Machine Learning: Supervised Methods NOTES

Kristian Bonnici

October 1, 2021

Contents

Part I

Theory

1 Introduction

1.1 Theoretical paradigms

Theoretical paradigms for machine learning **differ** mainly on what they assume about the process generating the data:

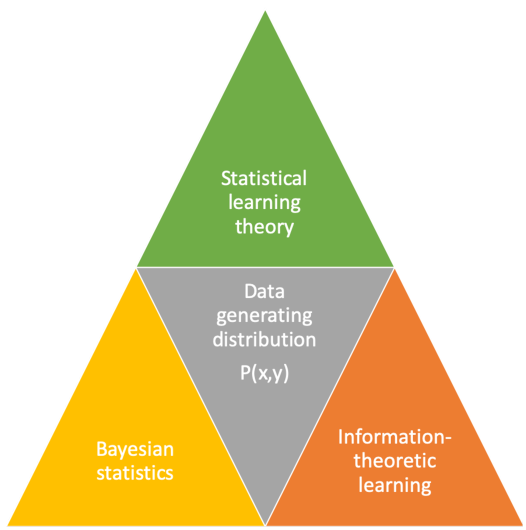

Figure 1: paradigms for data generation distributions.

- **Statistical learning theory (focus on this course):** assumes data is i.i.d from an unknown distribution P(x), does not estimate the distribution (directly)
- **Bayesian Statistics:** assumes prior information on P(x), estimates posterior probabilities

- **Information theoretic learning:** (e.g.Minimum Description Length principle, MDL): estimates distributions, but does not assume a prior on P(x)

1.2 Dimensions of a supervised learning algorithm

1. **Training sample:** $S = \{(x_i, y_i)\}_{i=1}^m$ the training examples $(x, y) \in X \times Y$ independently drawn from a identical distribution $(i.i.d)D$ defined on $X \times Y, X$ is a space of inputs, Y is the space of outputs.
2. **Model or hypothesis:** $h : X \to Y$ that we use to predict outputs given the inputs x.
3. **Loss function:** $L : Y \times Y \to \mathbb{R}, L(...) \geq 0, L(y, y')$ is the loss incurred when predicting y' when y is true.
4. **Optimization** procedure to find the hypothesis h that minimize the loss on the training sample.

1.3 Classification (Task 1/3)

Problem: partitioning the data into pre-defined classes by a *decision boundary* or *decision surface*.

Multi-class classification: more than two classes

- **Multi-label Classification:** An example can belong to multiple classes at the same time
- **Extreme classification:** Learning with thousands to hundreds of thousands of classes (Prof. Rohit Babbar @ Aalto)

1.3.1 Version space

Version space: the set of all consistent hypotheses of the hypothesis class

- **Consistent hypothesis:** if correctly classifies all training examples
- **In version space:**
  - **Most general hypothesis G:** cannot be expanded without including negative training examples
  - **Most specific hypothesis S:** cannot be made smaller without excluding positive training points

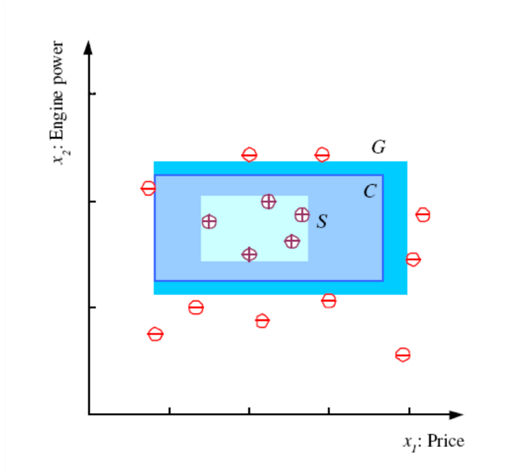

Figure 2: Illustration of a Version Space.

- Intuitively, the **"safest" hypothesis** to choose from the version space is the one that is furthers from the positive and negative training examples → maximum margin
  - Margin = minimum distance between the decision boundary and a training point

1.4 Regression (Task 2/3)

Problem: output variables which are numeric.

1.5 Ranking & preference learning (Task 3/3)

Problem: predict a ordered list of preferred objects.

Training data (typically): pairwise preferences.

- e.g. user x prefers movie y_i over movie y_j

Output: ranked list of elements.

1.6 Generalization

Aim: predict as well as possible the outputs of future examples, not only for training sample.

We would like to *minimize* the **generalization error**, or the **(true) risk**:

$$R(h) = E_{(x,y)\sim D}[L(h(x), y)] \tag{1}$$

Where:
D : Unknown distribution where from training and future examples are drawn from (i.i.d assumption)

What can we say about $R(h)$ based on training examples and the hypothesis class H alone? Two possibilities:

- Empirical evaluation by testing (Section 1.6.1)
- Statistical learning theory (Section 2)

1.6.1 Model evaluation by testing

What: estimate the model's ability to generalize on future data

How: approximating true risk by computing the empirical risk on a independent test sample:

$$R_{test}(h) = \sum_{(x_i,y_i)\in S_{test}}^{m} L(h(x_i), y_i)$$

- The expectation of $R_{test}(h)$ is the true risk $R(h)$

1.7 Hypothesis classes

There is a huge number of different **hypothesis classes** or **model families** in machine learning, **e.g:**

- **Linear models** such as logistic regression and perceptron
- **Neural networks:** compute non-linear input-output mappings through a network of simple computation units
- **Kernel methods:** implicitly compute non-linear mappings into high-dimensional feature spaces (e.g. SVMs)
- **Ensemble methods:** combine simpler models into powerful combined models (e.g. Random Forests)

Each have their different pros and cons in different dimensions (accuracy, efficiency, interpretability); No single best hypothesis class exists that would be superior to all others in all circumstances

---

2 Statistical Learning Theory

What: Theoretical background on machine learning.

Goal: Generalization (Section 1.6)

2.1 Probably Approximately Correct (PAC) learning

What: *Theoretical framework* that formalizes the notion of generalization in machine learning.

Ingredients:

- **input space** X containing all possible
- **inputs** x * set of possible **labels** Y (in binary classification $Y = \{0, 1\}$)

Goal: to learn a hypothesis with a low generalization error

$$R(h) = E_{x \sim D}[L_{0/1}(h(x), C(x))] = Pr_{x \sim D}(h(x) \neq C(x))$$