Optimization Methods In Machine Learning

Lecture 2: Finding a “good” hypothesis

Professor Katya Scheinberg

Lehigh University

Spring 2016

Outline

This lecture is taken from a short course at UT Austin
taught by N. Srebro and K. Scheinberg in 2011.

Outline

Ideal case

Consider we have sample space of X and space of all possible labels Y. We are trying to find a predictor such that minimize the expected loss.

- Assume that we have complete knowledge of the true source joint distribution $p_{X,Y}(x, y)$.
- Also, we have chosen loss function $\mathbf{loss}(\cdot, \cdot)$.

Question: How does actual practice differ from this ideal setting?

- We do not ever have complete knowledge of the true source joint distribution.

Outline

Sampling

In real world, we just can sample from our population $\mathcal{X}$. In order to move forward with the process of learning a predictor, we need to make some assumptions about how this sample data set is drawn.

- It is possible to make various assumptions about the sampling of the data set.
- We will consider **Statistical Learning Theory**.

We assume:

- we are given a particular observed sample data set s of m (input, label) pairs, written $s = \{(x_1, y_1), \ldots, (x_m, y_m)\}$.
- each point is an observation of the joint random variables (X_i, Y_i).
- They are independently and identically distributed (i.i.d.) according to the source joint distribution $p_{X,Y}(x, y)$.
- i.e. each random variable pair (X_i, Y_i) is sampled independently according to $p_{X,Y}$":

$$(X_i, Y_i) \underset{\text{ind.}}{\sim} p_{X,Y}$$

1

[1] Often, the sample data set is *not* drawn from the same distribution that we will measure our expected loss on. **But** any type of analysis of machine learning methods assumes that the sample data set *is* drawn from the same distribution that we use to measure the error

Choosing a predictor

- Assume that we have a loss function that characterizes what we care about.
- The process of learning from a sample data set is a mapping from a particular observed sample data set s and a loss function $\mathbf{loss}(\cdot,\cdot)$ to a predictor h.

$$[s, \mathbf{loss}(\cdot,\cdot)] \mapsto h.$$

- Learning algorithm takes a loss function and a sample data set of labeled examples and returns a predictor.[2]

[2] This is what we consider in this course. For this course, we will primarily be considering just simple supervised learning in which we would like to find a good predictor based on a labeled sample data set.

Empirical risk

- We want to find a predictor that minimizes the expected loss on the true source joint distribution, but ...
- We do not have complete knowledge of the true source joint distribution.
- We should choose our predictor to minimize the expected loss on what we do have access to.
- We hope that this predictor will do well on the true source joint distribution.
- The expected loss of a predictor h on a particular observed sample data set $s = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ could also be referred to as the *empirical risk* $\hat{R}_s[h(\cdot)]$

$$\hat{\mathbb{E}}_s[\mathbf{loss}(h(X), Y)] = \hat{R}_s[h(\cdot)] = \frac{1}{m}\sum_{i=1}^{m}\mathbf{loss}(h(x_i), y_i)$$

Outline

Empirical risk minimizer

- As the first step of Empirical Risk Minimization we choose some hypothesis class $\mathcal{H}$.
- We view $\mathcal{H}$ as a set of predictors, written as

$$\mathcal{H} = \{h \;:\; \mathcal{X} \to \mathcal{Y}\} \subseteq \mathcal{Y}^{\mathcal{X}}$$

- The Empirical Risk Minimization learning rule can be written as:

$$\mathbf{ERM}_{\mathcal{H}}(s) = \hat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_s(h) = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \mathbf{loss}(h(x_i), y_i)$$

Example of hypothesis classes

- Consider binary labels, so the label set is $\mathcal{Y} = \{+1, -1\}$. In addition, consider $\mathcal{X} = \mathbb{R}^2$.
- Some specific examples of hypothesis classes:

$$\begin{aligned}
\mathcal{H} &= \{x_i \geq \theta \mid i \in \{1,2\}, \theta \in \mathbb{R}\} \\
\mathcal{H} &= \left\{x \mapsto \mathbf{sign}\left(w^T x + b\right) \mid w \in \mathbb{R}^2, b \in \mathbb{R}\right\} \\
\mathcal{H} &= \left\{\sum_{i=1}^{2} x_i \leq \theta \mid \theta \in \mathbb{R}\right\}
\end{aligned} \tag{1}$$

- We can use any features of x_1 and x_2 to make a new class of hypothesis.

$$\mathcal{H} = \{\phi(x_i) \geq \theta \mid i \in \{1,2\}, \theta \in \mathbb{R}\} \tag{2}$$

- In a learning problem, we limit ourselves only to hypotheses in a certain class.
- We want to make the difference of expected risk and empirical risk, as small as possible.

Outline

Analysis of ERM

- Consider a specific predictor h.
- This predictor has an expected 01 loss $R_{01}(h)$ with respect to the true source joint distribution $p_{X,Y}(x, y)$.
- We only have access to an *estimate* of the expected 01 loss of the predictor h, $R_{01}(h)$, in the form of the sample average 01 loss $\hat{R}_{s,01}(h)$ of the predictor.
- There are many samples of size m that can be drawn from the true source distribution.
- If the sample average 01 loss of the predictor h is usually close to the expected 01 loss of the predictor h, then we can feel more confident about using the sample average 01 loss in place of the expected 01 loss.

Hoeffding's Inequality

- Let $T_1, \ldots, T_m$ be *independent* scalar random variables.
- Assume further that the T_i are *bounded* so that $T_i \in [a_i, b_i]$.
- Then, for the empirical mean of these m bounded variables,

$$\overline{T} = \frac{1}{m} \sum_{i=1}^{m} T_i,$$

- We have the inequality:

$$\mathbb{P}\left\{\left|\overline{T} - \mathbb{E}\left[\overline{T}\right]\right| \geq \varepsilon\right\} \leq 2\exp\left(-\frac{2\varepsilon^2 m^2}{\sum_{i=1}^{m}\left(b_i - a_i\right)^2}\right)$$

Comparing Expected Risk and Empirical Risk

- Let define the variables for our problem:

$$
\begin{aligned}
T_i &= \mathbf{loss}_{01}\left(h\left(X_i\right), Y_i\right) \\
a_i &= 0 \qquad\qquad \text{for } i \in \{1, 2\} \\
b_i &= 1
\end{aligned}
\tag{3}
$$

- Now we have:

$$
\begin{aligned}
\mathbb{E}(\hat{R}_{01}(h(x)) &= \mathbb{E}(\frac{1}{m}\sum_{i=1}^{m}\mathbf{loss}_{01}\left(h\left(X_i\right), Y_i\right)) \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}(\mathbf{loss}_{01}\left(h\left(X_i\right), Y_i\right)) \\
&= \frac{1}{m}\sum_{i=1}^{m}R_{01}(h(x)) \\
&= R_{01}(h(x)
\end{aligned}
\tag{4}
$$

- The expectation of empirical loss is expected loss.

Comparing Expected Risk and Empirical Risk

- Use the Hoeffding's Inequality:

$$\mathbb{P}\left\{\left|R_{01}(h(.)) - \hat{R}_{01}(h(.))\right| \geq \varepsilon\right\} \leq 2\exp\left(-2\varepsilon^2 m\right) \tag{5}$$

- If we have a sample with size m, the probability of the difference between expected loss and empirical loss be less than ϵ is:

$$\mathbb{P}\left\{\left|R_{01}(h(.)) - \hat{R}_{01}(h(.))\right| < \varepsilon\right\} \geq 1 - 2\exp\left(-2\varepsilon^2 m\right) = 1 - \delta \tag{6}$$

- So the sample size that grantees the accuracy of ϵ with probability of $1 - \delta$ can be found by considering this equation:

$$\epsilon = \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \tag{7}$$

Comparing Expected Risk and Empirical Risk (Cont.)

- So we can rewrite the inequality as follows:

$$\mathbb{P}\left\{\left|R_{01}(h(.)) - \hat{R}_{01}(h(.))\right| < \sqrt{\frac{\log \frac{2}{\delta}}{2m}}\right\} \geq 1 - \delta \tag{8}$$

- This inequality tells us just by changing the sample size, we can control the accuracy.

- What does this mean? What does the probability δ mean? How should we understand this inequality?

ERM (Empirical Risk Minimization).

- We want to find a hypothesis that minimizes the *expected risk*. However, as discussed, we can only find a hypothesis that minimizes empirical risk. Call such a hypothesis $\hat{h}$.

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \{\hat{R}\,(h(.)) = \frac{1}{m} \sum_i \mathbf{loss}\,(h(x_i), y_i)\} \tag{9}$$

- We produce a sample set and then search for the $\hat{h}$ that minimizes the empirical risk.
- What can we say about $\hat{h}$? How "good" is it?

Analysis of ERM. Flawed Version!

- Let's take a close look to the inequality (8). With probability of $1-\delta$ we have:

$$R_{01}\left(h(.)\right) < \hat{R}_{01}\left(h(.)\right) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \tag{10}$$

- This means the expected risk cannot be larger than the empirical risk plus an error. That is exactly what we want. Can we use it?

Analysis of ERM. Flawed Version!

- The best hypothesis of the class $\mathcal{H}$ is the following:

$$h^* = \arg\min_{h \in \mathcal{H}} \{R\left(h(.)\right) = E\left(\mathbf{loss}\left(h(x), y\right)\right)\} \tag{11}$$

- According to 10 and 11, we can find the following relationship:

$$R_{01}\left(h^*(.)\right) \leq R_{01}\left(h(.)\right) < \hat{R}_{01}\left(h(.)\right) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \tag{12}$$

- And based on (8) and (9), we have:

$$\hat{R}_{01}\left(\hat{h}(.)\right) \leq \hat{R}_{01}\left(h^*(.)\right) \leq R_{01}\left(h^*(.)\right) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \tag{13}$$

Analysis of ERM. Flawed Version!

- By using inequalities 10 and 13, following inequality will achieved:

$$R_{01}\left(\hat{h}(.)\right) < \hat{R}_{01}\left(\hat{h}(.)\right) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}} \leq R_{01}\left(h^*(.)\right) + 2\sqrt{\frac{\log \frac{2}{\delta}}{2m}} \tag{14}$$

- We now have shown that expected risk of our ERM $R_{01}\left(\hat{h}(.)\right)$ is not very different from expected risk of the expected risk minimizer h^*, which is the best we can hope for. The difference gets smaller if the sample set gets bigger.

- Sounds perfect!! But is it wrong!! Why?

Example

- Consider a problem with following data set X and labels Y:

$$X = \{\text{“People in the US”}\} \quad Y = \{\text{“Male”,”Female”}\} \tag{15}$$

- There is four kinds of different hypothesis classes:

$$\begin{aligned} H_b &= \{\text{“Predictor based only on month and day of birth”}\} \\ H_n &= \{\text{“Predictor based only on nationality}''\} \\ H_m &= \{\text{“Predictor based only on length of hair”}\} \\ H_p &= \{\text{“Predictor based only on last four digits of phone}''\} \end{aligned} \tag{16}$$

- Which of these hypothesis is the best?
- What is wrong with $2\sqrt{\frac{\log \frac{2}{\delta}}{2m}}$?