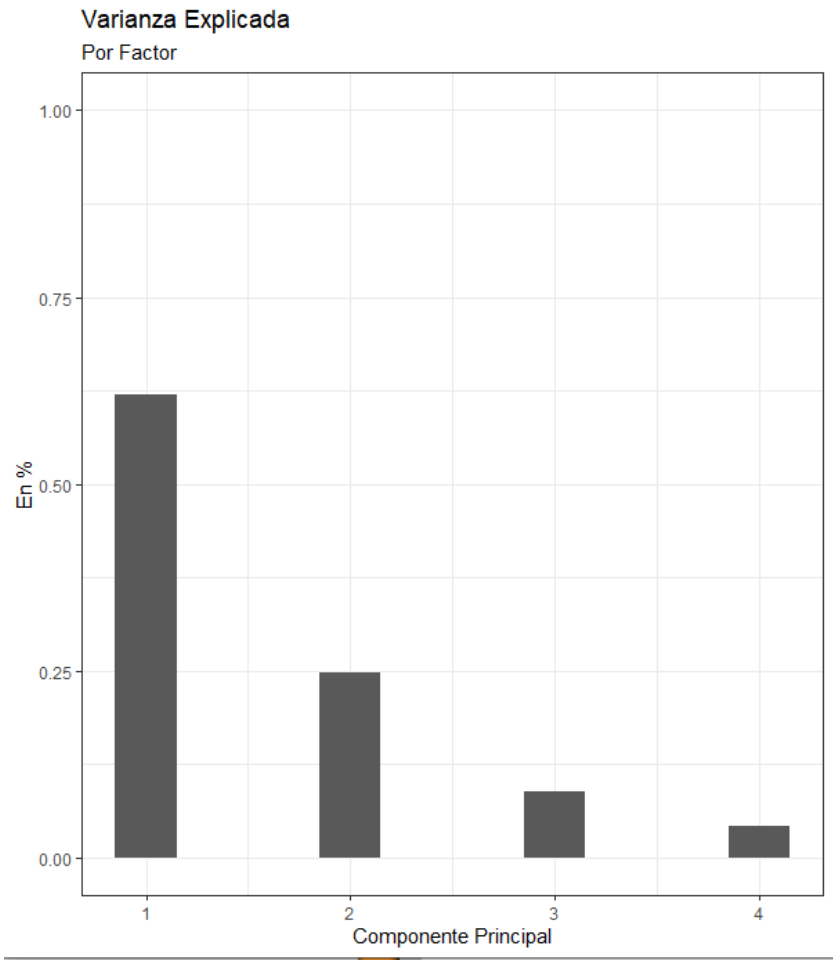
Número óptimo de componentes principales

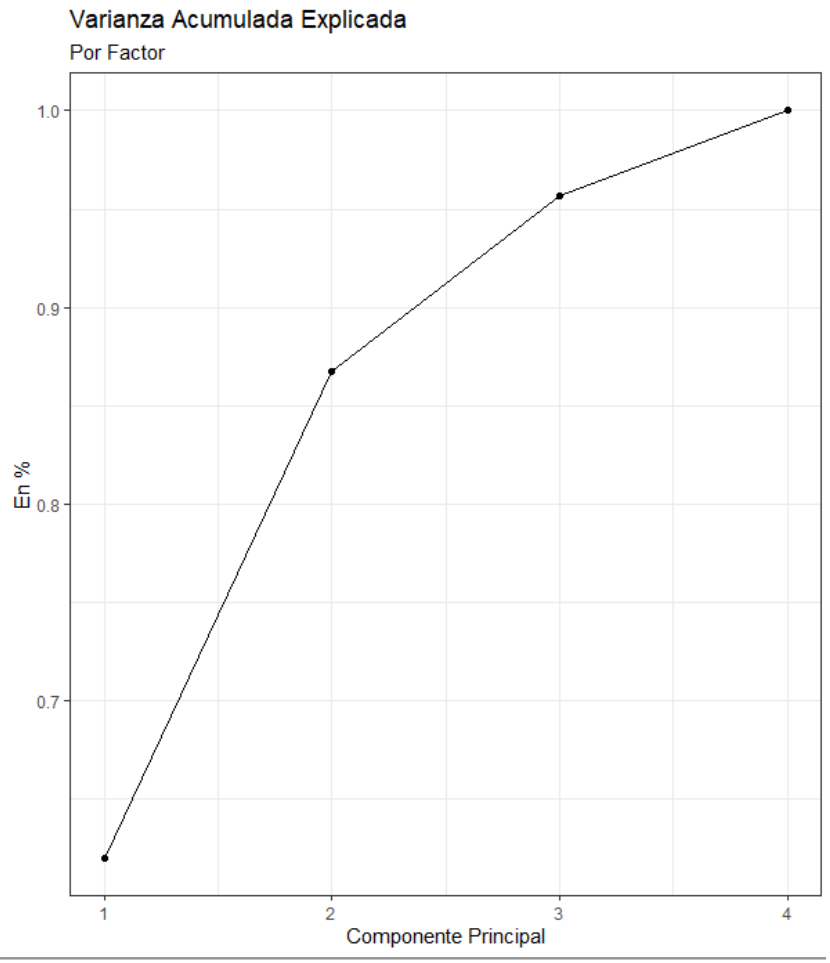
Por lo general, **dada una matriz de datos de dimensiones n x p, el número de componentes principales que se pueden calcular es como máximo de n-1 o p (el menor de los dos valores es el limitante).**

Sin embargo, siendo el objetivo del PCA reducir la dimensionalidad, suelen ser de interés utilizar el número mínimo de componentes que resultan suficientes para explicar los datos.

No existe una respuesta o método único que permita identificar cual es el número óptimo de componentes principales a utilizar.

PCA (11)


Varianza Explicada
Por Factor
1.00
0.75
0.50
0.25
0.00
En %
1
2
3
4
Componente Principal


Varianza Acumulada Explicada
Por Factor
1.0
0.9
0.8
0.7
En %
1
2
3
4
Componente Principal

Análisis Discriminante

Análisis Discriminante (1)

El Análisis Discriminante Lineal o *Linear Discriminant Analysis* (LDA) es un método de clasificación SUPERVISADO de variables cualitativas. Dos o más grupos son conocidos a priori y nuevas observaciones se clasifican en uno de ellos en función de sus características.

Mediante la utilización del **Teorema de Bayes**, LDA estima la probabilidad de que una observación, dado un determinado valor de los predictores, pertenezca a cada una de las clases de la variable cualitativa,

$$P(Y = k/X = x)$$

Finalmente **se asigna la observación a la clase k para la que la probabilidad predicha es mayor**.

Análisis Discriminante (2)

Es una **alternativa a la regresión logística** cuando la variable cualitativa tiene más de dos niveles.

Si bien existen extensiones de la regresión logística para múltiples clases, el LDA presenta una serie de **ventajas**:

- **Si las clases están bien separadas, los parámetros estimados en el modelo de regresión logística son inestables. El método de LDA no sufre este problema.**
- **Si el número de observaciones es bajo y la distribución de los predictores es aproximadamente normal en cada una de las clases, LDA es más estable que la regresión logística.**

Cuando se trata de un problema de clasificación con solo dos niveles, ambos métodos suelen llegar a resultados similares.

Análisis Discriminante (3)

El proceso de un análisis discriminante puede resumirse en 6 pasos:

1. Disponer de un conjunto de **datos de entrenamiento** (training data) en el que se conoce a que grupo pertenece cada observación.
2. Calcular las **probabilidades previas** (*prior probabilities*): la proporción esperada de observaciones que pertenecen a cada grupo.
3. **Determinar si la varianza o matriz de covarianzas es homogénea en todos los grupos**. De esto dependerá que se emplee LDA o QDA (cuadrática).
4. **Estimar los parámetros** necesarios para las funciones de probabilidad condicional, verificando que se cumplen las condiciones para hacerlo.
5. **Calcular el resultado de la función discriminante**. El resultado de esta determina a qué grupo se asigna cada observación.
6. Utilizar validación cruzada (***cross-validation)*** para estimar las probabilidades de clasificaciones erróneas.

Análisis Discriminante (4)

Las condiciones que se deben cumplir para que un Análisis Discriminante Lineal sea válido son:

- **Cada predictor que forma parte del modelo se distribuye de forma normal** en cada una de las clases de la variable respuesta.
- En el caso de múltiples predictores, las observaciones siguen una distribución normal multivariante en todas las clases.
- **La varianza del predictor es igual en todas las clases de la variable respuesta**. En el caso de múltiples predictores, la matriz de covarianza es igual en todas las clases.

Si esto no se cumple se recurre a Análisis Discriminante Cuadrático (QDA).

Cuando la condición de normalidad no se cumple, el LDA pierde precisión pero aun así puede llegar a clasificaciones relativamente buenas.

Análisis Discriminante (5)

Una vez que las normas de clasificación se han establecido, se tiene que evaluar la clasificación resultante y su eficacia.

En otras palabras, **evaluar el porcentaje de aciertos en las clasificaciones**. Las matrices de confusión son una de las mejores formas de evaluar la capacidad de acierto que tiene un modelo LDA. Muestran el número de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

El método LDA busca los límites de decisión que más se aproximan al clasificador de Bayes, que por definición, tiene el menor ratio de error total de entre todos los clasificadores (si se cumple la condición de normalidad).

Por lo tanto, el LDA intenta conseguir el menor número de clasificaciones erróneas posibles, pero no diferencia entre falsos positivos o falsos negativos.

Análisis Discriminante (6)

Si se quiere intentar reducir el número de errores de clasificación en una dirección determinada (por ejemplo, menos falsos negativos) se puede modificar el límite de decisión, aunque como consecuencia aumentará el número de falsos positivos.

Cuando para evaluar el error de clasificación se emplean las mismas observaciones con las que se ha creado el modelo, se obtiene lo que se denomina el training error.

Si bien esta es una forma sencilla de estimar la precisión en la clasificación, tiende a ser excesivamente optimista.

Es más adecuado evaluar el modelo empleando observaciones nuevas que el modelo no ha visto, obteniendo así el test error.

Algunos videos para mirar

PCA:
https://www.youtube.com/watch?v=6BeuHCo1gZQ
https://www.youtube.com/watch?v=oiR3k9H-7K0

LDA:
https://www.youtube.com/watch?v=hX6llK2mA3c

Clustering:
https://www.youtube.com/watch?v=w_aUCJHRv0Y&t=6s
https://www.youtube.com/watch?v=tkAJT8gWBSY

Mirada al futuro…

Charla con el CECE sobre Técnicas de Text Mining (junto a Rodrigo Del Rosso):

https://youtu.be/OyA2I9AoUpc

Síganme para más recetas

Material disponible en:

https://github.com/rodrigodelrosso/CECE-Nuevo-Espacio-Text-Mining

¿Dudas?

“Lo importante es no dejar de hacerse preguntas”

Albert Einstein

martinmasci@economicas.uba.ar