

BDI Python Code Clinic

17/05/2021

**Machine Learning in Python
with scikit-learn
Session 1**

Irina Chelysheva

Before we start...

List of the Python SciPy libraries required for this tutorial:

scipy

numpy

matplotlib

pandas

sklearn

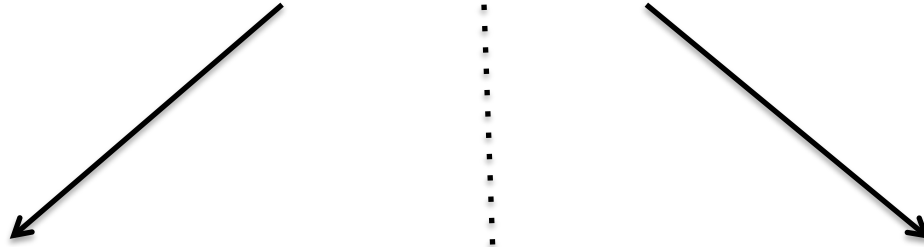
+my Python script downloaded from

https://github.com/Chelysheva/ML_Python_course

What is Machine Learning?

- Machine learning is all about extracting information from the data
- Intersection of statistics, artificial intelligence and computer science
- Key idea: The algorithm improves automatically through experience/training
- Based on input data (training data), the algorithm builds a model to make predictions for new data (test data)

ML problems



Supervised ML

input variables (x) + output variable (Y)

- Classification Ex. SVM, Random forest
- Regression Ex. Linear regression, Random forest

Unsupervised ML

only input data (X)

- Clustering Ex. K-means
- Association Ex. Apriori

Semi-supervised ML

*large amount of input data (X)
+ only some of the data is labeled (Y)*

Supervised vs Unsupervised Learning

- Supervised Learning:
 - Input and output values of the training data are being fed to the Algorithm
 - The presence of the output values guides the learning process
- Unsupervised Learning:
 - Only the input values of the training data are being fed to the algorithm
 - The algorithm has to find patterns within the data without a specified output goal

Two Types of Supervised Learning

- Regression Problems: How many/how much...?
 - The target variable is a real number
 - Typical methods: OLS regression, ridge/Lasso regression, regression trees, random forests, neural networks
- Classification Problems: Is this class A or B (or C)...?
 - The target variable is a categorical variable (binary dummy or an integer)
 - Typical methods: Logistic regression, k-nearest neighbors, regression trees, random forests, neural networks

Overview of ML algorithms



General ML project steps

- Defining Problem
- Preparing Data
- Evaluating Algorithms
- Improving Results
- Presenting Results

Problem statement

Using the UCI PIMA Indian Diabetes dataset to predict whether a person has diabetes or not based on the medical attributes provided.

-> 2 class classification problem

Assumptions:

This is enough data to split and reliably predict if the patient has diabetes.

Just these medical attributes are enough for diagnosis.

UCI PIMA Indian Diabetes dataset

Consists of several medical predictor variables and one target variable (class):

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)²)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 – no diabetes or 1 - diabetes)

Discover dataset

Load the dataset;
Summarize the dataset;
Clean-up dataset;
Visualize the dataset

Evaluate some algorithms

Separate out train and test datasets;
Set-up the test harness to use 10-fold cross validation;
Build multiple different models to predict diagnosis;
Select the best - finalized model

Make predictions

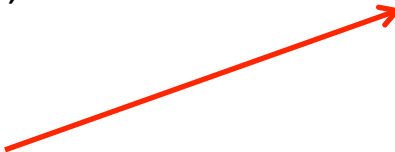
Make predictions on test set;
Evaluate the predictions:

- Accuracy (+null accuracy);
- Confusion matrix;
- Sensitivity and specificity

Export/import the model

Additionally improve the model

ROC curve



Logistic Regression (LR)
Linear Discriminant Analysis (LDA)
K-Nearest Neighbors (KNN)
Classification and Regression Trees (CART)
Gaussian Naive Bayes (NB)
Support Vector Machines (SVM)

- linear (LR and LDA)
- nonlinear (KNN, CART, NB and SVM)