Introduction to Machine Learning

Hossein Pourreza

November 2018

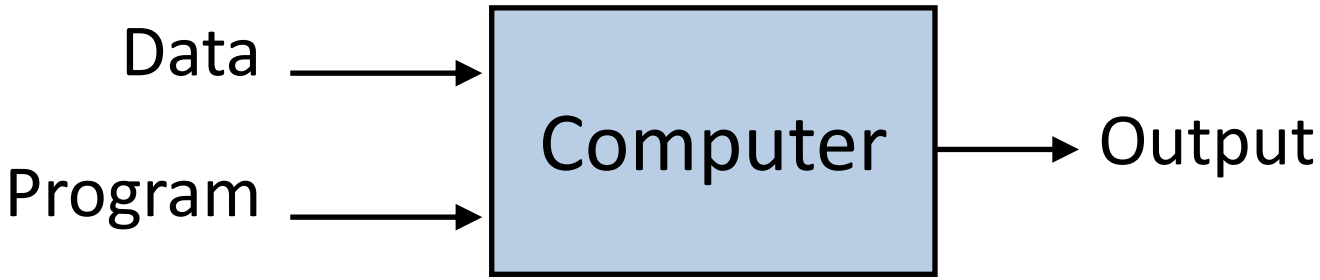
What is machine learning?

- A branch of artificial intelligence, concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data

- Every machine learning problem is basically an optimization problem
  - To find either a maximum or a minimum of a specific function

What is machine learning?

Traditional Programming



Machine Learning

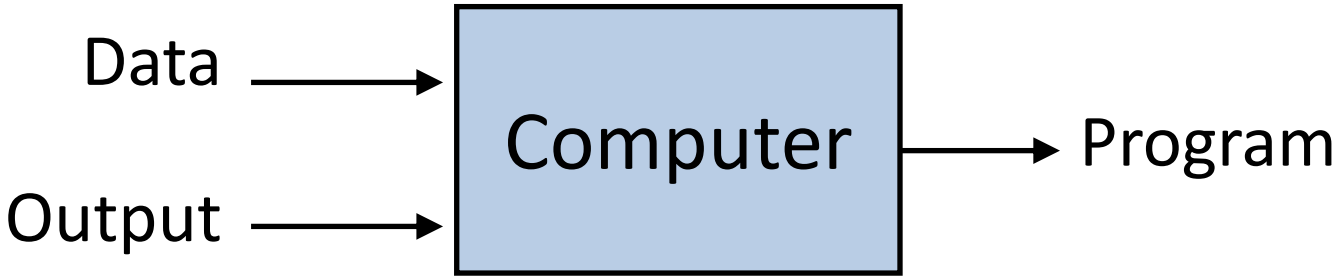


Figure is from https://courses.cs.washington.edu/courses/cse446/14wi/slides/intro.pdf

Machine learning applications

- Handwriting detection
- Image classification
- Spam filtering
- Fraud detection
- Market basket analysis

Data and machine learning

- In order to let a machine learn, you need to provide it with enough data
- Data has *features* used by the machine learning algorithm
  - E.g., columns of tabular data
- Selecting features correctly increases the learning accuracy

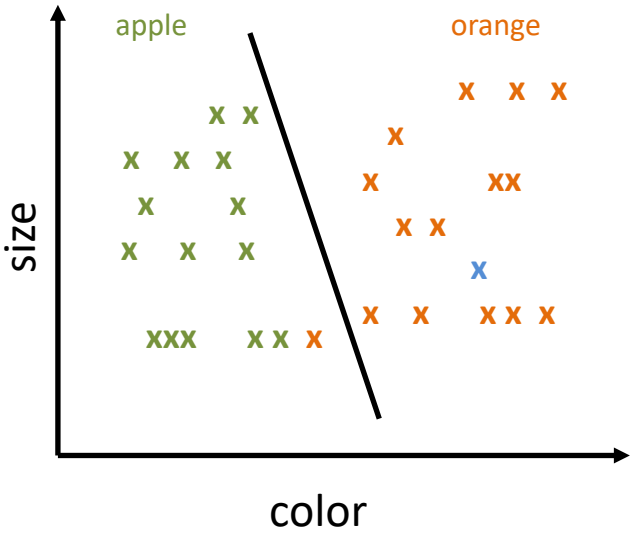
Types of learning

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Supervised learning

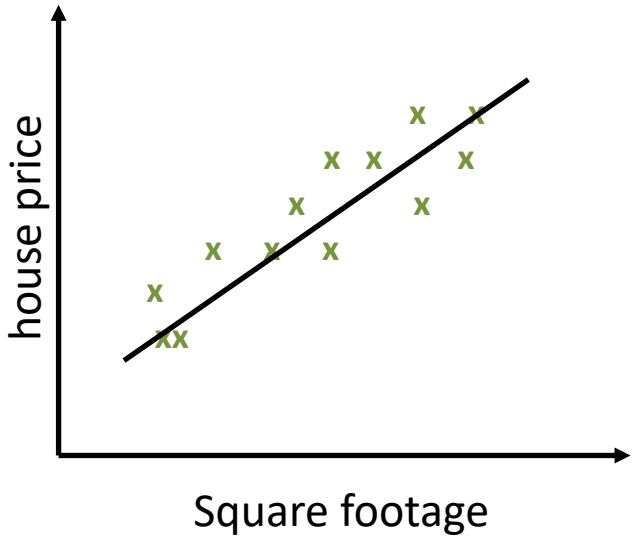
- Uses a *training set* including both features and desired output

Classification



Regression

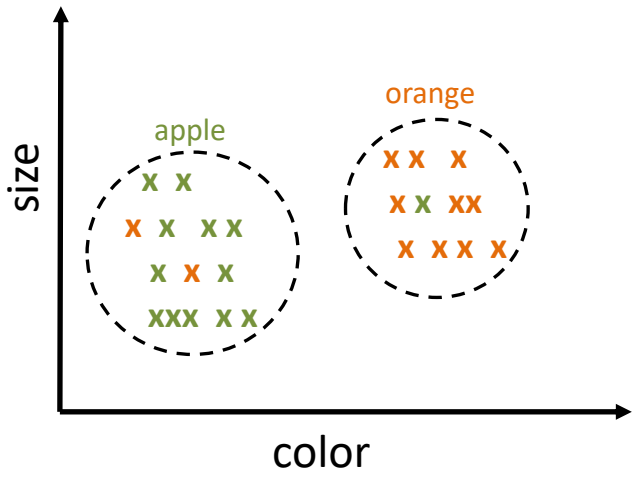
$y=A_0+A_1x$



Unsupervised learning

- There is no defined output and it learns what normally happens

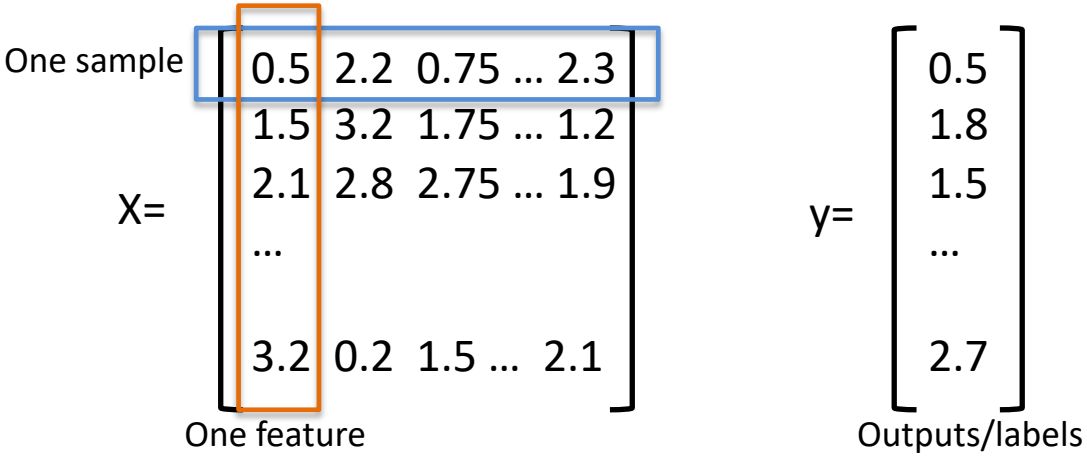
Clustering



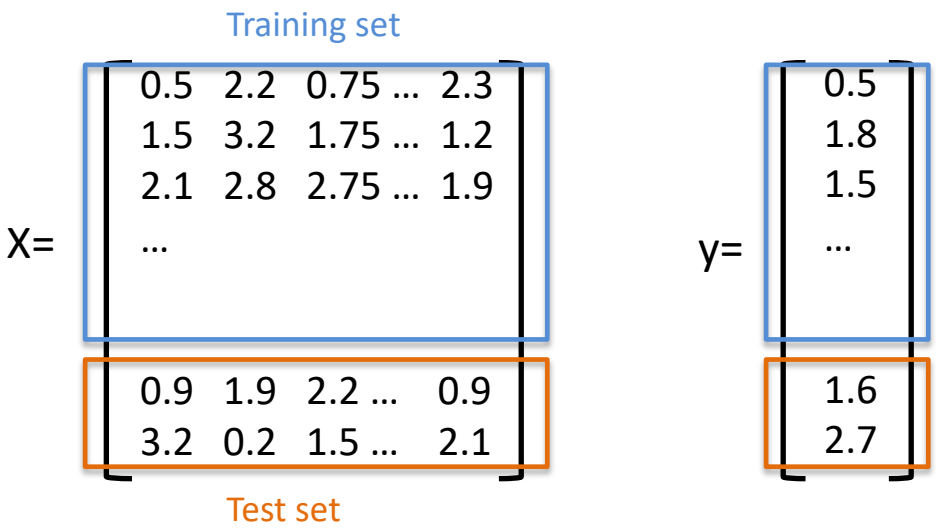
Representing data

- Data is usually represented as NxM matrix where N is the number of samples and M is the number of features

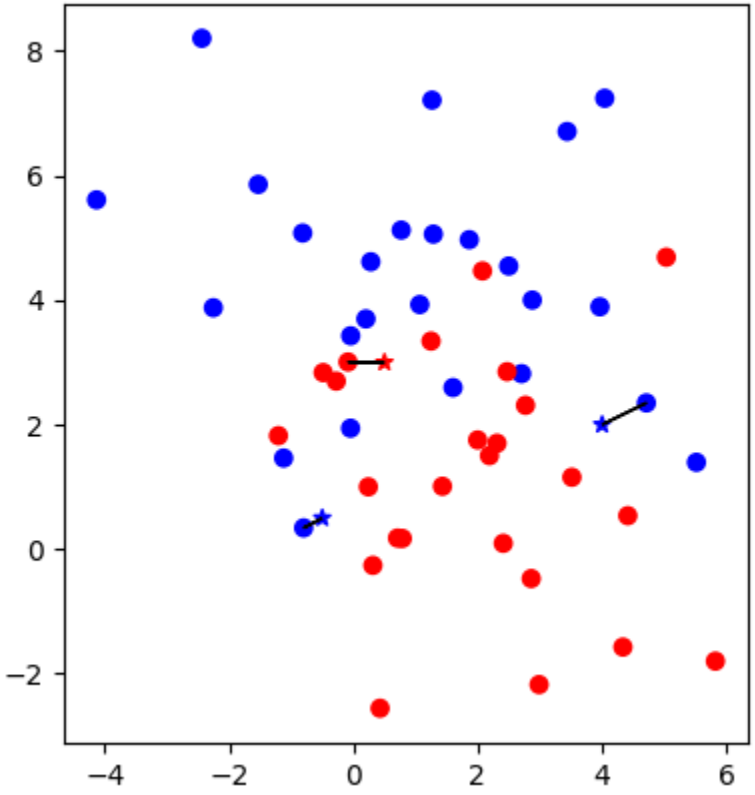
- Labels (outputs) are represented as a column vector



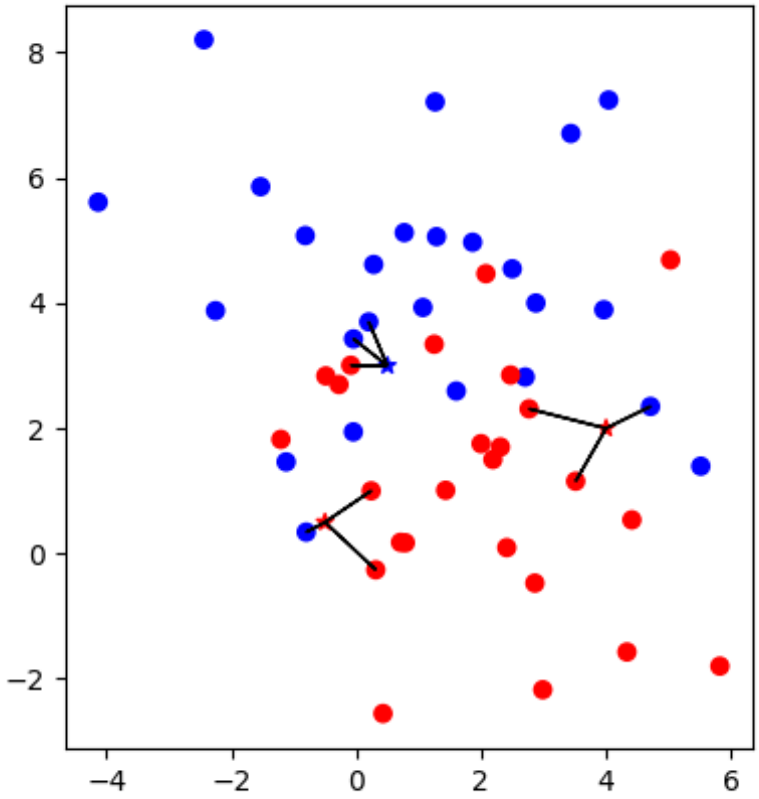
Training and test set



Nearest neighbor

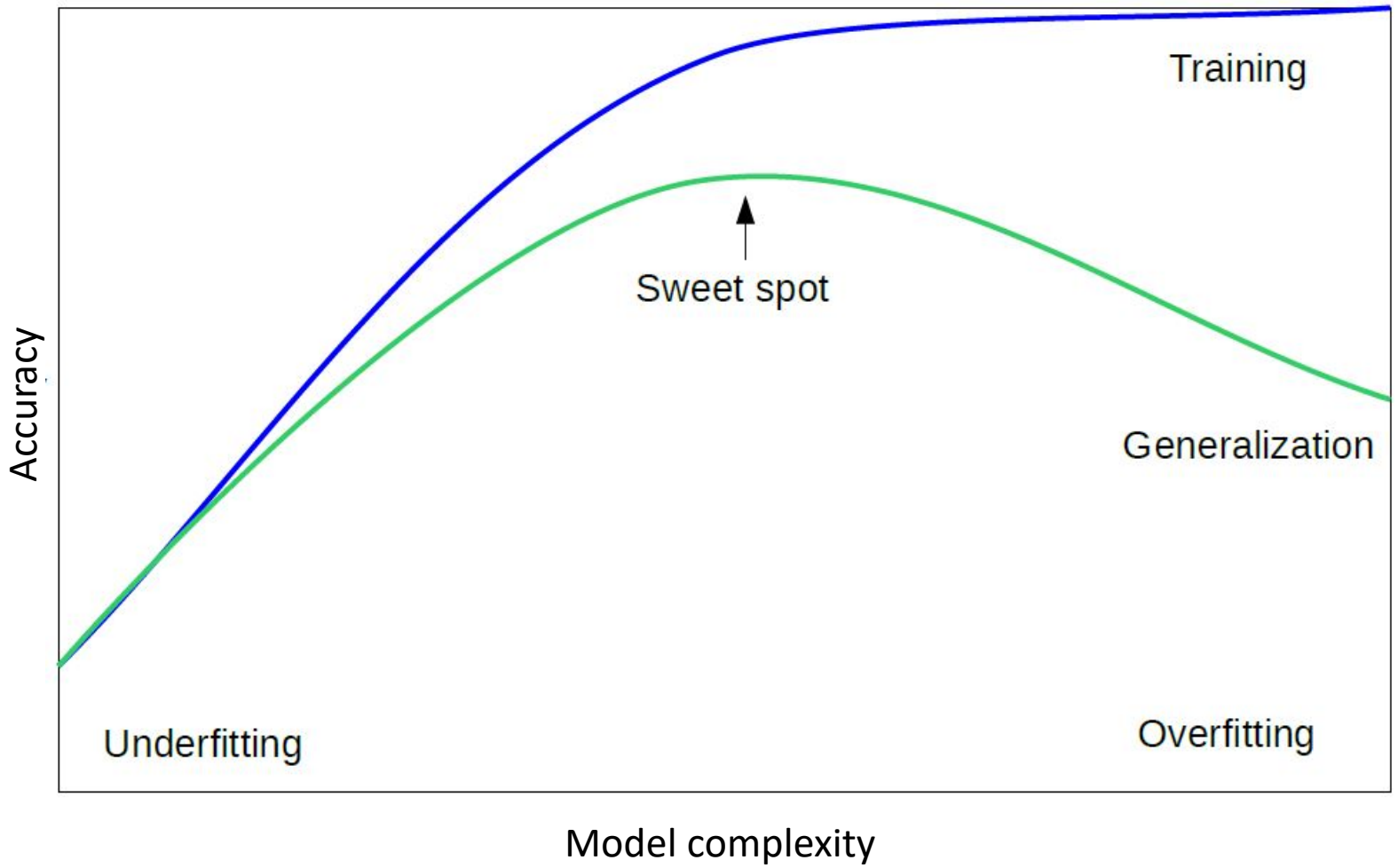


n_neighbors=1



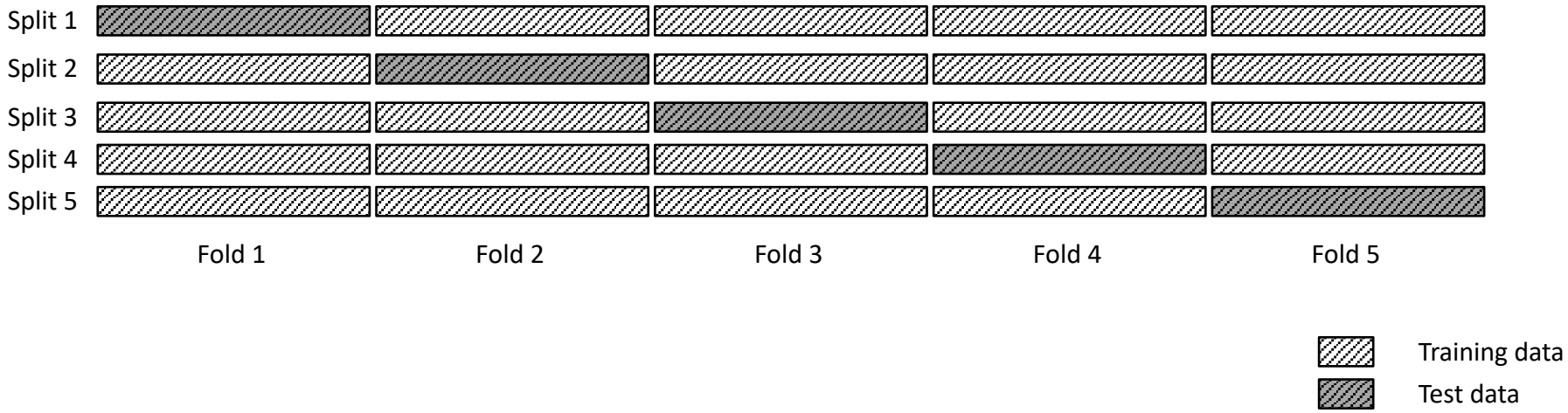
n_neighbors=3

Overfitting and underfitting



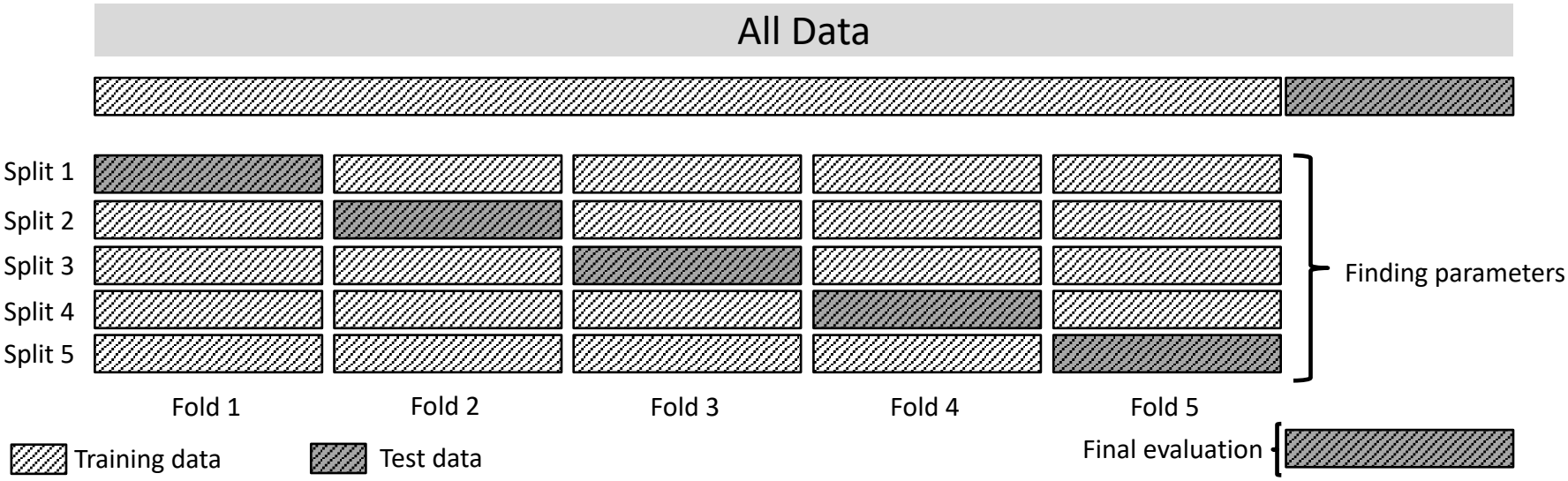
Cross validation

- In cross validation, you split your data into multiple folds, usually 5 or 10, and build multiple models
- For each of the splits of the data, you get a model evaluation and a score
- Better use of data but longer running time



Cross validation + test data

- Start out by splitting of the test data, then perform cross-validation on the training data
- Once the right setting of the parameters is found, re-train on the whole training set and evaluate on the test set
- The GridSearchCV function can perform CV+test



Preprocessing

- Consider the Boston housing dataset
  - The idea is to predict house prices based on a number of factors
  - Not all the factors have the same scale

- Some methods, e.g. KNeighborsRegressor, want data to be in the same scale

- Using StandardScaler, fit on training set, transform training set, fit KNeighborsRegressor on scaled data, transform test data, score scaled test data

- To scale, **always** fit on the training set and apply transform on both the training and the test set.

Categorical data

- Let's say you have three possible values for a given measurement for each setup
  - E.g., red, green, and blue

- You could try to encode these into a single real number, say 0, 1 and 2

- But, it imposes a linear relation between them, and in particular it defines an order between the categories

Categorical data

- A better way is to add one new feature for each category, and that feature encodes whether a sample belongs to this category or not.

- This method is called a **one-hot** encoding, because only one of the features is active at a time

	red	green	blue
Setup1	1	0	0
Setup2	0	1	0
Setup3	0	0	1

Machine learning in the cloud

- Companies such as Microsoft and IBM offer machine learning services in the cloud
- Easy to get started but works as a black box
- There could be some cost associated with using the model

Useful links and references

- https://github.com/rcc-uchicago/Workshops/tree/master/IntrotoML
- http://scikit-learn.org/stable/documentation.html
- https://github.com/amueller/ml-training-intro [1]
- http://www2.cs.uh.edu/~ceick/ML/ML09.html
- www.cs.washington.edu/446

[1] Slides are adopted from material in the repository