Foundations of Machine Learning

ION PETRE
DEPARTMENT OF MATHEMATICS AND STATISTICS
UNIVERSITY OF TURKU
SPRING 2020

http://etc.ch/ttKb

3

- **What is machine learning?**
  - Computational methods using experience to improve performance or to make accurate predictions
  - Experience: past information/data available to the learner
  - Crucial aspect: quantity and quality of the data

4

- **Examples of machine learning problems**
  - Predict the topic of an unseen document (e.g., Google News)
  - Predict if an email is spam or not
  - Predict the price of a house depending on its size, number of bedrooms, bathrooms, distance to the local school, distance to the nearest supermarket, and criminality rate of a 5km-radius neighborhood
  - Predict the grade of a breast cancer tumor based on a CT scan: size of tumor, thickness, but also age of patient, family cancer history
  - Predict a new product that a shopper may want to buy (customer segmentation, recommender systems)

5

- Machine learning
  - Design efficient and accurate predictions algorithms
  - Sample complexity: evaluate the sample size and quality required for a good prediction
  - Statistical tools for estimating the sample complexity and for estimating the accuracy of the algorithm predictions
  - Mathematical tools to solve the optimization problems behind machine learning

6

Standard learning tasks

- **Classification**: assign a category to each item
  - Examples: document classification, tumor classification, image recognition.
- **Regression**: predict a real value for each item
  - Examples: predict stock values, house prices, duration-to-relapse.
- **Ranking**: learn the order of items
  - Example: web search, online translations, spell-check correctors.
- **Clustering**: partition a set of items into homogenous subsets
  - Example: identify communities in social networks, group tumors together depending on their genetic profile
- **Dimensionality reduction (or manifold learning):** transform an initial representation of items into a lower-dimensional representation while preserving some properties of the initial representation
  - Example: feature extraction

7

Learning stages (1)

Running example: spam detection

- **Examples**: Instances of data used for learning
  - In our case: collection of emails categorized as spam or non-spam
- **Features**: the set of attributes associated to an example
  - In our case: length of message, name of sender, some characteristics of the header, certain links in the message, the nature of the attachments, etc.
- **Labels**: values or categories assigned to examples
  - In our case: SPAM and NO-SPAM
- **Hyperparameters**: free parameters not to be learned by the algorithm, but rather customizing the algorithm (e.g., model selection)

8

Learning stages (2)

Running example: spam detection

- **Training samples**: examples used to train a learning algorithm
  - In our case: a set of emails together with their spam/no-spam labels
- **Validation samples**: examples used to tune the hyperparameters of a learning algorithm when working with labeled data. Distinct from the training sample set
  - The prediction of the current learning algorithm is compared with the real value of the validation samples. This gives a suggestion on potentially better values for the hyperparameters
  - In our case: another set of labeled emails

9

Learning stages (3)

Running example: spam detection

- **Test samples**: examples used to evaluate the performance of the proposed learning algorithm. Distinct from the training and validation sample sets
  - The prediction of the final learning algorithm is compared with the real value of the validation samples. This gives a measure of the accuracy rate of the algorithm on new samples
  - In our case: yet another set of labeled emails

10

Learning stages (4)

Running example: spam detection

- **Loss function**: a function that measures the difference (or loss) between a predicted label and the true label of a sample.
  - In our case: a 0/1 (or true/false) function – was the prediction correct?
- **Hypothesis set**: a set of parametric functions mapping features to the set of labels
  - The purpose of the learning algorithm is to identify the best hypothesis in this set (based on the available samples), evaluate its accuracy, and offer it to be used on unseen samples

The machine learning process

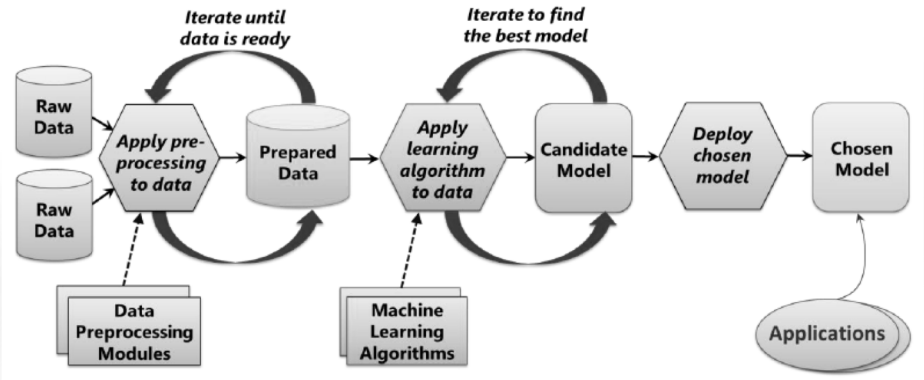

Figure 3: The machine learning process starts with raw data and ends up with a model derived from that data.

Source: D.Chappel, Introducing Azure Machine Learning. 2015

Learning scenarios (1)

- **Supervised learning**: the learner has access to a set of labeled examples as training data and makes predictions on unseen points
  - Classes of problems: classification, regression, ranking problems
  - Examples: spam detection, tumor grade prediction, predict house prices.
- **Unsupervised learning**: the learner has access only to unlabeled training data and makes predictions on unseen points
  - Classes of problems: clustering, dimensionality reduction
  - Examples: community detection, feature selection

Learning scenarios (2)

- **Semi-supervised learning**: the learner has access both to a (small) set of labeled examples and a (large) set of unlabeled examples as training data and makes predictions on unseen points
  - The hope is that the distribution of unlabeled data can help achieve a better performance than in supervised learning
  - Examples: speech analysis, protein sequence classification, web content classification.
- **Transductive learning**: the learner has access both to a (small) set of labeled examples and a (large) set of unlabeled examples as training data and aims to label the unlabeled data points
  - Examples: speech analysis, protein sequence classification, web content classification

Learning scenarios (3)

- **On-line learning**: multiple rounds where training and testing phases are intermixed.
  - Each round: the learner receives an unlabeled point, makes a prediction, receives the true label, and incurs a loss
  - Objective: minimize the cumulative loss over all rounds (or minimize the "regret")
  - Note: instances and their labels may be chosen adversarially in this scenario

Learning scenarios (4)

- **Reinforcement learning**: training and testing phases are intermixed.
  - To collect information, the learner actively interacts with the environment and in some cases it affects the environment, and receives an immediate reward for each action
  - Objective: maximize the reward over a course of actions and interactions with the environment
  - No long-term feedback is provided by the environment
  - **The exploration vs. exploitation dilemma**: choose between exploring unknown actions to gain more information vs. exploiting the information already collected

Learning scenarios (5)

- **Active learning**: the learner adaptively or interactively collects training examples by querying an oracle to request labels for new points
  - Objective: achieve a performance comparable to the standard supervised learning scenario (or passive learning), but with fewer labeled examples
  - Often used when labels are expensive to get, e.g., computational biomedicine

Generalizations (1)

- Machine learning is fundamentally about generalizations
- The typical formulation: select a function out of a hypothesis set, i.e., a subset of all functions. The selected function is proposed as the learned algorithm/predictor, to be used on all unseen points
- Question: how should a hypothesis set be chosen?
  - Rich/complex hypothesis set: the learner may choose a function that is highly consistent with the training samples
  - Less complex hypothesis set: more errors on the training samples are unavoidable
  - Which scenario is better?

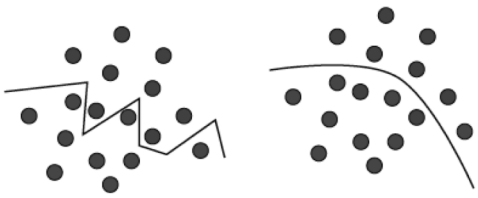

Figure 1.2
The zig-zag line on the left panel is consistent over the blue and red training sample, but it is a complex separation surface that is not likely to generalize well to unseen data. In contrast, the decision surface on the right panel is simpler and might generalize better in spite of its misclassification of a few points of the training sample.

Complex vs. simpler hypothesis set. Source: Mohri et al. Foundations of Machine Learning 2nd edition, MIT Press, 2018.

Generalizations (2)

19

- A predictor chosen from a very complex family can essentially memorize the data. The generalization on unseen points may be very poor
  - Lagrange interpolator: for a set of n+1 points, there is a polynomial of degree n going through each of those points.
- Trade-off between sample size and complexity
  - Small sample size and a predictor from a family that is too complex may lead to poor generalizations. **Overfitting**!
  - Larger sample size and a predictor from a family that is too simple may lead to poor accuracy. **Underfitting**!

This course: Foundations of Machine Learning

20

- **Learning objectives**. After the course the students will be familiar with the mathematical formulations of and the mathematical solutions to the following four central machine learning problems:
  - Regression
  - Dimensionality reduction
  - Density estimation
  - Classification
- **Target audience**
  - Mathematics students: open up the field of machine learning research in a way that is anchored in detailed mathematical setups
  - Computer science/engineering: unveil the mathematical setup behind some of the most popular machine learning techniques

Resources

- **Course website**: https://users.utu.fi/ionpet/foundations-of-machine-learning/
- **Course textbooks**
  - Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. Mathematics for Machine Learning. Cambridge University Press, 2020.
  - Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of Machine Learning. MIT Press, Second Edition, 2018.
- **Additional reading**
  - Avrim Blum, John Hopcroft, and Ravindran Kannan. Foundations of Data Science. Cambridge University Press, 2020.

Relation to some of the other courses in the university

- Background in linear algebra, multi-variate calculus, probabilities, statistics needed
  - For those who do not have it, or need a refresher, use as background reading the two course textbooks
  - The essential parts of it to be explained also in the lectures
- Courses on machine learning and data science in computer science
  - Typically those courses go relatively quickly over the mathematical details and focus more on the programming/engineering part
  - This course does the opposite: goes in details over the mathematical part of the field and only brushes on the engineering part
- Courses on statistical learning
  - This course is a broader presentation of the mathematics of machine learning

Course structure

- Lectures (14): 5.3 – 24.4.2020
  - Thursdays 14-16 room XVII
  - Fridays 12-14 room XVII
- Project
  - Optional: to carry up to 15 points towards the final exam
  - Goal: do a machine learning project on a real data set
  - Method: Microsoft Azure Machine Learning Studio (Classic)
  - Meetings: every week on Wednesdays 10-12, room M2
  - Presentations, project reports: compulsory; to be discussed later
- Final exams: May 4, May 25
  - Registration one week prior, at the latest