PCA y FA

Feature Engineering

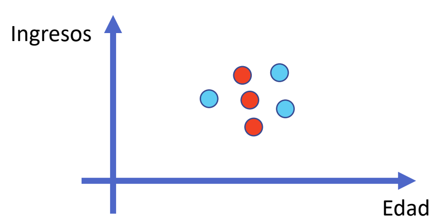
Dr. Víctor de la Cueva

1

Dimensiones

- En un set de datos a las variables (features) también se les conoce como dimensiones.
- De hecho, lo son, ya que cada punto (dato) está formado por N valores, correspondientes a las N variables, donde cada uno de estos valores se puede localizar en un eje cartesiano, lo que nos permite colocar el punto en un espacio N-dimensional:

2D



Edad	Ingresos	Tipo de Cel
25	27000	Android
30	32000	iPhone
…	…	…

2

Dimensionality Reduction

- En algunas tareas que se realizan con datos (análisis estadístico, machine learning, etc.) tener muchas variables causa algunos problemas:
  - Tiempo de cómputo
  - Overfitting
  - Espacio en memoria
- Por lo que es recomendable reducir el número de variables (dimensiones) al menor número posible sin perder información en forma significativa.
- Las técnicas que nos permiten hacerlo son conocidas como *Dimensionality Reduction.*

3

Reducción de dimensionalidad

- Es una rama de la Ingeniería de Características que se encarga de reducir el número de variables (dimensiones) sin perder (demasiada) información.
- Normalmente, la reducción se hace creando nuevas variables a partir de las ya existentes, de tal forma que conserven la información original proporcionada (al menos la mayoría).
- Hay un gran número de técnicas para esto, pero las dos más importantes son:
  - Principal Component Analysis (PCA)
  - Factor Analysis (FA)

4

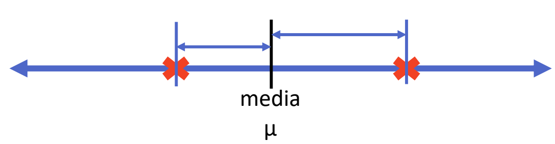
Datos categóricos

- PCA y FA sólo se aplican a datos numéricos (no categóricos).
- Se tiene que separar de las variables categóricas.
- Si después se considera que algunas de las variables categóricas son significativas o inclusive son nuestra variable a predecir (clasificación), se pueden volver a incluir en el conjunto de datos final.

5

Medición de la información

- Para poder seleccionar las mejores variables es necesario poder medir la cantidad de información que contienen.
- Una forma común de hacerlo es por medio de la varianza.
- La varianza es la medición de qué tanto se alejan los datos alrededor de la media.
- Se calcula como el promedio de los cuadrados de las desviaciones de cada dato con respecto a la media.
  - Se calcula en una sola dimensión, por lo que se tienen que proyectar.

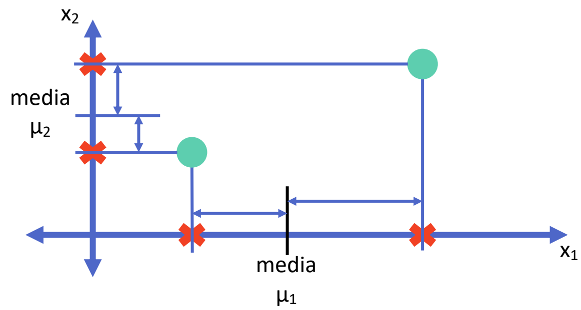


$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

6

Varianza en dos ejes

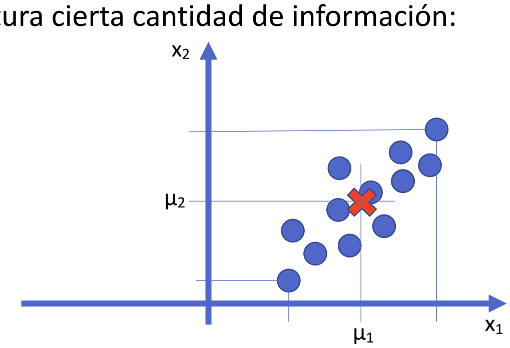
- Desde luego que se puede calcular la varianza para más de una dimensión, simplemente haciendo la proyección de los puntos en cada una de ellas:



7

Cada eje captura cierta información

- Cada eje captura cierta cantidad de información:



- La información total la define la suma de las varianzas capturadas por cada eje.
- Si borro alguna variable, pierdo información.

8

Principal Component Analysis PCA

9

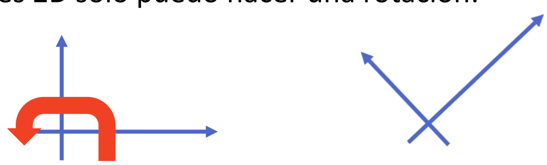
Análisis de componentes principales

- La técnica PCA de reducción de dimensionalidad, crea **k** nuevas variables (es decir, nuevos ejes), llamadas componentes, a partir de las **k** variables originales.
- Estos nuevos ejes se encuentran simplemente rotando el sistema de ejes originales hasta que uno de ellos, llamado componente principal 1 (PC1), logra maximizar la varianza explicada de los datos.
  - Los ejes siempre deben permanecer perpendiculares entre ellos.
- A continuación, el eje PC1 se queda fijo y se procede los siguientes con respecto a PC1, hasta lograr la siguiente (PC2) maximización de la varianza que queda, y así sucesivamente.
- Se hacen k-1 rotaciones, al final, el último eje queda ajustado automáticamente.
  - Porque se debe mantener la perpendicularidad entre ellos.

10

Encontrando los nuevos ejes

- Para encontrar los componentes se sigue un procedimiento matemático.
- Si lo hiciéramos físicamente, lo único que tenemos que hacer el rotarlos, garantizando que siempre se mantienen ortogonales (perpendiculares).
  - El número máximo de rotaciones que se realizan son n-1, donde n es el número de original de ejes, que es igual al número original del variables.
  - Por ejemplo, es 2D sólo puedo hacer una rotación:



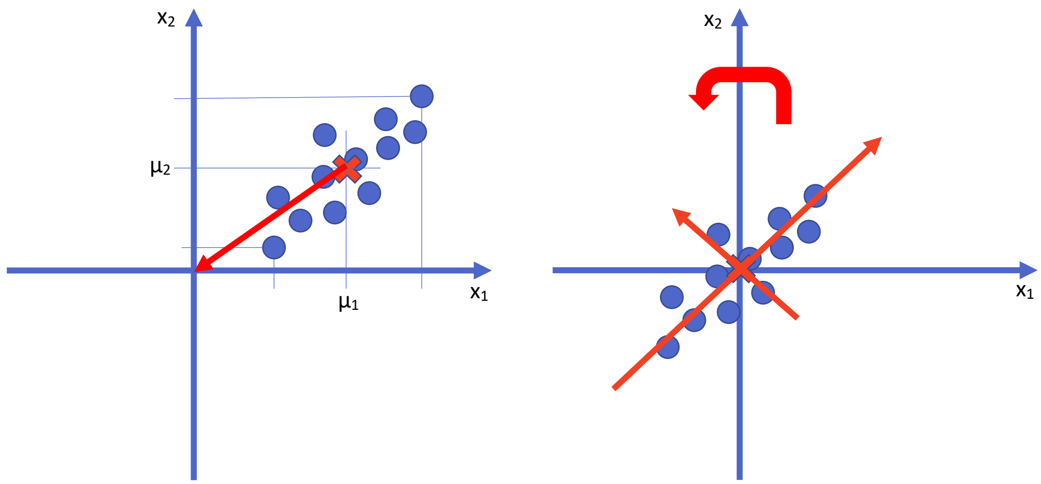
11

Datos centrados o escalados

- Como la varianza de los datos es muy sensible a la magnitud de los datos, para evitar que algún dato tenga más varianza por efecto de su magnitud, el primer paso, antes de calcular los PC, debe ser centrar los datos.
  - Esto se logra restando a cada dato la media, en cada una de las dimensiones originales.
- Algunas personas, hacen una estandarización de los datos para que su media sea cero y su varianza 1 en todas las dimensiones.
- Los outliers, como ya se explicó, también afectan mucho a la varianza y se deben tratar antes de iniciar con el proceso PCA.

12

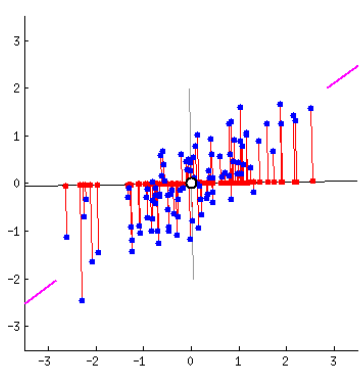
Nuevos ejes



13

El eje buscado

- El eje buscado es una recta que siempre pasa por (0,0), porque no tiene término independiente a_0.



https://builtin.com/data-science/step-step-explanation-principal-component-analysis

14

Los componentes

- Los componentes son los nuevos ejes.
- Los componentes encontrados deben tener la característica de ser independientes, lo cual se logra haciéndolos ortogonales (perpenduculares).
- Esto implica que, en realidad, lo que se está haciendo es rotar un sistema de ejes de m dimensiones, de tal forma que todos maximicen la varianza de los puntos proyectados sobre ellos.
- Primero se calcula el primer componente que cumpla con la condición especificada, luego el segundo, recordando que tiene que ser perpendicular al anterior, y así sucesivamente.
- El número máximo de componenetes que se pueden tener es el mínimo(n-1, m), donde n es el número de datos y m el número de variables.
  - En un data set el menor es casi siempre es m

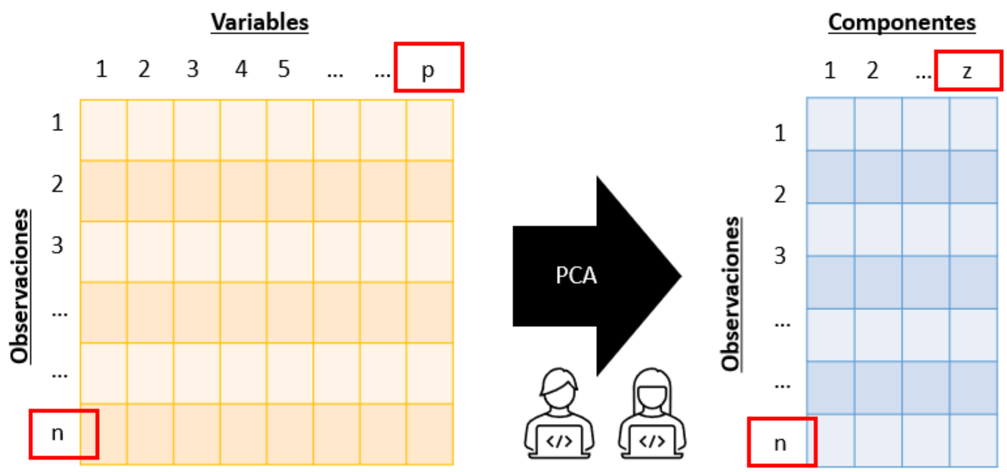
15

La varianza explicada

- Cada componente explica la varianza que capturó de los datos proyectados sobre él, de tal forma que, entre todos, explican el 100% de la variabilidad de los datos sobre cada eje.
- La varianza total (100%) capturada (explicada) es la suma de las varianzas de los datos sobre cada eje.
  - Entonces, cada componente explica un porcentaje de la varianza total.
- Los componentes se ordenan de mayor a menor de acuerdo a su varianza explicada.
- Finalmente, se seleccionan los componenetes cuya suma de variaza, es decir, la varianza explicada entre todos, pase un cierto límite (es común 70 u 80%, pero depende del problema).

16

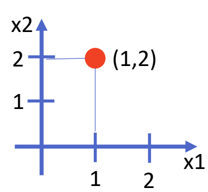
Gráficamente

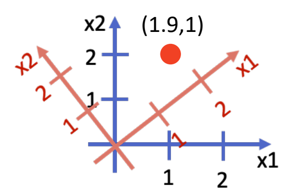


17

Nuevas coordenadas

- Los nuevos ejes ocasionan que los datos originales deban cambiar sus coordenadas, ahora con respecto a los nuevos ejes:





- Las nuevas coordenadas se obtienen proyectando el punto original sobre los nuevos ejes.

18

Los nuevos datos

- Los nuevos datos (mismos puntos con nuevas coordenadas) se obtienen simplemente proyectando los datos (puntos) originales sobre los nuevos ejes.
- Esto se logra con la simple multiplicación: $(CX^T)^T$.
  - X es (n x m): es la tabla de datos originales, de n datos y m variables.
  - C es (m x m): es la matriz de componentes
- $(CX^T)^T$ es (m x m)*(m x n) = (m x n) y al trasponerla da (n x m).

19

PCA en Python

- Todo lo anterior se puede hacer fácilmente en Python (R o Matlab también).
- Desde luego, hay una librería (`sklearn`) que nos ahorra todo eso y lo reduce a cinco simples instrucciones:

```
# 1. cargar la librería
from sklearn.decomposition import PCA
# 2. definir el número de componentes. Al inicio son el número de variables numéricas originales
pca = PCA(n_components=2)
# 3. obtener los componentes (fit) y aplicar a los datos (trasnform)
# fit obtiene los nuevos ejes (componentes) y transform las nuevas coordenas
# tambien se pueden aplicar por separado, primero pca.fit(df) y luego pca.transform(df)
principalComponents = pca.fit_transform(df)
# 4. cambiar de nombre a las columnas para mayor claridad nuestra
principalDf = pd.DataFrame(data = principalComponents, columns = ['PC1', 'PC2'])
# 5. imprimir las varianzas explicadas
print("var explicada: ",pca.explained_variance_ratio_)
# se seleccionan los componentes que sumen el límite de varianza que se desea explicar (~80%)
# estos (nuevas variables) sustituyen a las variables numéricas originales
# si se desea se pueden ver los datos con sus nuevas coordenadas
principalComponents
```

- Vamos a un ejemplo.

20

Ejemplo

- Un ejemplo pequeño para mostrar los conceptos:
  - x1 = [2.5, 0.5, 2.2, 1.9, 3.1, 2.3, 2.0, 1.0, 1.5, 1.1]
  - x2 = [2.4, 0.7, 2.9, 2.2, 3.0, 2.7, 1.6, 1.1, 1.6, 0.9]
- Hacer el PCA de la base de datos de los Iris de las plantas.

21

Factor Analysis (FA)

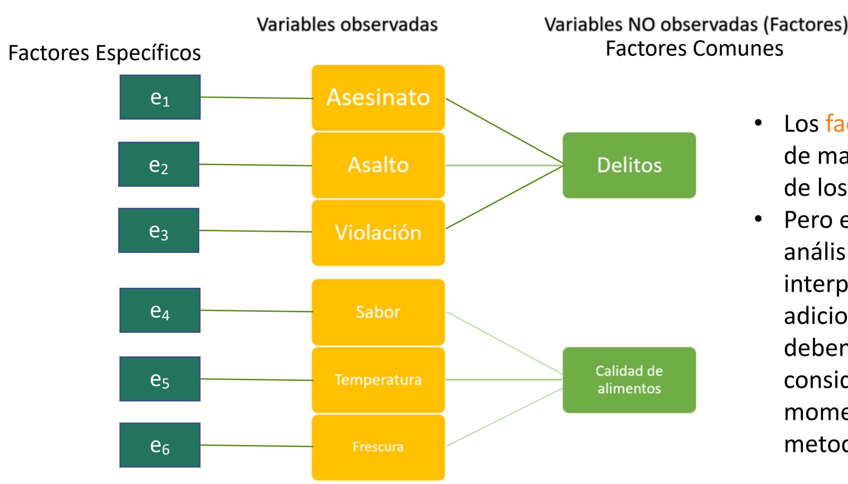
22

Factor

- Los factores son variables latentes, variables ocultas o variables no observadas, las cuales no se pueden medir.
- Al igual que los componentes, los factores son nuevos ejes, que se obtienen de forma diferente a la de los componentes, pero buscan el mismo objetivo.
- No todas las variables dependen de todos los factores.
  - Algunos factores tienen más peso que otros en ciertas variables

23

En forma gráfica



- Los factores se obtienen de manera similar a las de los PCA.
- Pero en el FA hay análisis, instrucciones e interpretaciones adicionales que se deben tomar en consideración al momento de usar esta metodología.

24

Puntos a considerar

- Crear como máximo m-1 factores, donde m es el número de variables.
  - Para establecer el número máximo de factores a formar se deben analizar los eigenvalores y aquellos que tengan un valor cercano a uno o superior nos ayudarán a determinar el número de factores a utilizar.
- Para saber si el FA es un buen enfoque para ser aplicado a mis variables, se debe realizar un análisis adicional:
  - La prueba de Barlett

25

La prueba de Barlett

- La prueba de Barlett, es una prueba de hipótesis para determinar si una matriz de correlación es significativa o no.
- En caso de no ser significativa, quiere decir que contamos con una matriz identidad (unos en la diagonal principal de la matriz y ceros en los demás campos).
  - Esto nos dice que no va a haber una relación oculta dentro de las variables, que nos permita agruparlas en nuevos factores.
- Al correr la prueba nosotros buscamos un valor de $p < 0.05$ (95% es el nivel de confianza estándar).
  - Si el valor p cumple esta condición, quiere decir que la matriz es significativa, esto es, que hay relación entre algunas variables y se puede proceder a la construcción de factores.

26

Recomendaciones al aplicar FA

- Es importante tomar en consideración algunas recomendaciones para que el modelo arroje resultados confiables:
  - Para poder aplicar FA es muy importante no contar con valores atípicos
  - El contar con datos atípicos en este tipo de análisis y transformación podría llevarnos a tener diferentes resultados con información sesgada
  - Los valores atípicos influyen en el cálculo directo de la varianza de un conjunto de datos, por lo tanto, su detección y manejo son fundamentales tanto para PCA como FA.
  - Tener un tamaño de muestra mayor al número de factores, es decir, contar con suficientes registros para que el modelo pueda determinar de forma confiable si efectivamente hay una relación subyacente entre cierta combinación de variables.

27

Nombres de los factores

- Una vez que la prueba de Barlett nos dijo que sí se puede hacer FA, se procede a realizarlo usando los eigenvalores que nos dirán cuántos factores obtener.
- El problema es cuando, una vez identificadas las características de los factores, se les trata de dar una interpretación para poderles asignar un nombre.
- Por ejemplo, si se lleva el registro de las calificaciones de un alumno en ciertas materias, los factores pueden ser:
  - Inteligencia
  - Dedicación

28

Pasos para aplicar FA

- Obtener media y desviación estándar de las variables
  - Para ver si es necesario escalarlas (si hay mucha variación en valores), si no, simplemente se centran (si su media está muy lejos de cero).
- Centrar y/o escalar los datos.
  - Tratar los valores atípicos.
- Hacer la prueba de Barlett.
  - Continuar su $p < 0.05$
- Obtener los eigenvalores de los datos escalados.
  - Quedarse con todos los que tengan un valor cercano o mayor a 1.
- Obtener los factores que se decidió encontrar.
- Los valores más grandes y parecidos en las columnas, son los que definen los factores.
  - Analizar las variables que involucra y ponerles nombre.
- Aplicar los factores a los datos para obtener los nuevos datos.
  - Pegarle las categóricas.
  - Guardarlos en un archivo.

29

Ejemplo

- Obtener los factores de la base de datos relacionada con los Iris de las flores.

30

Referencias

[1] S. Marsland. Machine Learning: An Algorithm Perspective. 2nd ed, Chapman & Hall/CRC (2015).

[2] Mathematical Approach to PCA. Geeks for Geeks. (https://www.geeksforgeeks.org/mathematical-approach-to-pca/). Consultado el 25-oct-2022.

[3] A Step-by-Step Explanation of Principal Component Analysis (PCA) (https://builtin.com/data-science/step-step-explanation-principal-component-analysis). Z. Jaadi. Consultado el 25-oct-2022.