UNIVERSITÄT BIELEFELD
Faculty of Technology

Programming

Applied Machine Learning

Luna Pianesi

Faculty of Technology, Bielefeld University


Machine Learning
Scikit-Learn
Applications

Machine Learning

- Branch of artificial intelligence
- Combination of statistics, optimization theory, computer science, information theory, ...

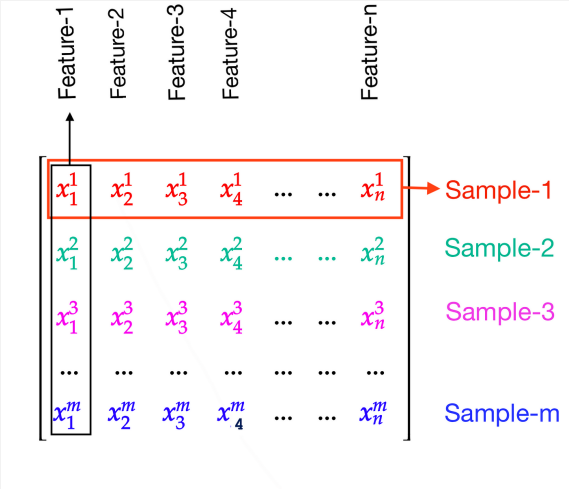
Unsupervised Learning

- Dimensionality reduction
- Clustering

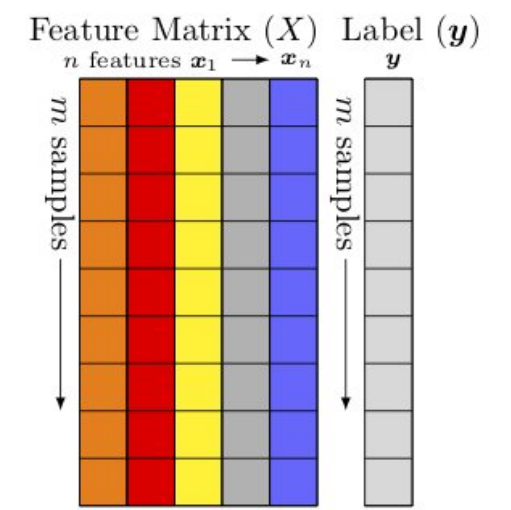
Supervised Learning

- Classification
- Regression

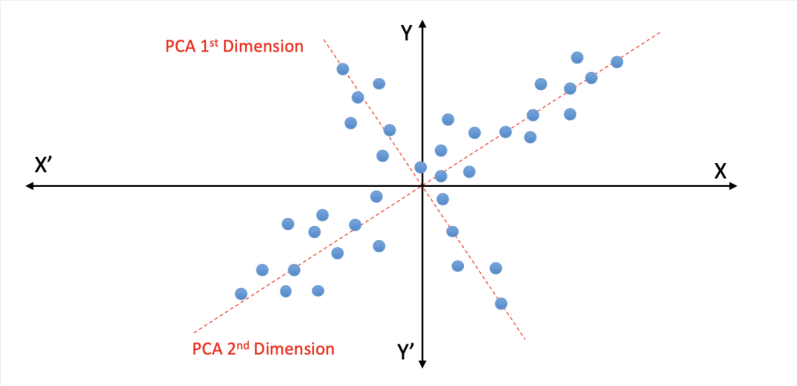
Data representation: feature matrix



Data representation: feature matrix



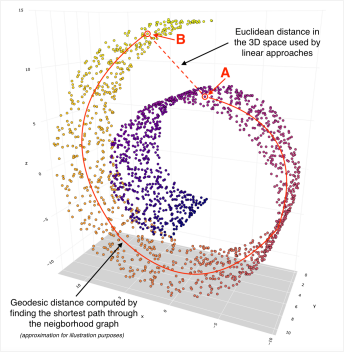
Dimensionality reduction



Some methods:

- Principal Component Analysis (PCA)
- Isomap

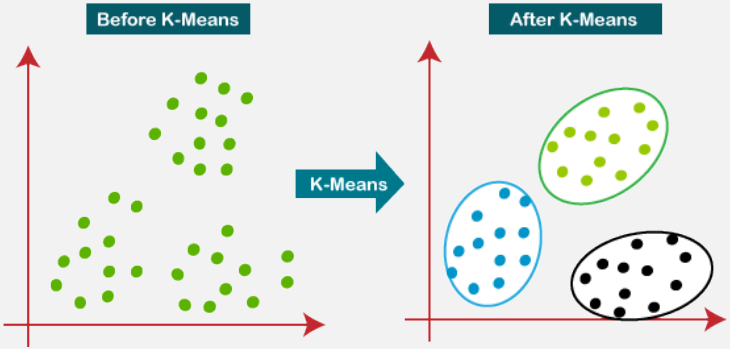
Dimensionality reduction



Some methods:

- Principal Component Analysis (PCA)
- Isomap

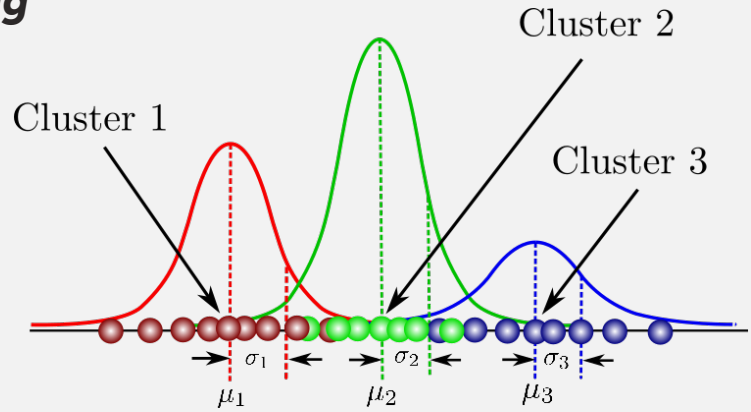
Clustering



Some methods:

- K-means
- Gaussian Mixture Models (GMM)
- Spectral Clustering

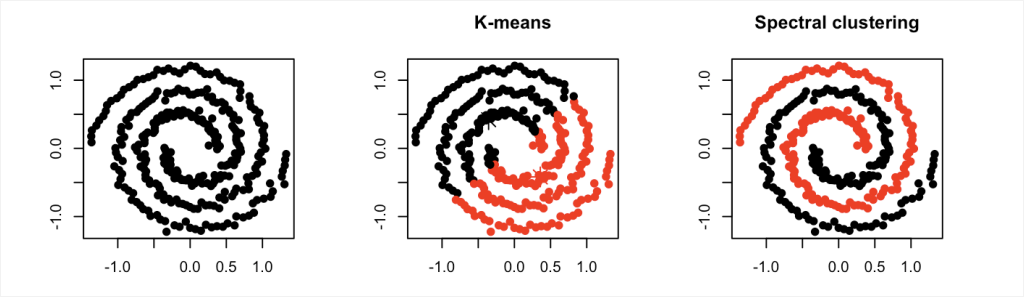
Clustering



Some methods:

- K-means
- Gaussian Mixture Models (GMM)
- Spectral Clustering

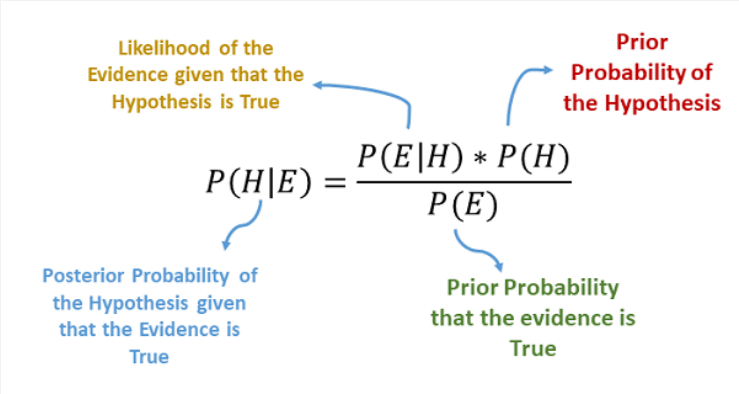
Clustering



Some methods:

- K-means
- Gaussian Mixture Models (GMM)
- Spectral Clustering

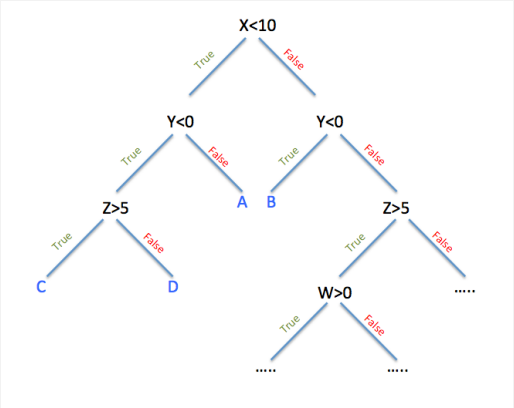
Classification



Some methods:

- Naive Bayes
- Decision Trees

Classification



Some methods:

- Naive Bayes
- Decision Trees

Regression

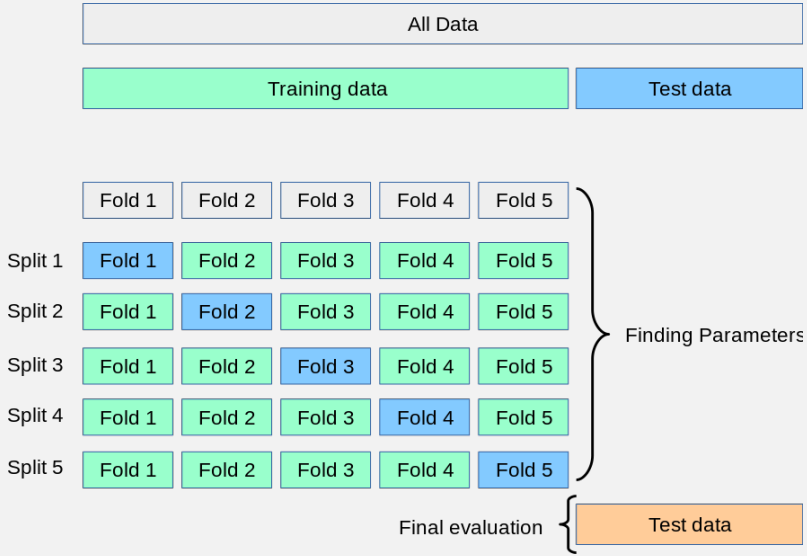
Some methods:

- Linear regression
- Ridge regression, Lasso regression
- Multiple regression
- Multivariate regression

Feature representation

- Categorical: can be stored and identified by names or labels
- Numerical: simply numbers

Cross-validation



Quiz

- Assign the following methods to their categories:
  - Naive Bayes
  - Kmeans
  - PCA
  - Decision Tree
  - Gaussian Mixture Models
  - Isomap
  - Spectral Clustering

- *True* or *false?*
  - Cross validation can only be performed on labeled data
  - Gaussian Mixture Models assumes that data points follow a normal distribution

Quiz

- Assign the following methods to their categories:
  - Naive Bayes — Classification
  - Kmeans — Clustering
  - PCA — Dimensionality red.
  - Decision Tree — Classification
  - Gaussian Mixture Models — Clustering
  - Isomap — Dimensionality red.
  - Spectral Clustering — Clustering

- *True* or *false?*
  - Cross validation can only be performed on labeled data — true
  - Gaussian Mixture Models assumes that data points follow a normal distribution — true


Machine Learning
Scikit-Learn
Applications

The Estimator API

Estimators of the Scikit-Learn package share a common API.

Use of estimators:

- Choose model (Estimator)
- Choose model hyperparameters
- Instantiate model with hyperparameters
- Call `fit()` to train the model on a given data set
- Apply model to new data:
  - Supervised learning: call `predict()`
  - Unsupervised learning: call `transform()` or `predict()` (depending on the estimator)

Quiz

- *True* or *false?*
  - The basic steps are *model*, *fit*, *predict/transform*
  - `LinearRegression.coef_` returns slope and intercept of line
  - Scikit-Learn can generate artificial datasets
  - Scikit-Learn doesn't provide real world data sets
  - transformers uses the `predict()` to transform data.
- Explain the function of the following estimators:
  - `OneHotEncoder`
  - `ColumnTransformer`
  - `DictVectorize`
  - `CountVectorizer`

Quiz

- *True* or *false?*
  - The basic steps are *model*, *fit*, *predict/transform* — true
  - `LinearRegression.coef_` returns slope and intercept of line — false
  - Scikit-Learn can generate artificial datasets — true
  - Scikit-Learn doesn't provide real world data sets — false
  - transformers uses the `predict()` to transform data. — false
- Explain the function of the following estimators:
  - `OneHotEncoder` Transforms one categorical feature with *n* possible values into *n* binary features
  - `ColumnTransformer` Transforms all columns of a `DataFrame`
  - `DictVectorize` Transforms `dict` with categorical variables into numeric features
  - `CountVectorizer` Tokenizes strings and constructs word count frequency matrix


Machine Learning
Scikit-Learn
Applications

Quiz

- In which order does function `train_test_split` return test/train data?
  - `Xtrain, Ytrain, Xtest, Ytrain`
  - `Xtest, Ytest, Xtrain, Ytrain`
  - `Xtrain, Xtest, Ytrain, Ytest`
  - `Xtest, Xtrain, Ytest, Ytrain`
- What data is stored in
  - `digits.images`
  - `digits.data`
  - `digits.target`

Quiz

- In which order does function `train_test_split` return test/train data?
  - `Xtrain, Ytrain, Xtest, Ytrain`
  - `Xtest, Ytest, Xtrain, Ytrain`
  - `Xtrain, Xtest, Ytrain, Ytest` ✔
  - `Xtest, Xtrain, Ytest, Ytrain`
- What data is stored in
  - `digits.images` bitmap data of all images
  - `digits.data` feature matrix
  - `digits.target` labels (ground truth digits)

Recap

Summary

- Machine Learning
  - Dimensionality reduction
  - Clustering
  - Classification
  - Regression
- Scikit-Learn
  - Estimator API
  - Feature representation
  - Crossvalidation
- Applications
  - Handwritten digits dataset
  - Text comparison

What comes next?

- Have a look at the Jupyter Notebook of this lecture
- Further reading about Pandas: Chapter 5 of the “Python Data Science Handbook”:
  `https://jakevdp.github.io/PythonDataScienceHandbook/`
- Have a look at the in-depth analyses that are provided in the handbook