CS861: Theoretical Foundations of Machine Learning Lecture 15 - 10/07/2024

University of Wisconsin–Madison, Fall 2024

Lecture 15: Density Estimation, Prediction Problems

Lecturer: Kirthevasan Kandasamy *Scribed by: Xuanfei Ren, Leitian Tao*

In this lecture, we will continue our discussion on density estimation, provide an upper bound via kernel density estimation, and introduce a framework for prediction problems.

1 Upper Bound via Kernel Density Estimation

Kernel Density Estimation (KDE) is a non-parametric technique to estimate the probability density function of a continuous random variable. The Kernel Density Estimation (KDE) has the following form:

$$\hat{p}(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{t-x_i}{h}\right),$$

where $\hat{p}(t)$ is the estimated density at point t, n is the number of data points, and x_i are the observed data points. h is the bandwidth parameter, which plays a critical role in KDE. K is a (smoothing) kernel with the following properties:

1. Normalization:

$$\int K(u)du = 1,$$

This ensures the result will integrate to 1 over its entire domain, maintaining the fundamental property of a probability density function.

2. Symmetry:

$$K(u) = K(-u),$$

This property ensures that the kernel is symmetric around zero. As a result, the estimated density will not be biased towards any direction from the point of estimation.

For the problem at hand, the kernel selected is

$$K(t) = \mathbb{I}(|t| \leq \frac{1}{2}),$$

which is sufficient for Lipschitz functions. This is a simple uniform kernel.

We can bound the risk as follows:

$$\begin{aligned}
\mathbb{E}[\|p-\hat{p}\|_2^2] &= \mathbb{E}\left[\int (p-\hat{p})^2 dt\right] \\
&= \mathbb{E}\left[\int (p-\mathbb{E}(\hat{p}))^2 dt + \int (\mathbb{E}(\hat{p})-\hat{p})^2 dt + 2\int (p-\mathbb{E}(\hat{p}))(\mathbb{E}(\hat{p})-\hat{p})dt\right] \\
&= \int_0^1 \underbrace{(p(t)-\mathbb{E}(\hat{p}(t)))^2}_{\textbf{bias}(t)} dt + \int_0^1 \underbrace{\mathbb{E}[(\hat{p}(t)-\mathbb{E}(\hat{p}(t)))^2]}_{\text{var}(t)} dt + 2\int_0^1 (p-\mathbb{E}(\hat{p}))\underbrace{\mathbb{E}[\mathbb{E}(\hat{p})-\hat{p}]}_{0} dt
\end{aligned}$$

The bias and variance terms can be written and bounded below. The bias of an estimator indicates how far on average the estimate is from the true value. In our context, the bias term is derived from:

$$\begin{aligned}
\text{bias}(t) &= \mathbb{E}[\hat{p}(t)] - p(t) \\
&= \mathbb{E}_{X\sim P}\left[\frac{1}{h}K\left(\frac{t-X}{h}\right)\right] - p(t) \quad (\text{Apply } \mathbb{E}\left[\frac{1}{n}\sum_i Z_i\right] = \mathbb{E}[Z]) \\
&= \int \frac{1}{h}K\left(\frac{t-x}{h}\right)p(x)dx - p(t) \\
&= \int K(u)\left(p(t+uh) - p(t)\right)du \\
&= \int K(u)(p(t+uh) - p(t))du
\end{aligned}$$

The bias depends on the choice of kernel and bandwidth h. Then we can bound the bias term:

$$\begin{aligned}
|\text{bias}(t)| &\le \int \left|K(u)\left(p(t+uh) - p(t)\right)\right| du \\
&\le \int K(u)L|uh|du \\
&= Lh\int K(u)|u|du \\
&= C_1 Lh
\end{aligned}$$

where $C_1 = \int K(u)|u|du = \frac{1}{4}$ is a constant that essentially depends on the kernel, and L is the Lipschitz constant, signifying the bound on the rate of change of our function.

Variance quantifies the dispersion of an estimator around its expected value. In the context of KDE, variance arises from the randomness in the sample. For our setup, the variance term is captured by:

$$\begin{aligned}
\text{Var}(t) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{t-X_i}{h}\right)\right) \\
&= \frac{1}{n}\text{Var}_{X\sim P}\left(\frac{1}{h}K\left(\frac{t-X}{h}\right)\right) \quad (\text{Apply Var}\left(\frac{1}{n}\sum_{i=1}^{n}Z_i\right) = \frac{1}{n}\text{Var}(Z_1)) \\
&\le \frac{1}{nh^2}\int K^2\left(\frac{t-x}{h}\right)p(x)dx \\
&= \frac{1}{nh}\int K^2(u)p(t+uh)du \\
&\le \frac{B}{nh}\int K^2(u)du = \frac{B}{nh}
\end{aligned}$$

Where B is a constant that bounds the product of the squared kernel and the true density. Combine these bounds for bias and variance terms, we have:

$$\begin{aligned}
\mathbb{E}_S[\|p-\hat{p}\|_2^2] &\le \int_0^1 \text{bias}^2(t)dt + \int_0^1 \text{Var}(t)dt \\
&\le \frac{1}{16}L^2h^2 + \frac{B}{nh} \\
&= O\left(\frac{L^{2/3}}{n^{2/3}}\right). \quad (\text{choose } h = \frac{1}{n^{1/3}L^{2/3}})
\end{aligned}$$

2 Lower Bounds for Prediction Problems

So far, we have talked about estimation in some metrics, for instance:

$$\rho(\theta_1, \theta_2) = |\theta_1 - \theta_2|, \quad \theta(P) \in \mathbb{R}$$
$$\rho(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_{L_2} = \sqrt{\int (\theta_1 - \theta_2)^2}, \quad \theta(P) \in \{f : \mathcal{X} \to \mathbb{R}\}$$

Next, we will develop a framework for proving lower bounds for prediction problems (among others). The framework is comprised of the following components:

1. Data space $\mathcal{Z}$. This is the space from which data samples arise.
2. A family of distributions $\mathcal{P}$, where $\forall P \in \mathcal{P}, \text{supp}(P) \subseteq \mathcal{Z}$. This represents a collection of probability distributions from which the data can be drawn.
3. A hypothesis/parameter space $\mathcal{H}$. This contains all the potential hypotheses or models that we might use to make predictions.
4. An *instance loss*, $f : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$, where $f(h, Z)$ is the loss of hypothesis h on instance Z. This measures how well a particular hypothesis h from $\mathcal{H}$ performs on an instance Z from $\mathcal{Z}$.
5. The *population loss*, $F(h, P) = \mathbb{E}_{Z \sim P}[f(h, Z)]$; *Excess population loss*, $L(h, P) = F(h, P) - \inf_{h' \in \mathcal{H}} F(h', P)$. These terms indicate how well our hypothesis does on average (under distribution P) and how it compares to the best possible hypothesis in $\mathcal{H}$, respectively.
6. A dataset S is drawn from some $P \in \mathcal{P}$.
7. An estimator $\hat{h}$, which maps the dataset S to a hypothesis in $\mathcal{H}$. Note that we overload notation here for $\hat{h}$:

$$\hat{h} : \text{a map from data to } \mathcal{H} \quad \text{(estimator)}$$
$$\hat{h} : \text{as the estimate} \quad (\hat{h} \in \mathcal{H})$$

8. Risk of estimator $\hat{h}$:

$$\begin{aligned} R(\hat{h}, P) &= \mathbb{E}[L(\hat{h}(S), P)] \\ &= \mathbb{E}[F(\hat{h}(S), P)] - \inf_{h \in \mathcal{H}} F(h, P) \end{aligned}$$

9. Minimax risk:

$$R^* = \inf_{\hat{h}} \sup_{P \in \mathcal{P}} R(\hat{h}, P)$$

Next, let us see an example.
Example 1. Excess risk in classification/regression

- Data space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}, \mathcal{H} : \{h : \mathcal{X} \to \mathcal{Y}\}$. This is the set of all possible prediction functions mapping from features $\mathcal{X}$ to outcomes $\mathcal{Y}$.
- Instance loss (in classification), $f(h, (X, Y)) = 1(h(X) \neq Y)$. This measures the discrepancy between a predicted and actual class label.

- $F(h, P) = \mathbb{E}_{X,Y\sim P}[1(h(X) \neq Y)]$ is the "risk". This is the expected value of the instance loss over the joint distribution of X and Y, and can be understood as the overall error rate of the classifier.
- $L(h, P) = F(h, P) - F(h^\star, P)$, where $h^\star$ is the Bayes optimal classifier, i.e.

$$h^\star = \arg\max_{y\in\mathcal{Y}} \mathbb{P}(Y = y|X = x),$$

  and $L(h, P)$ is the *excess risk* in classification. The excess risk quantifies how much worse our classifier h performs compared to the Bayes optimal classifier $h^\star$.
- In regression, we define $f(h, (X, Y)) = (h(X) - Y)^2$. The typical loss function used is the squared loss.
- $L(h, P) = \mathbb{E}_{X\sim P}[(h(X) - Y)^2] - \mathbb{E}_{X\sim P}[(h^\star(X) - Y)^2]$, where $h^\star(x) = \mathbb{E}[Y|X = x]$, $L(h, P)$ is the *excess risk* in regression. The excess risk measures how much worse our regression function performs compared to the optimal regression function $h^\star(x)$, which is the conditional expectation.

Remark Note that this is different to $\Phi \circ \rho(h_1, h_2) = \|h_1 - h_2\|^2_{L_2}$, which captures the squared difference between two hypotheses h_1 and h_2, and $\mathbb{E}_S[\Phi \circ \rho(\hat{h}, h^\star)]$ measures how much estimator $\hat{h}$ deviates from the optimal hypothesis $h^\star$.

Example 2 (Estimation error in a finite VC class). $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$

Our estimator $\hat{h}$ will choose some hypothesis in $\mathcal{H}$ using data. We can now view $L(h, P)$ as the estimation error. Recall, that letting $h^\star$ be the Bayes' optimal classifier, we can write

$$F(h, P) - F(h^\star, P) = \underbrace{F(h, P) - \inf_{h'\in\mathcal{H}} F(h', P)}_{\text{estimation error}=L(h,P)} + \underbrace{\inf_{h'\in\mathcal{H}} F(h', P) - F(h^\star, P)}_{\text{approximation error}}$$

In Homework 1, we saw that for ERM, when $\mathcal{H}$ has VC dimension $d_\mathcal{H}$, we have

$$R(\hat{h}_{ERM}, P) = \underbrace{\mathbb{E}_S[F(\hat{h}_{ERM}(S), P)]}_{\mathbb{E}_S[\mathbb{E}_{X,Y\sim P}[1(\hat{h}_{ERM}(S)(X)\neq Y)]]} - \inf_{h\in\mathcal{H}} F(h, P) \in \tilde{\mathcal{O}}\left(\sqrt{\frac{d_\mathcal{H}}{n}}\right)$$

We will use this framework to show a corresponding lower bound

$$\inf_{\hat{h}} \sup_{P\in\mathcal{P}} \left(\mathbb{E}_{S\sim\mathcal{P}}[F(\hat{h}(S), P)] - \inf_{h'\in\mathcal{H}} F(h', P)\right) \in \Omega\left(\sqrt{\frac{d_\mathcal{H}}{n}}\right)$$

Remark This framework applies for any statistical problem as long as you can define an instance loss $F(h, P)$.

To proceed, we will first define the separation of two distributions, with respect to a given hypothesis class and loss L.

Definition 1 (Separation). For two distributions P, Q, define the separation $\Delta(P, Q)$ as

$$\Delta(P, Q) = \sup\{\delta \geq 0; L(h, P) \leq \delta \Rightarrow L(h, Q) \geq \delta, \forall h \in \mathcal{H}, L(h, Q) \leq \delta \Rightarrow L(h, P) \geq \delta, \forall h \in \mathcal{H}\}$$

- P, Q are δ-separated if any hypothesis that does well on P (i.e. $L(h, P) \leq \delta$), does poorly on Q (i.e. $L(h, Q) \geq \delta$).
- We say a collection of distributions $\{P_1, \ldots, P_N\}$ are δ-separated if $\Delta(P_i, P_k) \geq \delta, \forall i \neq k$.

The following theorem can be proved using a similar technique to our previous theorem on reducing estimation to testing. You will do this in your homework.

Theorem 2 (Reduction to testing). Let $\{P_1, \dots, P_N\}$ be a δ-separated subset of $\mathcal{P}$. Let ψ be any test which maps the dataset to $[N]$. Then

$$R^* \geq \delta \inf_{\psi} \max_{j \in [N]} P_j(\psi \neq j)$$

We can then establish the following statements from the above result when S consists of n i.i.d data points.

Theorem 3 (Le Cam & Fano Method).

1. If $\{P_0, P_1\}$ are δ-separated,

$$R^* \geq \frac{\delta}{2} \|\mathbb{P}_0 \wedge \mathbb{P}_1\| \geq \frac{\delta}{4} e^{-\mathrm{KL}(P_0, P_1)}.$$

Hence, for i.i.d. data $S \sim P^n$, if $\mathrm{KL}(P_0, P_1) \leq \frac{\log(2)}{n}$, then $R^* \geq \frac{\delta}{8}$.

2. Local Fano Method: If $\{P_1, \dots, P_N\}$ are δ-separated, then

$$R^* \geq \delta \left(1 - \frac{\frac{1}{N^2} \sum_{j,k} \mathrm{KL}(P_j, P_k) + \log(2)}{\log(N)}\right).$$

Hence, for i.i.d. data $S \sim P^n$, if $\mathrm{KL}(P_j, P_k) \leq \frac{\log(N)}{4n}$, and $N \geq 16$, then $R^* \geq \frac{\delta}{8}$.

3 Application: Classification in a VC class

We will now use the above results to prove a lower bound for classification in a VC class.

Theorem 4. Let $\mathcal{P}$ be the set of all distributions supported on $\mathcal{X} \times \{0, 1\}$. Let $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$ be a hypothesis class with VC dimension $d \geq 8$. Let $S = \{(X_1, Y_1), \dots, (X_n, Y_n)\} \sim_{iid} P$, where $P \in \mathcal{P}$. Then, for any estimator $\hat{h}$ which maps the dataset S to a hypothesis in $\mathcal{H}$,

$$R^* = \inf_{\hat{h}} \sup_{P \in \mathcal{P}} \left(\mathbb{E}[F(\hat{h}, P)] - \inf_{h' \in \mathcal{H}} F(h', P)\right) \in \Omega\left(\sqrt{\frac{d}{n}}\right).$$

Acknowledgements

These notes are based on scribed lecture materials prepared in Fall 2023 by Haoyue Bai, Zexuan Sun, Ransheng Guan and Haoran Xiong.