ML in computational science applications

Claudio Schill

Overview

Supervised machine learning

- Using *labelled* training data to learn a model that is then able to make predictions based on unseen/future data
- Called supervised as the “true” values of the target variable are known during training of the model

Unsupervised machine learning

Finding structure in un-labelled data

- Clustering
- Dimensionality reduction
- Anomaly/Outlier detection

Reinforcement learning

Development of an agent that improves its performance based on interaction with the environment

When to use machine learning?

Data availability

Supervised:

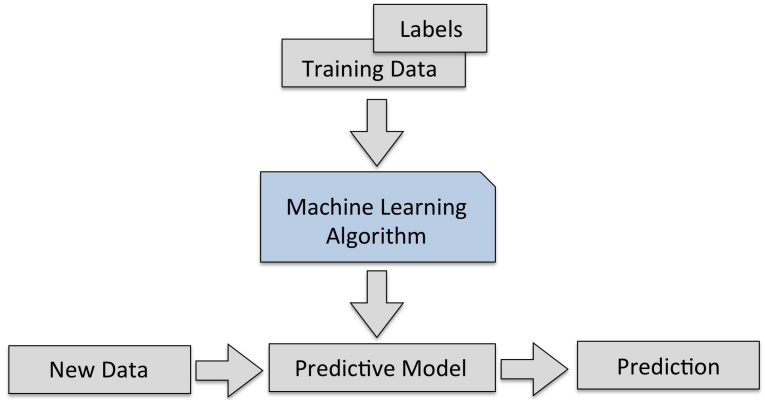
- Specific tasks/problems that are not solvable using mathematical models or rules based classification (e.g. image classification)
- Automation of tedious tasks, that are “simple” for a human

Unsupervised:

- Looking for structure in data, extracting meaningful information/features
- Dimensionality reduction for encoding, visualization of data, or reduction in number of features

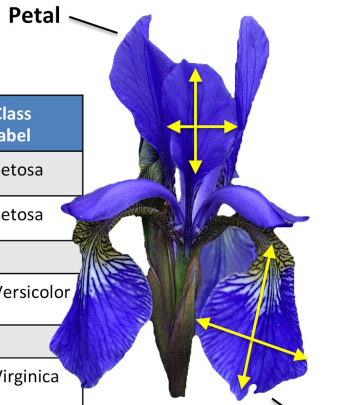
Supervised - Example

Aim: Predict iris flower species type from a set of measured attributes



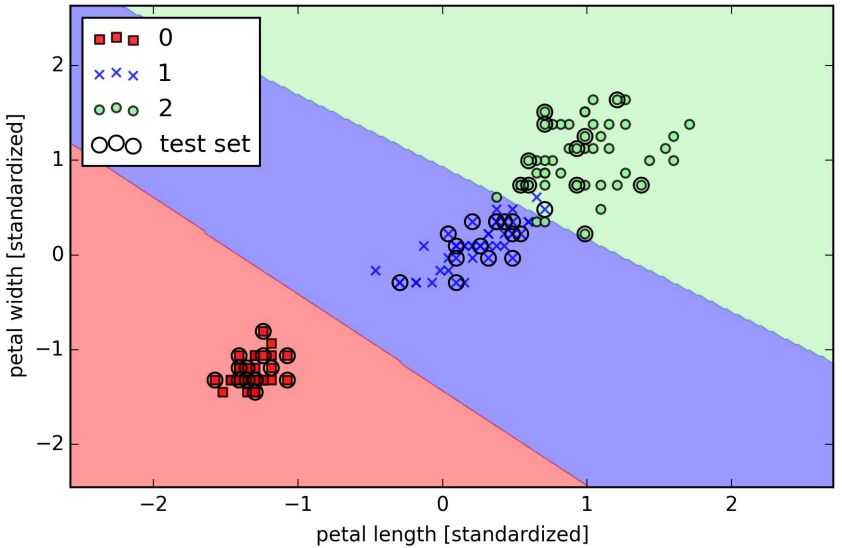
Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica



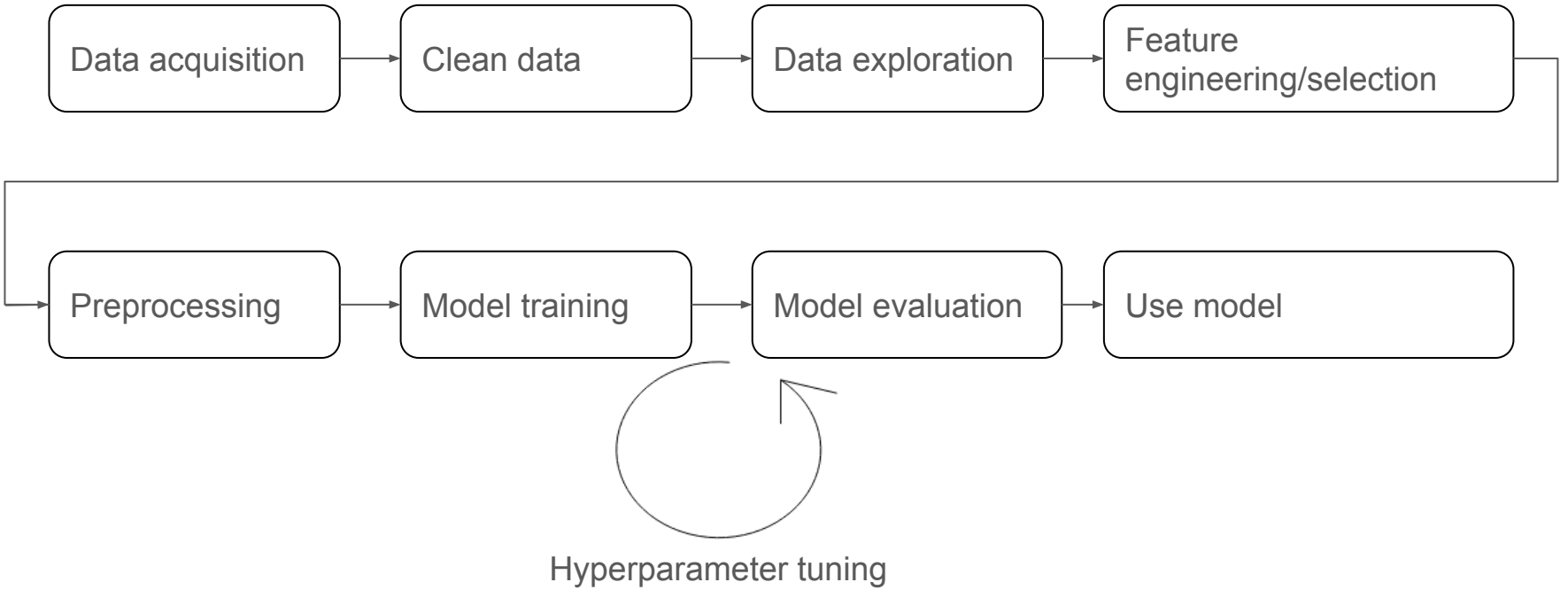
Features
(attributes, measurements, dimensions)

Class labels
(targets)



Figures from Sebastian Raschka, Python Machine Learning (2015)

Supervised - Workflow



A neural network for automated quality screening of ground motion records from small magnitude earthquakes

Xavier Bellagamba, Robin Lee, Brendon A. Bradley (2019)

Task/Problem: Automated screening and rating of the quality of ground motion records

- Manually classified ground motion records as either high or low quality
- Determined range of scalar features to use from the ground motion records
- Trained and evaluated a neural network

Example of automation of a task that can “easily” be done by a human, but difficult to do with rules based classification.

Supervised - Limitations

- Many machine learning algorithms are “black boxes”, making it difficult to understand/follow the decision making process

- Lack of (suitable) data
- Data bias
- Imbalanced dataset

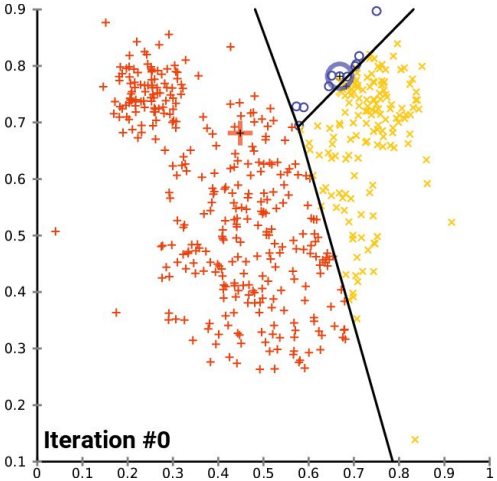
- Computation resource limitations

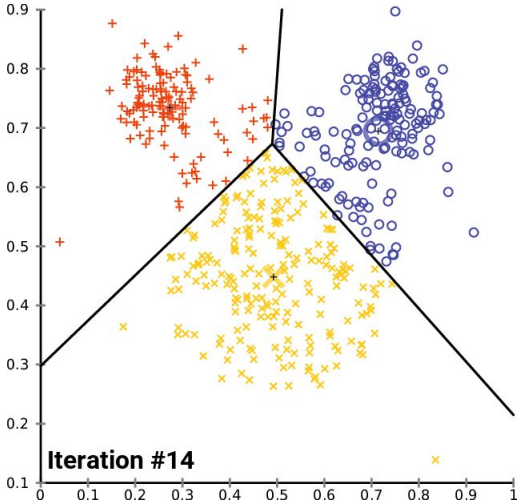
- Labelled data

Unsupervised - Example

K-means clustering:

- Splits data into k clusters, where k is set beforehand
- Places k centroids at random locations in the feature space
- Each iteration:
    - Compute distance of every sample to all k-centroids
    - Assign samples to their closest centroid
    - Update centroid locations using the mean of all samples assigned to the centroid





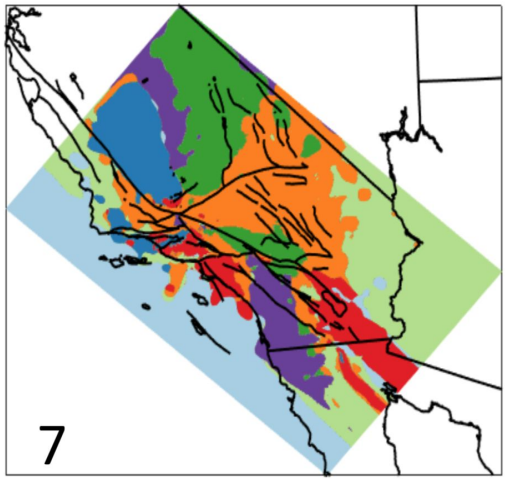
Objective tectonic regionalization of CVM-S4.26 using the k-means clustering algorithm

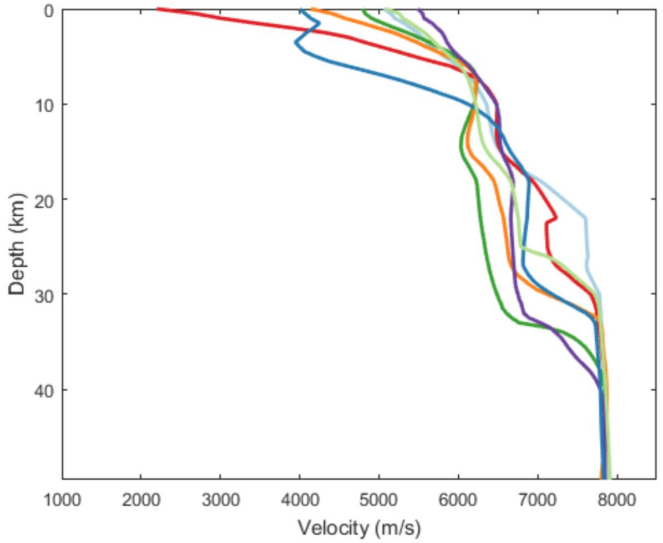
William K. Eymold and Thomas H. Jordan

Applied k-means clustering to the velocity profiles of the SCEC Community Velocity Model

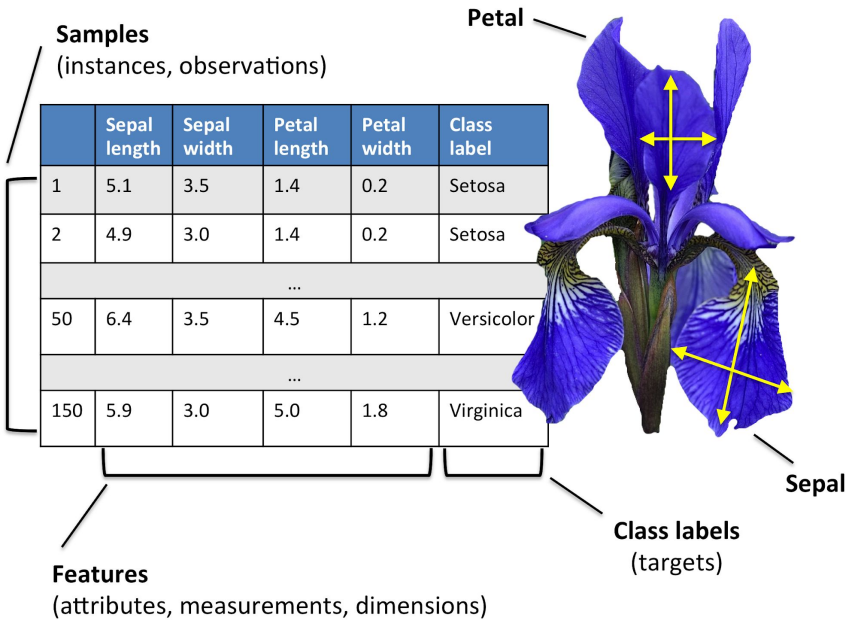
Each velocity profile consists of 100 velocities at a range of different depths

Evaluated at a range of different k-values

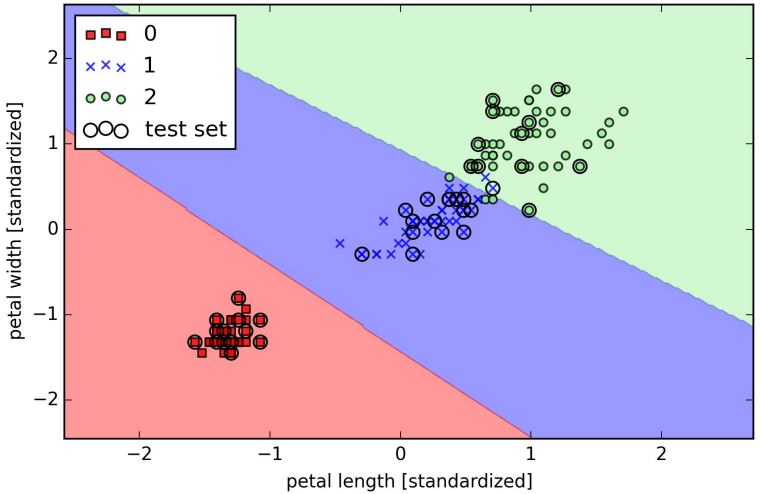




Supervised - Example



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica




```
#%% Pre=processing
X = scale(X)
print(np.mean(X, axis=0), np.std(X, axis=0))

#%% Split into train/test
X_train, X_val, y_train, y_val = \
    train_test_split(X, y, test_size=0.1)

#%% Fit the model
estimator = SVC()
estimator = estimator.fit(X_train, y_train)

#%% Evaluate
print(estimator.score(X_val, y_val))
```

Figures from Sebastian Raschka, Python Machine Learning (2015)

Python - Libraries

Relevant scientific libraries:

NumPy, Pandas, SciPy, Matplotlib - (https://www.scipy.org/)

Machine learning libraries:

Scikit-learn:

- Implementations of supervised learning models, clustering and dimensionality reduction algorithms
- Utility functions for preprocessing and model evaluation
- Documentation, user guides and examples

Keras:

- High-level neural network wrapper, running on top of either Tensorflow, CNTK or Theano
- Makes creation and training of neural networks simple and hassle-free

Questions?

List of python libraries

Scientific core libraries: Numpy, Pandas, Scipy (https://scipy.org/)

Visualisation: Matplotlib (https://matplotlib.org/), Plotly (https://plot.ly/python/)

Machine learning, Preprocessing, Utils: scikit-learn (https://scikit-learn.org/stable/index.html)

Gradient tree boosting: Xgboost (https://xgboost.readthedocs.io/en/latest/)

Deep Learning: Keras (https://keras.io/), Tensorflow (https://www.tensorflow.org/), Theano (http://www.deeplearning.net/software/theano/)

This is just a small list of the main libraries for machine learning in python, there are many others!