10-716: Advanced Machine Learning **Spring 2019**

Lecture 5: January 29

Lecturer: Pradeep Ravikumar *Scribes: Youngseog Chung, Aditya Galada, Vignesh Kannan*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

This lecture's notes illustrate some uses of various LaTeX macros. Take a look at this and imitate.

5.1 Continuing with loss functions

So far we have been discussing loss functions and various principles by which to make decisions based on the loss functions.

The Bayesian approach provided actionable decision rules but was computationally difficult to derive; meanwhile the minimax approach was not actionable. We continue our discussion on loss functions and decision making in this lecture.

5.1.1 Empirical Risk Minimization

We are still concerned with the statistical estimation setting where the action space $\mathcal{A}$ is identical to the parameter space, Θ, and our action corresponds to estimating the parameter based on the sample $X \in \mathcal{X}$

Suppose you have taken an action and chosen $\theta \in \Theta$ and let us consider the loss of having taken this action when the true parameter is θ^*, which we denote as $L(\theta^*, \theta)$

- Try to solve for $\inf_\theta L(\theta^*, \theta)$ i.e. choose θ s.t. it minimizes the loss.

  It would be ideal if we could solve for this parameter but we do not know θ^* hence we cannot solve this problem and thus it is not actionable

- Alternatively, define the loss as $L(\theta^*, \theta) = \mathbb{E}_{X \sim P(\cdot|\theta^*)} l(X, \theta)$, where $l(X, \theta)$ is another loss function defined on $\mathcal{X} \times \mathcal{A} \to \mathbb{R}$

  Here we can compute the loss $l(X, \theta)$ however we cannot compute the expectation, because once again, we do not know θ^*

  However, what we can do here is instead of the expectation, we can compute the empirical expectation with samples $X_1, ..., X_n \sim P(\cdot|\theta^*)$

  Hence we can approximate $L(\theta^*, \theta)$ with $L_n(\theta^*, \theta) = \frac{1}{n}\sum_{i=1}^n l(X, \theta)$

  $L_n(\theta^*, \theta)$ here is called the empirical risk.

Definition 5.1 *Emprical Risk Minimization*

$$ERM = \hat{\theta}_n \in arginf_{\theta \in \Theta_0 \subset \Theta} L_n(\theta^*, \theta)$$

- Here, we are optimizing over a subset of the whole parameter space Θ, as optimizing over the whole parameter space will always yield θ^*, which may not be attainable.
- We are effectively minimizing a surrogate loss, instead of the actual loss
- For any value of θ^*, we can point-wise minimize this estimator.

Since we are using a surrogate measure of loss $L_n(\theta^*, \theta)$ for the actual loss $L(\theta^*, \theta)$ we ideally want these two measures to be close, especially when for our ERM estimate, $\hat{\theta}_n$.

Later on, we will see that as $n \to \infty$, $L(\theta^*, \hat{\theta}_n) \to \inf_{\theta \in \Theta_0} L(\theta^*, \theta)$.

5.1.2 Example of ERM with the Likelihood Principle: MLE

Define the loss $l(X, \theta)$ as the likelihood ratio, $\log \frac{P_{\theta^*}(X)}{P_\theta(X)}$

$$\text{Risk} = \mathbb{E}_{X \sim P(\cdot|\theta^*)} \log \frac{P_{\theta^*}(X)}{P_\theta(X)} = \int P_{\theta^*}(X) \log \frac{P_{\theta^*}(X)}{P_\theta(X)} dx = KL(P_{\theta^*}, P_\theta)$$

Minimizing the risk, which is also the KL divergence between P_{θ^*} and P_θ w.r.t. θ would yield θ^*, however, yet again, we cannot do this optimization precisely because we don't know θ^*.

Therefore, we use the ERM estimate, $\hat{\theta}_n$.

$$\hat{\theta}_n \in \text{arginf}_\theta \{ \frac{1}{n} \sum_{i=1}^{n} \log \frac{P_{\theta^*}(X_i)}{P_\theta(X_i)} \} = \text{arginf}_\theta \{ -\frac{1}{n} \sum_{i=1}^{n} \log P_\theta(X_i) \} = \text{MLE of } \theta^*$$

5.2 Classification

We now consider the problem of classification as an application of Empirical Risk Minimization. We are given a sample $(X, Y) \sim P$ and observations $D = \{(X_i, Y_i)\}_{i=1}^n \underset{i.i.d}{\sim} P$ where $X \in \mathcal{X} = \mathbb{R}^d$ and $Y \in \mathcal{Y} = \{-1, 1\}$. The goal is to obtain a classifier $f : \mathcal{X} \to \mathcal{Y}$. A reasonable loss function is the 0-1 loss defined as follows,

$$l((X, Y), f) = \mathbb{I}(f(X) \neq Y)$$

Taking an expectation over the data $(X, Y) \sim P$ we get,

$$\begin{aligned}\mathbb{E}_{(X,Y)\sim P}[l((X, Y), f)] &= \mathbb{E}_{(X,Y)\sim P}[\mathbb{I}(f(X) \neq Y)] \\ &= \mathbb{P}_{(X,Y)\sim P}(f(X) \neq Y) \\ &= L(P, f)\end{aligned}$$

How do we fit this classification setting into our decision theory framework developed so far?
From a decision theoretic standpoint, P represents the state of nature (intuitively captures the relation between the covariates X and the labels Y) and f represents an action. Putting this into further perspective, the deterministic decision rule δ can be thought of as a mapping from the set of datasets, say $\mathcal{D} = \{D : D = \{(X_i, Y_i)\}_{i=1}^n \underset{i.i.d}{\sim} P\}$, to the set of classifiers $\mathcal{F} = f : \mathcal{X} \to \mathcal{Y}$. In machine learning, decision rules of the form $\delta : \mathcal{D} \to \mathcal{F}$ are learned using a learning algorithm (Eg: A neural network)

A nice way to choose a classifier f would be to minimize $L(P, f)$ over all possible f. Such a classifier f^* is called the *Bayes optimal classifier* and the value $L(P, f^*)$ associated with this classifier is called the *Bayes risk*.

Definition 5.2 *Bayes optimal classifier and Bayes risk*

$$f^* = \arg\inf_f L(P, f)$$
$$L(P, f^*) = \inf_f L(P, f)$$

But unfortunately, we don't know the true distribution P of the data to perform the minimization.
Here is where Empirical Risk Minimization is useful. We can use the ERM estimate $\widehat{f}_n$ to obtain an approximately good classifier given by,

$$\widehat{f}_n \in \arg\inf_f \ \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(f(X_i) \neq Y_i)$$

This gives us a very actionable estimator. However, there seems to be a problem in the terminology used in Definition 5.4. We have used the word *Bayes* but haven't really adopted a Bayesian treatment. Instead of defining a prior distribution π over the distribution P, we are performing a pointwise minimization for each P. To understand what a prior distribution π over distribution P means, we can think of P as a parametric model instead of a non-parametric model. With this simplifying assumption, the prior distribution π would just refer to the distribution over the finite number of parameters θ which define the parametric model P.

A completely Bayesian treatment would probably define the *Bayes optimal classifier* as follows,

$$f^* \in \arg\min_f \ \rho(\pi(P|D), f)$$

where,
$D = \{(X_i, Y_i)\}_{i=1}^n$ is the observed data and $\pi(P|D)$ is the posterior distribution of P having observed D. But this is quite difficult to solve for when compared to the ERM estimate $\widehat{f}_n$ (empirical estimate for f^*).

So in what sense is f^* Bayesian?
To answer this, we will look at f^* from a different perspective. Instead of taking P to be the state of nature, we will choose $Y \in \mathcal{Y}$ (which is after all another random variable like P) as the state of nature. The action space $\mathcal{A}$ will be the set $\mathcal{Y}$ itself. The decision rule will be $\delta : \mathcal{X} \to \mathcal{Y}$ a mapping from $\mathcal{X}$ to $\mathcal{Y}$. The loss for this setting is defined as below,

$$L(Y, a) = \mathbb{I}(Y \neq a)$$

We now describe both the frequentist and the Bayesian notions of risk with respect to the loss defined above. The frequentist risk fixes the state of nature Y and takes an expectation over X with a distribution conditioned (fixed) on Y for a given decision rule δ.

$$\begin{aligned} R(Y, \delta) &= \mathbb{E}_{X\sim P(\cdot|Y)}[L(Y, \delta(X))] \\ &= \mathbb{E}_{X\sim P(\cdot|Y)}[\mathbb{I}(Y \neq \delta(X))] \\ &= \mathbb{E}[\mathbb{I}(Y \neq \delta(X))|Y] \end{aligned}$$

The Bayes risk computes the expectation of $R(Y, \delta)$ using a prior distribution π over the state of Y for a given decision rule δ.

$$\begin{aligned}
r(\pi, \delta) &= \mathbb{E}_{Y\sim\pi}[R(Y, \delta)] \\
&= \mathbb{E}_{Y\sim\pi}\Big[\mathbb{E}[\mathbb{I}(Y \neq \delta(X))|Y]\Big] \\
&= \mathbb{E}_{Y\sim\pi, X\sim P(\cdot|Y)}[\mathbb{I}(Y \neq \delta(X))] && \text{by the Law of Total Expectation} \\
&= \mathbb{E}_{(X,Y)\sim P}[\mathbb{I}(Y \neq \delta(X))] && P(X, Y) = \pi(Y)P(X|Y) \\
&= \mathbb{P}_{(X,Y)\sim P}(Y \neq \delta(X))
\end{aligned}$$

We see that in the above setting, the decision rule $\delta : \mathcal{X} \to \mathcal{Y}$ plays the role of the classifier $f : \mathcal{X} \to \mathcal{Y}$. Recalling the definition of the *Bayes rule* δ^π, the following relation between δ^π and f^* can be derived.

$$\begin{aligned}
\delta^\pi &= \arg\min_{\delta}\ r(\pi, \delta) \\
&= \arg\min_{\delta}\ \mathbb{P}_{(X,Y)\sim P}(Y \neq \delta(X)) \\
&= \arg\min_{\delta}\ L(P, \delta) \\
&= f^*
\end{aligned}$$

Hence, when we treat the classifier f itself as the decision rule δ, f^* turns out to be the *Bayes rule* δ^π and is rightly called the *Bayes optimal classifier* and its associated loss value of $L(P, f^*) = r(\pi, \delta^\pi)$ is accordingly called the Bayes risk.

But unlike the Bayesian decision theory we have looked at so far, this estimator $\delta^\pi = f^*$ is not actionable. This is because in a usual Bayesian setting, we have access to both the prior distribution π over the state of nature θ as well as the distribution $P(\cdot|\theta)$ over samples X given the state of nature θ. However, in our Bayesian framework for the problem of classification, we neither know the prior distribution π over the labels Y (the state of nature) nor do we know the distribution $P(\cdot|Y)$ over the covariates X given the class label Y. Essentially, it boils down to not knowing the true joint distribution P over (X, Y) (since if we knew π and $P(\cdot|Y)$ we could compute the joint distribution $P = \pi P(\cdot|Y)$).

This problem of not knowing the true joint distribution P can be handled by using a plug-in distribution as a replacement for P in the following 2 ways.

- **Method 1**: Empirical Risk Minimization
  As explained earlier, this approach approximates the *Bayes optimal classifier* f^* (or δ^π) with the ERM estimate $\widehat{f}_n$ by minimizing the empirical expectation of loss instead of the true expectation of loss.
- **Method 2**: Use the following decision rule (or more precisely, classifier in this context of classification).

  $$\delta_P(X) = \text{sgn}(\mathbb{P}_P(Y = 1|X) - 1/2)$$

  It can be shown that the above classifier δ_P is the *Bayes optimal classifier* with respect to a 0-1 loss and a uniform prior π over the state of nature Y, i.e. $\pi(Y = 1) = \pi(Y = -1) = 1/2$. This classifier essentially ouputs the label which has the higher probability given the covariate X.

  But computing $\mathbb{P}(Y = 1|X)$ again demands the knowledge of P so we resort to using an empirical distribution P_n instead.

  $$\delta_{P_n}(X) = \text{sgn}(\mathbb{P}_{P_n}(Y = 1|X) - 1/2)$$

  where $\mathbb{P}_{P_n}(Y = 1|X) - 1/2)$ can be estimated using techniques like Logistic Regression on an observed dataset $D = \{(X_i, Y_i)\}_{i=1}^n$.

To summarize, though we intended to use a Bayesian treatment we did not have direct access to the prior π over the state of nature Y and the distribution $P(\cdot|Y)$ over the covariates X given the state of nature Y. But we did have this information in an "indirect" way through the observed samples $D = \{(X_i, Y_i)\}_{i=1}^n$. So we resorted to approximating the *Bayes optimal classifier* f^* using this noisy information about the true distribution P.

The next section describes a more general setting of where Empirical Risk Minimization can be applied. It turns out that there are a whole class loss functions called *Decomposable losses* which can use Empirical Risk Minimization.

5.3 Decomposable Loss Functions

In the previous section we discussed that decision theoretic optimization can be performed by approximating the loss functions and minimizing the surrogate losses. Decomposable losses are a special class of loss functions which can be written as a function of the examples.

Definition 5.3 *Decomposable Loss function*

$$L(\theta^*, \theta) = \mathbb{E}_{x \sim P(\cdot|\theta*)}[l(x, \theta)]$$

Most rational loss functions can be expressed in the above form. The 0-1 loss function used in the classification example is a decomposable loss.The overall loss depends on a loss that is incurred on an individual sample.

$$L(P, f) = \mathbb{P}[y \neq f(x)] = \mathbb{E}_{(x,y)\sim P}\mathbb{I}(y \neq f(x))$$

However certain loss functions cannot we written in the above form. For example, the precision loss do not decompose as a function of individual examples. Precision refers to the number of true positives divided by the total number of samples labelled as positive. In terms of probability it can be written as :

$$L(P, f) = \mathbb{P}[y = 1|f(x) = 1]$$
$$L(P, f) = \frac{\mathbb{P}[y = f(x) = 1]}{\mathbb{P}[f(x) = 1]} \neq \mathbb{E}_{(x,y)\sim P}l((x, y), f)$$

It is impossible to decompose the precision loss as function of examples due to the presence of the probability ratios in the above equation.

Theorem 5.4 *Rationality of Preference Order : A loss function $L(P, f)$ where P is the underlying distribution of samples and f is a fixed decision rule can be decomposed as $L(P, f) = R(P) = \mathbb{E}_{z\sim P}l(z)$ if the preference order is rational. If any distribution P follows these 4 axioms, then the preference order is said to be rational :*

1. *: If $P_1, P_2 \in \mathcal{P}$, then $P_1 = P_2$ or $P_1 > P_2$ or $P_1 < P_2$ [Preference Order]*
2. *: If $P_1 \leq P_2$, $P_2 \leq P_3 \implies P_1 \leq P_3$ [Transitivity Property]*
3. *: If $P_1 \leq P_2 \quad \implies \alpha P_1 + (1-\alpha)P_3 \leq \alpha P_2 + (1-\alpha)P_3$ [Mixture Model]*
4. *: No infinitely desirable or undesirable distributions [Bounded]*

The preference order corresponds to the risk over distributions. If these 4 axioms over the preference orders over distributions are satisfied, then the risk can be written as an expectation of some loss function over individual samples.

When performing ERM, we optimize with respect to the surrogate loss and obtain $\hat{\theta}_n$

$$\hat{\theta}_n \in \arg\inf_{\theta} l_n(\theta^*, \theta) = \frac{1}{n}\sum_{i=1}^{n} l(x_i, \theta)$$

We interested in knowing when does $\hat{\theta}_n$ converge to θ^* and the rate of pointwise convergence. Rates of converges depends on number of samples and the structural complexity.

$$l_n(\theta^*, \theta) - L(\theta^*, \theta) = \frac{1}{n}\sum_{i=1}^{n} l(x_i, \theta) - \mathbb{E}[l(x, \theta)]$$

In other words, we want to find out how far the empirical expectation is from the true expectation. If we regard $l(x_i, \theta)$ as z_i, the task essential reduces to calculating how far the sample average is from the true expectation.

With an aim to obtain a quantitative measure for how good the surrogate loss is, we continue our discussion to tail bounds.

5.4 Tail Bounds

Tail bounds address the question of how far is a random variable from its expectation.

5.4.1 Markov Inequality

Markov's Inequality is a classical tail bound. It serves as a basis for deriving advanced tail bounds.
Markov's Inequality : If X is a nonnegative random variable, then the probability that x is at least t is at most the expectation of X divided by t:

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

This can be rewritten in the following way for monotonically increasing functions.
If g is a nonnegative and strictly increasing function of random variable x, then we can consider g(x) to be a random variable and apply Markov's Inequality to obtain:

$$P(x > t) = P(g(x) > g(t)) \leq \frac{\mathbb{E}[g(x)]}{g(t)}$$

The motivation behind choosing a strictly increasing function is that $g(x) > g(t) \implies x > t$
A family of powerful bounds can be obtained by varying g.

5.4.2 Chebyshev Inequality

Chebyshev's Inequality Applying Markov's Inequality with $g(t) = t^k$ where $k \geq 2$

$$\begin{aligned} P(|x-\mu| > t) &\leq \frac{\mathbb{E}[|x-\mu|^k]}{t^k} \\ P(|x-\mu| > t) &\leq \frac{\mathbb{E}[|x-\mu|^2]}{t^2} \text{ if k=2} \end{aligned}$$

The term $\mathbb{E}[|x-\mu|^k]$ is actually the k^{th} central moment. The case of k=2 is most often used in practice. The Chebyshev Bound is an extremely poweful bound. For large enough values of k no other bound is strictly better than the Chebyshev's bound.

5.4.3 Chernoff Inequality

Chernoff's Inequality Applying Markov's Inequality with $g(t) = e^{\lambda t}$

$$P(x-\mu > t) \leq \frac{\mathbb{E}e^{\lambda(x-\mu)}}{e^{\lambda t}}$$

The term $\mathbb{E}e^{\lambda(x-\mu)}$ is known as the moment generating function (MGF). This is because its derivatives with respect to lambda at $x = 0$ yield the central moments. Assuming that the MGF exists for some $\lambda \in [0, b]$

$$\begin{aligned} P(x-\mu > t) &\leq \inf_{\lambda \in [0,b]} \frac{\mathbb{E}e^{\lambda(x-\mu)}}{e^{\lambda t}} \\ \log P(x-\mu > t) &\leq \inf_{\lambda \in [0,b]} \log \frac{\mathbb{E}e^{\lambda(x-\mu)}}{e^{\lambda t}} \\ \log P(x-\mu > t) &\leq \inf_{\lambda \in [0,b]} \log \mathbb{E}e^{\lambda(x-\mu)} - \lambda t \end{aligned}$$

The bound obtained by Chernoff's Inequality depends upon how fast the MGF decays as a function of λ. Chernoff's Inequality gives exponential bounds on the tail distributions of random variables. There exist some distributions where this bound is an equality.

For example, consider the Gaussian distribution $X \sim \mathcal{N}(\mu, \sigma^2)$
The Moment Generating Function (MGF) of a Gaussian distribution is

$$\begin{aligned} \mathbb{E}[e^{\lambda x}] &= e^{\mu\lambda + \frac{\lambda^2\sigma^2}{2}} \\ \mathbb{E}[e^{\lambda(x-\mu)}] &= e^{\frac{\lambda^2\sigma^2}{2}} \end{aligned}$$

The MGF exists for $\lambda \geq 0$. Plugging in the value of MGF in to the Chernoff bound

$$\inf_{\lambda \geq 0} \left\{ \frac{\lambda^2\sigma^2}{2} - \lambda t \right\}$$

Solving for optimal λ, we get

$$\begin{aligned} \lambda &= \frac{t}{\sigma^2} \\ \log P(x-\mu > t) \leq \inf_{\lambda \geq 0} \left\{ \frac{\lambda^2\sigma^2}{2} - \lambda t \right\} &= \frac{-t^2}{2\sigma^2} \\ \log P(x-\mu > t) &\leq \frac{-t^2}{2\sigma^2} \\ P(x-\mu > t) &\leq e^{\frac{-t^2}{2\sigma^2}} \end{aligned}$$

From the above equation we observe that the probability decreases exponentially as t increases. Most of the density is centered around the mean and rate of decay is controlled by the variance σ^2.

The only property of the Gaussian variable used was MGF. We can apply this property to Sub-gaussian random variable x is such that $\mathbb{E}[e^{\lambda(x-\mu)}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$ for some $\sigma > 0$ and all $\lambda \in \mathbb{R}$. Then $P(x-\mu > t) \leq e^{\frac{-t^2}{2\sigma^2}}$

Until now we have discussed a right side tail bound, i.e. bounding the probability $\mathbb{P}(x-\mu > t)$. By performing the following manipulation, it is possible to obtain a two sided tail bound on $\mathbb{P}(|x-\mu| > t)$

Consider x drawn from a Sub-Gaussian distribution with parameter σ, $X \sim SG(\sigma)$. By property of sub-gaussian random variables

$$\mathbb{E}[e^{\lambda(x-\mu)}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$$

Now consider the random variable $-x$

$$\mathbb{E}[e^{-\lambda(x-\mu)}] \leq e^{\frac{(-\lambda)^2\sigma^2}{2}} = e^{\frac{(\lambda)^2\sigma^2}{2}}$$

Therefore $-x$ is also Sub-Gaussian with parameter σ, $-x \sim SG(\sigma)$. The lower tail bound is :

$$P(-x+\mu > t) \leq e^{\frac{-t^2}{2\sigma^2}}$$

Combining the left and right tail bounds :

$$P(|x+\mu| > t) \leq 2e^{\frac{-t^2}{2\sigma^2}}$$

The above bounds hold for any Sub-Gaussian random variable.

Rademacher Random Variable Rademacher distribution is a discrete probability distribution where a random variable X has a 50% chance of being +1 and a 50% chance of being -1.

$$P(x) = \left\{ \begin{array}{ll} \frac{1}{2}, & \text{for } x = -1 \\ \frac{1}{2}, & \text{for } x = +1 \end{array} \right\}$$

From the above definition it is clear that the mean of this distribution is 0. It also can be shown that the Radamacher random variable is subgaussian with parameter 1.

$$\mathbb{E}[e^{\lambda(x-0)}] \leq e^{\frac{\lambda^2}{2}} \implies x \sim SG(1)f$$