Lecture 7: Nonparametric Classification and Regression

Akshay Krishnamurthy
akshay@cs.umass.edu

September 26, 2017

1 Recap

So far in the course we have mostly been talking about how to prove generalization bounds. We saw a bunch of concentration inequalities, talked about uniform convergence, and proved generalization bounds in terms of VC-dimension, Rademacher complexity, and covering numbers. This has mostly focused on statistical aspects of machine learning. In the next part of the course, we will turn to more algorithmic aspects of machine learning.

We will start with a class of algorithms that are more popular in the statistics community, but have a clean analysis and allow us to dig deeper into the bias/variance or approximation/estimation tradeoff that we have seen several times in the course. Specifically here, we will make a fairly weak assumption on true distribution, and we will use this to bound the overall error, rather than just the excess risk.

2 Nonparametric Regression

We consider the regression setting, where $\mathcal{X}$ is a compact subset of $\mathbb{R}^d, \mathcal{Y} = \mathbb{R}$ and the loss function is the square loss $\ell(y, y') = \frac{1}{2}(y - y')^2$. The basic ideas also apply for classification. Let $\eta(x) = \mathbb{E}[Y|X = x]$ which is the true regression function. In Homework 1 and briefly in Lecture 1, we saw how to estimate η when we made the linearity assumption that $\eta(x) = \langle \beta^\star, x\rangle$. Today we'll see how to estimate η under much weaker assumptions.

Observe first that optimal predictor for the square loss is $\eta(x)$ in the sense

$$\eta = \operatorname*{argmin}_{f:\mathcal{X}\to\mathcal{Y}} R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \frac{1}{2}(y - f(x))^2.$$

Moreover, if we find an estimator $\hat{\eta}$, we get

$$\begin{aligned} R(\hat{\eta}) &= \mathbb{E}_{\mathcal{D}} \frac{1}{2}(y - \hat{\eta}(x))^2 = \mathbb{E}_{\mathcal{D}} \frac{1}{2}(y - \eta(x) + \eta(x) - \hat{\eta}(x))^2 \\ &= \mathbb{E}_{\mathcal{D}} \frac{1}{2}(y - \eta(x))^2 + \mathbb{E}_{\mathcal{D}} \frac{1}{2}(\eta(x) - \hat{\eta}(x))^2 \\ &= R(\eta) + \frac{1}{2} \underbrace{\int_{\mathcal{X}} (\eta(x) - \hat{\eta}(x))^2 d\mathcal{D}(x)}_{\|\eta - \hat{\eta}\|^2_{L_2(\mathcal{D})}}. \end{aligned}$$

So to analyze an estimator $\hat{\eta}$ we need to bound the integrated squared error to the true regression function η. We actually kind of did this already for the linear case in Homework 1.

Using essentially the same argument, we can further decompose the integrated squared error into two terms:

$$\|\eta - \hat{\eta}\|^2_{L_2(\mathcal{D})} = \|\hat{\eta} - \mathbb{E}\hat{\eta}\|^2_{L_2(\mathcal{D})} + \|\eta - \mathbb{E}\hat{\eta}\|^2_{L_2(\mathcal{D})} = \int \mathrm{Var}(\hat{\eta}(x)) d\mathcal{D}(x) + \int \mathrm{Bias}^2(\hat{\eta}(x)) d\mathcal{D}(x)$$

This decomposition, which applies to any regression problem demonstrates a bias-variance tradeoff, which is similar to the approximation/estimation tradeoff. If our estimator belongs to a simple model class, then it will have high bias, but the variance will be small. On the other hand, if the estimator belongs to a rich model class, then the bias will be smaller, but it is harder to estimate something from a rich class, so the variance will be higher. Today we'll toggle between these with a single parameter, which will help us manage this tradeoff.

In most simpler statistics problems, we would use an unbiased estimator so the second term will be zero. For example, if we are just interested in estimating the mean of a distribution, we often use the sample mean, which is unbiased. However, using regularization (or shrinkage), you can actually do better by introducing some bias. So in another way, you can think of regularization as helping manage the bias/variance tradeoff.

If we don't make any assumptions about the regression function m then we're trying to estimate an entire function, which in a sense is an infinite-dimensional parameter, and it seems inevitable that we must introduce some bias if we want any non-trivial performance. The function could just be too wiggly for us to fit it well at all. However, once we do introduce bias, there are several estimators that make sense. We've mostly seen parametric estimators that use particular representations, like linear functions. Today we'll see non-parametric estimators that are much more flexible.

Example 1 (Histogram estimator)**.** *Let $\mathcal{X} = [0,1]^d$ and choose a small number $h \in (0,1)$. Partition the cube into $N = (1/h)^d$ bins $B_1, \ldots, B_N$ of side-length h and in each bin use the average of the Ys as the prediction. More formally*

$$\hat{\eta}(x) = \sum_{j=1}^{N} \mathbf{1}\{x \in B_j\} \frac{\sum_{i=1}^n y_i \mathbf{1}\{x_i \in B_j\}}{\sum_{i=1}^n \mathbf{1}\{x_i \in B_j\}}$$

This estimator is clearly biased for η, since it is always piecewise constant, but we made no such restriction on η itself. In this sense, if h is small the bias is smaller, since we are approximate a function by a piecewise function with more and more pieces. On the other hand, if h is small, then the variance is larger since we are taking sample averages with less and less data. In the homework you'll analyze a histogram estimator for a slightly different problem.

We'll study a slightly different estimator from the histogram, which is called **kernel regression**. The idea is pretty similar though. Instead of taking average within bins, we'll take *locally weighted averages* to estimate the regression function at a single point. Once you've seen the histogram estimator, the following should feel not-too-different

$$\hat{\eta}(x) = \frac{\sum_{i=1}^n y_i \mathbf{1}\{\|x_i - x\| \le h\}}{\sum_{i=1}^n \mathbf{1}\{\|x_i - x\| \le h\}}$$

If the denominator is zero for any x, then we simply set $\eta(x) = 0$. This is the estimator we'll analyze today, but essentially the same analysis applies to other "weighting functions," which are called **smoothing kernels** or just **kernels**.

Definition 1. *A one-dimensional **smoothing kernel** is any function K such that $K(x) \ge 0$ and*

$$\int K(x)dx = 1, \quad \int xK(x)dx = 0, \quad \textit{and} \quad \sigma_K^2 = \int x^2 K(x)dx > 0.$$

In our case, we will take $K(x) = \frac{1}{2}\mathbf{1}\{|x| \le 1\}$ which clearly satisfies the properties. Then in the estimator we replace the indicator with $K(\|x_i - x\|/h)$.

2.1 Analysis

We will analyze the estimator under a fairly weak Lipschitz-continuity assumption.

Assumption 2. *We assume that η is L-Lipschitz continuous, which means that $|\eta(x) - \eta(z)| \le L\|x - z\|$ for all $x, z \in \mathbb{R}^d$.*

Theorem 3. *Assume that* $\mathrm{Var}(Y|X = x) \le \sigma^2 < \infty$*. Then*

$$\mathbb{E}_{S^n \sim \mathcal{D}^n} \|\hat{\eta} - \eta\|_{L_2(\mathcal{D})}^2 \le c_1 h^2 + \frac{c_2}{nh^d}.$$

Hence, if we choose $h = cn^{-1/(d+2)}$ then the integrated MSE is $O(n^{-2/(d+2)})$.

Proof. We'll do a bias/variance-type decomposition like the one above, but for the bias term we'll only take expectation over Y. Define

$$\bar{\eta}(x) = \frac{\sum_{i=1}^n \eta(X_i)\mathbf{1}\{\|X_i - x\| \le h\}}{\sum_{i=1}^n \mathbf{1}\{\|X_i - x\| \le h\}}$$

Then if the denominator is non-zero, we can use the Lipschitz property to bound the difference to $\eta(x)$. If the denominator is zero, we can just bound with $\eta(x)$. Or formally, letting $A(x) = \mathbf{1}\{\sum_{i=1}^n \mathbf{1}\{\|X_i - x\| \le h\} > 0\}$, we have

$$|\bar{\eta}(x) - \eta(x)|^2 \le \left|\frac{\sum_{i=1}^n (\eta(X_i) - \eta(x))\mathbf{1}\{\|X_i - x\| \le h\}}{\sum_{i=1}^n \mathbf{1}\{\|X_i - x\| \le h\}} A(x) + \eta(x)(1 - A(x))\right|^2 \le L^2h^2 + \eta(x)^2(1 - A(x))$$

For the variance, let us condition on $X_1, \ldots, X_n$ and just take expectation with respect to $Y_{1:n}$, under the assumption that $A(x)$ holds.

$$\mathbb{E}\left((\hat{\eta}(x) - \bar{\eta}(x))^2 | X_{1:n}\right) = \frac{\sum_{i=1}^n \mathrm{Var}(Y_i|X_i)\mathbf{1}\{\|X_i - x\| \le h\}}{(\sum_{i=1}^n \mathbf{1}\{\|X_i - x\| \le h\})^2} \le \frac{\sigma^2}{\sum_{i=1}^n \mathbf{1}\{\|X_i - x\| \le h\}}$$

Of course if $A(x) = 0$ then the conditional variance is zero. So now we have to control the denominator here, but only when $A(x) = 1$. Let $Z = \sum_{i=1}^n \mathbf{1}\{\|X_i - x\| \le h\}$ and observe that Z is distributed like a binomial random variable with n draws and with parameter $q = \mathcal{D}(X \in B(x, h))$.

Lemma 4. *Let $Z \sim Bin(n, q)$. Then,*

$$\mathbb{E}\frac{\mathbf{1}\{Z > 0\}}{Z} \le \frac{2}{nq}$$

Proof.

$$\begin{aligned}
\mathbb{E}\frac{\mathbf{1}\{Z > 0\}}{Z} &\le \mathbb{E}\frac{2}{1 + Z} = \sum_{k=0}^{n} \frac{2}{k+1}\binom{n}{k} q^k(1-q)^{n-k} \\
&= \frac{2}{(n+1)q} \sum_{k=0}^{n} \binom{n+1}{k+1} q^{k+1}(1-q)^{n-k} \\
&\le \frac{2}{(n+1)q} \sum_{t=0}^{n+1} \binom{n+1}{t} q^{t}(1-q)^{n-t+1} \\
&= \frac{2}{(n+1)q}(q + (1-q))^{n+1} = \frac{2}{(n+1)q} \le \frac{2}{nq}.
\end{aligned}$$

The first inequality is based on a pointwise upper bound. The first equality just expands the expectation using the Binomial distribution. The second inequality changes does a change of variables $t = k + 1$, which ends up adding the term for $t = 0$. Finally we use the binomial theorem and simple approximations. □

The lemma reveals that

$$\mathbb{E}\left(\frac{A(x)}{\sum_{i=1}^n \mathbf{1}\{\|X_i - x\| \le h}\right) \le \frac{2}{n\mathcal{D}(B(x, h))}$$

So we can upper bound the variance by

$$\mathrm{Var}(\hat{\eta}) = \mathbb{E}_{X_1^n} \int \mathrm{Var}(\hat{\eta}(x) \mid X_1^n) d\mathcal{D}(x) \le \frac{2\sigma^2}{n} \int \frac{d\mathcal{D}(x)}{\mathcal{D}(B(x, h))}$$

Now cover the support of $\mathcal{D}$ at scale $h/2$ with points $z_1, \ldots, z_M$, so that $\bigcup B(z_j, h/2) \supset \mathrm{supp}(\mathcal{D})$. A volumetric argument shows that $M \le O(1/h^d)$, which should be familiar from the number of bins in the histogram estimator.

Lemma 5 (Volumetric argument)**.** *Let supp($\mathcal{D}$) be contained in the euclidean ball of some radius B then there exists a set of $M \leq O(1/h^d)$ points that cover supp($\mathcal{D}$) at scale $h/2$.*

Proof. Let $z_1, \ldots, z_M$ be the largest set of points such that for each pair $\|z_i - z_j\| \geq \epsilon$ (this is called a packing). Since it is of maximal size, no other point can be added and hence this point set is a cover at scale ϵ. Moreover, since all points in the packing have distance at least ϵ from each other, the balls $B(z_i, \epsilon/2$ are disjoint and they are all contained in a ball of radius at most $B + \epsilon/2$. Thus

$$M\text{vol}(B(\epsilon/2)) \leq \text{vol}(B(B + \epsilon/2))$$

which implies

$$M(\epsilon/2)^d\text{vol}(B(1)) \leq (B + \epsilon/2)^d\text{vol}(B(1))$$

So that $M \leq (2B/\epsilon + 1)^d \leq O(1/\epsilon^d)$. □

Now we can finish the bound for the variance

$$\text{Var}(\hat{\eta}) \leq \frac{2\sigma^2}{n}\sum_{j=1}^{M}\int\frac{\mathbf{1}\{x \in B(z_j, h/2)\}d\mathcal{D}(x)}{\mathcal{D}(B(x,h))} \leq \frac{2\sigma^2}{n}\sum_{j=1}^{M}\int\frac{\mathbf{1}\{x \in B(z_j, h/2)\}d\mathcal{D}(x)}{\mathcal{D}(B(z_j,h/2))} \leq \frac{2\sigma^2 M}{n} \leq \frac{c_1}{nh^d}.$$

The last term that we need to deal with is the extra part in the bias.

$$\begin{aligned}\mathbb{E}_{S^n\sim\mathcal{D}^n}\int\eta(x)^2(1 - A(x))d\mathcal{D}(x) &= \sup_x \eta^2(x)\int\mathbb{E}_{S^n\sim\mathcal{D}^n}[1 - A(x)]d\mathcal{D}(x) = \sup_x \eta^2(x)\int(1 - \mathcal{D}(B(x,h))^n d\mathcal{D}(x)\\ &\leq \sup_x \eta^2(x)\int\exp(-n\mathcal{D}(B(x,h)))\frac{n\mathcal{D}(B(x,h))}{n\mathcal{D}(B(x,h))}d\mathcal{D}(x)\\ &\leq \sup_x \eta^2(x)\sup_u u\exp(-u)\int\frac{1}{n\mathcal{D}(B(x,h))}d\mathcal{D}(x)\\ &\leq \sup_x \eta^2(x)\sup_u u\exp(-u)\frac{c_1}{nh^d} \leq \frac{c^2}{nh^d}.\end{aligned}$$

Collecting terms, we get $O(1/nh^d + h^2)$ ignoring lots of constants. □

3 Remarks and extensions

1. **Other non-parametric estimators.** We studied only the kernel regression estimator, but there are other non-parametric methods that are worth briefly mentioning. The histogram we saw above and you'll study in the homework, and the other popular one is the nearest-neighbor estimator. Here, at every point x you estimate $\eta(x)$ by taking average of the labels for the k-nearest points. All of these work fairly well although for computational reasons the k-NN method is often the most preferred, since you can build data structures to find nearest neighbors in sublinear time.

2. **Convergence rate.** While the kernel regression estimator is consistent under very weak assumptions, namely just smoothness, the convergence rate of $n^{-\frac{2}{d+2}}$ scales quite poorly with the dimension. This is known as the *curse of dimensionality*, which I'm sure many of you have heard about. Unfortunately without making stronger assumptions this is unavoidable.

3. **Higher order smoothness.** You can also analyze the estimator under stronger smoothness assumptions like Lipschitz continuous derivatives, which is called a Holder class. If s derivatives are continuous, then the rate is $O(n^{\frac{-2s}{2s+d}})$ which shows that more smoothness helps in terms of estimation. Lipschitz continuity is a special case where $s = 1$.

4. **Bandwidth selection.** All of these estimators have a bandwidth type parameter, which in our case was h. We have to choose h optimally for our problem, in terms of the dimension, the smoothness properties, etc. This is a fairly well-studied topic and there are many heuristic and principled methods, for example cross validation.

5. **Adaptivity.** While the convergence rate can be quite poor, there are many situations in which these estimators can adapt to favorable problems to achieve much faster convergence rates. For example there are ways to adapt to local smoothness properties. In classification, the convergence rate can be much faster if the noise level in the problem is small.

6. **Nonparametric density estimation.** Another important problem in statistics is density estimation, where given samples from a distribution with density p, we want to produce an estimate of the density $\hat{p}$. A similar estimator works here and actually the analysis is much cleaner since there is no randomness in the denominator. Most of the difficulty with the kernel regression proof was in controling the term from the denominator.

7. **Nonparametric classification.** Essentially the same estimator can be used for nonparametric classification. Just take a locally weighted average of the labels. It might be a good exercise to try to analyze the misclassification rate.