DATA SCIENCE

Federico Baiocco
baioccofede@gmail.com
3512075440

Clase 13 - Agenda

MODULO 2: FEATURE ENGINEERING

- Feature engineering
- Sklearn
- Imputación de valores nulos
- Práctica con sklearn

¿ Dudas de la clase pasada ?

Feature engineering

En esta etapa los datos son analizados y refinados para que estén listos para la preparación o entrenamiento de un modelo de machine learning.

Feature engineering

En la ciencia de datos, los datos son nuestra "materia prima" y como toda materia prima, los datos no suelen llegarnos "listos" para utilizar. Tenemos variables de distintos tipos, valores faltantes, outliers, etc.

Al final de esta etapa, nuestros datos deben quedar "limpios" y listos para entrenar nuestro modelo de machine learning o hacer el análisis que necesitemos.


scikit
learn

SCIKIT learn

Scikit learn es otra de las librerías más utilizadas en el proceso de data science.

Muy buena documentación: https://scikit-learn.org/stable/user_guide.html

6. Dataset transformations

- ▶ 6.1. Pipelines and composite estimators
- ▶ 6.2. Feature extraction
- ▶ 6.3. Preprocessing data
- ▶ 6.4. Imputation of missing values
- ▶ 6.5. Unsupervised dimensionality reduction
- ▶ 6.6. Random Projection
- ▶ 6.7. Kernel Approximation

SCIKIT learn

Podemos usarla para:

- Preprocesamiento de datos
- Modelos de aprendizaje supervisado
- Modelos de aprendizaje no supervisado
- Tunning de hiper-parámetros de modelos
- Métricas de Evaluación

Todo esto vamos a ir viendo ...

SCIKIT learn

¿ Cómo se usa ?

La clase pasada vimos programación orientada objetos. En Scikit learn, hay muchas clases definidas a partir de las cuales podemos instanciar objetos. Cada una de estas clases esta documentada y sirve para un propósito específico.

Los principales objetos son:
- Estimadores (estimators) → Tienen un método .fit
- Predictores (predictors) → Tienen un método .predict
- Transformadores (transformers) → Tienen un método .transform
- Modelos (model) → Tienen un método .score

SCIKIT learn

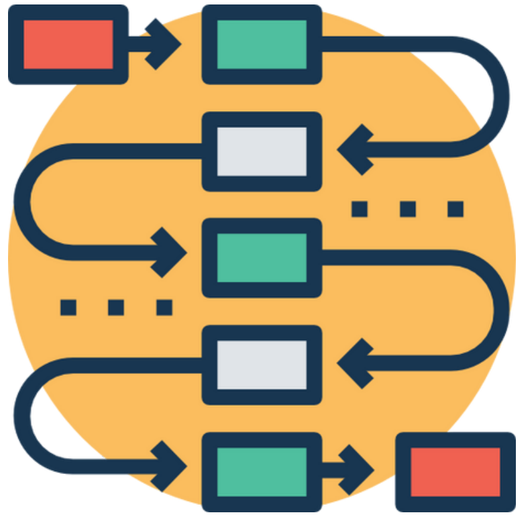
En general:

1- Se crea un objeto con ciertos parámetros.
2- Se implementa un método .fit que aprende de los datos
3- Se implementa un método .transform que transforma los datos o un .predict que predice sobre los datos

1- objeto = ClaseDeSklearn()

2- objeto.fit(datos)

3- objeto.transform(datos)



Imputers

¿ Qué hacemos con los valores faltantes ?

Imputers

Vimos que es muy común tener valores faltantes en un dataset. Estos valores faltantes pueden estar representados por espacios en blanco, NaN, o algún otro valor.
Cuando queremos entrenar un modelo de machine learning, no podemos pasarle valores nulos, por lo tanto, tenemos que decidir que hacer con los mismos.
- Imputers?
- Descartar filas?
- Descartar la columna entera?

Al descartar sin criterio podemos perder datos que podrían haber sido muy útiles, por lo tanto siempre hay que analizar bien cada problema.

¿ Cuándo descartarían una fila ?
¿ Cuándo una columna ?

Imputers
Univariante/multivariante

- Univariante: Miramos de a una variable (columna) a la vez. Los datos faltantes de esa columna se rellenan en base a los valores no faltantes que hay en la misma columna. Por ej:
    - Completamos los faltantes con la media/mediana de los no faltantes de la columna

- Multivariante: No nos fijamos de a una columna a la vez, sino que miramos todas las columnas y tratamos de buscar relaciones entre las mismas para decidir con que valor rellenar los faltantes. Por ejemplo:
    - Supongamos que tenemos un dataset con datos de departamentos. Una columna nos dice la superficie total y la otra la superficie cubierta. Si tenemos datos faltantes en la columna superficie cubierta, no podemos completar con la media de esta columna ya que tenemos la restricción: superficie cubierta <= superficie total. En este caso, podríamos por ejemplo completar el dato con el valor de la columna superficie total

Imputers
Vecinos más cercanos

Scikit learn también nos provee un imputer llamado KNNImputer (k nearest neighbors imputer). Como su nombre lo indica, lo que hace es medir la distancia que hay entre cada una de las filas del dataset, cuando encuentra un valor faltante, se fija en los valores que tienen las filas más cercanas para el dato faltante y en base a eso le asigna un valor.
Para asignar el valor puede utilizar la media de los vecinos mas cercanos, o calcular una media ponderada en base a qué tan lejos está cada vecino.

En este imputer, debemos especificar un valor para "K" (el imputer va a fijarse en los K vecinos más cercanos).

Abrimos notebook feature eng

Preguntas?

Para la próxima clase:

- Terminar de completar los nulos del dataset!