SYS 6016/4582: Machine Learning		**Spring 2017**
Lecture 2		
Instructor: Quanquan Gu	*Date: Jan 24th, 2017*	*Scriber: Xiyuan Ge*

Topics covered in the last class:
1) What is machine learning?
2) The difference and relationship between statistics and machine learning.
3) Three foundations of machine learning: statistics, optimization and information theory.
4) The concepts of expected risk, empirical risk as well as empirical risk minimization.
5) Decomposition of excess risk into estimation error and approximation error terms.

Before continuing with the topic of empirical risk minimization framework, we would like to discuss some subtle cultural difference between statistics and machine learning. In statistics, researchers typically start from looking at a dataset from other scientific disciplines such as biology and neuroscience. With the dataset, statisticians continue to abstract out the mathematical/statistical problem, and then invent new statistical models/procedures to analyze/interpret the underlying meaning in the dataset.

Regarding publications, statisticians generally emphasize on journal publications. Top-tier journals include Annals of Statistics, Journal of the American Statistical Association and Journal of the Royal Statistical Society (series A and B).

Originally invented by computer scientists who have their publication venues in conference proceedings, Machine Learning follows the same fashion. Top machine learning conferences include Conference on Learning Theory (COLT), Conference on Neural Information Processing Systems (NIPS) and International Conference on Machine Learning (ICML).

Back to the topic of empirical risk minimization, we draw n i.i.d. training examples which are pairs of feature vector and label (X_i, Y_i)'s, and then define expected risk as given in Lecture 1.

Definition 1 (Expected Risk) *For any given classifier h, the expected risk of h is defined as*

$$R(h) = \mathbb{P}(h(X) \neq Y) = \mathbb{E}[\mathbb{1}\{h(X) \neq Y\}].$$

In practice, since we do not know the joint distribution of X and Y, it is difficult to evaluate the expected risk. Therefore, in real applications, we use samples of X and Y to approximate the expected risk, which gives the empirical risk based on the i.i.d. training data.

Definition 2 (Empirical Risk) *For any given classifier h, its empirical risk is defined as*

$$\hat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{h(X_i) \neq Y_i\}.$$

In many occasions, a machine learning problem is translated into the empirical risk minimization problem, stated as

$$\hat{h}_n = \arg\min_{h \in \mathcal{H}} \hat{R}_n(h),$$

where $\mathcal{H}$ is a hypothesis class. In optimization, since it is often impossible to search for the optimal h in the entire function space, we restrain $\mathcal{H}$ to be some specific function classes, such as linear functions, neural networks, etc.

Definition 3 (Bayes Classifier) *We define Bayes classifier as*

$$h^* = \arg\min R(h).$$

Note that Bayes classifier minimizes the probability of misclassification, and there is no constraint in the definition of Bayes Classifier.

Definition 4 (Bayes Risk) *We call the expected risk with respect to Bayes classifier as Bayes Risk, denoted by $R(h^*)$.*

Definition 5 (Regression Function) *We define the regression function of Y on X as*

$$\eta(X) = \mathbb{P}(Y = 1|X) = \mathbb{E}[Y|X].$$

It is easy to show that the Bayes classifier for binary classification has the following form

$$h^*(X) = \begin{cases} 1, & \text{if } \eta > \frac{1}{2} \\ 0, & \text{if } \eta \leq \frac{1}{2} \end{cases}$$

Definition 6 (Excess Risk) *The excess risk of a classifier h is defined as*

$$\mathcal{E}(h) := R(h) - R(h^*).$$

Note that we are interested in the excess risk of the empirical risk minimizer, i.e., $\mathcal{E}(\hat{h}_n) = R(\hat{h}_n) - R(h^*)$, which is the difference between the expected risk of empirical risk minimizer and the expected risk of Bayes classifier.

Excess risk $\mathcal{E}(\hat{h}_n)$ can be decomposed as follows

$$R(\hat{h}_n) - R(h^*) = \underbrace{R(\hat{h}_n) - \min_{h \in \mathcal{H}} R(h)}_{I_1} + \underbrace{\min_{h \in \mathcal{H}} R(h) - R(h^*)}_{I_2},$$

where term I_1 is called estimation error and term I_2 is approximation error. They were discussed in Lecture 1 as well.

Note that the approximation error only depends on the choice of $\mathcal{H}$ and risk function but not sample size. However, the estimation error shrinks as sample size increases: as training sample size $n \to \infty$, the empirical risk minimizer will converge to the expected risk minimizer in $\mathcal{H}$. For the rest of this class, we will concentrate on bounding the estimation error, and the approximation error will be discussed in the section of kernel method and deep learning.

Now we could focus on how to bound the estimation error. We decompose it into three terms:

$$R(\hat{h}_n) - R(\tilde{h}) = \underbrace{R(\hat{h}_n) - \hat{R}_n(\hat{h}_n)}_{(i)} + \underbrace{\hat{R}_n(\hat{h}_n) - \hat{R}_n(\tilde{h})}_{(ii)} + \underbrace{\hat{R}_n(\tilde{h}) - R(\tilde{h})}_{(iii)},$$

where $\tilde{h} = \arg\min_{h \in \mathcal{H}} R(h)$ is referred to as oracle classifier, which is the minimizer of expected risk in $\mathcal{H}$. The purpose of adding/substracting and grouping here is to bound each of the three terms individually, and then sum up to get an upper bound for the estimation error.

The easiest term to bound is (ii). Observe that $\hat{R}_n(\hat{h}_n) - \hat{R}_n(\tilde{h}) \leq 0$ given the optimality $\hat{h}_n = \arg\min_{h \in \mathcal{H}} \hat{R}_n(h)$.

For term (iii), we have

$$\hat{R}_n(\tilde{h}) - R(\tilde{h}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\tilde{h}(X_i) \neq Y_i\} - \mathbb{E}[\mathbb{1}\{\tilde{h}(X) \neq Y\}].$$

Note that $\mathbb{1}\{\tilde{h}(X_i) \neq Y_i\}$'s are independent random variables, therefore the sample mean $\hat{R}_n(\tilde{h})$ asymptotically converge to the population mean $R(\tilde{h})$. However, this asymptotic argument is not powerful enough in practice, as machine learning researchers usually concern about how large the sample size is needed in order to precisely construct a model to certain accuracy, which is usually called the sample complexity of a machine learning algorithm. Here, with solely a finite sample, imposing a bound on term (iii) is done by using the technique of concentration inequality, which will be introduced later.

Next, on term (i),

$$R(\hat{h}_n) - \hat{R}_n(\hat{h}_n) = \mathbb{E}[\mathbb{1}\{\hat{h}_n(X \neq Y\}] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{h}_n(X_i) \neq Y_i\}.$$

Observe that in term (iii), we have the oracle minimizer $\tilde{h}$ which is independent of the training examples. However, note that the second term in (i) $\sum_{i=1}^{n} \mathbb{1}\{\hat{h}_n(X_i) \neq Y_i\}/n$ is not a summation of independent random variables because $\hat{h}_n$ depends on the training examples (X_i, Y_i)'s. As a result, to bound term (i), it requires a technique from the empirical process theory [1].

In detail, we use a uniform convergence argument to get rid of the dependency in $\sum_{i=1}^{n} \mathbb{1}\{\hat{h}_n(X_i) \neq Y_i\}/n$. The idea is to use the supremum in the hypothesis class $\mathcal{H}$,

$$|R(\hat{h}_n) - \hat{R}_n(\hat{h}_n)| \leq \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_n(h)|.$$

It is worth noting that, in many machine learning papers, a term mentioned often is Generalization Error Bound. Here we define it as following.

Definition 7 (Generalization Error Bound) *The generalization error bound of an empirical risk minimizer $\hat{h}_n$ is*

$$R(\hat{h}_n) \leq \hat{R}_n(\hat{h}_n) + \Delta(n, \mathcal{H}, \delta)$$

where $\Delta(n, \mathcal{H}, \delta)$ depends on the sample size n, hypothesis class $\mathcal{H}$ and the confidence level δ. Note that the $R(\hat{h}_n)$ is called testing error of the empirical risk minimizer, and the $\hat{R}_n(\hat{h}_n)$ is its training error. The idea of using generalization error bound is that we could then bound the testing error by the training error plus an additional term. The generalization error bound will be a byproduct when deriving the estimation error in detail later in this semester.

Now we move on to concentration inequalities which is the second major topic of this class. First let's recall some probability inequalities:

Theorem 1 (Markov Inequality) *Let X be a nonnegative random variable with finite $\mathbb{E}(X)$, for any $t > 0$, we have*

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t}.$$

Lemma 1 (Hoeffding's Lemma) *Suppose that a random variable X is bounded between a and b, then the moment generating function satisfies*

$$\mathbb{E}\{\exp[t(X - \mathbb{E}(X))]\} \leq \exp\left(\frac{t^2(b-a)^2}{8}\right).$$

Next, we are going to present some concentration inequalities. The first one is Hoeffding's inequality, which can be derived from Hoeffding's Lemma.

Theorem 2 (Hoeffding's Inequality) *Let $Z_1 \ldots . Z_n$ be independent (but not necessarily identical) random variables with $a_i \leq Z_i \leq b_i$, then for any $\epsilon > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} Z_i\Big]\right| > \epsilon\right) \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right),$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} Z_i\Big] > \epsilon\right) \leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right),$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} Z_i\Big] < -\epsilon\right) \leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

References

[1] Sara A Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.