CMSC 35900 (Spring 2008) Learning Theory **Lecture: 7**

Probabilistic Setup and Empirical Risk Minimization

Instructors: Sham Kakade and Ambuj Tewari

1 Probabilistic Setup

So far we have refrained from making any assumptions about how the data is generated. We have seen some algorithms with worst case guarantees. For the next few lectures, we will assume that the data set

$$(X_1, Y_1), \ldots, (X_m, Y_m)$$

consists of independent and identically distributed random variable pairs from an unknown probability distribution P on $\mathcal{X} \times \mathcal{Y}$.

As in the online model, we use a loss function $\phi : \mathcal{D} \times \mathcal{Y} \to [0, 1]$ to measure the quality of our predictions. Here $\mathcal{D}$ is some appropriate superset of $\mathcal{Y}$ in which our predictions always lie. For binary classification, $\mathcal{Y} = \{-1, +1\}$. If $\mathcal{D} = \mathbb{R}$ then a very natural choice of the loss function is the *0-1 loss function*,

$$\phi(y', y) = \mathbf{1}\left[\operatorname{sgn}(y') \neq y\right] .$$

A few other popular loss functions are given below. The notation $(z)_+ = \max\{z, 0\}$ denotes the postive part of z.

$\phi(y', y) = (1 - y'y)_+$	hinge loss
$\phi(y', y) = \exp(-y'y)$	exponential loss
$\phi(y', y) = \ln(1 + \exp(-y'y))$	logistic loss

For regression, $\mathcal{Y} = \mathcal{D} = \mathbb{R}$. Some commonly used loss functions for regression are given below.

$\phi(y', y) = (y' - y)^2$	squared loss
$\phi(y', y) = \lvert y' - y \rvert$	absolute-value loss
$\phi(y', y) = (\lvert y' - y \rvert - \epsilon)_+$	ϵ-insensitive loss
$\phi(y', y) = \begin{cases} (y' - y)^2, & \lvert y' - y \rvert \leq \delta \\ 2\delta \lvert y' - y \rvert - \delta^2, & \text{otherwise} \end{cases}$	Huber's loss

Suppose we have fixed a hypothesis class $\mathcal{F} \subseteq \mathcal{D}^{\mathcal{X}}$ we want to work with. For example, $\mathcal{F}$ could be the set of polynomials of degree less than d, decision trees of depth less than D or neural networks with at most H hidden units. Now there might not be a function $f : \mathcal{X} \to \mathcal{D}$ in $\mathcal{F}$ such that $Y = f(X)$ with probability 1. In this so called *agnostic* setting, our goal is simply to perform as well as the "best" function in the class $\mathcal{F}$. The best function depends on our choice of the loss function and we denote it by f^*_ϕ,

$$f^*_\phi := \operatorname{argmin}_{f \in \mathcal{F}} L_\phi(f) ,$$

where $L_\phi(f)$ is the *ϕ-risk* of f,

$$L_\phi(f) := \mathbb{E}\left[\phi(f(X), Y)\right] .$$

We denote the minimum possible (over *all* measurable functions) ϕ-risk by L_ϕ^*,

$$L_\phi^* := \min_f L_\phi(f) \ .$$

Suppose $\hat{f} \in \mathcal{F}$ is a function that we pick based on our sample. Then we can write,

$$L_\phi(\hat{f}) - L_\phi^* = \underbrace{L_\phi(\hat{f}) - L_\phi(f_\phi^*)}_{\text{estimation error}} + \underbrace{L_\phi(f_\phi^*) - L_\phi^*}_{\text{approximation error}} \ .$$

Once the class $\mathcal{F}$ is fixed, the approximation error is a fixed deterministic quantity. If the class $\mathcal{F}$ is too "small", the estimation error will be small but the approximation error will be large. On the other hand, if $\mathcal{F}$ is too "big", the approximation error will be small but the estimation error will be large. Therefore, there is a trade-off between these two errors.

Let us focus on the estimation error. How can we choose $\hat{f} \in \mathcal{F}$ based on our data such that $L_\phi(\hat{f}) \approx L_\phi(f_\phi^*)$? Since the underlying distribution P generating (X, Y) is unknown to us, we cannot compute $L_\phi(f)$ and hence f_ϕ^* is unknown to us. We can, however, compute the *empirical ϕ-risk* of any $f \in \mathcal{F}$,

$$\hat{L}_\phi(f) := \frac{1}{m} \sum_{i=1}^{m} \phi(f(X_i), Y_i) \ .$$

By the law of large numbers, $\hat{L}_\phi(f) \to L_\phi(f)$ as $n \to \infty$. So, a natural thing to do is to work with empirical risks and define the *empirical risk minimizer*,

$$\hat{f}_\phi^* := \operatorname{argmin}_{f \in \mathcal{F}} \hat{L}_\phi(f) \ .$$

Unfortunately, the strong law of large number by itself is not sufficient to guarantee that the estimation error will be small. We need a stronger condition: the law of large number should hold *uniformly* over the class $\phi \circ \mathcal{F}$. That is,

$$\sup_{f \in \mathcal{F}} \left| \hat{L}_\phi(f) - L_\phi(f) \right| \to 0 \text{ as } n \to \infty \ .$$

The following theorem then shows how to control the estimation error.

Theorem 1.1. *Suppose*

$$\sup_{f \in \mathcal{F}} \left| \hat{L}_\phi(f) - L_\phi(f) \right| \leq \epsilon_m \ . \tag{1}$$

Then, we have

$$L_\phi(\hat{f}_\phi^*) - L_\phi(f_\phi^*) \leq 2\epsilon_m \ .$$

Proof. We have,

$$\begin{aligned} L_\phi(\hat{f}_\phi^*) &\leq \hat{L}_\phi(\hat{f}_\phi^*) + \epsilon_m && [\because \hat{f}_\phi^* \in \mathcal{F},\ (1)] \\ &\leq \hat{L}_\phi(f_\phi^*) + \epsilon_m && [\because \hat{f}_\phi^* \text{ minimizes } \hat{L}_\phi] \\ &\leq L_\phi(f_\phi^*) + 2\epsilon_m && [\because f_\phi^* \in \mathcal{F},\ (1)] \ . \end{aligned}$$

□

2 Classification

Binary classification ($\mathcal{Y} = \{-1, +1\}$) is a very important special case of the general prediction problem we considered above. For $x \in \mathcal{X}$, define

$$\eta(x) := \mathbb{P}\left(Y = +1 | X = x\right) \ .$$

Fix ϕ to be the 0-1 loss for now. So, we can drop the subscript in $L_\phi(f)$ and refer to it simply as the risk of f. Note that $R(f)$ is simply the probability of making a prediction error. What is the function that minimizes $R(f)$? Intuitively, to minimize the probability of error, we should predict 1 if $\eta(x) \geq 1/2$ and 0 otherwise. Define

$$f_\eta(x) := \begin{cases} +1\,, & \eta(x) \geq 1/2 \\ -1\,, & \text{otherwise}\,. \end{cases}$$

The theorem below say that $R(f) - R(f_\eta)$ is non-negative for any f and hence f_η minimizes the risk. Note that changing the definition of f_η on the set $\{x \in \mathcal{X} \mid \eta(x) = 1/2\}$ does not affect its risk. We therefore arbitrarily defined it to be $+1$ there.

Theorem 2.1. *For any $f : \mathcal{X} \to \mathbb{R}$, we have*

$$R(f) - R(f_\eta) = \mathbb{E}\left[\mathbf{1}\left[\operatorname{sgn}(f(X)) \neq f_\eta(X)\right] \cdot |2\eta(X) - 1|\right] \geq 0\,.$$

Proof. We have,

$$\begin{aligned} R(f) - R(f_\eta) &= \mathbb{E}\left[\mathbf{1}\left[\operatorname{sgn}(f(X)) \neq Y\right]\right] - \mathbb{E}\left[\mathbf{1}\left[f_\eta(X) \neq Y\right]\right] \\ &= \mathbb{E}\left[\mathbf{1}\left[\operatorname{sgn}(f(X)) \neq Y\right] - \mathbf{1}\left[f_\eta(X) \neq Y\right]\right] \\ &= \mathbb{E}\left[\mathbf{1}\left[\operatorname{sgn}(f(X)) \neq f_\eta(X)\right] \cdot \left(\mathbf{1}\left[\operatorname{sgn}(f) \neq Y\right] - \mathbf{1}\left[f_\eta(X) \neq Y\right]\right)\right] \\ &= \mathbb{E}\left[\mathbf{1}\left[\operatorname{sgn}(f(X)) \neq f_\eta(X)\right] \cdot \left(\mathbf{1}\left[f_\eta(X) = Y\right] - \mathbf{1}\left[f_\eta(X) \neq Y\right]\right)\right] \\ &= \mathbb{E}\left[\mathbf{1}\left[\operatorname{sgn}(f(X)) \neq f_\eta(X)\right] \cdot \left(2 \cdot \mathbf{1}\left[f_\eta(X) = Y\right] - 1\right)\right] \\ &= \mathbb{E}\left[\mathbf{1}\left[\operatorname{sgn}(f(X)) \neq f_\eta(X)\right] \cdot \left(2 \cdot \mathbb{E}\left[\mathbf{1}\left[f_\eta(X) = Y\right] | X\right] - 1\right)\right] \\ &= \mathbb{E}\left[\mathbf{1}\left[\operatorname{sgn}(f(X)) \neq f_\eta(X)\right] \cdot \left(2 \cdot \max\{\eta(X), 1 - \eta(X)\} - 1\right)\right]\,, \end{aligned}$$

where the last line follows because $\mathbb{P}\left(f_\eta(X) = Y | X = x\right) = \max\{\eta(x), 1-\eta(x)\}$. Noting that $2\max\{\eta, 1-\eta\} - 1 = |2\eta - 1|$ finishes the proof. □

3 Regression

Another important special case of the prediction problem is when $\mathcal{Y} = \mathbb{R}$. Fix ϕ to be the squared loss for now. Let us see which function minimizes the risk,

$$R(f) = \mathbb{E}\left[(f(X) - Y)^2\right]\,.$$

Define

$$f_\rho(x) = \mathbb{E}\left[Y | X = x\right]\,,$$

to be the conditional mean of Y given $X = x$. The following theorem shows that $R(f) \geq R(f_\rho)$ for any $f : \mathcal{X} \to \mathbb{R}$ and gives an explicit formula for the difference.

Theorem 3.1. *For any $f : \mathcal{X} \to \mathbb{R}$, we have*

$$R(f) - R(f_\rho) = \mathbb{E}\left[(f(X) - f_\rho(X))^2\right] \geq 0\,.$$

Proof. First of all, note that for any $g : \mathcal{X} \to \mathbb{R}$, we have

$$\mathbb{E}\left[g(X)(f_\rho(X) - Y)\right] = 0\,. \tag{2}$$

This is easily by taking conditional expectations and using $f_\rho(X) = \mathbb{E}\left[Y|X\right]$. Then we have,

$$\begin{aligned} R(f) - R(f_\rho) &= \mathbb{E}\left[(f(X) - Y)^2\right] - \mathbb{E}\left[(f_\rho(X) - Y)^2\right] \\ &= \mathbb{E}\left[(f(X) - Y)^2 - (f_\rho(X) - Y)^2\right] \\ &= \mathbb{E}\left[(f(X) - f_\rho(X) + f_\rho(X) - Y)^2 - (f_\rho(X) - Y)^2\right] \\ &= \mathbb{E}\left[(f(X) - f_\rho(X))^2\right] + 2\mathbb{E}\left[(f(X) - f_\rho(X))(f_\rho(X) - Y)\right] \\ &= \mathbb{E}\left[(f(X) - f_\rho(X))^2\right]\,, \end{aligned}$$

where the last line follows from (2) by taking g to be $f - f_\rho$. □