# O'Reilly Hands-On Machine Learnign with Scikit-Learn and TensorFlow

Philip Tracton

October 13, 2018

# Contents

# 1 Chapter 01 The Machine Learning Landscape

Machine Learning is the science and art of programming computers so they can *learn from data*

## 1.1 Definitions

- **training sets** are the examples the system uses to learn from.
- **training samples** are the samples or data in the training set
- **training data** is the new data used after training to see if it worked
- **data mining** is applying the techniques of machine learning to large amounts of data to discover patterns that were not immediately apparent
- **labels** training data fed to your algorithm that includes desired solutions
- **features** are an attribute and its value
- **agent** a reinforcement learning system it can observe environment, take actions, get rewards for good actions and penalties for bad ones
- **learning rate** how fast an algorithm adapts to changing data
- **similarity measure** is a method of seeing how close to samples are to each other.
- **utility function** is a measure of how good your functon is
- **cost function** is a measure of how bad your function is.
- **sampling noise** happens when there is too little data and you get non-representative data as chance
- **sampling bias** if the sampling method is flawed and leads to non-representative data.
- **Feature selection** is selecting the most useful feature of those available to train on
- **Feature Extraction** is combining one or more existing features into a single more useful feature (dimensionality reduction)

- **regularization** is constraining a model to make it simpler and reduce the risk of overfitting
- **hyperparameter** is a parameter of the learning algorithm, not the model, so it is not affected by training and must be set prior to training
- **generalization error** is the error rate on new cases. Done by evaluating model on the test set
- **cross validation** split the training set into complimentary subsets and each model is trained against a different combination of sets and validated against the remaining parts.
- **No Free Lunch** if you make no assumptions about the data there is no reason to prefer one model over others.

## 1.2 Concepts

### 1.2.1 Types of Machine Learning Systems

- Trained with human supervision
- Learn incrementally on the fly
- compare new data points to old data points and predict.

1. Supervised Learning
    - Used labelled training data
    - Typically used for classification tasks
    - Typically used for predicting target number. Given *features* it can go through a regression to predict new values.
    - Some Supervised Learning Algorithms in book
        - k-Neareset Neighbor
        - Linear Regression
        - Logisitic Regression
        - Support Vector Machines
        - Decision Trees and Random Forests
        - Neural Networks
2. Unsupervised Training

- Training data is unlabelled.
- Some important unsupervised learning algorithms
- Detect groups via clustering
- Reduce dimensionality to simplify data without loosing information
- Anomaly detection of finding outliers in data sets
- association rule learning is to dig into a large data set and discover interesting relations between attributes
- Visualization generate 2d or 3d representation of the data you feed it
  - Clustering
    - k-Means
    - Hierarchical Cluster Analysis
    - Expectation Maximization
  - Visualization and Dimensionality Reduction
    - Principal Component Analysis
    - Kernal PCA
    - Locally Linear Embedding
  - Association Rule Learning
    - Apriori
    - Eclat

3. Semi-Supervised Learning

   - partially labelled training data. Usually mostly unlabelled with some labelled data
   - Use a combination of supervised and unsupervised algorithms

4. Reinforcement Learning System

   - The agent (learning system) observes the environment and gets awards or penalties.
   - It must learn on its own the best strategy to maximize rewards and minimize penalties over time.

5. Batch Learning

   - System is incapable of learning over time and must be trained with with all available data

- This is called offline learning.

6. Online Learning
    - Incremental trainig by feeding it sequential data in small groups
    - Good for systems that receive a continuous flow of data
    - Can be used to train systems of huge data that do not fit into memory (out of core learning)
    - Incremental learning is a better name for this
    - bad data will cause system performance to decline over time
    - must manage learning rate, too fast will forget old information and too slow will be hard to adapt

### 1.2.2 Instance Based vs Model Based

- Good performance on training data is nice but true goal is good performance on new instances

1. Instance Based
    - The system learns examples by heart and generalizes to new cases using a similarity measure.
2. Model Based
    - Make a model from the examples and use that to make a prediction on new data samples.
    - Model selection can be a challenge.

### 1.2.3 Main Challenge of Machine Learning

1. Insufficient Quantity of Training Data
    - Need many thousands of examples to do this correctly.
2. Nonrepresentative Training Data
    - Model will behave based on training data. If it is not similar to production data, then the model will give poor results.
    - Be aware of sampling noise and sampling bias
    - Leads to inaccurate predictions

3. Poor Quality Data
    - If data is full of outliers, errors and noise, the algorithm will fail to detect underlying patterns
    - Worth time and effort of cleaning up data
        - May help to remove outliers
        - Fill in missing data? Fill in with what? Mean, Median, 0?
4. Irrelevant Features
    - Garbage In Garbage Out
    - This only works if you have enough relevant features and not too many irrelevant features.
    - Do feature selection
    - Do feature extration
5. Overfitting the Training Data
    - Model performs well on training data but fails to generalize on other data
    - This is over generalizing.
    - Happens when the model is too complex compared to the amount and noise of training data
    - Regularize your model
6. Underfitting the Training Data
    - Opposite of over fitting. Happens when you algorithm is too simple for the data
    - Fix with
        - More powerful algorithm
        - Better features
        - Reducing constraints

## 1.3 Testing and Validating

- Only way to tell if this works is to try on new cases
- Split your data into 2 set, training and testing.

- Common fix is to have a 3rd set of data, validation set.
- Process
  - Train many models an hyperparameters on the training data
  - Select the ones that perform the best for running with the validation set
  - Run a final test with the test data
- Use cross-validation

## 1.4 Exercises

### 1.4.1 How would you define Machine Learning?

### 1.4.2 Can you name 4 types of problems where it shines?

### 1.4.3 What is a labelled training set?

### 1.4.4 What are 2 most common supervised tasks?

### 1.4.5 Can you name 4 common unsupervised tasks?

### 1.4.6 What type of Machine Learning Algorithm would you use to allow a robot to walk in various unknown terrains?

### 1.4.7 What type of algorithm would you use to segment your customers into multiple groups?

### 1.4.8 Would you frame the problem of spam detection as a supervised or unsupervised learning problem?

### 1.4.9 What is online learning?

### 1.4.10 What is out of core learning?

### 1.4.11 What type of algorithm relies on a similarity measurement to make predicitions?

### 1.4.12 What is the difference between a model parameter and a learning algorithm's hyperparameter?

### 1.4.13 What do model based learning algorithms search for? What is the most common strategy they use to succeed? How do they make predictions?

### 1.4.14 Can you name 4 of the main challenges in Machine Learning?

### 1.4.15 If your model performs great on training data, but generalizes poorly to new instances what is happening? Can you name 3 possible solutions?

### 1.4.16 What is a testing set and why would you want to use it?

### 1.4.17 What is the purpose of a validation set?

### 1.4.18 What can go wrong if you tune hyperparameters using the test set?

### 1.4.19 What is cross-validation and why would you prefer it to a validation set?