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

Lecture 15: May 19

Lecturer: Lecturer: Pradeep Ravikumar *Scribes: Scribes: Tolani Olarinre, Srinivas R,3*

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

In the previous lecture, we motivated the problem of estimating the parameter $\theta^* \in \Theta$ given some samples Z drawn from the distribution $\mathbb{P}$. We formulated this as a hypothesis testing problem on the 2δ-separated set $\{\theta_1, \ldots, \theta_M\} \subseteq \theta(\mathcal{P})$ of hypothesis for θ^*. We defined the hypothesis test as a function $\psi : \mathcal{Z} \to \{1, \ldots, M\}$, with error probability $[\psi(Z) \neq J]$, where $J \sim \text{UNIF}\{1, \ldots, M\}$. We ended the lecture by lower bounding the minimax risk for the estimation problem with the infimum (over hypothesis test functions) error probability $\inf_\psi[\psi(Z) \neq J]$. In today's lecture, we look at ways of lower bounding this inf error probability.

15.1 Binary classification and LeCam's method

We begin by looking at the simplest case of hypothesis testing: binary classification. Here we have two possible distributions

$$\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P} \;\text{ s.t }\; \rho(\theta(\mathbb{P}_0), \theta(\mathbb{P}_1)) \geq 2\delta$$

which could have generated some data

$$(Z|J = j) \sim \mathbb{P}_j,\ j \in \{0, 1\}$$

and

$$P(J = 0) \;=\; P(J = 1) \;=\; \frac{1}{2}$$

The goal is to lower bound the misclassification error over all binary classifiers $\psi : \mathcal{Z} \to \{0, 1\}$, i.e. lower bound $\inf_\psi[\psi(Z) \neq J]$ (called the *Bayes risk*). We can derive an expression for the inf misclassification error in terms of total variation distance $||\mathbb{P}_0 - \mathbb{P}_1||_{TV}$ as follows:

$$\inf_\psi \mathbb{Q}[\psi(Z) \neq J] = \frac{1}{2}\Big\{1 - ||\mathbb{P}_0 - \mathbb{P}_1||_{TV}\Big\} \tag{15.1}$$

Observe that the worst case bayes risk is $\frac{1}{2}$, which occurs when the total variation distance between $\mathbb{P}_0$ and $\mathbb{P}_1$ is 0, at which point the classifier is random guessing. Conversely, when the data is linearly separable, the total variation between the two distributions is 1, the bayes risk is 0.

To derive equation (15.1), we let $A = \{Z \in \mathcal{Z} | \psi(Z) = 1\}$. For a binary classifier, we have that

$$\begin{aligned}
\mathbb{Q}[\psi(Z) = J] &= \frac{1}{2}\mathbb{P}_1(A) + \frac{1}{2}\mathbb{P}_0(A^c) \\
&= \frac{1}{2}\mathbb{P}_1(A) + \frac{1}{2}\big(1 - \mathbb{P}_0(A)\big) \\
&= \frac{1}{2} + \frac{1}{2}\big(\mathbb{P}_1(A) - \mathbb{P}_0(A)\big) \\
\sup_{\psi} \mathbb{Q}[\psi(Z) = J] &= \frac{1}{2} + \frac{1}{2}\sup_{A \subseteq \mathcal{Z}}\big(\mathbb{P}_1(A) - \mathbb{P}_0(A)\big) \\
&= \frac{1}{2} + \frac{1}{2}||\mathbb{P}_0 - \mathbb{P}_1||_{TV} \\
\inf_{\psi} \mathbb{Q}[\psi(Z) \neq J] &= 1 - \sup_{\psi}[\psi(Z) = J] \\
&= \frac{1}{2} - \frac{1}{2}||\mathbb{P}_0 - \mathbb{P}_1||_{TV}
\end{aligned}$$

Putting this together with the result from the last lecture, we have the lower bound on minimax risk for binary classification as

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_{Z \sim \mathbb{P}_\theta}[\Phi \circ \rho(\hat{\theta}, \theta)] \geq \frac{\Phi(\delta)}{2}\left\{1 - ||\mathbb{P}_0 - \mathbb{P}_1||_{TV}\right\} \tag{15.2}$$

LeCam's method seeks to choose $\mathbb{P}_0$ and $\mathbb{P}_1$ such that their TV distance is as small but their corresponding parameters $\theta(\mathbb{P}_0)$ and $\theta(\mathbb{P}_1)$ are far (at least 2δ) apart. This amounts to choosing the two distributions which have high misclassification error. We can then choose δ to lower bound the misclassification error of a classifier on data from these distributions.

15.1.1 Example

We define the gaussian location family as $\{\mathbb{P}_\theta \equiv \mathcal{N}(\theta, \sigma^2) \mid \theta \in \mathbb{R}\}$.

We're given $Z = (Y_1, ..., Y_n)$ where $Y_i \overset{iid}{\sim} \mathcal{N}(\theta^*, \sigma^2)$.

We want a lower bound on the error in estimating θ^* with $\hat{\theta}$ under absolute error $\rho(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$.

We can apply LeCam's method by selecting two distributions $\mathbb{P}_0$ and $\mathbb{P}_{2\delta}$, so that their parameters $\theta(\mathbb{P}_0) = 0$ and $\theta(\mathbb{P}_{2\delta}) = 2\delta$ are 2δ separated. For n iid samples, we represent the product distribution as $\mathbb{P}_\theta^n$. It can be shown that the total variation between two distributions is bounded as

$$||\mathbb{P}_0^n - \mathbb{P}_\theta^n||_{TV}^2 \leq \frac{1}{4}\left\{e^{\frac{n\theta^2}{\sigma^2}} - 1\right\} = \frac{1}{4}\left\{e^{\frac{4n\delta^2}{\sigma^2}} - 1\right\} \text{ when } \theta = 2\delta$$

Setting $\delta = \frac{\sigma}{2\sqrt{n}}$ and substituting in for the minimax lower bound (15.2), we get

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_{Z \sim \mathbb{P}_\theta}[\,|\hat{\theta} - \theta|\,] \geq \frac{\delta}{2}\left\{1 - \frac{1}{2}\sqrt{e-1}\right\} \geq \frac{\delta}{6} = \frac{1}{12}\frac{\sigma}{\sqrt{n}}$$

So the sample mean $\hat{\theta}_n := \frac{1}{n}\sum_{i=1}^n Y_i$ is an estimator that achieves the LeCam bound (ignoring constant factors). It's worse case risk is $\sup_{\theta \in \mathbb{R}} \mathbb{E}_{Z \sim \mathbb{P}_\theta}[\,|\hat{\theta} - \theta|\,] \asymp \frac{\sigma}{\sqrt{n}}$

15.2 Generalized Lecam's Method

We should choose our hypotheses and classification problem such that the mis-classification error is high. This allows us to get tighter lower bounds on the minimax risk.

To increase the misclassification error, we shall consider the case where we have multiple hypotheses in place of 2. The idea is to consider a series of binary classification between hypotheses. We can still partition the distributions associated with these hypotheses into two groups:

$$\mathcal{P}_0, \mathcal{P}_1 \subseteq \mathcal{P}$$

If

$$\sup_{\mathbb{P}_0 \in \mathcal{P}_0, \mathbb{P}_1 \in \mathcal{P}_1} \rho(\theta(\mathbb{P}_0), \theta(\mathbb{P}_1)) \geq 2\delta$$

then the following inequality holds:

$$\inf_{\hat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}[\rho(\hat{\theta}, \theta^*)] \geq \delta \sup_{\mathbb{P}_0 \in \text{CONV}(\mathcal{P}_0), \mathbb{P}_1 \in \text{CONV}(\mathcal{P}_1)} (1 - ||\mathbb{P}_1 - \mathbb{P}_0||_{TV})$$

Note that $\mathbb{P}_0$ and $\mathbb{P}_1$ can be from the convex hull of $\mathcal{P}_0$ and $\mathcal{P}_1$ respectively. This allows us to make the distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ close, even though the distributions in $\mathcal{P}_0$ and $\mathcal{P}_1$ are at a distance of at least 2δ. If they are close, the classification problem is intuitively harder.

15.2.1 Example

We return to our example of Gaussian Location to illustrate the above point. Let $Y_i \overset{iid}{\sim} \mathcal{N}(\theta^*, \sigma^2)$ and $Z = (Y_1, Y_2 ... Y_n)$

Let $\mathcal{P}_0 = \{\mathbb{P}_0\}$ and $\mathcal{P}_1 = \{\mathbb{P}_{2\delta}, \mathbb{P}_{-2\delta}\}$

Notice that though any 2 distributions chosen from $\mathcal{P}_0$ and $\mathcal{P}_1$ respectively are at least 2δ apart, intuitively we expect $\mathbb{P}_0$ and $\bar{\mathbb{P}}_1 = \frac{1}{2}\mathbb{P}_{2\delta} + \frac{1}{2}\mathbb{P}_{-2\delta}$ to be quite similar. Distinguishing between them is a harder problem and leads to a higher misclassification error, which allows us to get a tighter bound.

We proceed by first upper bounding the TV distance between the distributions:

We can show that:

$$||\mathbb{P}_0 - \bar{\mathbb{P}}_1||_{TV} \leq \frac{1}{4}\left(e^{\frac{1}{2}\left(\frac{2\sqrt{n}\delta}{\sigma}\right)^4} - 1\right)$$

Setting $\delta = \frac{\sigma t}{2\sqrt{n}}$, and using the Generalized Lecam's Inequality:

$$\begin{aligned}\inf_{\hat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}[\rho(\hat{\theta}, \theta^*)] &\geq \delta(1 - ||\mathbb{P}_0 - \bar{\mathbb{P}}_1||_{TV}) \\ &\geq \sup_t \frac{\sigma}{4\sqrt{n}} t \left(1 - \frac{1}{2}\sqrt{e^{t^2/2} - 1}\right) \\ &= \left(\frac{3}{20}\right) \frac{\sigma}{\sqrt{n}}\end{aligned}$$

Note that the constant is higher than what we get using the simpler Lecam's method.

15.3 Fano's Method

In this section we talk about the problem of multi-class classification. We introduce Fano's method which builds on a classical result from information theory namely Fano's inequality.

For the following we assume:

$$\begin{aligned}J &\sim \text{Uniform}(1, 2...M) \\ (Z|J = j) &\sim \mathbb{P}_{\theta_j}\end{aligned}$$

Our goal is to determine the index J of the probability distribution from which a given sample has been drawn. Intuitively, the difficulty of this depends on the amount of dependence between Z and index J.

Let $\mathbb{Q}_{\mathbb{Z},\mathbb{J}}$ denote the joint distribution of inputs and labels, and $\mathbb{Q}_{\mathbb{Z}}\mathbb{Q}_{\mathbb{J}}$ denote the product of the marginal distributions. If we have $\mathbb{Q}_{\mathbb{Z},\mathbb{J}} = \mathbb{Q}_{\mathbb{Z}}\mathbb{Q}_{\mathbb{J}}$, Z and J are statistically independent, which means knowing one of them does not let us infer anything about the other. In this case, the best we can do is random guessing. To quantify the amount of dependence of Z and J we define mutual information I as the Kullback-Leibler divergence:

$$I(Z, J) = D_{KL}(\mathbb{Q}_{Z,J}|\mathbb{Q}_Z\mathbb{Q}_J) \geq 0$$

Now Fano's inequality gives us a lower bound on the classification error:

$$\mathbb{Q}(\psi(\mathcal{P}) \neq J) \geq 1 - \frac{I(Z, J) + \log 2}{\log M}$$

Proposition: Let $\{\theta_1, \ldots, \theta_M\}$ be a 2δ-separated set in the ρ semi-metric on $\Theta(\mathcal{P})$, and suppose that J is uniformly distributed over the index set $\{1, \ldots, M\}$, and $(Z|J = j) \sim \mathbb{P}_{\theta_j}$. Then for any increasing function $\Phi : [0, \infty) \to [0, \infty)$, the minimax risk is lower bound as

$$m(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta)(1 - \frac{I(Z, J) + \log 2}{\log M})$$

where $I(Z, J)$ is the mutual information between Z and J.

If we look at the above we can see that with decreasing δ the separation criterion becomes milder, because $M \equiv M(2\delta)$ increases. At the same time the mutual information $I(Z, J)$ will decrease, because J can take a larger number of potential numbers. By decreasing δ enough we can ensure that:

$$\frac{I(Z,J) + \log 2}{\log M} \leq \frac{1}{2}$$

There are various ways to upper bound the mutual information $I(Z, J)$ in this setting. On way is to use the convexity of the KL-divergence which leads us the following inequality.

$$I(Z,J) \leq \frac{1}{M^2} \sum_{j,k=1}^{M} \mathcal{D}_{KL}(\mathbb{P}^n_{\theta_j}, \mathbb{P}^n_{\theta_k})$$

15.3.1 Example: Normal location model

Assume $J \sim \text{Uniform}(\{0, 2\delta, -2\delta\} = \text{Uniform}(\{\theta_1, \theta_2, \theta_3\})$, let $(Z|J = j) \sim \mathcal{N}(\theta_j, \sigma^2) = \mathbb{P}_{\theta_j}$ and $Z = (y_1, \ldots, y_n)$ with ${}_i \sim \mathcal{N}(\theta^*, \sigma^2)$. The pairwise KL-divergence of the Gaussian distributions is.

$$\begin{aligned}
\mathcal{D}_{KL}(\mathcal{N}(\theta, \sigma^2 I), \mathcal{N}(\theta', \sigma^2 I)) &= \frac{||\theta - \theta'||_2^2}{2\sigma^2} \\
\implies \mathcal{D}_{KL}(\mathbb{P}^n_{\theta_j}, \mathbb{P}^n_{\theta_k}) = \frac{n}{2\sigma^2}(\theta_i - \theta_k)^2 &\leq \frac{2n\delta^2}{\sigma^2} \quad (2\delta\text{-seperated}) \\
\implies I(Z,J) &\leq \frac{2n\delta^2}{\sigma^2}
\end{aligned}$$

By picking $\delta^2 = \sigma^2/20$ we can upper bound:

$$\begin{aligned}
\frac{I(Z,J) + \log 2}{\log M} \leq \frac{\frac{2n\delta^2}{\sigma^2} + \log 2}{\log 3} &\leq \frac{3}{4} \\
\implies \inf_{\psi} \mathbb{Q}(\psi(\mathcal{P}) \neq J) &\geq 1 - \frac{1}{4} = \frac{1}{4} \\
\implies m(\theta(\mathcal{P}), \rho) \geq \frac{\delta}{4} = \frac{\sigma}{4\sqrt{20}\sqrt{n}}
\end{aligned}$$

15.3.2 Generalized Fano Method

We generalize the previous approach. Suppose we can construct a local packing $\{\theta_1, \ldots, \theta_M\}$ of the parameter space Ω such that $\rho(\theta_i, \theta_j) \geq 2\delta$ for $i \neq j$ and for some $c \geq 0$, the KL-divergence is bounded:

$$\mathcal{D}_{KL}(\mathbb{P}^n_{\theta_j}, \mathbb{P}^n_{\theta_k}) \leq cn\delta^2$$

This implies $I(Z, J) \leq cn\delta^2$ and lets us bound:

$$\log M \geq 2\{cn\delta^2 + \log 2\}$$
$$\implies m(\theta(\mathcal{P}), \rho) \geq \frac{\delta}{2}$$

15.3.3 Example: Linear Regression

Assume a standard linear regression model where $y = X\theta^* + w \in \mathbb{R}$, where $\theta^* \in \mathbb{R}^d$, $X \in \mathbb{R}^{n \times d}$ and $w \sim \mathcal{N}(0, \sigma^2)$. Let design matrix X be fixed such that only w is random. In order to find a bound on the classification error we try to find a packing $\{\theta_1, \ldots, \theta_M\}$ with $\rho(\theta_i, \theta_j) \geq 2\delta$ defining $\mathbb{P}^m_{\theta_i} \sim \mathcal{N}(X\theta_i, \sigma^2 I)$. As in the previous example the KL-divergence for Gaussian distributions is:

$$\mathcal{D}_{KL}(\mathbb{P}^n_{\theta_j}, \mathbb{P}^n_{\theta_k}) = \frac{||X(\theta - \theta')||_2^2}{2\sigma^2}$$
$$\implies \frac{||X(\theta - \theta')||_2}{\sqrt{2}\sigma} \leq \frac{||X\theta||}{\sqrt{2\sigma^2}} + \frac{||X\theta'||}{\sqrt{2\sigma^2}} \leq \frac{8\delta\sqrt{n}}{\sqrt{2\sigma^2}}$$

Now we set $\delta = \{X\theta : ||x\theta||_2 \leq 4\delta\sqrt{n}\}$ and derive:

$$\mathcal{D}_{KL}(\mathbb{P}^n_{\theta_j}, \mathbb{P}^n_{\theta_k}) \leq \frac{64n\delta^2}{2\sigma^2} = \frac{32n\delta^2}{\sigma^2}$$

References

[wainwright] M. Wainwright, "High Dimensional Statistics," *Prerelease*, 2019

[CW87] D. Coppersmith and S. Winograd, "Matrix multiplication via arithmetic progressions," *Proceedings of the 19th ACM Symposium on Theory of Computing*, 1987, pp. 1–6.