Introduction

In this notebook, we will:

- Explain the main machine learning concepts
- Used scikit-learn to build a first model
- Learn our first algorithm (kNN)

Types of machine learning

We often distinguish 3 `types` of machine learning:

- **Supervised Learning**: learn a model from labeled *training data,* then make predictions
- **Unsupervised Learning**: explore the structure of the data to extract meaningful information
- **Reinforcement Learning**: develop an agent that improves its performance based on interactions with the environment

Note:

- Semi-supervised methods combine the first two.
- ML systems can combine many types in one system.

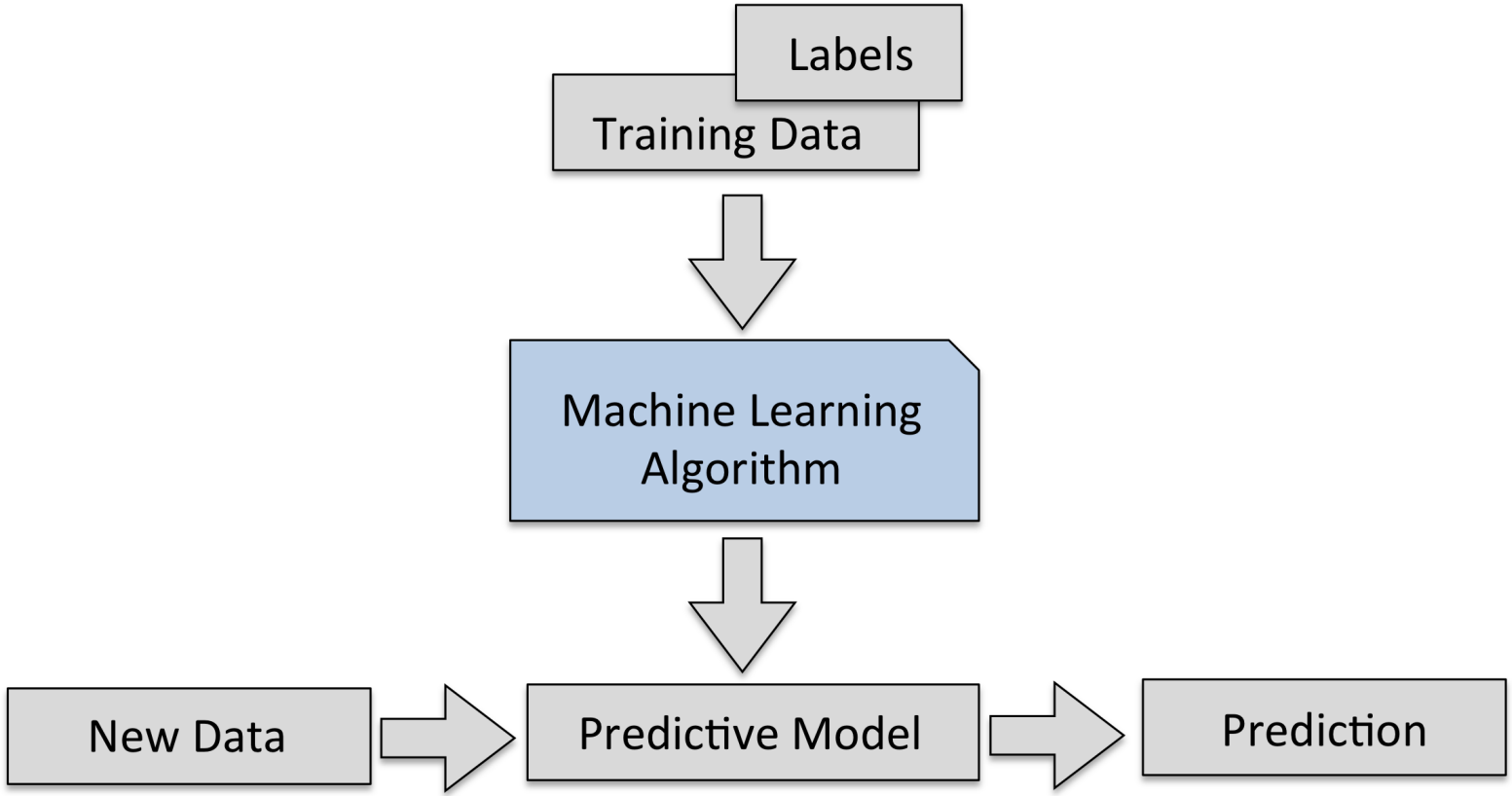
Supervised Machine Learning

- Learn a model from labeled training data, then make predictions
- Supervised: we know the correct/desired outcome (label)

2 subtypes:

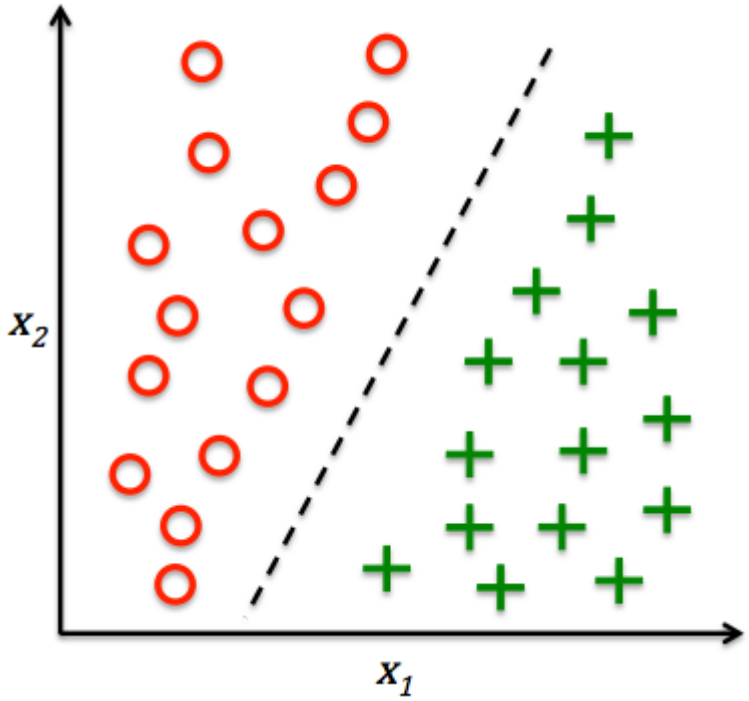
- Classification: predict a *class label* (category), e.g. spam/not spam
  - Many classifiers can also return a *confidence* per class
- Regression: predict a continuous value, e.g. temperature
  - Some algorithms can return a *confidence interval*

Most supervised algorithms that we will see can do both.


Labels
Training Data
Machine Learning Algorithm
New Data
Predictive Model
Prediction

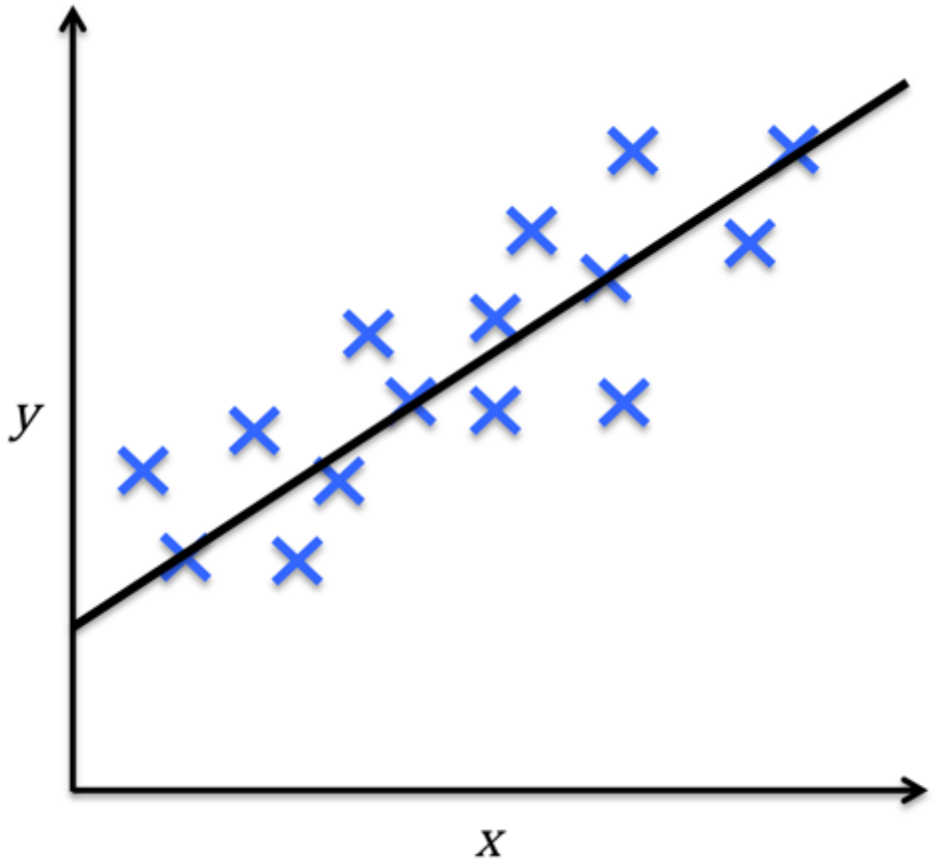
Classification

- Class labels are discrete, unordered
- Can be *binary* (2 classes) or *multi-class* (e.g. letter recognition)
- Dataset can have any number of predictive variables (predictors)
  - Also known as the dimensionality of the dataset
- The predictions of the model yield a *decision boundary* separating the classes



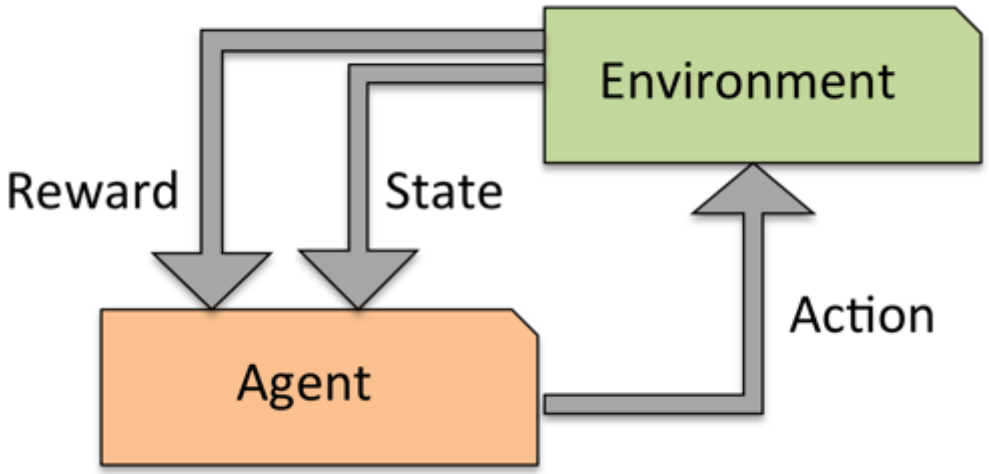
Regression

- Target variable is numeric
- Find the relationship between predictors and the target.
  - E.g. relationship between hours studied and final grade
- Example: Linear regression (fits a straight line)



Reinforcement learning

- Develop an agent that improves its performance based on interactions with the environment
  - Example: games like Chess, Go,...
- *Reward function* defines how well a (series of) actions works
- Learn a series of actions that maximizes reward through exploration

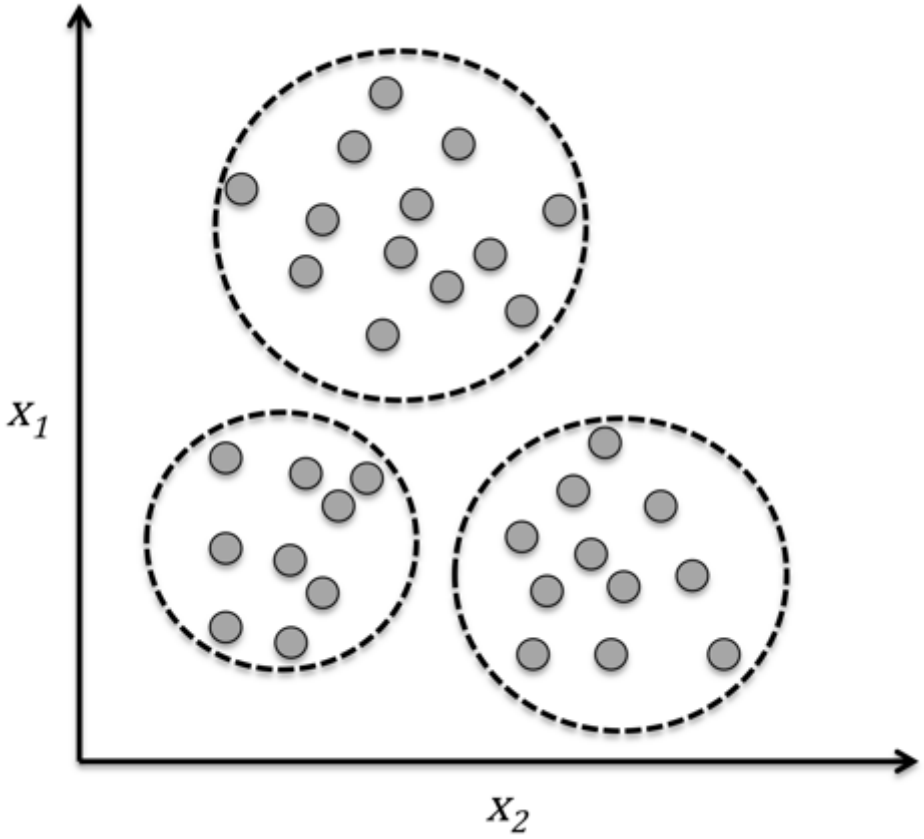


Unsupervised Machine Learning

- Unlabeled data, or data with unknown structure
- Explore the structure of the data to extract information
- Many types, we'll just discuss two.

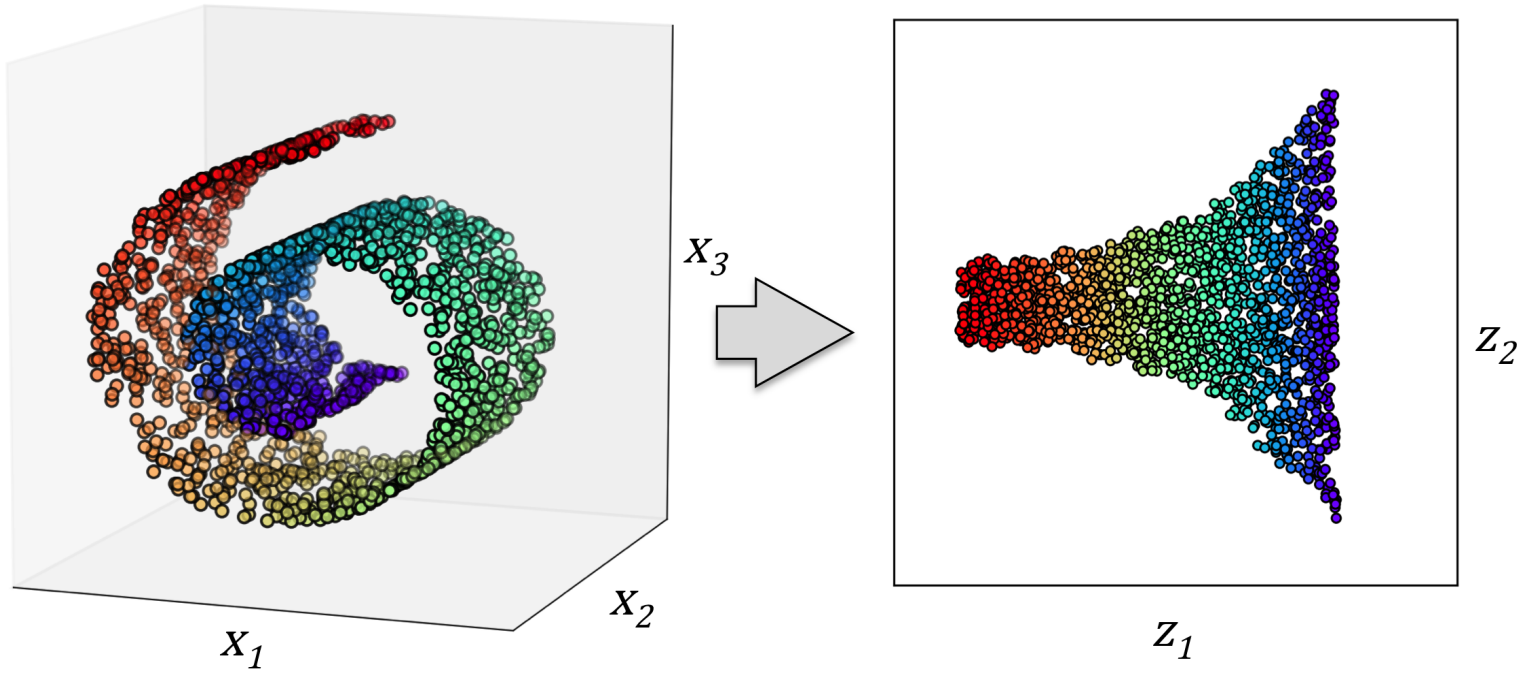
Clustering

- Organize information into meaningful subgroups (clusters)
- Objects in cluster share certain degree of similarity (and dissimilarity to other clusters)
- Example: distinguish different types of customers



Dimensionality reduction

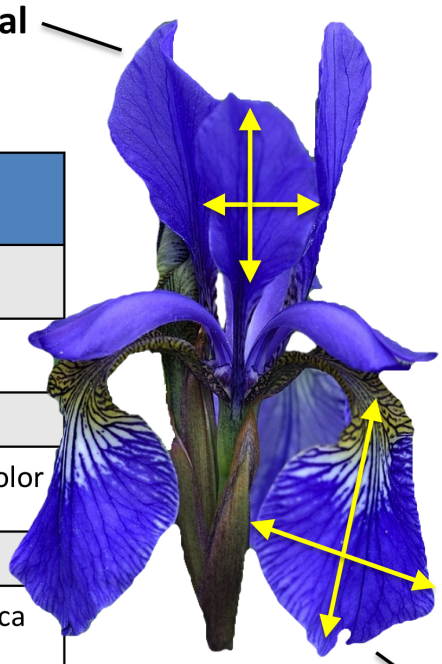
- Data can be very high-dimensional and difficult to understand, learn from, store,...
- Dimensionality reduction can compress the data into fewer dimensions, while retaining most of the information
- Contrary to feature selection, the new features lose their (original) meaning
- Is often useful for visualization (e.g. compress to 2D)


x_3
x_2
x_1
z_2
z_1

Basic Terminology (on Iris 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)

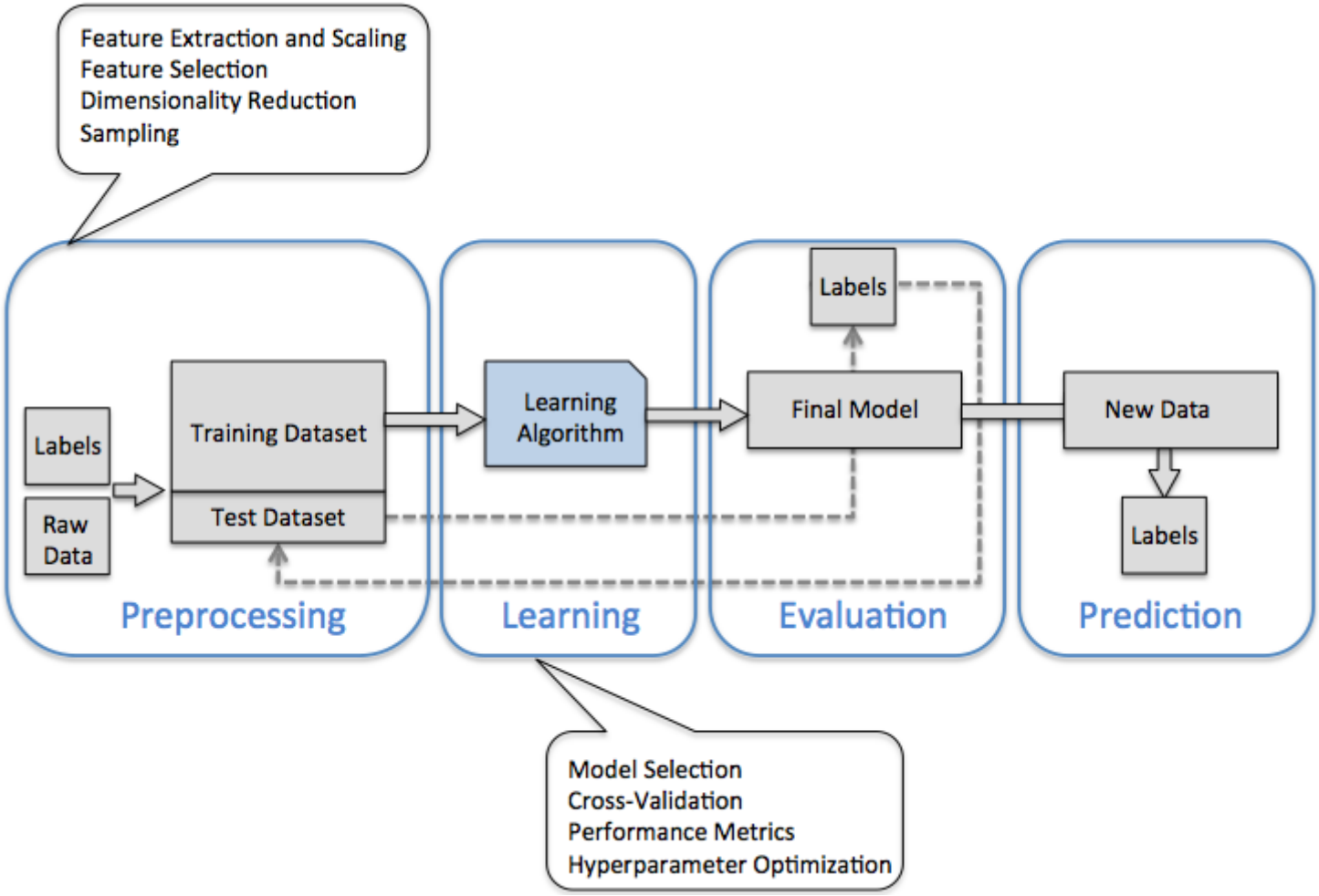
Building machine learning systems

A typical machine learning system has multiple components:

- Preprocessing: Raw data is rarely ideal for learning
  - Feature scaling: bring values in same range
  - Encoding: make categorical features numeric
  - Discretization: make numeric features categorical
  - Feature selection: remove uninteresting/correlated features
  - Dimensionality reduction can also make data easier to learn

- Learning and model selection
  - Every algorithm has its own biases
  - No single algorithm is always best (No Free Lunch)
  - *Model selection* compares and selects the best models
    - Different algorithms
    - Every algorithm has different options (hyperparameters)
  - Split data in training and test sets

- Together they form a *workflow* of *pipeline*



scikit-learn

One of the most prominent Python libraries for machine learning:

- Contains many state-of-the-art machine learning algorithms
- Offers comprehensive documentation (http://scikit-learn.org/stable/documentation) about each algorithm
- Widely used, and a wealth of tutorials (http://scikit-learn.org/stable/user_guide.html) and code snippets are available
- scikit-learn works well with numpy, scipy, pandas, matplotlib,...

Algorithms

See the Reference (http://scikit-learn.org/dev/modules/classes.html)

Supervised learning:

- Linear models (Ridge, Lasso, Elastic Net, ...)
- Support Vector Machines
- Tree-based methods (Classification/Regression Trees, Random Forests,...)
- Nearest neighbors
- Neural networks
- Gaussian Processes
- Feature selection

Unsupervised learning:

- Clustering (KMeans, ...)
- Matrix Decomposition (PCA, ...)
- Manifold Learning (Embeddings)
- Density estimation
- Outlier detection

Model selection and evaluation:

- Cross-validation
- Grid-search
- Lots of metrics

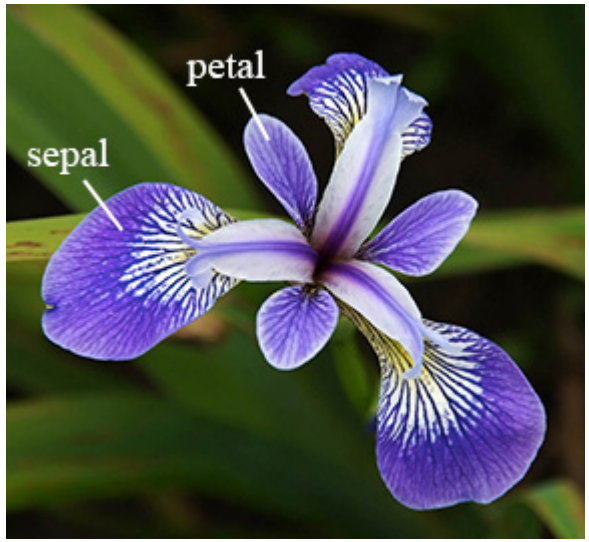
Data import

Multiple options:

- A few toy datasets are included in `sklearn.datasets`
- You can import data files (CSV) with `pandas` or `numpy`
- You can import 1000s of machine learning datasets from OpenML

Example: classification

Classify types of Iris flowers (setosa, versicolor, or virginica) based on the flower sepals and petal leave sizes.



Iris is included in scikitlearn, we can just load it.
This will return a `Bunch` object (similar to a `dict`)

```
iris_dataset = load_iris()
print("Keys of iris_dataset: {}".format(iris_dataset.keys()))
print(iris_dataset['DESCR'][:193] + "\n...")
```

```
Keys of iris_dataset: dict_keys(['data', 'target', 'target_names', 'DESC
R', 'feature_names', 'filename'])
.. _iris_dataset:

Iris plants dataset
--------------------

**Data Set Characteristics:**

    :Number of Instances: 150 (50 in each of three classes)
    :Number of Attributes: 4 numeric, pre
...
```

The targets (classes) and features are stored as `lists`, the data as an `ndarray`

```
print("Targets: {}".format(iris_dataset['target_names']))
print("Features: {}".format(iris_dataset['feature_names']))
print("Shape of data: {}".format(iris_dataset['data'].shape))
print("First 5 rows:\n{}".format(iris_dataset['data'][:5]))
```

```
Targets: ['setosa' 'versicolor' 'virginica']
Features: ['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)',
'petal width (cm)']
Shape of data: (150, 4)
First 5 rows:
[[5.1 3.5 1.4 0.2]
 [4.9 3.  1.4 0.2]
 [4.7 3.2 1.3 0.2]
 [4.6 3.1 1.5 0.2]
 [5.  3.6 1.4 0.2]]
```

The targets are stored separately as an `ndarray`, with indices pointing to the features

```
print("Target names: {}".format(iris_dataset['target_names']))
print("Targets:\n{}".format(iris_dataset['target']))
```

```
Target names: ['setosa' 'versicolor' 'virginica']
Targets:
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2]
```

Building your first model

All scikitlearn classifiers follow the same interface

```
class SupervisedEstimator(...):
    def __init__(self, hyperparam, ...):

    def fit(self, X, y):   # Fit/model the training data
        ...                # given data X and targets y
        return self

    def predict(self, X):  # Make predictions
        ...                # on unseen data X
        return y_pred

    def score(self, X, y): # Predict and compare to true
        ...                # labels y
        return score
```

Training and testing data

To evaluate our classifier, we need to test it on unseen data.
`train_test_split`: splits data randomly in 75% training and 25% test data.

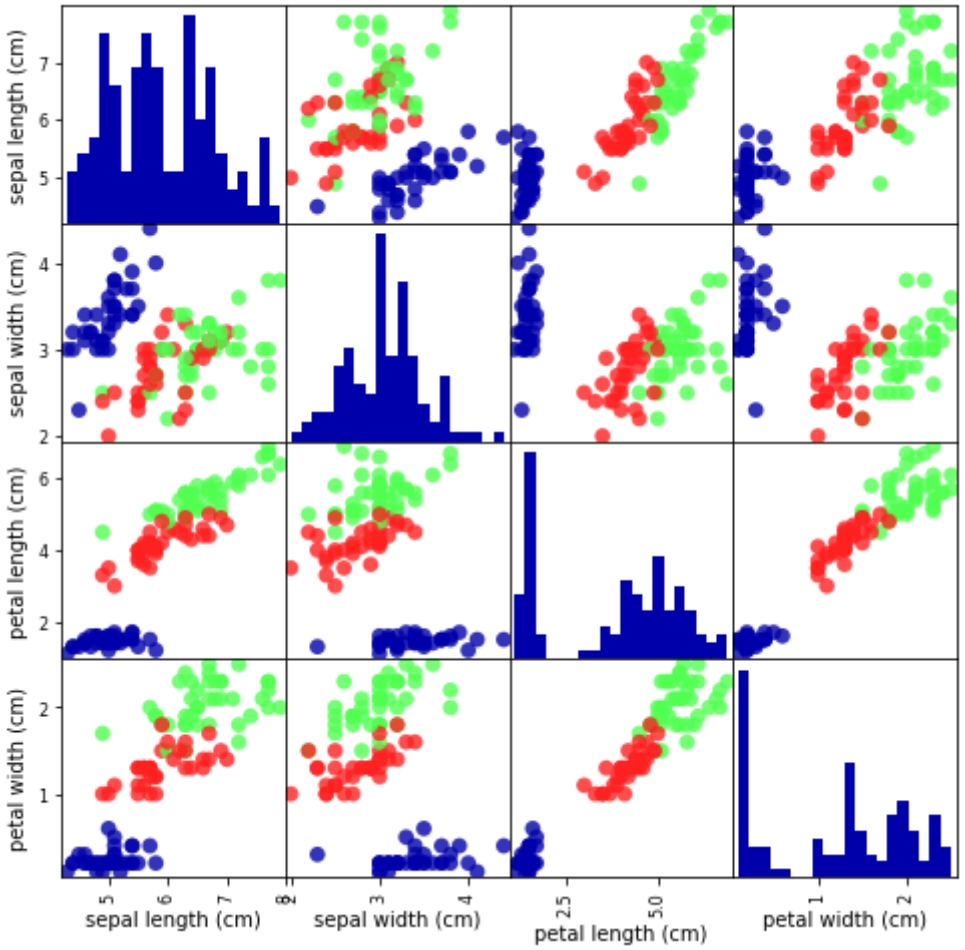
```
X_train, X_test, y_train, y_test = train_test_split(
    iris_dataset['data'], iris_dataset['target'], random_state=0)
```

```
X_train shape: (112, 4)
y_train shape: (112,)
X_test shape: (38, 4)
y_test shape: (38,)
```

Note: there are several problems with this approach that we will discuss later:

- Why 75%? Are there better ways to split?
- What if one random split yields different models than another?
- What if all examples of one class all end up in the training/test set?

Looking at your data (with pandas)



Fitting a model

The first model we'll build is called k-Nearest Neighbor, or kNN. More about that soon.
kNN is included in `sklearn.neighbors`, so let's build our first model

```
knn = KNeighborsClassifier(n_neighbors=1)
knn.fit(X_train, y_train)
```

Out[7]:
```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=None, n_neighbors=1, p=2,
           weights='uniform')
```

Making predictions

Let's create a new example and ask the kNN model to classify it

```
X_new = np.array([[5, 2.9, 1, 0.2]])
prediction = knn.predict(X_new)
```

```
Prediction: [0]
Predicted target name: ['setosa']
```

Evaluating the model

Feeding all test examples to the model yields all predictions

```
y_pred = knn.predict(X_test)
```

```
Test set predictions:
 [2 1 0 2 0 2 0 1 1 1 2 1 1 1 1 0 1 1 0 0 2 1 0 0 2 0 0 1 1 0 2 1 0 2 2 1
0
 2]
```

We can now just count what percentage was correct, or use `score`

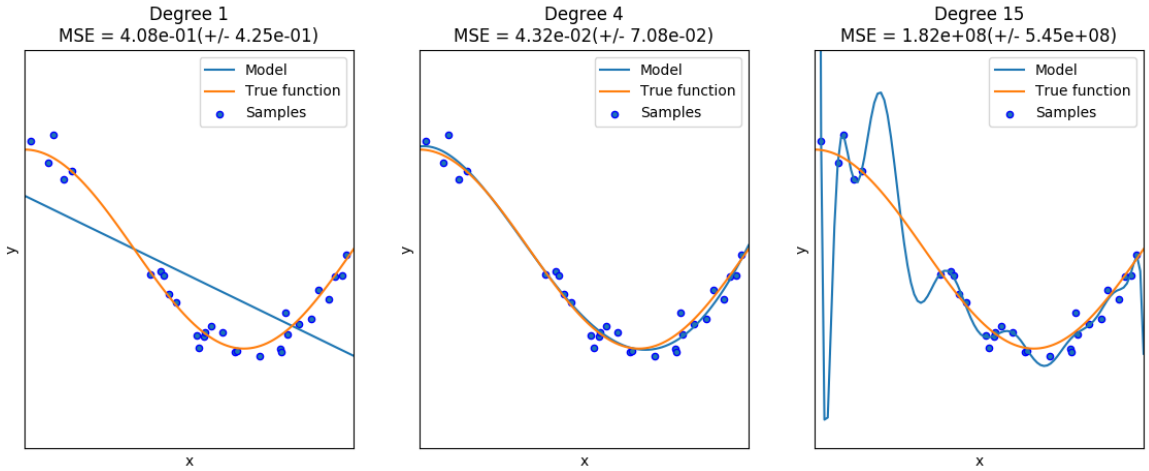
```
print("Score: {:.2f}".format(np.mean(y_pred == y_test)))
print("Score: {:.2f}".format(knn.score(X_test, y_test) ))
```

```
Score: 0.97
Score: 0.97
```

Generalization, Overfitting and Underfitting

- We **hope** that the model can *generalize* from the training to the test data: make accurate predictions on unseen data
- It's easy to build a complex model that is 100% accurate on the training data, but very bad on the test data
- Overfitting: building a model that is *too complex for the amount of data* that we have
  - You model peculiarities in your data (noise, biases,...)
  - Solve by making model simpler (regularization), or getting more data
- Underfitting: building a model that is *too simple given the complexity of the data*
  - Use a more complex model

- There is often a sweet spot that you need to find by optimizing the choice of algorithms and hyperparameters, or using more data.



In all supervised algorithms that we will discuss, we'll cover:

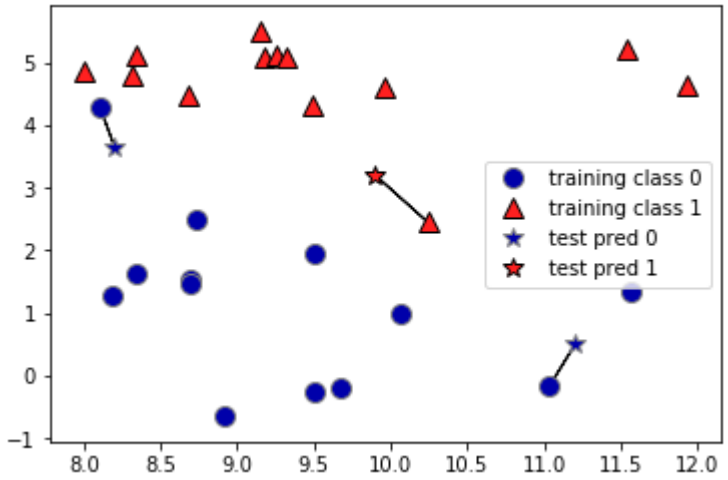
- How do they work
- How to control complexity
- Hyperparameters (user-controlled parameters)
- Strengths and weaknesses

k-Nearest Neighbor

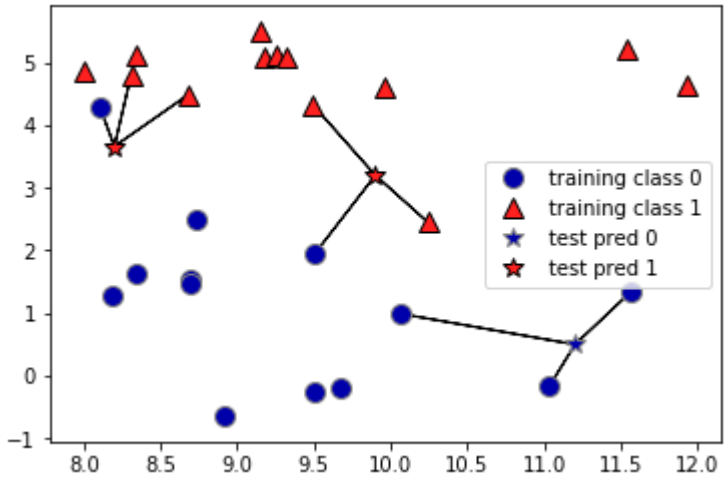
- Building the model consists only of storing the training dataset.
- To make a prediction, the algorithm finds the *k* closest data points in the training dataset

k-Nearest Neighbor Classification

for k=1: return the class of the nearest neighbor



for k>1: do a vote and return the majority (or a confidence value for each class)



Let's build a kNN model for this dataset (called 'Forge')

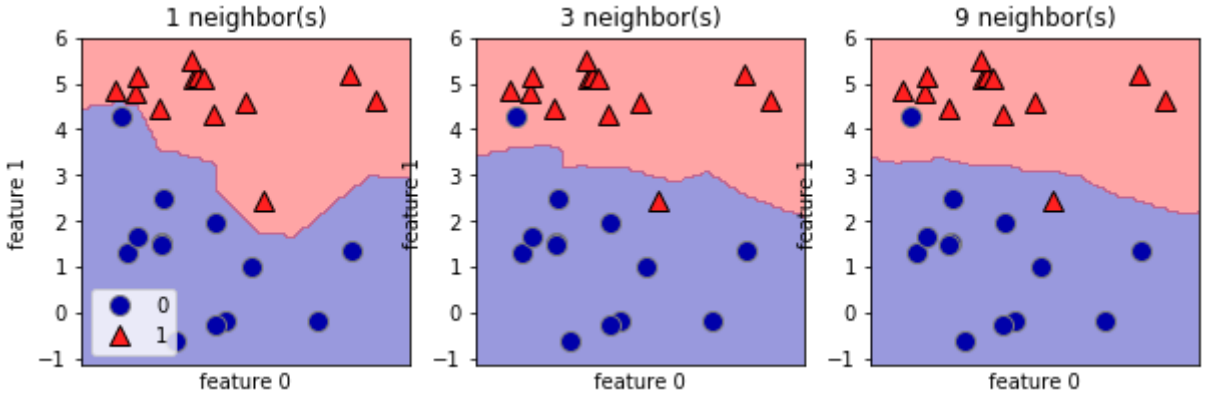
```
X, y = mglearn.datasets.make_forge()
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0
)
clf = KNeighborsClassifier(n_neighbors=3)
clf.fit(X_train, y_train)
```

Out[13]:
```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=None, n_neighbors=3, p=2,
           weights='uniform')
```

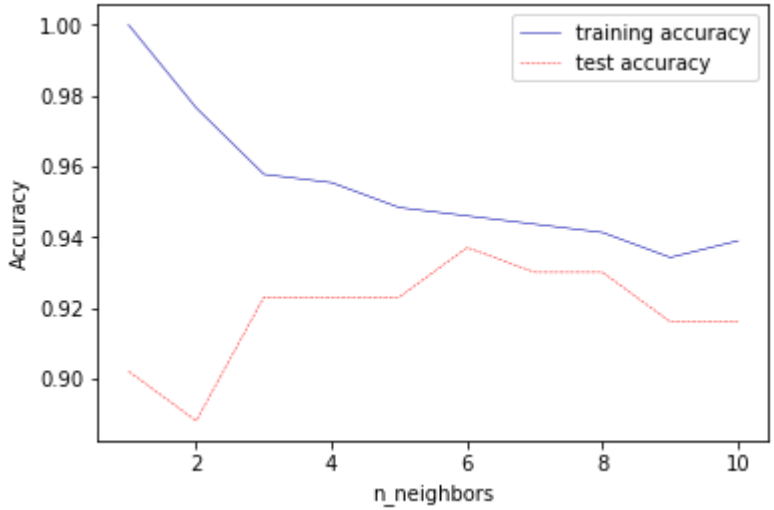
```
Test set accuracy: 0.86
```

Analysis

We can plot the prediction for each possible input to see the *decision boundary*

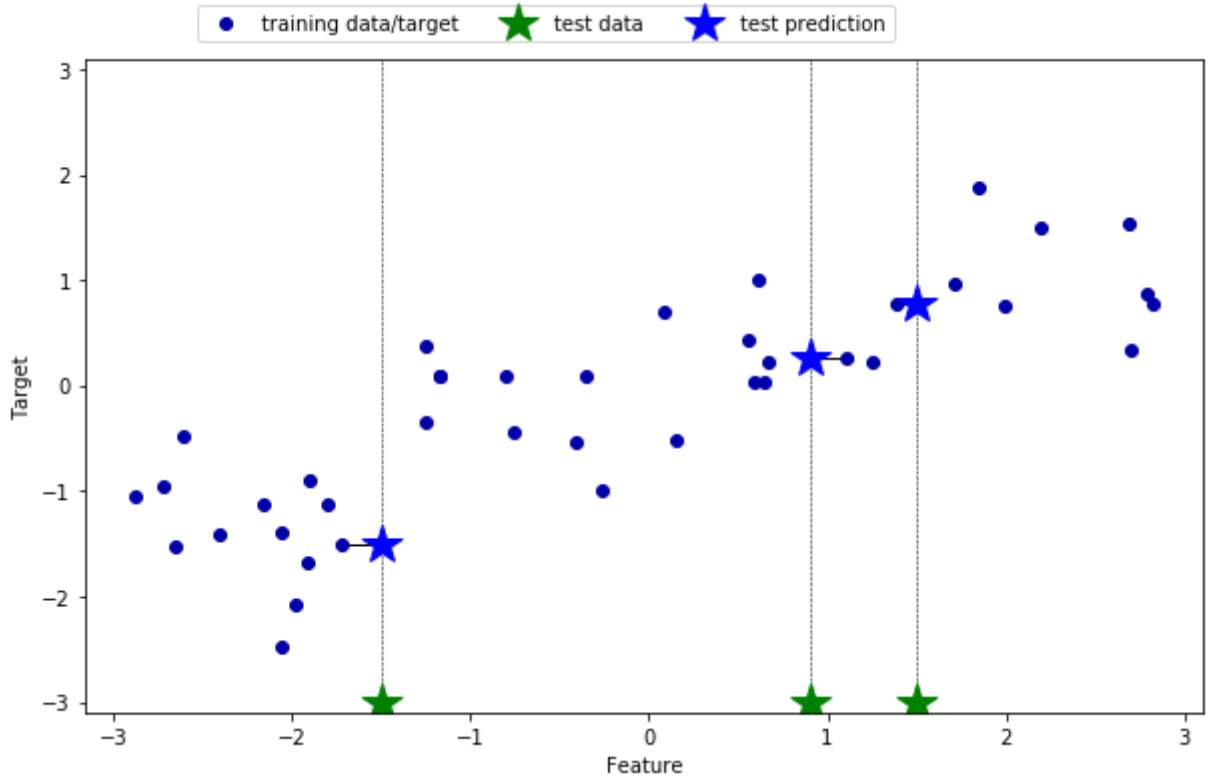


We can more directly measure the effect on the training and test error on a larger dataset (breast_cancer)

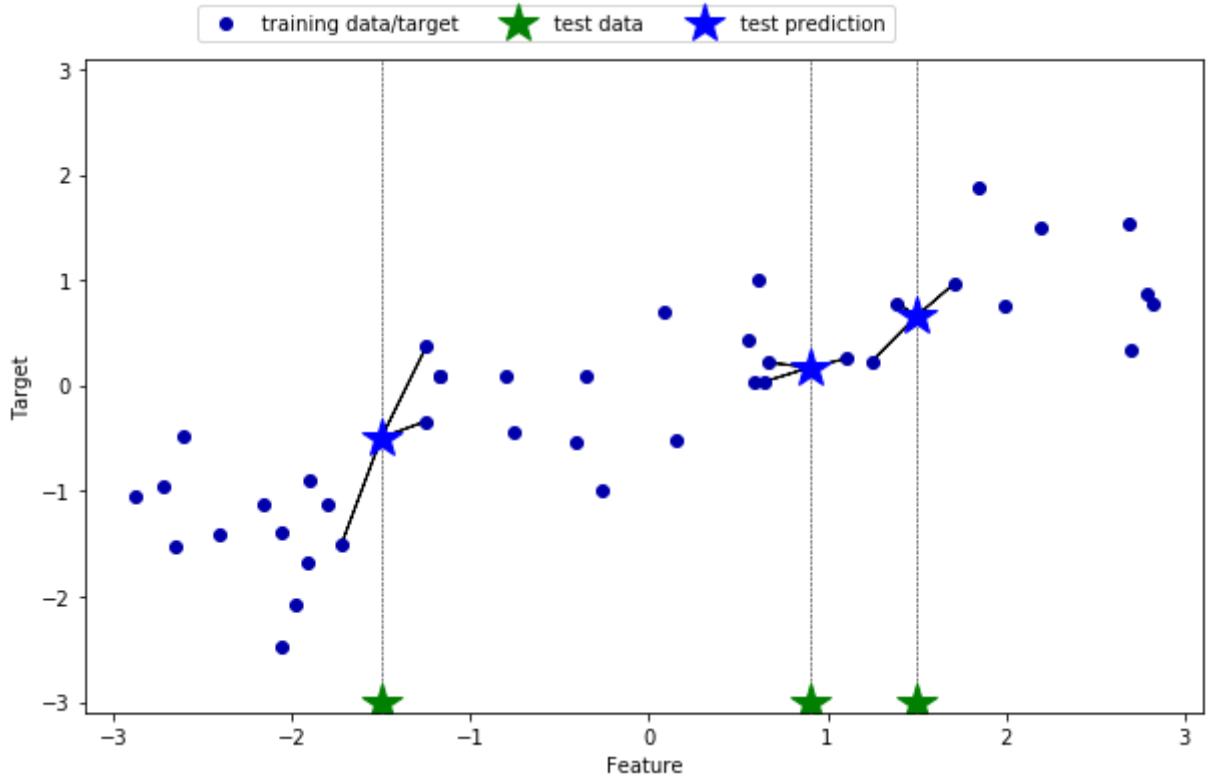


k-Neighbors Regression

for k=1: return the target value of the nearest neighbor



for k>1: return the *mean* of the target values of the *k* nearest neighbors



To do regression, simply use `KNeighborsRegressor` instead

```
X, y = mglearn.datasets.make_wave(n_samples=40)
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0
)
reg = KNeighborsRegressor(n_neighbors=3)
reg.fit(X_train, y_train)
```

Out[19]:

```
KNeighborsRegressor(algorithm='auto', leaf_size=30, metric='minkowski',
          metric_params=None, n_jobs=None, n_neighbors=3, p=2,
          weights='uniform')
```

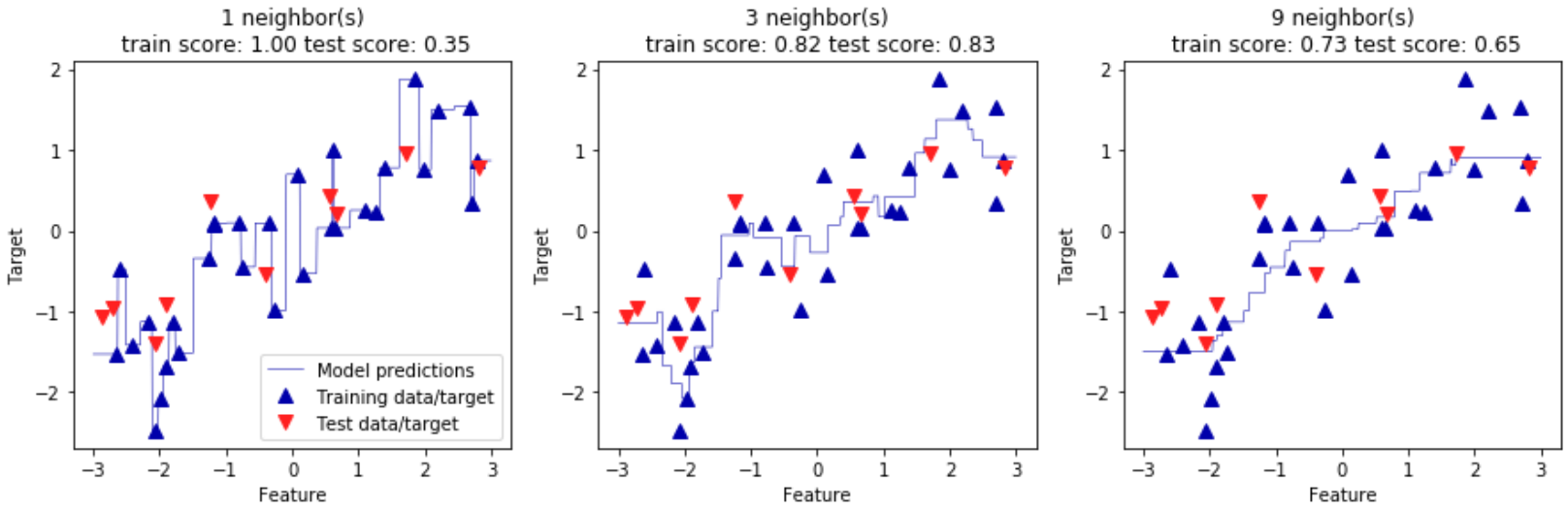
The default scoring function for regression models is R^2. It will be discussed later. the optimal value is 1. Negative values mean the predictions are worse than just predicting the mean.

```
Test set predictions:
[-0.054  0.357  1.137 -1.894 -1.139 -1.631  0.357  0.912 -0.447 -1.139]

Test set R^2: 0.83
```

Analysis

We can again output the predictions for each possible input, for different values of k.



We see that again, a small k leads to an overly complex (overfitting) model, while a larger k yields a smoother fit.

kNN: Strengths, weaknesses and parameters

- There are two important hyperparameters:
  - n_neighbors: the number of neighbors used
  - metric: the distance measure used
    - Default is Minkowski (generalized Euclidean) distance.
- Easy to understand, works well in many settings
- Training is very fast, predicting is slow for large datasets
- Bad at high-dimensional and sparse data (curse of dimensionality)

Summary

- We've covered the main machine learning concepts
- We used scikit-learn to build a first model
- We met our first algorithm (kNN)
- We learned to split the data in training and test sets
- We observed that there is always a balance between underfitting and overfitting
- In kNN, we can control and optimize this using the number of neighbors k