36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 21: October 19

Lecturer: Siva Balakrishnan

21.1 Review and Outline

In the last class we discussed the kernel density estimator:

1. Motivation, estimator
2. Bias and Variance of KDEs for twice-differentiable densities

Today we will discuss cross-validation, and then start discussing hypothesis testing.

21.2 Cross-validation

In both non-parametric regression and in density estimation we had an unknown bandwidth parameter. In practice, this tuning parameter is chosen using cross-validation. We will only go over this briefly in this class. Lets first discuss this in the context of regression, and then we will discuss density estimation.

Before we discuss cross-validation lets understand a train-test split, i.e. suppose we split the data into two parts we can estimate our regression function for a grid of bandwidths $\{h_1, \ldots, h_M\}$. Now, we want to pick one of these bandwidths.

In this case, we could simply check how well we can predict on the test set, i.e.,

$$\widehat{R}(\widehat{f}_{h_1}, f) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (Y_i - \widehat{f}_{h_1}(X_i))^2,$$

and repeat this for each of the bandwidths and then pick the bandwidth that minimizes this. Why does this work? Essentially, train-test splits allow us to *estimate* the risk. We should be a little bit careful about what risk we are estimating:

$$\mathbb{E}(\widehat{R}(\widehat{f}_{h_1}, f)) = \mathbb{E}(f(X) + \epsilon - f_{h_1}(X))^2 = \mathbb{E}(f(X) - f_{h_1}(X))^2 + \sigma^2.$$

We can of course ignore the σ^2, but one should notice that the risk we are estimating:

$$\mathbb{E}(f(X) - f_{h_1}(X))^2 = \int (f(x) - f_{h_1}(x))^2 p(x) dx,$$

where p is the density of the covariates. This is sometimes called the $L_2(\mathbb{P})$-risk, as opposed to the L_2-risk:

$$R = \int (f(x) - f_{h_1}(x))^2 dx.$$

Most people would consider the $L_2(\mathbb{P})$-risk to be more natural, since it puts less weight in places where you have less data.

Practitioners might be concerned that this the train-test split is wasteful of the data: you might need a pretty large test set to get a good estimate and this is data that you might have instead used to estimate the model. Also, train-test splits have a "lottery" effect: you might get unlucky in the way you split the data and this could affect results.

K-fold cross-validation tries to get around this by splitting the data into K pieces (think of K as a small number like 5). Now, we repeat the train-test split K times, each time we use $K-1$ pieces for training and the K-th piece for testing. In this way we get, K estimates of the error for each value of the bandwidth. We average these K numbers to get our risk estimate. Finally, we choose the value of the bandwidth that minimizes the risk. The extreme case of K-fold cross-validation is called leave-one-out or n-fold cross-validation. Here we leave out one observation, and try to predict it and then cycle through the observations.

Regression is a case when its relatively clear how to estimate the risk. It is much less clear in the context of density estimation. In density estimation, we want to estimate:

$$R(\widehat{f}_h, f) = \int (\widehat{f}_h(x) - f(x))^2 dx,$$

and pick an h that (approximately) minimizes this. We can expand this out:

$$R(\widehat{f}_h, f) = \int \widehat{f}_h^2(x) dx + \int f^2(x) dx - 2 \int \widehat{f}_h(x) f(x) dx,$$

the second term does not depend on h so we can ignore it while selecting h, the first term of course we can calculate, but the third term depends on the unknown true density. The standard procedure is to estimate this using the samples in a leave-one-out fashion. The estimator of the risk is then:

$$\widehat{J}(\widehat{f}_h, f) = \int \widehat{f}_h^2(x) dx - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_{-i,h}(X_i).$$

Here $\widehat{f}_{-i,h}(X_i)$ is the KDE without the ith observation, with bandwidth h, evaluated at the ith observation. Finally, we choose h to minimize this quantity.

21.3 A simple analysis of the train-test split

Lets try to understand cross-validation in a simple scenario. This is an example of something that we refer to as model selection in statistics. We will do this in the context of point estimation, but one could use *exactly* the same argument for bandwidth selection.

Say we have models $\mathcal{M}_1, \ldots, \mathcal{M}_M$. These are different models that we think might be reasonable fits to the data. Now, we observe our data $(X_1, \ldots, X_{2n})$ and randomly split it into train and test sets of size n each. We really should refer to the test set as a validation set but we will ignore this for today.

On the train set, we fit our models (say using the MLE), and compute point estimates $\widehat{\theta}_1, \ldots, \widehat{\theta}_M$. Now, suppose that we want to select the model/estimate that fits the data well. We will use the negative log-likelihood as our measure, i.e., we want an estimate that has low negative log-likelihood. This is the same as using the KL divergence as our loss function.

We can use the test set to estimate the negative log-likelihood:

$$R_i = \frac{-1}{n} \sum_{i=1}^{n} \log f_{\widehat{\theta}_i}(X_{n+i}).$$

Note that:

$$\mathbb{E}(R_i) = -\mathbb{E}_{f_{\theta^*}} \log f_{\widehat{\theta}_i}(X) = KL(f_{\theta^*} || f_{\widehat{\theta}_i}) - \mathbb{E}_{f_{\theta^*}} \log f_{\theta^*}(X),$$

so we are estimating the KL divergence upto some term that does not depend on $\widehat{\theta}_i$. So minimizing $\mathbb{E}(R_i)$ is equivalent to minimizing the KL divergence.

We can now use the LLN to argue that if the test-set size goes to ∞ then our risk estimates converge to their expectations, and then we will find the model/estimate with the lowest KL to the true model.

Suppose however we wanted to be more precise, and try to understand the role of the test set size and the number of models M?

We could use Hoeffding's inequality. This will need an assumption that $|\log f_\theta(X)| \leq B$ for every θ and X that we care about (this can be relaxed using more complex techniques). Now, notice that the following is an important but straightforward consequence of Hoeffding's inequality:

$$\mathbb{P}(\max |R_i - \mathbb{E}(R_i)| \geq \epsilon) \leq 2M \exp(-2n\epsilon^2/(4B^2)).$$

This is true since for each i we know that

$$\mathbb{P}(|R_i - \mathbb{E}(R_i)| \geq \epsilon) \leq 2 \exp(-2n\epsilon^2/(4B^2)).$$

so we can obtain the desired inequality via a union bound (if the max exceeds ϵ at least one of the terms must exceed ϵ).

Define,

$$\epsilon_n = \sqrt{\frac{4B^2 \log(2M/\alpha)}{n}},$$

then we know that

$$\mathbb{P}(\max |R_i - \mathbb{E}(R_i)| \geq \epsilon_n) \leq \alpha.$$

Suppose we select the model $\widehat{i} = \arg\min_i R_i$, and let $i^* = \arg\min_i \mathbb{E}(R_i)$, then we have that with probability at least $1 - \alpha$:

$$\mathbb{E}(R_{\widehat{i}}) \leq R_{\widehat{i}} + \epsilon_n \leq R_{i^*} + \epsilon_n \leq \mathbb{E}(R_{i^*}) + 2\epsilon_n.$$

So the model we select will be sub-optimal by at most $2\epsilon_n$. In regression, we would use exactly the same reasoning, but just replace the risk with the squared loss. Similar logic is helpful in density estimation as well.

Reasoning about K-fold cross-validation turns out to be much more challenging, because the data re-use breaks independence assumptions.

21.4 Hypothesis Testing

Today we will just talk about the basic terminology. In the next few lectures we will go into much more detail. The typical (and most basic) setting is that we observe:

$$X_1, \ldots, X_n \sim f_\theta$$

and want to test if $\theta = \theta_0$ or not. A typical example is where we have a coin and would like to know if the coin is fair or not. In a clinical trial we might have a control group and a group taking the drug, and we would like to know if the difference in some health outcome is 0 or not.

The way we formalize this is by defining a *null hypothesis* H_0 and an *alternative hypothesis* H_1.

So we would say:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0.$$

The more general case is that we have two sets of parameters Θ_0 and Θ_1 which are non-overlapping, i.e. $\Theta_0 \cap \Theta_1 = \emptyset$ and would like to test the hypothesis:

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1.$$

We will refer to the case when Θ_0 is a single point as a *simple* null versus the more general case of a *composite* null.

In hypothesis testing, the question is never if the null hypothesis is true or not. Rather the question of interest is whether we have sufficient evidence to reject the null hypothesis or not. So in hypothesis testing, there are two possibilities you reject the null hypothesis or you retain it. To reiterate, retaining the null hypothesis is not a statement about whether it is true or not.

There are two types of errors one might make in hypothesis testing: a *Type I* error is when the null hypothesis is true but was incorrectly rejected, and a *Type II* error is when the alternate hypothesis was true but we failed to reject the null.