**CS229T/STATS231: Statistical Learning Theory**

Lecturer: Tengyu Ma
Scribe: Zach Izzo

Lecture #1
September 24, 2018

# 1 Review and Overview

In this lecture we delineate a mathematical framework for supervised learning. We focus on regression problems and define the notion of the loss/risk associated with a model. We then analyze a particular loss function (the squared loss) in a general setting, then specialize our result to the case of linear models. We next define the notion of parameterized families of hypotheses and the maximum likelihood estimate, and we conclude with an asymptotic result relating the training MLE to the true maximum likelihood parameter.

# 2 Formulation of supervised learning

We begin by constructing a mathematical framework for prediction problems. Our framework consists of the following elements:

1. A space of possible data points $\mathcal{X}$.
2. A space of possible labels $\mathcal{Y}$.
3. A joint probability distribution $P$ on $\mathcal{X} \times \mathcal{Y}$. We assume that our training data consists of $n$ points
$$(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \overset{\text{i.i.d.}}{\sim} P$$
each drawn independently from $P$.
4. A prediction function/model $f : \mathcal{X} \to \mathcal{Y}$.
5. A loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. We will usually assume that $\ell$ is bounded below by some constant, typically 0.

Given the prediction function $f$ and the loss function $\ell$, the loss of an example is $\ell(f(x), y)$. We can then define the *expected risk* (or *expected loss*, or *population risk*)

$$L(f) \triangleq \mathbb{E}_{(x,y)\sim P}[\ell(f(x), y)].$$

Our goal will be to obtain a small expected loss. Often it will be infeasible to consider all possible models $f$, so we may restrict ourselves to a certain family of hypotheses $\mathcal{F}$. In this case, we define the *excess risk* of a model $f$ as

$$L(f) - \inf_{g \in \mathcal{F}} L(g).$$

This gives us a measure of how well our model fits the data relative to the best we can hope to do within our set of options $\mathcal{F}$.

Within this framework, there are two main types of problems we will consider: *regression* problems, where the set of labels is $\mathcal{Y} = \mathbb{R}$; and *classification* problems, where the set of labels is some finite set $\mathcal{Y} = \{1, \ldots, k\}$. We will focus on regression problems in this lecture.

## 3 Regression problems and squared loss

We consider the regression problem of predicting $y$ given $x$. We take as our loss function the *squared loss*

$$\ell(\hat{y}, y) = (\hat{y} - y)^2, \qquad L(f) = \mathbb{E}_{(x,y)\sim P}[(f(x) - y)^2].$$

In this setting, we can decompose the risk in a very informative way.

**Lemma 1** (Decomposition of loss)**.** *Under the squared loss, we have the decomposition*

$$L(f) = \mathbb{E}_{x\sim P_x}[(f(x) - \mathbb{E}[y \mid x])^2] + \mathbb{E}_{x\sim P_x}[\mathrm{Var}(y \mid x)]$$

*where $P_x$ is the marginal distribution of $x$.*

The second term in this expansion is the intrisic variable of the label; it gives a lower bound on the loss we can achieve. Since the first term in the decomposition is nonnegative, it is an immediate corrolary that the optimal model is $f(x) = \mathbb{E}[y \mid x]$.

In order to prove Lemma 1, we make use of the following claim.

**Claim 2.** *If $Z$ is a random variable and $a$ is a constant, then*

$$\mathbb{E}[(Z - a)^2] = (\mathbb{E}[Z] - a)^2 + \mathrm{Var}(Z).$$

The proof of this claim is left as an exercise on HW 0. We are now ready to prove Lemma 1.

*Proof of Lemma 1.* We have

$$\begin{aligned} L(f) &= \mathbb{E}[(f(x) - y)^2] \\ &= \mathbb{E}_{x\sim P_x}[\mathbb{E}_{P_{y\mid x}}[(f(x) - y)^2 \mid x]] && \text{(Law of total expectation)} \\ &= \mathbb{E}_{x\sim P_x}[(f(x) - \mathbb{E}[y \mid x])^2 + \mathrm{Var}(y \mid x)]. && \text{(Claim 2)} \end{aligned}$$

Note that Claim 2 holds in the third equation since $f(x)$ is a constant when we have conditioned on $x$. The desired result follows from linearity of expectation. □

Lemma 1 gives us a general lower bound on risk under squared loss. If we impose more structure on the set of hypotheses $\mathcal{F}$ from which we can select $f$, we can gain more information on the risk.

## 4 Linear regression under squared loss

A commonly used choice of hypotheses is the set of linear functions:

$$\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R} \mid f(x) = w^\top x,\ w \in \mathbb{R}^d\}.$$

For $f \in \mathcal{F}$, we then have

$$L(f) = L(w) = \mathbb{E}[(w^\top x - y)^2].$$

Henceforth, we will denote $w^* \in \mathrm{argmin}_{w\in\mathbb{R}^d} L(w)$ and $\hat{w}$ will denote a model learned from training data.

One may ask why we have only allowed linear functions with 0 instead of allowing a nonzero intercept. Actually, the framework we have outlined above is enough to accomodate nonzero intercepts. If we wish to analyze the function $w^\top x + b$, we can simply set $\tilde{x} = (x, 1)$ and $\tilde{w} = (w, b)$. Then $\tilde{w}^\top \tilde{x} = w^\top x + b$ and we have reduced to the case of 0 intercept.

When we restrict to linear models, we can further decompose the risk under squared loss.

**Lemma 3.** *With $w^* \in \operatorname{argmin}_{w \in \mathbb{R}^d} L(w)$, we have*

$$L(\hat{w}) = \mathbb{E}_x[\mathrm{Var}(y \mid x)] + \mathbb{E}_x[(\mathbb{E}[y \mid x] - w^{*\top} x)^2] + \mathbb{E}_x[(w^{*\top} x - \hat{w}^\top x)^2]. \tag{1}$$

The second term in equation (1) can be thought of as the approximation error incurred by linear models. The third term can be interpreted as the estimation error we incur from having only a finite sample.

*Proof.* Define $g(\hat{w}) \triangleq \mathbb{E}[(\mathbb{E}[y \mid x] - \hat{w}^\top x)^2]$. By Lemma 1,

$$L(\hat{w}) = \mathbb{E}[\mathrm{Var}(y \mid x)] + g(\hat{w}). \tag{2}$$

Observe that since $w^* \in \operatorname{argmin} L(w)$, $\nabla L(w^*) = 0$. Furthermore, since $\mathbb{E}_x[\mathrm{Var}(y \mid x)]$ is a constant with respect to $w$, we have

$$\begin{aligned} \nabla L(w) &= \nabla g(w) \\ &= \mathbb{E}[\nabla_w (\mathbb{E}(y \mid x) - w^\top x)^2] \\ &= 2\mathbb{E}[(\mathbb{E}(y \mid x) - w^\top x)x]. \end{aligned}$$

Since $\nabla L(w^*) = 0$ we have

$$\mathbb{E}[(\mathbb{E}[y \mid x] - w^{*\top} x)x] = 0. \tag{3}$$

Next, we expand:

$$\begin{aligned} g(\hat{w}) &= \mathbb{E}[(\mathbb{E}[y \mid x] - \hat{w}^\top x)^2] \\ &= \mathbb{E}[((\mathbb{E}[y \mid x] - w^{*\top} x) - (\hat{w}^\top x - w^{*\top} x))^2] \\ &= \mathbb{E}[(\mathbb{E}[y \mid x] - w^{*\top} x)^2 + (\hat{w}^\top x - w^{*\top} x)^2] \\ &\quad - 2\mathbb{E}[(\mathbb{E}[y \mid x] - w^{*\top} x)(\hat{w}^\top x - w^{*\top} x)]. \end{aligned}$$

Finally, observe that

$$\mathbb{E}[(\mathbb{E}[y \mid x] - w^{*\top} x)(\hat{w}^\top x - w^{*\top} x)] = (\hat{w}^\top - w^{*\top})\mathbb{E}[(\mathbb{E}[y \mid x] - w^{*\top} x)x].$$

By equation (3), this quantity vanishes and it follows that

$$g(\hat{w}) = \mathbb{E}[(\mathbb{E}[y \mid x] - w^{*\top} x)^2] + \mathbb{E}[(\hat{w}^\top x - w^{*\top} x)^2]. \tag{4}$$

Combining equations (2) and (4) gives the desired result. □

# 5 Parameterized families of hypotheses

Linear models are one type of *parameterized family* of hypotheses. In general, a parameterized family is given by a parameter space $\Theta$. For each $\theta \in \Theta$ there is a hypothesis $f_\theta(x)$, sometimes written $f(\theta; x)$. In this case we may write the loss function as

$$\ell(f_\theta(x), y) = \ell((x, y), \theta).$$

In the special case of linear functions, our parameter space is $\Theta = \mathbb{R}^d$ and for $\theta \in \Theta$ we have $f_\theta(x) = \theta^\top x$.

### 5.1 Well-specified case and maximum likelihood

In the well-specified case, $P_\theta(y \mid x)$ is a family of distributions parameterized by $\theta \in \Theta$, and $y \mid x \sim P_{\theta^*}(y \mid x)$ is distributed according to some ground truth parameter $\theta^*$. We define the *maximum likelihood* loss function by

$$\ell((x, y), \theta) = -\log P_\theta(y \mid x),$$

so that minimizing the loss function is equivalent to maximizing the likelihood of the data.

For example, suppose that $y \mid x$ is Gaussian distributed with mean ${\theta^*}^\top x$ and variance 1, i.e. $y \mid x \sim N({\theta^*}^\top x, 1)$. The likelihood is then

$$\begin{aligned}
\ell((x, y), \theta) &= -\log P_\theta(y \mid x) \\
&= -\log \exp\left(-\frac{(y - \theta^\top x)^2}{2}\right) + c \\
&= \frac{(y - \theta^\top x)^2}{2} + c
\end{aligned}$$

where $c$ is the log of the normalizing constant. This computation shows that in the Gaussian setting, minimizing the squared loss actually recovers the MLE.

## 6 Training loss

Often we do not know the true underlying distribution $P$ with which to compute the expected loss. In these cases we need to use an approximation based on the data we do have. This motivates our definition of the *training loss*

$$\hat{L}(\theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell((x^{(i)}, y^{(i)}), \theta).$$

In the special case of maximum likelihood, we have $\hat{L}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log p_\theta(y^{(i)} \mid x^{(i)})$. We define the *maximum likelihood estimator*

$$\hat{\theta}_{\text{MLE}} \in \operatorname{argmin}_{\theta \in \Theta} \hat{L}(\theta).$$

This approximation is "good" in the sense that as $n \to \infty$, the minimizer of the training loss $\hat{\theta}_{\text{MLE}}$ approaches the true maximum likelihood parameter $\theta^*$. The following theorem quantifies this fact.

**Theorem 4** (Asymptotic of MLE)**.** *Assume $\nabla^2 L(\theta^*)$ is full rank. Let $\hat{\theta} = \hat{\theta}_{\text{MLE}}$ and*

$$Q \triangleq \mathbb{E}_{(x,y) \sim P}[\nabla_\theta(\log p_\theta(y \mid x))(\theta^*) \nabla_\theta(\log p_\theta(y \mid x)(\theta^*)^\top].$$

*Assuming that $\hat{\theta} = \hat{\theta}_n \xrightarrow{p} \theta^\star$ (i.e. consistency) and under appropriate regularity conditions,*

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, Q^{-1}) \text{ and } n(L(\hat{\theta}) - L(\theta^*)) \xrightarrow{d} \frac{1}{2}\chi^2(p).$$

*as $n \to \infty$, where $p$ is the dimension of $\theta$ and $\chi^2(p)$ is the distribution of the sum of the squares of $p$ i.i.d. standard Gaussian random variables.*

**Remark.** *The matrix $Q$ referenced in Theorem 4 is known as the Fisher information matrix.*

**Corollary 5.** *The first result of Theorem 4 implies that as $n \to \infty$, $\hat{\theta} - \theta^* \to 0$. The second result implies that $L(\hat{\theta}) - L(\theta^*) \approx p/2n$ since $\mathbb{E}[Z] \approx p$ when $Z \sim \chi^2(p)$.*