EECS 598-005: Theoretical Foundations of Machine Learning **Fall 2015**

Lecture 12: Noisy Setting and Rademacher Complexity

Lecturer: Jacob Abernethy *Scribes: Aniket Deshmukh*

12.1 Noisy Setting

In the most general setting, the hypothesis class $\mathcal{H}$ may contain no consistent hypotheses due to noise. Namely, there might not be a hypothesis in $\mathcal{H}$ that correctly labels every point. This chapter and the following two will address learning guarantees in this setting.

We model this case by supposing that points $x \in \mathcal{X}$ and corresponding labels $y \in \mathcal{Y}$ are drawn at random according to some joint distribution $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$. Unlike the noiseless setting, y is not necessarily a function of x; they could even be independent!

12.1.1 Loss functions

We use loss functions to measure the error of a predicted label.

Definition 12.1 (Loss function)**.** *A **loss function** is any function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$. The first argument is the predicted label and the second argument is the true label.*

Loss functions are typically nonnegative and it is often desirable for them to be convex in the first argument.

Examples: Some common examples for $\mathcal{Y} = \{-1, 1\}$ are shown below. Notably, all the loss functions can be written in terms of $y\hat{y}$ and so we plot them as univariate functions of $y\hat{y}$.

algebraic,linewidth=1pt,xunit=1.5cm,yunit=1cm,plotpoints=500 (-2.75,-1.25)(2.25,4.25)
[axesstyle=frame,tickstyle=top,Ox=-2,Dy=1](-2,0)(2,4) [linecolor=red]-22(1-x/abs(x))/2
[linecolor=blue]-211-x[linecolor=blue]120
[linecolor=black]-12(1-x)$^2[linecolor = magenta] - 22ln(1 + 2.718^(- x))[linecolor = red](0.8, 3.6)(1, 3.6)[l](1.1, 3.6)$0-1$[linecolor = blue](0.8, 3.2)(1, 3.2)[l](1.1, 3.2)Hinge[linecolor = black](0.8, 2.8)(1, 2.8)[l](1.1, 2.8)Square[linecolor = magenta](0.8, 2.4)(1, 2.4)[l](1.1, 2.4)Logistic[t](0, -0.75)y[b]90(-2.5, 2)Lossfunction\ell(y\hat{y})$

Loss function	$\ell(\hat{y}, y) = \ell(y\hat{y})$
$0-1$ loss	$\mathbb{1}[y \neq \hat{y}] = \mathbb{1}[y\hat{y} \neq 1]$ $(= \mathbb{1}[y\hat{y} < 0]$ if $\hat{y} \in \{-1, 1\})$
Hinge loss	$\max(0, 1 - y\hat{y})$
Square loss	$(\hat{y} - y)^2 = (1 - y\hat{y})^2$
Logistic loss	$\log\Big(1 + \exp(-y\hat{y})\Big)$

Figure 12.1: Plot of loss functions with expressions

The $0-1$ loss is a natural loss function for binary classification but it is non-convex in its first argument, making optimization difficult. Hence it is often replaced by a convex surrogate. For example, support vector machines use the hinge loss as a convex surrogate.

12.1.2 Risk

We use the expected loss, called *risk*, to measure the error of a hypothesis.

Definition 12.2 (Risk)**.** *The* ***risk*** *of a hypothesis h is*

$$R(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)].$$

Definition 12.3 (Bayes Risk)**.** *The* ***Bayes risk*** *is the minimum risk possible among measurable hypotheses*

$$R^* = \inf_{h\in\mathcal{H}^*} R(h).$$

Note that H^* here is the set of all measurable functions.

Note that the Bayes risk R^* could be large; in the noisy setting there need not be any hypothesis that predicts the labels reliably. However, it is the best we can do, and so our goal will be to find hypotheses with risk close to R^*.

12.2 Empirical Risk Minimization (ERM)

Empirical Risk Minimization tries to find a minimum risk hypothesis by choosing a hypothesis from a given class $\mathcal{H}$ that minimizes the empirical risk, defined as follows.

Definition 12.4 (Empirical Risk)**.** *The* ***empirical risk*** *of a hypothesis h on samples $S = ((x_1, y_1), ..., (x_m, y_m))$ is given by*

$$\hat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h(x_i), y_i). \tag{12.1}$$

Namely, ERM chooses the hypothesis

$$h_S^{ERM} = \arg\min_{h\in\mathcal{H}} \hat{R}_S(h).$$

Note that the minimization is carried out over a chosen hypothesis class $\mathcal{H}$. ERM chooses a hypothesis from the class that has minimum empirical risk.

12.3 Bias-Variance Trade Off

One of the big goals in learning theory is to bound how much larger the risk of ERM is than the Bayes risk. Namely, we want to bound

$$R\left(h_S^{ERM}\right) - R^* = \underbrace{\left(R\left(h_S^{ERM}\right) - \inf_{h\in\mathcal{H}} R(h)\right)}_{\text{estimation error}} + \underbrace{\left(\inf_{h\in\mathcal{H}} R(h) - R^*\right)}_{\text{approximation error}} \tag{12.2}$$

The **estimation error** measures how well ERM was able to find a minimum risk hypothesis in the class. Namely, it measures how much risk was introduced by minimizing the empirical risk instead of the true risk. The **approximation error** measures how well the hypothesis class is able to approximate a minimum risk hypothesis overall. Namely, it measures how much risk was introduced by restricting the hypotheses to be in the class $\mathcal{H}$.

Note that algorithms in general first choose some model class $\mathcal{H}$ (i.e., commit to a particular type of hypothesis) and then try to fit the model (i.e., identify a hypothesis in the class that fits the data best). If the model is rich (i.e., has lots of hypotheses) one would expect good approximation error but potentially at the cost of estimation error. On the other hand a restrictive model (with only a few hypotheses) would be easier to estimate/learn but this might come at the cost of approximation error. Statisticians call this a bias-variance trade off.

12.4 Uniform Deviation Bound

Estimating the approximation error is generally difficult (even with various assumptions) but we can get an initial bound on the estimation error. For convenience, let $h^* = \arg\min_{h\in\mathcal{H}} R(h)$ be a minimum risk hypothesis in the class. Then we can write the estimation error as

$$\begin{aligned}
R\left(h_S^{ERM}\right) - \inf_{h\in\mathcal{H}} R(h) &= \left[R\left(h_S^{ERM}\right) - \hat{R}_S\left(h_S^{ERM}\right)\right] + \left[\hat{R}_S\left(h_S^{ERM}\right) - \hat{R}_S(h^*)\right] + \left[\hat{R}_S(h^*) - R(h^*)\right] \\
&\leq \left[R\left(h_S^{ERM}\right) - \hat{R}_S\left(h_S^{ERM}\right)\right] + \left[\hat{R}_S(h^*) - R(h^*)\right] \\
&\leq 2\sup_{h\in\mathcal{H}} |\hat{R}_S(h) - R(h)|
\end{aligned}$$

where the second line follows from the fact that

$$\hat{R}_S(h_S^{ERM}) = \inf_{h\in\mathcal{H}} \hat{R}_S(h) \leq \hat{R}_S(h^*).$$

Controlling the term

$$2\sup_{h\in\mathcal{H}} |\hat{R}_S(h) - R(h)|$$

is called a uniform deviation bound.

12.5 Rademacher Complexity

The Rademacher Complexity of a class of functions measures how rich the class is. It does so by measuring how well the class can fit random noise. In particular, it uses Rademacher random variables.

Definition 12.5 (Rademacher Random Variable)**.** *A* ***Rademacher Random Variable*** *takes on values* ±1 *and is defined by the Rademacher distribution*

$$\sigma = \begin{cases} +1 & w.p.\ 1/2 \\ -1 & w.p.\ 1/2 \end{cases}. \tag{12.3}$$

Using Rademacher random variables, we then define the empirical Rademacher complexity and use that to define the Rademacher complexity.

Definition 12.6 (Empirical Rademacher Complexity)**.** *The* ***Empirical Rademacher Complexity*** *of a class G of functions $g : \mathcal{X} \to \mathbb{R}$ with respect to a sample $S = (x_1, ..., x_m)$ is*

$$\hat{\mathfrak{R}}_S(G) := \mathop{\mathbb{E}}_{\{\sigma_i\}}\left[\sup_{g\in G} \frac{1}{m}\sum_{i=1}^m g(x_i)\sigma_i\right].$$

where $\sigma_1, \ldots, \sigma_m$ are independent Rademacher random variables.

Definition 12.7 (Rademacher Complexity)**.** *The* ***Rademacher Complexity*** *of a class G of functions $g : \mathcal{X} \to \mathbb{R}$ with respect to a distribution $\mathcal{D} \in \Delta(\mathcal{X})$ is*

$$\mathfrak{R}_m(G) := \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\hat{\mathfrak{R}}_S(G)\right]$$

where m is the sample size.

Note that:

1. $\mathfrak{R}_m(\{g\}) = 0$ for any function g. Namely, the Rademacher complexity of a single function class is zero.

2. $\mathfrak{R}_m$(“all functions with output in $[-1,1]$”) $= 1$ when $\mathbb{P}_{S\sim\mathcal{D}^m}\{$“$S$ has duplicates”$\} = 0$. Namely, the Rademacher complexity of a class with all possible functions is one so long as S almost surely does not contain duplicates.

Our goal will be to choose a function class with $\mathfrak{R}_m(G) = O\left(1/\sqrt{m}\right)$. The following theorem connects the Rademacher complexity with the discussion of uniform deviation bounds above.

Theorem 12.8. *Let G be a class of functions $g : \mathcal{X} \to [0,1]$ and $\mathcal{D} \in \Delta(X)$ be a distribution. Then with probability at least $1-\delta$ we have that*

$$\sup_{g\in G}(\mathbb{E}g - \hat{\mathbb{E}}_S g) \le 2\mathfrak{R}_m(G) + \sqrt{\frac{log\frac{1}{\delta}}{2m}}$$

where $S \sim \mathcal{D}^m$ and

$$\mathbb{E}g = \mathop{\mathbb{E}}_{x\sim\mathcal{D}} g(x) \qquad\qquad \hat{\mathbb{E}}_S g = \frac{1}{m}\sum_{i=1}^m g(x_i).$$

Proof: For convenience, define

$$\Phi(S) = \sup_{g\in G}(\mathbb{E}g - \hat{\mathbb{E}}_S g).$$

We first bound $\mathbb{E}_{S\sim\mathcal{D}^m}[\Phi(S)]$. To do so we introduce another sample $S' = (x'_1,\dots,x'_m) \sim \mathcal{D}^m$ and note that

$$\mathbb{E}g = \mathop{\mathbb{E}}_{x\sim\mathcal{D}} g(x) = \frac{1}{m}\sum_{i=1}^m \mathop{\mathbb{E}}_{x\sim\mathcal{D}} g(x) = \frac{1}{m}\sum_{i=1}^m \mathop{\mathbb{E}}_{x'_i\sim\mathcal{D}} g(x'_i) = \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m} \frac{1}{m}\sum_{i=1}^m g(x'_i) = \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m} \hat{\mathbb{E}}_{S'} g.$$

Thus we can write

$$\begin{aligned}
\mathbb{E}[\Phi(S)] &= \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{g\in G}\Big(\mathbb{E}g - \hat{\mathbb{E}}_S g\Big)\right] = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\sup_{g\in G}\mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\Big(\hat{\mathbb{E}}_{S'}g - \hat{\mathbb{E}}_S g\Big)\right] \\
&\le \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\left[\sup_{g\in G}\Big(\hat{\mathbb{E}}_{S'}g - \hat{\mathbb{E}}_S g\Big)\right] = \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m}\sup_{g\in G}\left[\frac{1}{m}\sum_{i=1}^m\Big(g(x'_i) - g(x_i)\Big)\right] \\
&= \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m,\{\sigma_i\}}\sup_{g\in G}\left[\frac{1}{m}\sum_{i=1}^m \sigma_i\Big(g(x'_i) - g(x_i)\Big)\right] \\
&\le \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m,\{\sigma_i\}}\sup_{g\in G}\frac{1}{m}\sum_{i=1}^m \sigma_i g(x'_i) \; + \mathop{\mathbb{E}}_{S,S'\sim\mathcal{D}^m,\{\sigma_i\}}\sup_{g\in G}\frac{1}{m}\sum_{i=1}^m -\sigma_i g(x_i) \\
&= 2\mathfrak{R}_m(G).
\end{aligned}$$

We finish the rest of the proof in the next lecture. ■