INFSCI 2915: Machine Learning

Implementation Workflow and K-Nearest Neighbor Classification in Python

Mai Abdelhakim

School of Computing and Information

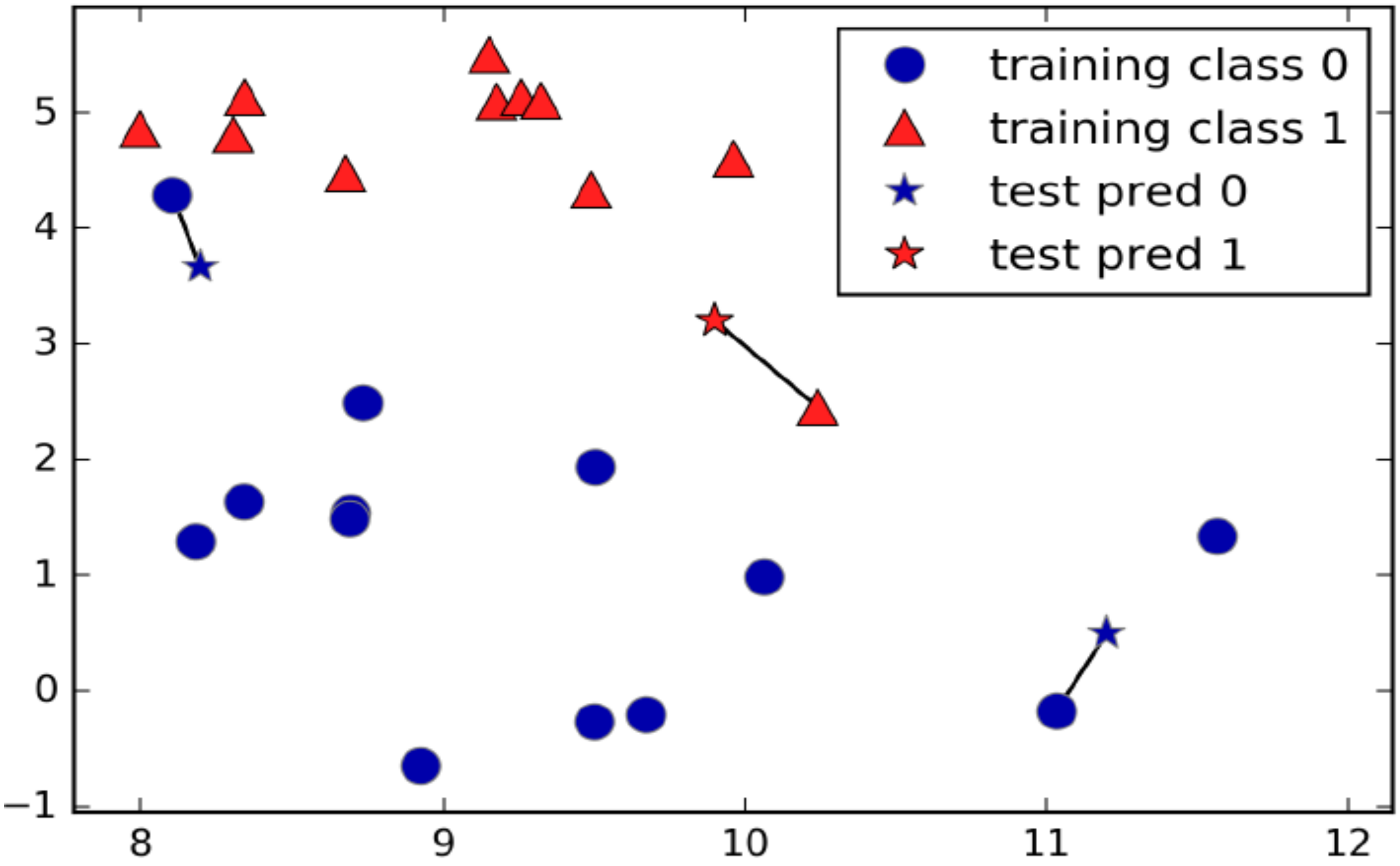
610 IS Building

Spring 2018

Recall K-Nearest Neighbor (KNN)

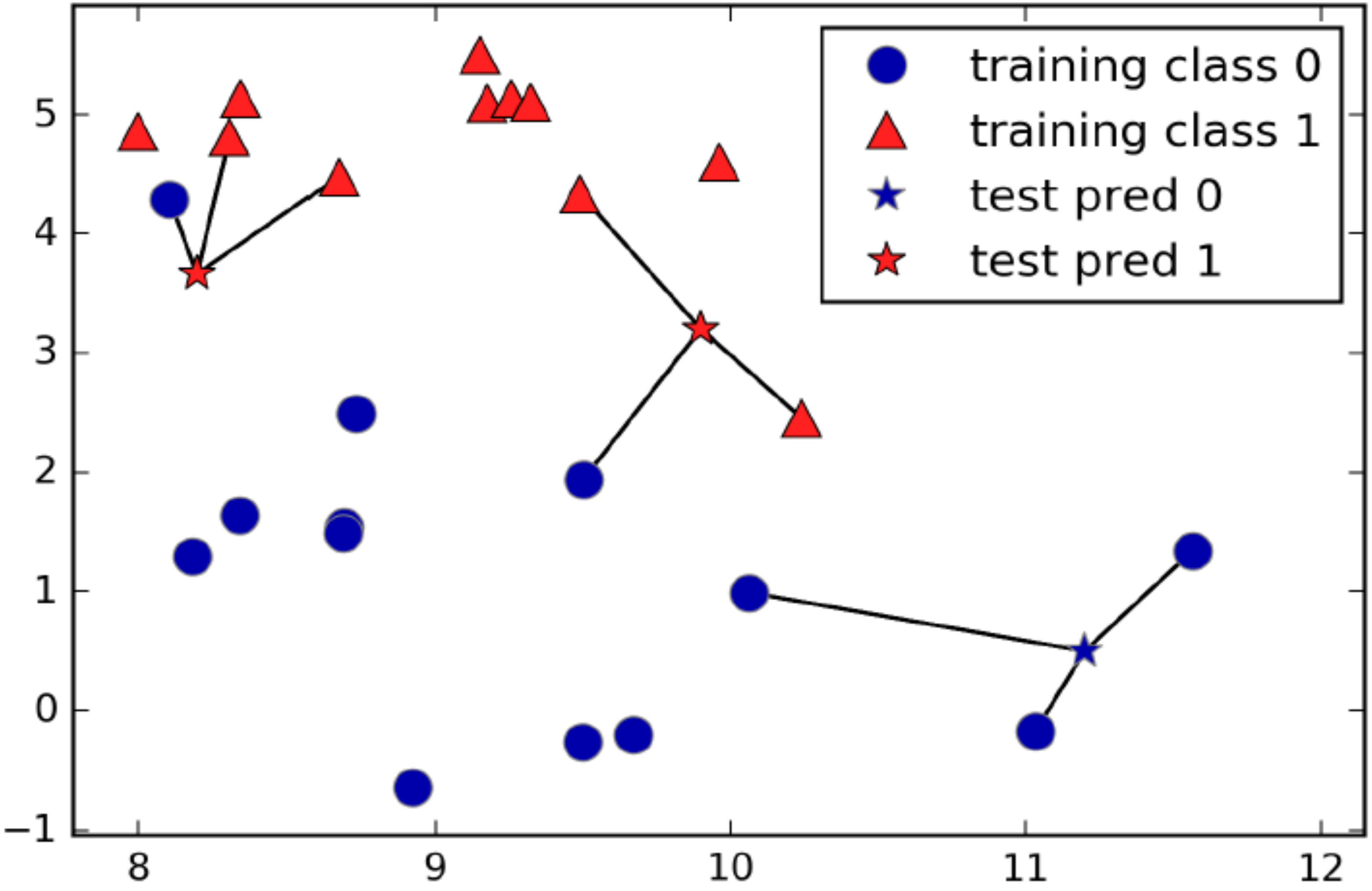
- Store the training set
- Prediction for a new data: algorithm finds K points in the training set that are closest to the new data point
  - Typically **Euclidean distance is used** to find close neighbors
  - Assume Point 1: with feature vector $P_1 = \{x_{11}, x_{12}, \ldots, x_{1p}\}$ $X_{i,j}$: the *j*th feature of *i*th data point
    Point 2, with feature vector $P_2 = \{x_{21}, x_{22}, \ldots, x_{2p}\}$
    Then the Euclidean distance between the two samples is:
    $d(P_1,P_2)=\sqrt{\sum_{i=1}^{p}(x_{1i}-x_{2i})^2}$
- Then make prediction using **majority** class among the neighbors

KNN with K=1



training class 0
training class 1
test pred 0
test pred 1
5
4
3
2
1
0
−1
8
9
10
11
12

KNN with K=3



training class 0
training class 1
test pred 0
test pred 1
5
4
3
2
1
0
−1
8
9
10
11
12

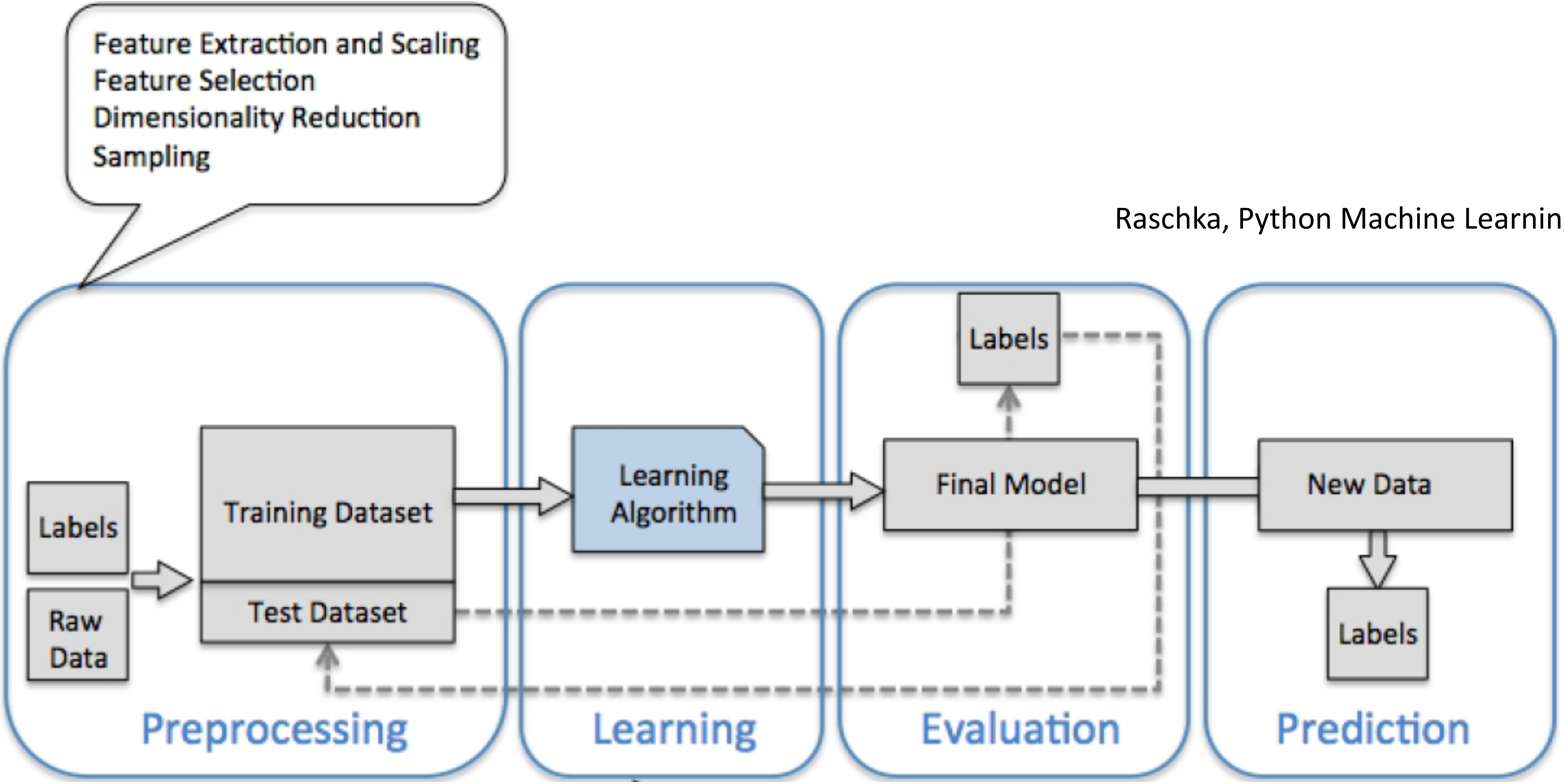
Keep the big picture in mind!

- Understand what is the **problem** you are trying to solve
- Can the collected **data solve** the problem?
- Have you collected **enough data** to represent the problem?
- What **features** of data to extract to enable accurate predictions?
- Which **machine learning algorithm and parameters** to use?
- How to access the **accuracy** and **success of my application**?

Implementation Steps

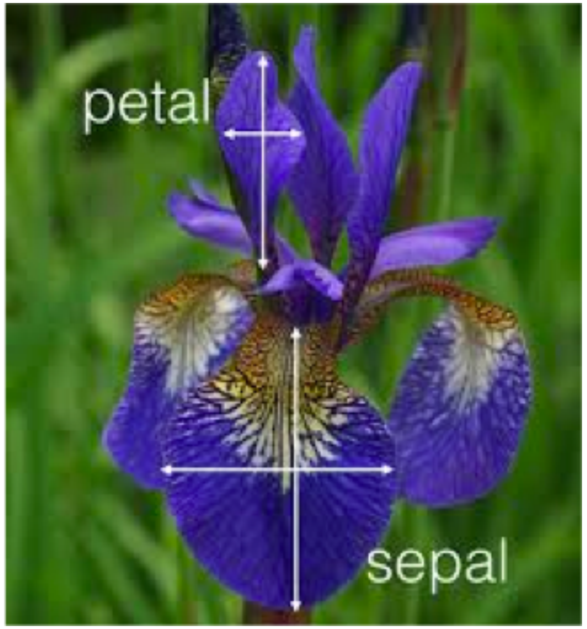
- Step 1: Get the data, and represent information by features (feature extraction) --- *preprocessing*
- Step 2: Split the Data to Training and Test Set (*Preprocessing*)
- Step 3: Define your Model
- Step 4: Fit (Train) your Model using training data (*Learning*)
- Step 5: Performance Evaluation using test data (*Evaluation*)
  You can then use for *prediction*

Raschka, Python Machine Learning



Example: Classify Iris Species

- Goal: Distinguish species of different Iris flowers
  - Species: (0) Setosa, (1) Versicolor, (2) Virginica
- Features:
  - length and width of sepal
  - Width and length of petals
- From measurements of Iris flowers whose species are known, develop a model that predict the species with new measurements
- **Supervised learning:** we know correct species of training data – called **label**
- **Classification problem**

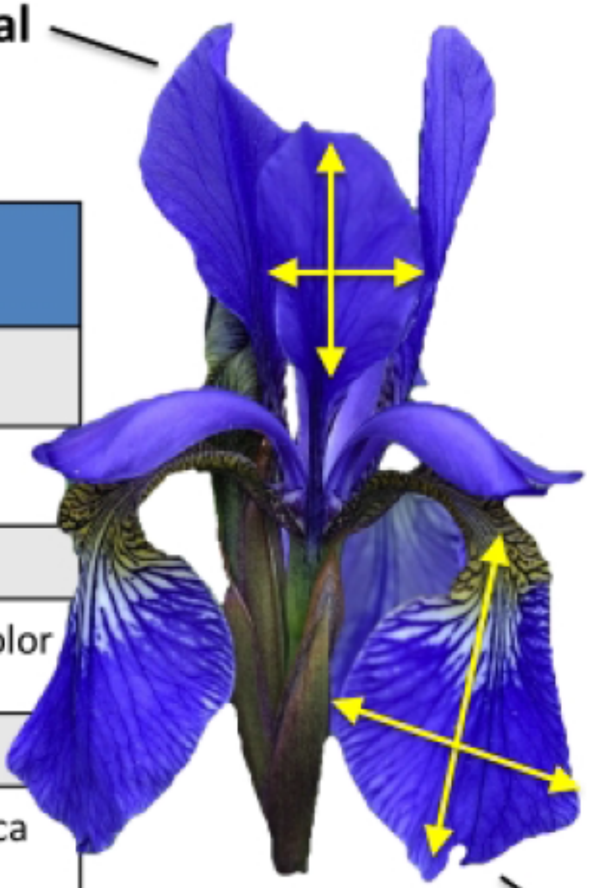


Step 1: Get the Data

- Classical dataset in machine learning, included in Scikit-learn – **datasets module**
- In python type the following to import the dataset and save it in an object:
  from sklearn.datasets import load_iris #load the dataset
  iris_dataset=load_iris() #this is an object similar to a dictionary
- There are 150 samples in the dataset

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)

Raschka, Python Machine Learning

- Find the keys of the iris_datasets: ['target_names', 'data', 'target', 'DESCR', 'feature_names'])
    - DESCR: contains description of the dataset
    - feature_names: array of strings containing the species
    - **target**: 0, 1, 2 corresponding to species (Setosa, Versicolor, Virginica)
        - np.unique(iris_dataset['target'])
    - **data**: contains the measurement of data (features/predictors)

- Find type and shape of ‘data’

- Print the first five rows of ‘data’ (use: iris_dataset[‘data’][:5])

Step 2: Split the Data to Training and Test Set

- Split the data to training (75%) and test (25%)
- Scikit-learn has a function that **shuffles** and **splits** the data
- In python:
  from **sklearn.model_selection** import **train_test_split**
  X_train, X_test, Y_train, Y_test =
  train_test_split(iris_dataset['data'],iris_dataset['target'],random_state=100)
  - random_state is a seed to the random generator to ensure that same sequence is output every run
- Find the shape of text and train samples:
  - 112 samples (75% of data points) and 38 samples

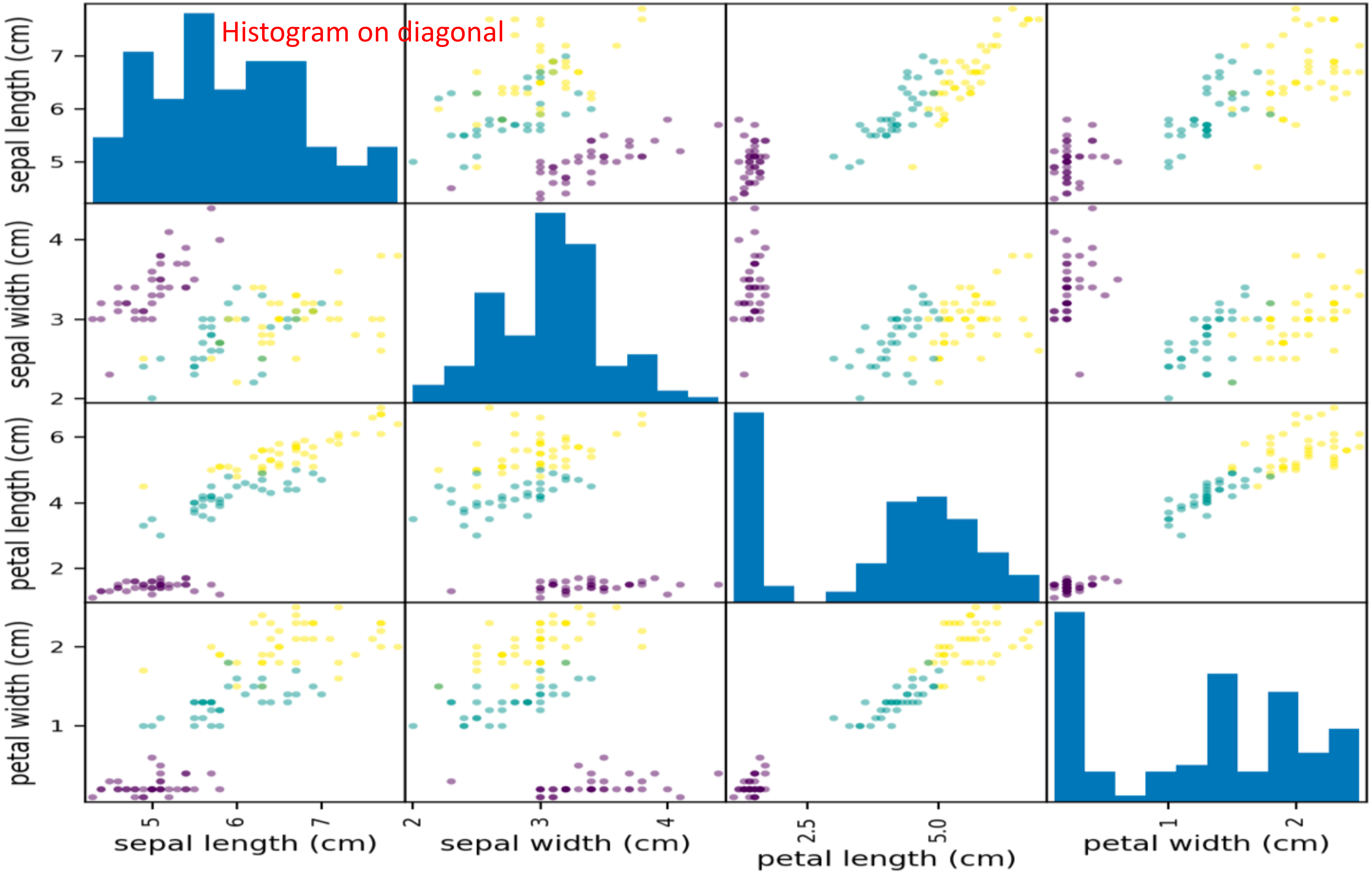
Inspect the data

- In general, look into your data to make sure that there is nothing missing and you have sufficient number of samples from each class

- Use **scatter_plot** which is a function supported by **pandas** whose input has to be a **DataFrame**

- Python:

```
%matplotlib inline
import pandas as pd
iris_dataFrame=pd.DataFrame(X_train, columns=iris_dataset.feature_names)
#create a scatter_matrix for the dataframe
sm=pd.plotting.scatter_matrix(iris_dataFrame,c=Y_train, figsize=(15,15))
```



Different colors = different classes (species) Data can be distinguished through the features!

- You can draw pairplots with Seaborn as well

```
import seaborn as sns
sns.pairplot(iris_dataFrame)
```

Step 3: Define your Model
K-Nearest Neighbor

- In scikit-learn: KNeighborClassifier class in neighbors module
  from sklearn.neighbors import KNeighborsClassifier
- Then, **initiate class into an object**
  knn=**KNeighborsClassifier**(n_neighbors=k)

Step 4: Fit (Train) your Model

- **Fit the model using the training data:** by calling the fit object
  knn.fit(X_train, Y_train)

Step 5: Performance Evaluation

- Accuracy: the fraction of flowers for which the correct species **was predicted correctly**

  Accuracy=**knn.score**(X_test,Y_test)

- Alternatively, you can make predictions and then accuracy using the test set as follows:
  Y_predict=knn.predict(X_test)
  np.mean(Y_predict==Y_test)

Make Predictions

- **Make predictions:** predict label of new data
  - Define a new sample with: sepal length=5cm, sepal width=2.9cm , petal length: 1cm, petal width= 0.2cm.
    - X_new=np.array([5,2.9,1,0.2])
  - Call predict method:
    prediction=**knn.predict**(X_new)

  - Print the predictions
    print('Predictions is:', prediction, '\n') # this is number 0 or 1 or 2 (corresponding to a particular species). \n is for new line
    print('The prediction is:', iris_dataset['target_names'][prediction]) # this prints the name of this species

Feature Scaling

- Assume two features: one in the range 0 – 1, and another in the range of 100-10000.
  - The contribution to the Euclidean distance will be different!
- Thus, feature scaling is recommended in practice and would improve performance
- Done in the preprocessing step

Feature Scaling - MinMaxScaler

- MinMaxScaler: scales features to be in range 0 -1

```
from sklearn import preprocessing
scaler=preprocessing.MinMaxScaler().fit(X_train) #define scaler depending on the features in training data
X_train_transformed=scaler.transform(X_train) #apply scaling on training set
X_test_transformed=scaler.transform(X_test) #apply scaling on test set
```

Feature Scaling - StandardScaler

- StandardScaler: scales features so that they are all with zero mean and unit variance

```
from sklearn import preprocessing
scaler=preprocessing.StandardScaler().fit(X_train) #define scaler depending on the features in training data
X_train_transformed=scaler.transform(X_train) #apply scaling on training set
X_test_transformed=scaler.transform(X_test) #apply scaling on test set
```

Exercise

A) Classify the Iris species with KNN approach using the **first two feature only** (X_train[:, : 2], X_test[:, : 2]), and check the accuracy as **K changes**. Let K takes the values [1, 5, 10, 15]. No need to scale features.

In the code, use **random_state=100** in **train_test_split**

- Plot the accuracy and comment on your result

B) Use the Iris example, and find the accuracy of the KNN approach with K=5 when **different number of features** is used **without scaling**

  - To use N_features only from training data use: X_train[:, : N_features]

- Write down the accuracy when using one, two, three, and the four features
- Repeat when feature scaling with MinMaxScaler is used