

Práctica 2.

Feature Selection

*Máster en Big Data*

*Curso 2019-2020*

***Data Science & Advanced Analytics***

*Montse Llos i Bombardó*

*Feature Selection*

**Es el proceso de reducir el número de atributos de entrada para:**

* **Reducir uso de memoria**
* **Reducir tiempo computacional**
* **Reducir nombre de variables**

**¿Cómo? Existen diversas soluciones**

**Correlaciones**

**Hacer una matriz de correlación de todas las variables vs todas para eliminar aquellas que tengan una alta correlación entre ellas (una ya representa al conjunto). Ejemplo: Edad y Año de Nacimiento**

**Forward Selection**

**Uso de un modelo que va aumentando el número de variables input hasta que el rendimiento del modelo baje**

**Se usa un modelo de bajo coste computacional**

**Backward Selection**

**Uso de un modelo que parte del total de atributos y los va disminuyendo hasta que el rendimiento baja**

**Se usa un modelo de bajo coste computacional**

**PCA – Principal Component Analysis**

**Generación de nuevos atributos que son un compendio de los atributos input.**

**Cada nuevo atributo está compuesto de un conjunto de atributos input que no están correlacionados entre ellos**

**Normalmente en PCA se acostumbra a quedarse con los nuevos atributos creados y que representan el 95% de la varianza del resultado**

**La varianza representa el cuadrado de la desviación estandar**

**Hay dos matrices de salida de PCA interesantes a estudiar:**

* **Eigenvalues**
  + **Nos da como interfieren los nuevos atributos en la varianza de la salida**
* **Eigenvectors**
  + **Nos da el peso de cada atributo a los nuevos atributos que se están creando**

**Problemas PCA**

* **Como se fija en la varianza el ruido afecta mucho (son los que tienen una mayor variabilidad)**
* **Es importante tener en cuenta la desviación estándar. Los que tienen más proporción de varianza sobre la salida son los que tienen una desviación estándar mayor. Por tanto, es importante normalizar antes los datos de entrada para tener una variabilidad sobre los mismos márgenes. Puede ser que igualmente así no descartemos ningún atributo.**

### 

*Práctica 2.1*

La empresa de e-commerce VamosBien quiere mejorar aún más la detección de abandono de clientes con lo cual nos pasa más información ligada al cliente como el importe total que ha transaccionado con ellos, la media, número de transacciones y cómo han sido pagados con los distintos tipos de forma de pago que ofrecen: cash, tarjeta o cheque.

Para llevar a cabo la práctica, debemos bajarnos del campo virtual la práctica 2, abrirla en RapidMiner y seguir los pasos que se nos van especificando.