Stat 928: Statistical Learning Theory **Lecture: 1**

Introduction and the Bias-Variance Tradeoff

Instructor: Sham Kakade

1 Goal of Statistical Learning Theory

The goal of statistical learning theory is to study the statistical behavior of statistical machine learning algorithms, and understand their theoretical properties such as rate of convergence (upper bound), optimality (whether upper bound matches lower bound), computational efficiency, under different conditions. Some questions to address:

This class focuses mostly on the analysis of supervised learning (prediction) algorithms.

- Bias Variance
- Linear Regression: fixed design (when X's are fixed and Y's are random); random design (when X's and Y's are random); feature selection (more dimensions than points); ridge regression
- Classification
- online learning
  - obtains points in a sequential manner
  - easy optimization algorithms
- convex losses
- PCA
- Assume using L_1 regularization for feature selection. Under what conditions it can recover the correct feature set?
- *sharp analysis and lower bounds, where possible*

Some of the tools that we will utilize are:

- concentration of measure (central limit theorem; tail bounds)
- empirical process theory
- oracle inequalities
- covering numbers (fundamental), and some Rademacher and VC tools
- some martingale tools
- (some) convex analysis
- (some) random matrix analysis

2 Goal of Class and Requirements

Goals:

1. understand theoretical analysis and basic techniques
2. analysis of common algorithms
3. ability to read theoretical papers
4. basic theoretical analysis
5. intuition for performance of algorithms

Assignments:

1. occasional HWs
2. reading papers with theoretical analysis
3. project?

3 Example Supervised Learning Problems

Two basic paradigms we will focus on are regression and classification.

3.1 Linear Least Squares Regression

The Input X is a p-dimensional real valued vector in R^p. The output Y is a real-valued number (e.g. return of a particular stock). The function class C consists of linear functions, parameterized by linear weight (coefficient) vector $w \in R^p$. That is, $f_w \in C$ as a linear function $f_w(x) = w \cdot x$. The quality measure $L(f) = \mathbb{E}(f(X) - Y)^2$ is the squared error, under some distribution on X and Y.

Empirical risk minimization is the least squares estimator:

$$\hat{w} = \frac{1}{n} \arg\min_{w \in R^p} \sum_i (w^\top X_i - Y_i)^2,$$

and the generalization error of $\hat{w}$ is

$$E_{X,Y}(\hat{w}^\top X - Y)^2.$$

3.2 Binary Linear Classification

The Input X is a p-dimensional real valued vector in R^p (e.g., representing information of an email). The output Y is a binary-valued number (whether the email is a spam or not), assume the binary values are $\{-1, 1\}$. The function class C consists of linear functions, parameterized by linear weight (coefficient) vector $w \in R^p$. That is, $f_w \in C$ as a linear function $f_w(x) = w^\top x$. The quality measure is $\ell(f(x), y) = I(f(x) \neq y)$, where $I(\cdot)$ is the 0-1 valued indicator function, so that $\ell(f(x), y) = 0$ if $f(x) = y$ (prediction is correct), or $\ell(f(x), y) = 1$ if $f(x) \neq y$ (prediction is incorrect). This loss is called classification error loss.

Empirical risk minimization is the least squares estimator:

$$\hat{w} = \frac{1}{n} \arg\min_{w \in R^p} \sum_i I(w^\top X_i \neq Y_i)$$

and the generalization error of $\hat{w}$ is

$$L(f) = E_{X,Y} I(w^\top X \neq Y).$$

4 The Squared Error; Linear Regression; and Bias-Variance

The (generalization) squared error of $f : \mathbb{R}^p \to \mathbb{R}$ is

$$L(f) = \mathbb{E}_{X,Y}(f(X) - Y)^2.$$

which we are interested in optimizing.

Note that the Bayes optimal function is the conditional expectation. To see this, first observe:

Lemma 4.1.
We have that:

$$L(f) = \mathbb{E}_X(f(X) - \mathbb{E}[Y|X])^2 + \mathbb{E}_X[\mathit{VAR}(Y|X)]$$

Proof. First, note that:

$$\begin{aligned} L(f) &= \mathbb{E}_{X,Y}(f(X) - Y)^2 \\ &= \mathbb{E}_{X,Y}(f(X) - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - Y)^2 \\ &= \mathbb{E}_{X,Y}(f(X) - \mathbb{E}[Y|X])^2 + \mathbb{E}_{X,Y}(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y) + \mathbb{E}_{X,Y}(\mathbb{E}[Y|X] - Y)^2 \\ &= \mathbb{E}_{X,Y}(f(X) - \mathbb{E}[Y|X])^2 + \mathbb{E}_X(\ (f(X) - \mathbb{E}[Y|X])\mathbb{E}_Y(\mathbb{E}[Y|X] - Y)\ ) + \mathbb{E}_{X,Y}(\mathbb{E}[Y|X] - Y)^2 \\ &= \mathbb{E}_{X,Y}(f(X) - \mathbb{E}[Y|X])^2 + 0 + \mathbb{E}(\mathbb{E}[Y|X] - Y)^2 \\ &= \mathbb{E}(f(X) - \mathbb{E}[Y|X])^2 + \mathbb{E}[\text{VAR}(Y|X)] \qquad (1) \end{aligned}$$

which completes the proof. □

Corollary 4.2. *The Bayes optimal predictor (that which minimizes the squared loss) is* $\mathbb{E}[Y|X]$.

Aside: Defining the conditional expectation is delicate issue in measure theory (involving the Radon-Nikodyn derivative). From a functional analysis point of view, one can actually define the conditional expectation as any function $f(x)$ which achieves the infimum loss. There are many such functions (though they may only disagree on sets of measure 0) — all such functions are considered versions of the conditional expectation.

The square loss is actually quite natural when dealing with $Y \in \{0, 1\}$ if we seek to model probabilities, which is what the following corollary observes.

Corollary 4.3. *If* $Y \in \{0, 1\}$*, then the Bayes optimal predictor is the conditional probability* $Pr(Y = 1|X)$*, since* $Pr(Y = 1|X) = \mathbb{E}[Y|X]$.

4.1 Linear Least Squares Regression and the Error Decomposition

We are typically provided with some training set $\mathcal{T}$ of the form $(X_1, Y_1), \ldots (X_n, Y_n)$. The are two natural sampling processes for this set.

- Fixed Design: Here we consider X_1 to X_n as *fixed* (e.g. not random variables). Our goal is to estimate the function $\mathbb{E}[Y|X_i]$. This is sometimes called signal reconstruction. The loss considered uses the uniform distribution over these X_i's, e.g.:

$$L(2) = \frac{1}{n}\sum_i \mathbb{E}[(w^\top X_i - Y_i)^2|X_i]$$

- Random Design: Both X and Y are random.

Let $\hat{w}_\mathcal{T}$ be the linear function constructed using the training set (we drop the $\mathcal{T}$ subscript, when clear from context).

The following lemma characterizes the *expected loss* of $\hat{w}_\mathcal{T}$

Lemma 4.4. *Let w^* be the best linear predictor, e.g.*

$$w^* \in \arg\min L(w)$$

Let $\hat{w}_\mathcal{T}$ be any estimator based on $\mathcal{T}$. For any distribution over (X, Y), and any distribution over $\mathcal{T}$, we have that:

$$\begin{aligned}
&\mathbb{E}_\mathcal{T} L(w_\mathcal{T})\\
=&\mathbb{E}_X[\mathit{VAR}(Y|X)] + \mathbb{E}_X(\mathbb{E}[Y|X] - \mathbb{E}_\mathcal{T}[w_\mathcal{T}]\cdot X)^2 + \mathbb{E}_{X,\mathcal{T}}(\mathbb{E}_\mathcal{T}[w_\mathcal{T}]\cdot X - w_\mathcal{T}\cdot X)^2\\
=&\mathbb{E}_X[\mathit{VAR}(Y|X)] + \mathbb{E}_X(\mathbb{E}[Y|X] - w^*\cdot X)^2 + \mathbb{E}_X(w^*\cdot X - \mathbb{E}[w_\mathcal{T}]\cdot X)^2 + \mathbb{E}_{X,\mathcal{T}}(\mathbb{E}[w_\mathcal{T}]\cdot X - w_\mathcal{T}\cdot X)^2\\
=&\text{"noise variance + "approximation error of function class" + "estimation bias"+ "estimation variance"}
\end{aligned}$$

(note we have made no assumptions about $\mathcal{T}$).

Proof. For the first equality, (using equation 1)

$$\begin{aligned}
&\mathbb{E}_\mathcal{T} L(w_\mathcal{T})\\
=&\mathbb{E}[\mathrm{VAR}(Y|X)] + \mathbb{E}_{X,\mathcal{T}}(w_\mathcal{T}\cdot X - \mathbb{E}[Y|X])^2
\end{aligned}$$

The last term is equal to:

$$\begin{aligned}
&\mathbb{E}_{X,\mathcal{T}}(w_\mathcal{T}\cdot X - \mathbb{E}[Y|X])^2\\
=&\mathbb{E}_{X,\mathcal{T}}(w_\mathcal{T}\cdot X - \mathbb{E}[w_\mathcal{T}]\cdot X + \mathbb{E}[w_\mathcal{T}]\cdot X - \mathbb{E}[Y|X])^2
\end{aligned}$$

Now observe that:

$$\begin{aligned}
&\mathbb{E}_{X,\mathcal{T}}(w_\mathcal{T}\cdot X - \mathbb{E}[w_\mathcal{T}]\cdot X)(\mathbb{E}[w_\mathcal{T}]\cdot X - \mathbb{E}[Y|X])\\
=&\mathbb{E}_X(\mathbb{E}_\mathcal{T}[w_\mathcal{T}\cdot]X - \mathbb{E}[w_\mathcal{T}]\cdot X)(\mathbb{E}[w_\mathcal{T}]\cdot X - \mathbb{E}[Y|X])\\
=&0
\end{aligned}$$

The final equation is a HW problem. □

Let us make the following observations:

- the first term is referred to as the noise (note that the noise at a given X may depend on X, e.g. it may be heteroskedastic)
- the second term (in the second equation) is referred to as the "bias", which decomposes into two terms, the approximation error of the class and the bias of the algorithm

- the final term is the variance
- note that in expectation the approximation error term

Remember that our goal is to find $\hat{f}$ that predicts well on unseen data (test data). However, we only observe prediction accuracy of $\hat{f}$ on the training data. In order to obtain highly accurate classifier, we have to balance the following two aspects:

- prediction rule should fit the training data well; that is, achieving small training error
  - requires a more expressive model.
- performance of prediction rule on test data should match that on training data
  - requires a less expressive (more stable) model.

There are various related theoretical concepts: training versus test error, bias variance trade-off, overfitting, model complexity, generalization performance, regularization.

4.2 Regret and Risk

It is often useful to compare to the best thing we could hope for in our class. In other words, we may be interested in:

$$L(w_{\mathcal{T}}) - L(w^*)$$

which is the *regret*.

The risk is the expected value:

$$E_{\mathcal{T}}(L(w_{\mathcal{T}}) - L(w^*))$$

Note that:

$$E_{\mathcal{T}}(L(w_{\mathcal{T}}) - L(w^*)) = \mathbb{E}_X(w^* \cdot X - \mathbb{E}[w_{\mathcal{T}}] \cdot X)^2 + \mathbb{E}_{X,\mathcal{T}}(\mathbb{E}[w_{\mathcal{T}}] \cdot X - w_{\mathcal{T}} \cdot X)^2$$

Note that the “approximation error” is not explicitly in the risk.