Python Supervised Learning with scikit-learn

Integrated Master's in Informatics Engineering

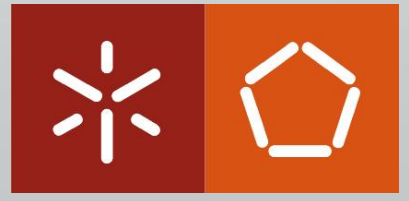
Learning and Extraction of Knowledge

2018/2019

Synthetic Intelligence Lab

Filipe Gonçalves

César Analide



What is machine learning?

- The art *and* science of:
  - Giving computers the ability to learn to make decisions from data
  - … without being explicitly programmed!
- Examples:
  - Learning to predict whether an email is spam or not
  - Clustering wikipedia entries into different categories

Reinforcement learning

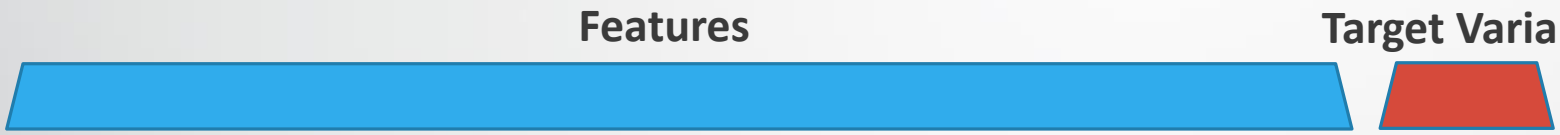
- Software agents interact with an environment
  - Learn how to optimize their behavior
  - Given a system of rewards and punishments
  - Draws inspiration from behavioral psychology
- Applications
  - Economics
  - Genetics
  - Game playing
- AlphaGo: First computer to defeat the world champion in Go

Supervised/Unsupervised learning

- Supervised learning: Uses labeled data
  - Predictor variables/features and a target variable
  - Aim: Predict the target variable, given the predictor variables
    - Classification: Target variable consists of categories
    - Regression: Target variable is continuous
- Unsupervised learning: Uses unlabeled data
  - Uncovering hidden patterns from unlabeled data
  - Example:
    - Grouping customers into distinct categories (Clustering)

Naming conventions

- Features = predictor variables = independent variables
- Target variable = response variable = dependent variable
- Example:



	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

Supervised Learning

- Automate time-consuming or expensive manual tasks
  - Example: Doctor's diagnosis
- Make predictions about the future
  - Example: Will a customer click on an ad or not?
- Need labeled data
  - Historical data with labels
  - Experiments to get labeled data
  - Crowd-sourcing labeled data

Exploratory data analysis

- **The Iris dataset**
  - Features:
    - Petal length
    - Petal width
    - Sepal length
    - Sepal width
  - Target variable: Species
    - Versicolor
    - Virginica
    - Setosa

The Iris dataset in scikit-learn

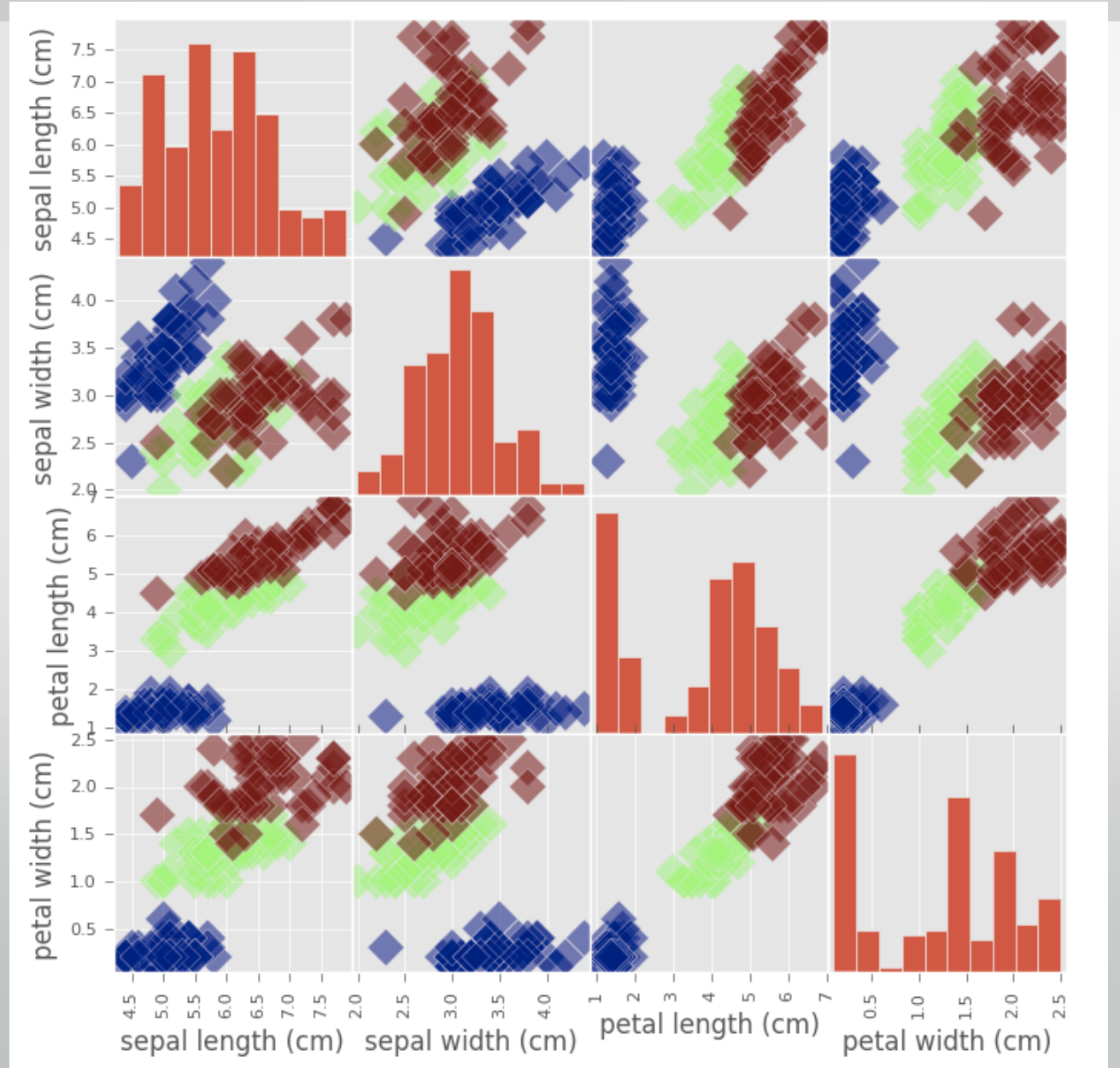
```
In [1]: from sklearn import datasets

In [2]: import pandas as pd

In [3]: import numpy as np

In [4]: import matplotlib.pyplot as plt

In [5]: plt.style.use('ggplot')

In [6]: iris = datasets.load_iris()

In [7]: type(iris)
Out[7]: sklearn.datasets.base.Bunch

In [8]: print(iris.keys())
dict_keys(['data', 'target_names', 'DESCR', 'feature_names', 'target'])
In [9]: type(iris.data), type(iris.target)
Out[9]: (numpy.ndarray, numpy.ndarray)

In [10]: iris.data.shape
Out[10]: (150, 4)

In [11]: iris.target_names
Out[11]: array(['setosa', 'versicolor', 'virginica'], dtype='<U10')
```

Exploratory data analysis

```
In [12]: X = iris.data

In [13]: y = iris.target

In [14]: df = pd.DataFrame(X, columns=iris.feature_names)

In [15]: print(df.head())
   sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)
0                5.1               3.5                1.4               0.2
1                4.9               3.0                1.4               0.2
2                4.7               3.2                1.3               0.2
3                4.6               3.1                1.5               0.2
4                5.0               3.6                1.4               0.2
```

```
In [16]: _ = pd.scatter_matrix(df, c = y, figsize = [8, 8],
    ...:                                s=150, marker = 'D')
```

Visual Exploratory Data Analysis

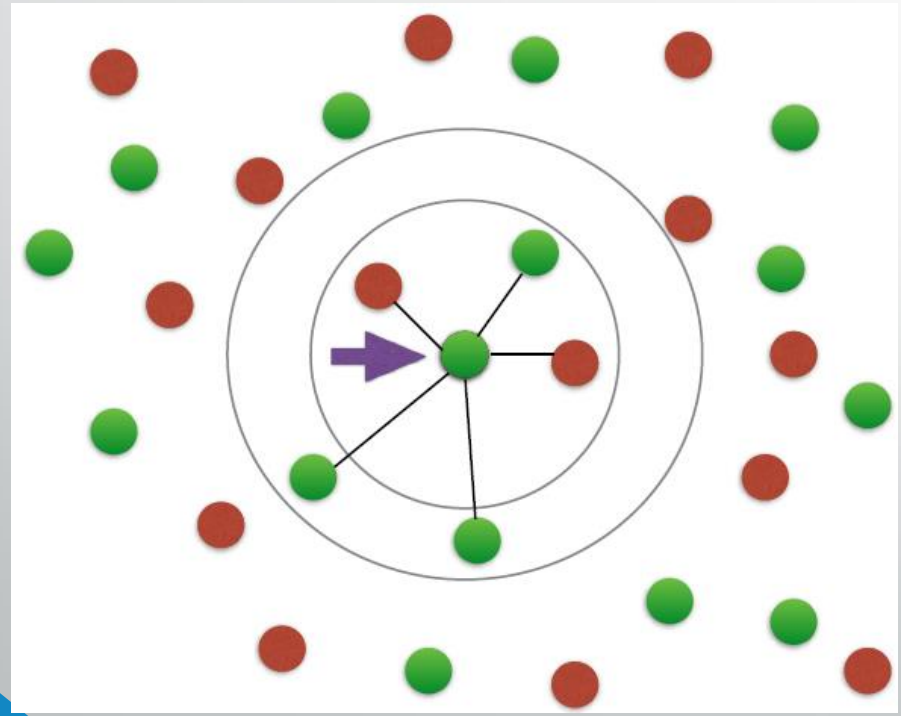


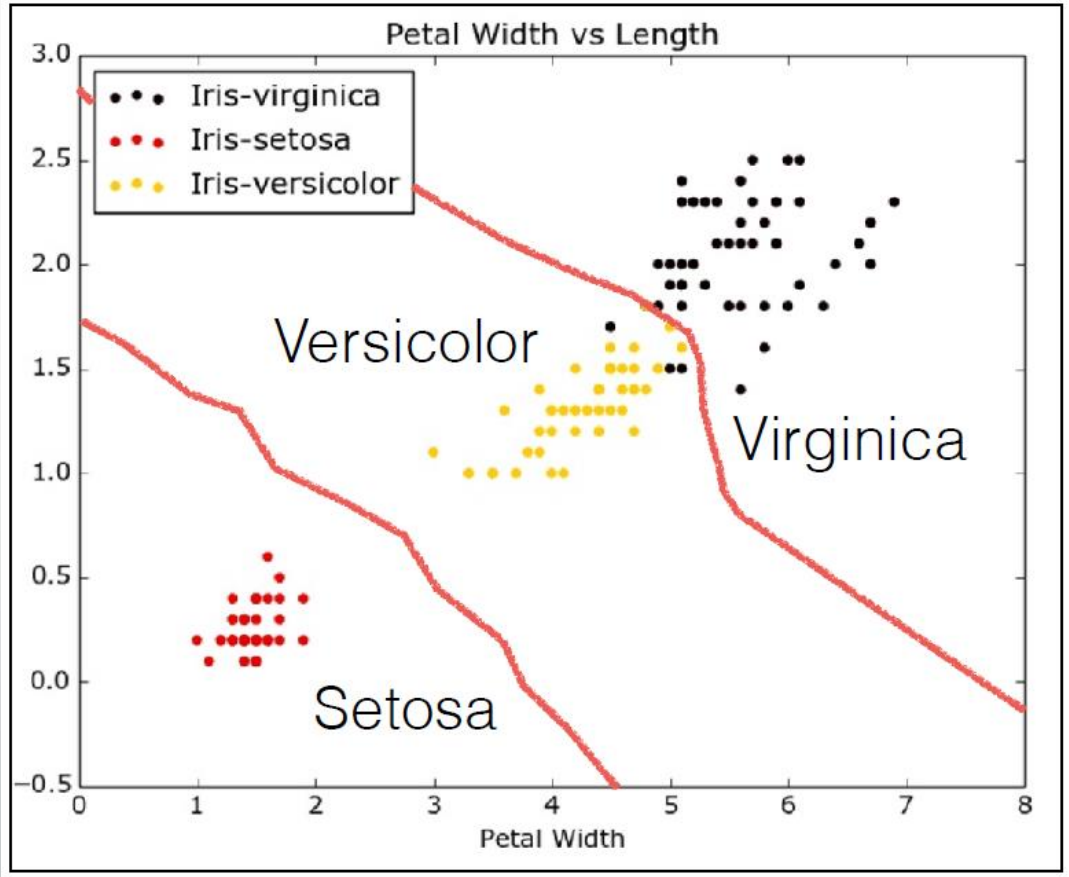
Scikit-learn fit and predict

- All machine learning models implemented as Python classes
  - They implement the algorithms for learning and predicting
  - Store the information learned from the data
- Training a model on the data = 'fitting' a model to the data
  - .fit() method
- To predict the labels of new data: .predict() method

Clustering using KNN

- Basic idea: Predict the label of a data point by
  - Looking at the 'k' closest labeled data points
  - Taking a majority vote





Using scikit-learn to fit a classifier

```
In [1]: from sklearn.neighbors import KNeighborsClassifier

In [2]: knn = KNeighborsClassifier(n_neighbors=6)

In [3]: knn.fit(iris['data'], iris['target'])
Out[3]: KNeighborsClassifier(algorithm='auto', leaf_size=30,
   ...: metric='minkowski',metric_params=None, n_jobs=1,
   ...: n_neighbors=6, p=2,weights='uniform')

In [4]: iris['data'].shape
Out[4]: (150, 4)

In [5]: iris['target'].shape
Out[5]: (150,)
```

Predicting on unlabeled data

```
In [6]: prediction = knn.predict(X_new)

In [7]: X_new.shape
Out[7]: (3, 4)

In [8]: print('Prediction {}'.format(prediction))
Prediction: [1 1 0]
```

Using Logistic Regression

```
In [1]: from sklearn.linear_model import LogisticRegression

In [2]: lr = LogisticRegression()

In [3]: lr.fit(X_train, y_train)

In [4]: lr.predict(X_test)

In [5]: lr.score(X_test, y_test)
```

Using Support Vector Machines

```
In [1]: import sklearn.datasets

In [2]: wine = sklearn.datasets.load_wine()

In [3]: from sklearn.svm import SVC

In [4]: svm = SVC() # default hyperparameters

In [5]: svm.fit(wine.data, wine.target);

In [6]: svm.score(wine.data, wine.target)
Out[6]: 1.
```

Supervised learning in Python

- We will use scikit-learn/sklearn (Python3 + Pip required)
  - Integrates well with the SciPy stack
- Links (Windows)
  - Python (download & install) - https://www.python.org/downloads/
    - During install, activate option "Add Python to environment variables"
  - Pip (download) - https://bootstrap.pypa.io/get-pip.py
  - Open the Command Prompt and navigate to the **get-pip.py** file.
  - Run the following command: **python get-pip.py**
- Linux
  - Terminal -> sudo apt-get install python3 python3-pip
- Mac
  - Terminal -> sudo easy_install pip

Python Supervised Learning with scikit-learn

Integrated Master's in Informatics Engineering

Learning and Extraction of Knowledge

2018/2019

Synthetic Intelligence Lab

Filipe Gonçalves

César Analide