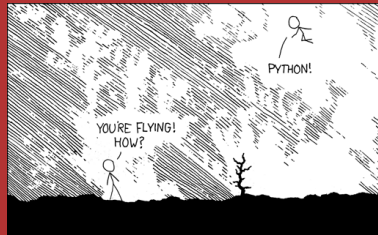


Curso - Python orientado a Machine Learning - Parte 3: Introd. a ML



Erick Castillo / @ecastillo (slack IV Devs)

Enero - 2018

Indice

- 1 Juegos previos
- 2 Machine Learning, teoría
 - A very (very) brief introduction
 - Tipos de aprendizaje
 - Algoritmos
 - Machine Learning, Minería de Datos y KDD
 - Terminología
- 3 Machine Learning, práctica
 - scikit-learn

Juegos previos

- lápices
- puntos en plano cartesiano
- playtennis



Tipos de aprendizaje

- supervisado: clasificación, regresión
- no supervisado: agrupamiento
- refuerzo

Algoritmos

Algunos algoritmos populares:

- regresión lineal
- regresión logística
- kNN
- naive bayes
- SVM
- árboles de decisión
- random forest
- K-means
- redes neuronales
- algoritmos genéticos



Machine Learning, Minería de Datos y KDD

- preprocesamiento de los datos
 - OLTP y OLAP
 - Datamart
 - ETL
- procesamiento de los datos
 - la idea básica es entrenar el modelo con una porción de los datos (training dataset) y luego validarlo con el resto de los datos (test dataset).
- post-procesamiento de los datos
 - idealmente, los patrones descubiertos deben tener 3 cualidades:
 - 1. ser precisos
 - 2. comprensibles (inteligibles)
 - 3. interesantes (útiles y novedosos)

scikit-learn

- "Simple and efficient tools for data mining and data analysis"
- Python 2 y 3
- muy bien documentada
- dependencias: numpy, scipy y matplotlib
- última versión estable a la fecha: 0.19.1
- sitio web oficial: <http://scikit-learn.org>

¿Qué trae?

- datasets
 - cada columna es un atributo
 - cada fila es una instancia
- métodos de preprocesamiento
- modelos de procesamiento
- métodos de postprocesamiento

Esquema de trabajo

Requerimientos para trabajar con datos en scikit-learn:

- features & response son objetos separados (ej: `iris.data` & `iris.target`)
- features & response deben ser numéricos
- features & response deben ser arrays Numpy (`<class 'numpy.ndarray'>`)
- features & response deben tener dimensiones específicas (ej: `iris.data.shape = (150, 4)`, `iris.target.shape = (150,)`)

Convenciones

- X : matriz de características
- y : vector de respuesta
- X_{train} , X_{test} , y_{train} , y_{test}