STAT 425: Introduction to Nonparametric Statistics **Winter 2018**

Lecture 10: Regression: Basis Approach and Regression Tree

Instructor: Yen-Chi Chen

Reference: page 108–111 and Chapter 8 of *All of nonparametric statistics.*

10.1 Basis Approach

Recall that we observes pairs $(X_1, Y_1), \cdots, (X_n, Y_n)$ and we are interested in the regression function $m(x) = \mathbb{E}(Y_1|X_1 = x)$. In this section, we will make the following two assumptions:

- $Y_i = m(X_i) + \sigma \cdot \epsilon_i$, where $\epsilon_i \sim N(0, 1)$ is the noise. Moreover, $\epsilon_1, \cdots, \epsilon_n$ are IID.
- $X_i = \frac{i}{n}$. Namely, the covariates consist a uniform grid over $[0, 1]$ and is non-random.

Similar to the basis approach for the density estimation problem where we approximate the density function by the sum of coefficients and basis, we will approximate the regression function by a basis:

$$m(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x),$$

where $\{\phi_1, \phi_2, \cdots\}$ is an orthonormal basis and $\theta_1, \theta_2, \cdots$ are the coefficients.

Again, here we consider the cosine basis:

$$\phi_1(x) = 1, \quad \phi_j(x) = \sqrt{2}\cos((j-1)\pi x), j = 2, 3, \cdots.$$

As is done in the density estimation, we will use only the top M basis to form our estimator. Namely,

$$\widehat{m}_M(x) = \sum_{j=1}^{M} \widehat{\theta}_j \phi_j(x),$$

for some coefficient estimates $\widehat{\theta}_1, \cdots$. Again, M is the tuning parameter in our estimator.

Here is a simple choice of the coefficient estimates that we will be using:

$$\widehat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i \phi_j(X_i) = \frac{1}{n}\sum_{i=1}^{n} Y_i \phi_j\left(\frac{i}{n}\right).$$

To determine the tuning parameter M, we analyze the MISE. We start with analyzing the bias and variance of $\widehat{\theta}_j$.

10.1.1 Asymptotic theory

Asymptotic normality. Note that the estimator can be rewritten as

$$\begin{aligned}\widehat{m}_M(x) &= \sum_{j=1}^{M} \widehat{\theta}_j \phi_j(x)\\ &= \sum_{j=1}^{M} \frac{1}{n}\sum_{i=1}^{n} Y_i \phi_j\left(\frac{i}{n}\right)\phi_j(x)\\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i \sum_{j=1}^{M}\phi_j\left(\frac{i}{n}\right)\phi_j(x).\end{aligned}$$

Thus, for M being fixed, we have

$$\sqrt{n}\left(\widehat{m}_M(x) - \mathbb{E}(\widehat{m}_M(x))\right) \overset{D}{\to} N(0,\sigma_M^2)$$

for some σ_M^2. Note that later our analysis will demonstrate

$$\mathbb{E}(\widehat{m}_M(x)) = \sum_{j=1}^{M}\theta_j\phi_j(x), \quad \sigma_M^2 = \sigma^2\sum_{j=1}^{M}\phi_j^2(x).$$

Bias.

$$\begin{aligned}\mathbf{bias}(\widehat{\theta}_j) &= \mathbb{E}(\widehat{\theta}_j) - \theta_j\\ &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\phi_j\left(\frac{i}{n}\right)|X_i = \frac{i}{n}\right) - \theta_j\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(Y_i|X_i = \frac{i}{n}\right)\phi_j\left(\frac{i}{n}\right) - \theta_j\\ &= \frac{1}{n}\sum_{i=1}^{n} m\left(\frac{i}{n}\right)\phi_j\left(\frac{i}{n}\right) - \theta_j\\ &= \frac{1}{n}\sum_{i=1}^{n} m\left(\frac{i}{n}\right)\phi_j\left(\frac{i}{n}\right) - \int_0^1 m(x)\phi_j(x)dx.\end{aligned}$$

Namely, the bias is the difference between actual integration and a discretized version of integration. We know that when n is large, the two integrations are almost the same so we can ignore the bias. Thus, we will write

$$\mathbf{bias}(\widehat{\theta}_j) = 0$$

for simplicity.

Variance.

$$
\begin{aligned}
\mathsf{Var}(\widehat{\theta}_j) &= \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(m\left(\frac{i}{n}\right)+\sigma\cdot\epsilon_i\right)}_{=Y_i}\phi_j\left(\frac{i}{n}\right)\right)\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\phi_j^2\left(\frac{i}{n}\right)\mathsf{Var}\left(\epsilon_i\right)\\
&= \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\phi_j^2\left(\frac{i}{n}\right).
\end{aligned}
$$

Note that $\frac{1}{n}\sum_{i=1}^{n}\phi_j^2\left(\frac{i}{n}\right)\approx\int_0^1\phi_j^2(x)dx=1$. For simplicity, we just write

$$
\mathsf{Var}(\widehat{\theta}_j)=\frac{\sigma^2}{n}.
$$

MISE. To analyze the MISE, we first note that the bias of $\widehat{m}_M(x)$ is

$$
\mathbf{bias}(\widehat{m}_M(x))=\mathbb{E}(\widehat{m}_M(x))-m(x)=\sum_{j=1}^{M}\theta_j\phi_j(x)-\sum_{j=1}^{\infty}\theta_j\phi_j(x)=\sum_{j=M+1}^{\infty}\theta_j\phi_j(x).
$$

This further implies that the integrated sqaured bias

$$
\begin{aligned}
\int_0^1\mathbf{bias}^2(\widehat{m}_M(x))dx &= \int_0^1\sum_{j=M+1}^{\infty}\theta_j\phi_j(x)\sum_{\ell=M+1}^{\infty}\theta_\ell\phi_\ell(x)dx\\
&= \sum_{j=M+1}^{\infty}\theta_j\sum_{\ell=M+1}^{\infty}\theta_\ell\underbrace{\int_0^1\phi_j(x)\phi_\ell(x)dx}_{=I(j=\ell)}\\
&= \sum_{j=M+1}^{\infty}\theta_j^2.
\end{aligned}
$$

Again, if we assume that m satisfies $\int_0^1|m''(x)|^2dx<\infty$, we have

$$
\sum_{j=M+1}^{\infty}\theta_j^2=O(M^{-4}).
$$

Now we turn to the analysis of variance.

$$
\begin{aligned}
\mathsf{Var}(\widehat{m}_M(x)) &= \mathsf{Var}\left(\sum_{j=1}^{M}\widehat{\theta}_j\phi_j(x)\right)\\
&= \sum_{j=1}^{M}\mathsf{Var}(\widehat{\theta}_j)\phi_j^2(x)\\
&= \frac{\sigma^2}{n}\sum_{j=1}^{M}\phi_j^2(x).
\end{aligned}
$$

Thus, the integrated variance is

$$\int_0^1 \mathsf{Var}(\widehat{m}_M(x))dx = \frac{\sigma^2}{n}\sum_{j=1}^M \int_0^1 \phi_j^2(x)dx = \frac{\sigma^2 M}{n} = O\left(\frac{M}{n}\right).$$

Recall that the MISE is just the sum of integrated bias and integrated variance, we obtain

$$\textbf{MISE}(\widehat{m}_M) = \int_0^1 \textbf{bias}^2(\widehat{m}_M(x))dx + \int_0^1 \mathsf{Var}(\widehat{m}_M(x))dx = O(M^{-4}) + O\left(\frac{M}{n}\right).$$

Thus, the optimal choice is

$$M^* \asymp n^{1/5}.$$

10.1.2 Basis approach as a linear smoother

The basis estimator is another linear smoother. To see this, we use the follow expansion:

$$\begin{aligned}\widehat{m}_M(x) &= \sum_{j=1}^M \widehat{\theta}_j \phi_j(x)\\ &= \sum_{j=1}^M \frac{1}{n}\sum_{i=1}^n Y_i\phi_j(X_i)\phi_j(x)\\ &= \sum_{i=1}^n \left(\sum_{j=1}^M \frac{1}{n}\phi_j(X_i)\phi_j(x)\right) Y_i\\ &= \sum_{i=1}^n \ell_i(x)Y_i,\end{aligned}$$

where $\ell_i(x) = \sum_{j=1}^M \frac{1}{n}\phi_j(X_i)\phi_j(x)$.

Recall that from the linear smoother theory, we can estimate σ^2 using the residuals and the degree of freedom:

$$\widehat{\sigma}^2 = \frac{1}{n-2\nu+\widetilde{\nu}}\sum_{i=1}^n e_i^2,$$

where $e_i = \widehat{Y}_i - Y_i = \widehat{m}_M(X_i) - Y_i$ and $\nu, \widetilde{\nu}$ are the degree of freedoms (see the previous lecture note).

With this variance estimator and the fact that $\mathsf{Var}(\widehat{m}_M(x)) = \frac{\sigma^2}{n}\sum_{j=1}^M \phi_j^2(x)$ and the asymptotic normality, we can construct a confidence interval (band) of m using

$$\widehat{m}_M(x) \pm z_{1-\alpha/2}\frac{\widehat{\sigma}^2}{n}\sum_{j=1}^M \phi_j^2(x).$$

Note that this confidence interval is valid for $\mathbb{E}(\widehat{m}_M(x)) = \sum_{j=1}^M \theta_j\phi_j(x)$, not the actual $m(x)$. The difference between them is the bias of our estimator.

10.2 Regression Tree

In this section, we assume that the covariate may have multiple dimensions, i.e., $x = (x_1, \cdots, x_d)$. And our data are $(X_1, Y_1), \cdots, (X_n, Y_n) \sim P$ for some CDF P. Again, we are interested in the regression function $m(x) = \mathbb{E}(Y_1|X_1 = x)$.

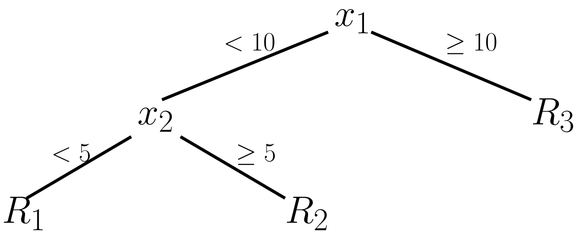
Regression tree constructs an estimator of the form:

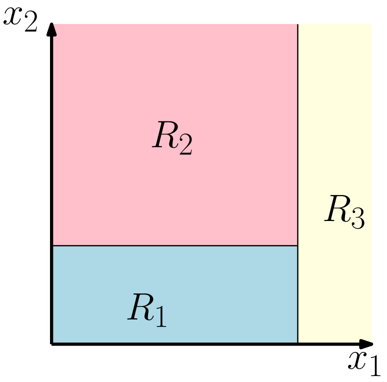
$$m(x) = \sum_{\ell=1}^{M} c_\ell I(x \in R_\ell),$$

where R_ℓ is some rectangle partition of the space of covariates.

Here is an example of a regression tree and its splits. In this example, there are two covariates (namely, $d = 2$) and we have 3 regions R_1, R_2, R_3:

$$R_1 = \{(x_1, x_2) : x_1 < 10, x_2 < 5\}, \quad R_2 = \{(x_1, x_2) : x_1 < 10, x_2 \geq 5\}, R_3 = \{(x_1, x_2) : x_1 \geq 10\}.$$





A regression tree estimator will predict the same value of the response Y within the same area of the covariate. Namely, $m(x)$ will be the same when x is within the same area.

To use a regression tree, there are $2M$ quantities to be determined: the regions $R_1, \cdots, R_M$ and the predicted values $c_1, \cdots, c_M$. When $R_1, \cdots, R_M$ are given, $c_1, \cdots, c_M$ can be simply estimated by the average within each region, i.e.,

$$\widehat{c}_\ell = \frac{\sum_{i=1}^n Y_i I(X_i \in R_\ell)}{\sum_{i=1}^n I(X_i \in R_\ell)}.$$

Thus, the difficult part is the determination of $R_1, \cdots, R_M$.

Unfortunately, there is no simple closed form solution to these regions. We only have a procedure for computing it. Here is what we will do in practice. Let X_{ij} be the j-th coordinate of the i-th observation (X_i).

1. For a given j, we define
$$R_a(j,s) = \{x : x < s\}, \quad R_b(j,s) = \{x : x \geq s\}.$$
2. Find c_a and c_b that minimizes
$$\sum_{X_i \in R_a} (Y_i - c_a)^2, \quad \sum_{X_i \in R_b} (Y_i - c_b)^2,$$
respectively.
3. Compute the score
$$S(j,s) = \sum_{X_i \in R_a} (Y_i - c_a)^2 + \sum_{X_i \in R_b} (Y_i - c_b)^2.$$
4. Change s and repeat the same calculation until we find the minimizer of $S(j,s)$, denoted the minimal score as $S^*(j)$.
5. Compute the score $S^*(j)$ for $j = 1, \cdots, d$.
6. Pick the dimension (coordinate) and the corresponding split point s that has the minimal score $S^*(j)$. Partition the space into two parts according to this split.
7. Repeat the above procedure for each partition until certain stopping criterion is satisfied.

Using the above procedure, we will eventually end up with a collection of rectangle partitions $\widehat{R}_1, \cdots, \widehat{R}_M$. Then the final estimator is
$$\widehat{m}(x) = \sum_{\ell=1}^{M} \widehat{c}_\ell I(x \in \widehat{R}_\ell).$$

For the stopping criterion, sometimes people will pick the number of M so as long as we obtain M regions, the splitting procedure will stop. However, such a choice M is rather arbitrary. A popular alternative is to top the criterion based on minimizing some score that balances the fitting quality and the complexity of the tree. For instance, we may stop the criterion if the following score is no longer decreasing:
$$C_{\lambda,n}(M) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}(X_i))^2 + \lambda M,$$
where $\lambda > 0$ is a tuning parameter that determines the 'penalty' for having a complex tree. In the next lecture, we will talk more about this penalty type tuning parameter.

Remark.

- **Interpreation.** Regression tree has a powerful feature that it is easy to interpret. Even without much training, a practitioner can use the output from a regression tree very easily. A limitation of the regression tree is that it partitions the space of covariates into rectangle regions, which may be unrealistic for the actual regression model.
- **Cross-validation.** How to choose the tuning parameter λ? There is a simple approach called the cross-validation[1] that can compute a good choice of this quantity. Not only λ, other tuning parameters such as the number of basis M, the smoothing bandwidth h, the bin size b, can be chosen using the cross-validation.

[1] https://en.wikipedia.org/wiki/Cross-validation_(statistics)

- **MARS (multivariate adaptive regression splines).** The regression tree has another limitation that it predicts the same value within the same region. This creates a jump on the boundary of two consecutive regions. There is a modified regression tree called MARS (multivariate adaptive regression splines) that allows a continuous (and possibly smooth) changes over two regions. See `https://en.wikipedia.org/wiki/Multivariate_adaptive_regression_splines`.