DATA SCIENCE

Leandro Ferrado, Javier Lezama, Valentina Rubiolo

ACÁMICA

14 de Febrero de 2019

Plan de estudio

4	proyectos reales
6	meses presenciales
2	clases por semana
3	horas por clase
14	horas semanales de trabajo en casa

Tématicas que vamos a aprender

- Machine Learning.

Tématicas que vamos a aprender

- Machine Learning.
- NLP, (Siri, Alexia, etc).

Tématicas que vamos a aprender

- Machine Learning.
- NLP, (Siri, Alexia, etc).
- Clustering.

Tématicas que vamos a aprender

- Machine Learning.
- NLP, (Siri, Alexia, etc).
- Clustering.
- Deep Learning.

Tématicas que vamos a aprender

- Machine Learning.
- NLP, (Siri, Alexia, etc).
- Clustering.
- Deep Learning.
- Cloud A.i, (Watson de IBM).

Tématicas que vamos a aprender

- Machine Learning.
- NLP, (Siri, Alexia, etc).
- Clustering.
- Deep Learning.
- Cloud A.i, (Watson de IBM).
- Deployment.

Proyectos

PROYECTO 1

EXPLORACIÓN DE DATOS

- Intro: ¿Qué es un dataset?, tipos de datos, tipos de problemas.
- Introducción Data Science: workflow.
- Presentación Jupyter Notebooks y bibliotecas NumPy, Pandas, Scikit-Learn
- Breve introducción NumPy: tipos de datos y operaciones.
- Exploración de datos: Pandas.

FEATURE ENGINEERING

- Feature engineering
- Outliers.
- Missings values.
- Variables categóricas, dummies, nuevas variables.

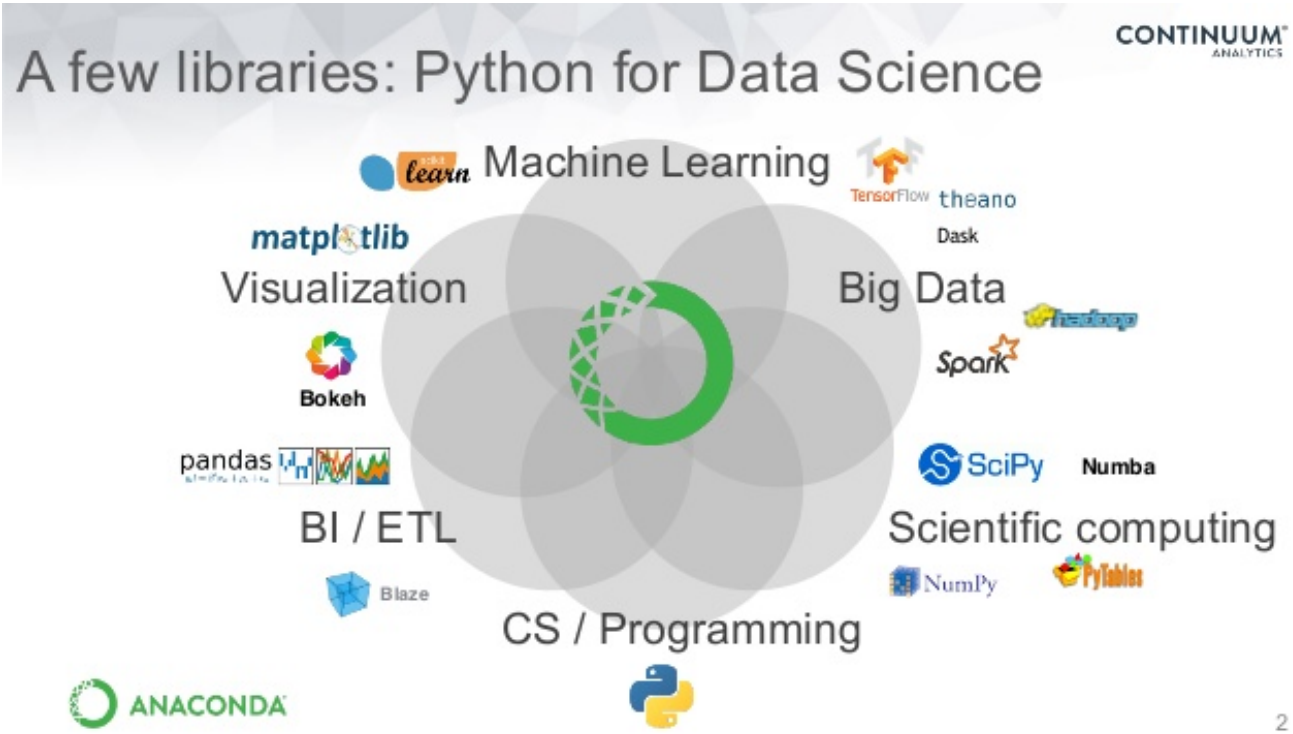
REGRESIÓN

- Evaluación de modelos: training/testing, matriz de confusión.
- Algoritmo KNN.
- Algoritmo "Decision Trees".
- Clasificación y regresión con estos algoritmos.
- Overfitting/Underfitting
- Cross Validation
- Pipelines en scikit-learn

OPTIMIZACIÓN DE MODELOS

- Tradeoff bias/variance.
- Optimización de parámetros: GridSearch y hyperopt
- Selección stepwise de variables: fordward, backward.
- Otras métricas: AUC, F1, kappa.

Algunas librerías de python



Experiencias previas

- Lenguaje de programación: Python, R, Visualbasic, Java, C/C++, R, MatLab/Octave, Envi, Fortran, etc.

Experiencias previas

- Lenguaje de programación: Python, R, Visualbasic, Java, C/C++, R, MatLab/Octave, Envi, Fortran, etc.
- Estadistíca: Probabilidad, variables independientes/dependientes, Descriptores de distribuciones, histogramas, procesos de Markov, etc.

Instalación de entorno

- Python 3.6 y librerías compatibles.

Instalación de entorno

- Python 3.6 y librerías compatibles.
- Cuestiones de Setup y buenas prácticas.

Manos a la obra.

- Operaciones básicas en Python.

Manos a la obra.

- Operaciones básicas en Python.
- https://github.com/leferrad/acamica-ds-cor

Explicación

- Python.

Explicación

- Python.
- Jupyter.

Explicación

- Python.
- Jupyter.
- NumPy.

Explicación

- Python.
- Jupyter.
- NumPy.
- Matplotlib.

Explicación

- Python.
- Jupyter.
- NumPy.
- Matplotlib.
- Pandas.

Python

- Python es un lenguaje de programación multiparadigma.

Python

- Python es un lenguaje de programación multiparadigma.
- Apareció en 1991.

Python

- Python es un lenguaje de programación multiparadigma.
- Apareció en 1991.
- Amigable.

Python

- Python es un lenguaje de programación multiparadigma.
- Apareció en 1991.
- Amigable.
- Principales utilización: Big Data, Data Mining, Machine Learning, NLP, etc.

Jupyter

- Entorno de trabajo que soporta Python, entre otros lenguajes.

Jupyter

- Entorno de trabajo que soporta Python, entre otros lenguajes.
- Surge en 2014 como una evolución de iPython.

Jupyter

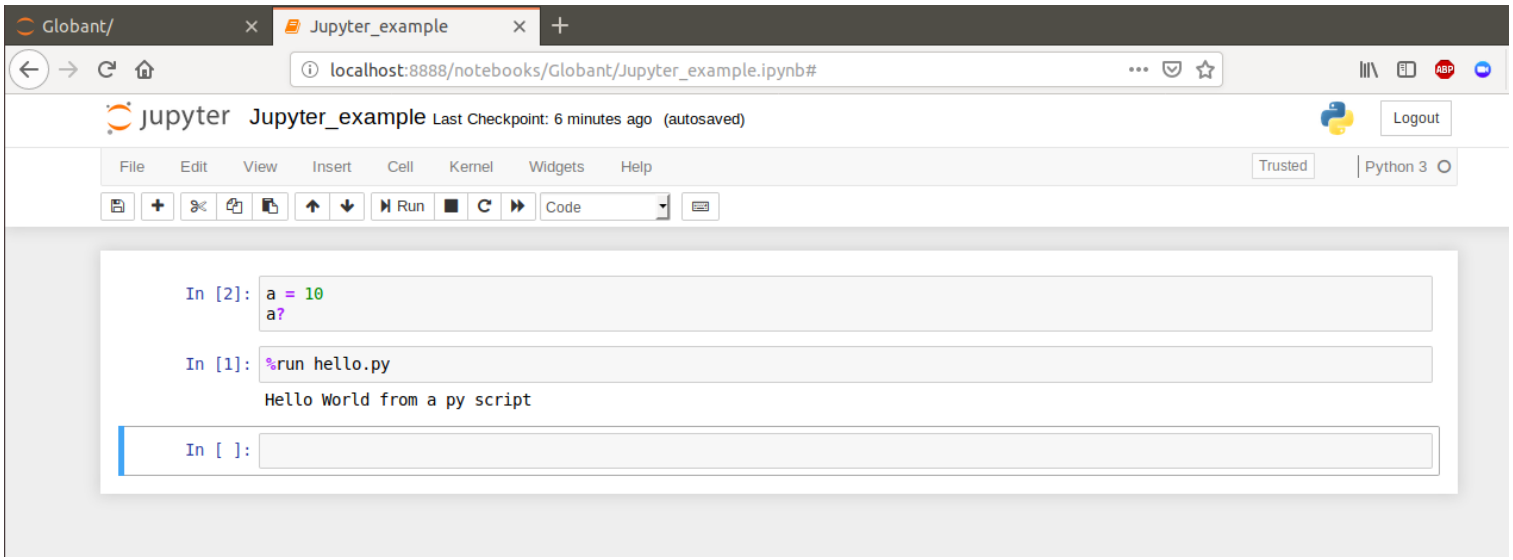
- Entorno de trabajo que soporta Python, entre otros lenguajes.
- Surge en 2014 como una evolución de iPython.
- Principales objetivos: fomentar y simplificar la compartición de conocimientos y resultados a través de las notebook.

Jupyter

- Entorno de trabajo que soporta Python, entre otros lenguajes.
- Surge en 2014 como una evolución de iPython.
- Principales objetivos: fomentar y simplificar la compartición de conocimientos y resultados a través de las notebook.

Jupyter

- Entorno de trabajo que soporta Python, entre otros lenguajes.
- Surge en 2014 como una evolución de iPython.
- Principales objetivos: fomentar y simplificar la compartición de conocimientos y resultados a través de las notebook.



NumPy

- A powerful N-dimensional array object.

NumPy

- A powerful N-dimensional array object.
- Vectorización.

NumPy

- A powerful N-dimensional array object.
- Vectorización.
- Sophisticated (broadcasting) functions.

NumPy

- A powerful N-dimensional array object.
- Vectorización.
- Sophisticated (broadcasting) functions.
- Tools for integrating C/C++ and Fortran code.

NumPy

- A powerful N-dimensional array object.
- Vectorización.
- Sophisticated (broadcasting) functions.
- Tools for integrating C/C++ and Fortran code.
- Useful linear algebra, Fourier transform, and random number capabilities.

Matplotlib

- Es una libreria de Python de dibujos 2D.

Matplotlib

- Es una libreria de Python de dibujos 2D.
- Funciones.

Matplotlib

- Es una libreria de Python de dibujos 2D.
- Funciones.
- Histogramas.

Matplotlib

- Es una libreria de Python de dibujos 2D.
- Funciones.
- Histogramas.
- Gráficas de barras.

Matplotlib

- Es una libreria de Python de dibujos 2D.
- Funciones.
- Histogramas.
- Gráficas de barras.
- Gráficos de dispersión, etc.

Pandas

- Python Data Analisis Library.

Pandas

- Python Data Analisis Library.
- Herramienta de manipulación de datos de alto nivel.

Pandas

- Python Data Analisis Library.
- Herramienta de manipulación de datos de alto nivel.
- Estructura de datos clave: DataFrame.

Pandas

- Python Data Analisis Library.
- Herramienta de manipulación de datos de alto nivel.
- Estructura de datos clave: DataFrame.
- DataFrame: permite almacenar y manipular datos tabulados en filas de observaciones y columnas de variables.

Pandas

- Python Data Analisis Library.
- Herramienta de manipulación de datos de alto nivel.
- Estructura de datos clave: DataFrame.
- DataFrame: permite almacenar y manipular datos tabulados en filas de observaciones y columnas de variables.
- https://pandas.pydata.org/pandas-docs/stable/getting$_s$*tarted*/10*min*.*html*.

¿Dudas? ¿Comentarios?