Week 8 Notes: Some Nonparametric Methods

Antonio R. Linero
University of Texas at Austin

Goals

1. Be able to perform linear smoothing, including selecting bandwidth parameters.
2. Know different linear smoothing techniques (local constant, local linear, basis function expansions).
3. Be able to do Bayesian inference with Gaussian processes.
4. Know the fundamental difficulties with nonparametric estimation.

Nonparametrics

Recall the differences between parametric, semiparametric, and nonparametric methods.

Nonparametrics

Recall the differences between parametric, semiparametric, and nonparametric methods.

- **Parametric:** DGP indexed by finite dimensional θ. (Normal linear regression, most GLMs).
- **Semiparametric:** DGP indexed by finite-dimensional parameter of interest θ a an infinite-dimensional nuisance parameter η. (GEEs, Quasi-Likelihood, Empirical Likelihood)
- **Nonparametric:** DGP indexed by infinite-dimensional θ of interest.

Note: some flexibility in the terms.

Bias-Variance Tradeoff

Exercise: Bias-Variance Tradeoff

Show that $\text{MSE}(\widehat{\mu}, \mu)$ decomposes into a *bias* term and a *variance* term $\text{MSE} = B^2 + V$:

$$B = \mathbb{E}_\theta\{\widehat{\mu}(x)\} - \mu(x) \qquad \text{and} \qquad V = \text{Var}\{\widehat{\mu}(x)\}.$$

Typical situation: in nonparametric settings:

- Estimation error is balanced when *these terms are of the same magnitude.*
- Uncertainty quantification requires *the variance term to dominate.* (Called *undersmoothing*.)

Density Estimation With Histograms

Exercise: Density Estimation With Histograms

Some people refer to the decomposition above as the *bias-variance tradeoff*. Why is this a tradeoff? Here's a simple example to convey the intuition.

Suppose we observe $Z_1, \ldots, Z_N$ from some distribution F and want to estimate $f(0)$, the value of the probability density at 0. Let h be a small positive number, called the *bandwidth*, and define the quantity

$$\pi_h = \Pr\left(-\frac{h}{2} < Z < \frac{h}{2}\right) = \int_{-\frac{h}{2}}^{\frac{h}{2}} f(z)\, dz.$$

For small h we have $\pi_h \approx h\, f(0)$ provided that $f(x)$ is continuous at 0.

a. Let M be the number of observations in a sample of size N that fall within the interval $(-h/2, h/2)$. What is the distribution of M. What are its mean and variance in terms of N and π_h? Propose a simple estimator $\widehat{f}(0)$ of $f(0)$ based on M.

b. Suppose we expand $f(z)$ in a second-order Taylor series about 0:

$$f(z) \approx f(0) + f'(0)\, z + \frac{1}{2} f''(0)\, z^2.$$

Use this, together with the bias-variance decomposition, to show that

$$\text{MSE}\{\widehat{f}(0), f(0)\} \approx A h^4 + \frac{B}{N h}$$

for constants A and B that you should (approximately) specify. What happens to the bias and variance when you make h small? When you make h big?

Curve Fitting by Linear Smoothing

Exercise: Linear Smoothing

Consider a nonlinear regression problem with one predictor and one response: $Y_i = \mu(X_i) + \epsilon_i$ where the ϵ_i's are mean-zero random variables.

a. Suppose you want to estimate the value of the regression function $\mu(x^\star)$ at some new point $x^\star$. Assume for the moment that $\mu(x)$ is linear and that both Y_i and X_i are mean 0, in which case $Y_i = \beta\, X_i + \epsilon_i$.

   Recall the least-squares estimator for multiple regression. Show that for the one-predictor case, your prediction $\mu(x^\star) = \widehat{\beta}\, x^\star$ can be expressed as a *linear smoother* of the form

$$\widehat{\mu}(x^\star) = \sum_{i=1}^{N} w(X_i, x^\star)\, Y_i$$

   for any $x^\star$. Inspect the weighting function you derived. Briefly describe your understanding of how the resulting smoother behaves, compared with the smoother that arises from an alternate form of the weight function $w(X_i, x^\star)$:

$$w_K(X_i, x^\star) = \begin{cases} 1/K & \text{if } X_i \text{ is one of the } K \text{ closest sample points to } x^\star, \\ 0 & \text{otherwise.} \end{cases}$$

   This is referred to as the *K-nearest neighbor* smoother.

b. A *kernel function* $K(x)$ is a smooth function satisfying

$$\int_{-\infty}^{\infty} K(x)\ dx = 1, \qquad \int_{-\infty}^{\infty} x\, K(x)\ dx = 0. \qquad \int_{-\infty}^{\infty} x^2\, K(x)\ dx < \infty.$$

Cross-Validation for Selecting Bandwidths

Exercise: Cross Validation

Left unanswered so far in our previous study of kernel regression is the question: how does one choose the bandwidth h used for the kernel? Assume for now that the goal is to predict well, not necessarily to recover the truth. (These are related but distinct goals.)

a. Presumably a good choice of h would be one that led to smaller predictive errors on fresh data. Write a function or script that will: (1) accept an old ("training") data set and a new ("testing") data set as inputs; (2) fit the kernel-regression estimator to the training data for specified choices of h; and (3) return the estimated functions and the realized prediction error on the testing data for each value of h. This should involve a fairly straightforward "wrapper" of the function you've already written.

b. Imagine a conceptual two-by-two table for the unknown, true state of affairs. The rows of the table are "wiggly function" and "smooth function," and the columns are "highly noisy observations" and "not so noisy observations." Simulate one data set (say, 500 points) for each of the four cells of this table, where the X_i's take values in the unit interval. Then split each data set into training and testing subsets. You choose the functions.[a] Apply your method to each case, using the testing data to select a bandwidth parameter. Choose the estimate that minimizes the average squared error in prediction, which estimates the mean-squared error:

$$L_N(\widehat{\mu}) = \frac{1}{N^\star} \sum_{i=1}^{N^\star} (Y_i^\star - \widehat{Y}_i^\star)^2 \,,$$

where $(Y_i^\star, X_i^\star)$ are the points in the test set, and $\widehat{Y}_i^\star$ is your predicted value arising from the model you fit using only the training data. Does your out-of-sample predictive validation method lead to reasonable choices of h for each case?

Heteroskedasticity

Exercise: Heteroskedasticity

In this exercise we will consider linear smoothing when (potentially) we are concerned that the errors in the model $Y_i = r(X_i) + \epsilon_i$ do not have constant variance.

a. Suppose that the ϵ_i's have constant variance σ^2 (that is, the spread of the residuals does not depend on x). Derive the mean and variance of the sampling distribution for the locally constant linear smoother. Note: the random variable $\widehat{\mu}(x)$ is just a scalar quantity at x, not the whole function.

b. We don't know the residual variance, but we can estimate it. A basic fact is that if X is a random vector with mean μ and covariance matrix Σ, then for any symmetric matrix Q of appropriate dimension, the quadratic form $X^\top QX$ has expectation
$$E(X^\top QX) = \mathrm{tr}(Q\Sigma) + \mu^\top Q\mu\,.$$
Consider an arbitrary linear smoother (i.e., one with $\widehat{Y} = HY$ for some smoothing matrix H). Write the vector of residuals as $R = Y - \widehat{Y} = Y - HY$, where H is the smoothing matrix. Compute the expected value of the estimator
$$\widehat{\sigma}^2 = \frac{\|R\|^2}{n - 2\mathrm{tr}(H) + \mathrm{tr}(H^\top H)}\,,$$
and simplify things as much as possible. Roughly under what circumstances will this estimator be nearly unbiased for large N? Note: the quantity $2\mathrm{tr}(H) - \mathrm{tr}(H^\top H)$ is often referred to as the "effective degrees of freedom" in such problems.

c. Load `utilities.csv` located at https://raw.githubusercontent.com/theodds/StatModelingNotes/master/datasets/utilities.csv into `R`. This data set shows the monthly gas bill (in dollars) for a single-family home in Minnesota, along with

Locally Constant Regression

- Kernel smoothing: **locally constant**, solves

$$\widehat{\mu}(x) = a = \arg\min_{\mathbb{R}} \sum_{i=1}^{N} w_i(x)(Y_i - a)^2 \, ,$$

https://rafalab.dfci.harvard.edu/dsbook/ml/img/binsmoother-animation.gif

Locally Constant Regression

- Kernel smoothing: **locally constant**, solves

$$\widehat{\mu}(x) = a = \arg\min_{\mathbb{R}} \sum_{i=1}^{N} w_i(x)(Y_i - a)^2 \,,$$

https://rafalab.dfci.harvard.edu/dsbook/ml/img/binsmoother-animation.gif

- Local linear smoothing: use a **local polynomial** instead

$$\widehat{\mu}(x) = x^\top \beta(x) = x^\top \arg\min_{\beta} \sum_{i=1}^{N} w_i(x)(Y_i - X_i^\top \beta)^2$$

https://rafalab.dfci.harvard.edu/dsbook/ml/img/loess-animation.gif

Loess

- Tricube weighting function

$$w_i(x) = \left\{ 1 - \left(\frac{|x - X_i|}{h_\alpha(x)} \right)^3 \right\}_+^3$$

- Data-adaptive bandwidth:

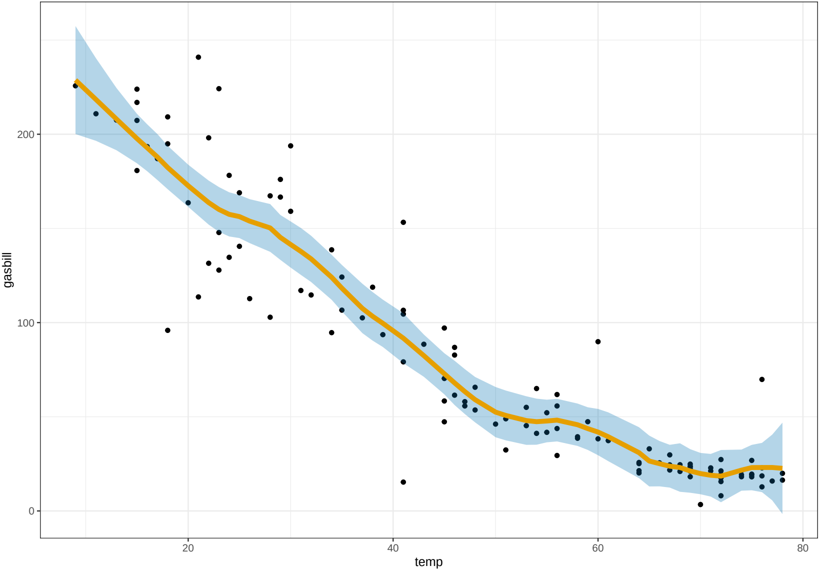
$$h_\alpha(x) = \max_{j:X_j \in B_x} \{|x - X_j|\}$$

where B_x is a neighborhood of x chosen so that $100\alpha\%$ of the X_i's are in B_x.

Example

```r
utilities_file <- str_c("https://raw.githubusercontent.com/theodds",
                        "/StatModelingNotes/master/datasets/utilities.csv")
utilities <- readr::read_csv(utilities_file)
loess_util <- loess(gasbill ~ temp, data = utilities, span = 0.2, degree = 1)
loess_preds <- predict(loess_util, utilities, se = TRUE)
utilities %>%
  mutate(fit = loess_preds$fit, se = loess_preds$se.fit) %>%
  ggplot(aes(x = temp, y = gasbill, ymin = fit - 2 * se, ymax = fit + 2 * se)) +
  geom_point() +
  geom_ribbon(alpha = 0.3, fill =   "#0072B2") +
  geom_line(aes(y = fit), color = "#E69F00", size = 2)
```

Example



Bayesian Variant: Gaussian Processes

Choose a prior on a function space $\mathscr{F}$, with large support:

$$\Pi(\sup_{x \in \mathcal{X}} |\mu_0(x) - \mu(x)| < \epsilon) > 0$$

Bayesian Variant: Gaussian Processes

Choose a prior on a function space $\mathscr{F}$, with large support:

$$\Pi(\sup_{x \in \mathcal{X}} |\mu_0(x) - \mu(x)| < \epsilon) > 0$$

Examples of spaces:

- $C^0(\mathcal{X})$
- $C^2(\mathcal{X})$
- $\mathscr{L}_2(\mathcal{X})$

Gaussian Processes

Definition (Gaussian Process)

Let $m : \mathcal{X} \to \mathbb{R}$ and $K : \mathcal{X}^2 \to \mathbb{R}$. A random function $\mu : \mathcal{X} \to \mathbb{R}$ is said to be a *Gaussian process* if, for any *finite* set $D = \{x_1, \dots x_D\}$ we have

$$\mu(\mathbf{x}) = \text{Normal}\{m(\mathbf{x}), K(\mathbf{x}, \mathbf{x})\}$$

where $\mathbf{x} = (x_1, \dots, x_D)^\top$, $\mu(\mathbf{x}) = (\mu(x_1), \dots, \mu(x_D))^\top$, $m(\mathbf{x}) = (m(x_1), \dots, m(x_D))^\top$, and $K(\mathbf{x}, \mathbf{x}')$ is a covariance matrix with $(i, j)^{\text{th}}$ entry $K(x_i, x_j')$. The function $K(\cdot, \cdot)$ is referred to as a *covariance function*. To denote this fact, we write $\mu \sim \text{GP}(m, K)$.

When is $K(\mathbf{x}, \mathbf{x}')$ valid?

Review of MVN

Definition (Multivariate Normal Distribution)

We say that X has an *n-dimensional multivariate normal* distribution with mean vector μ and covariance matrix Σ if, or every $\lambda \in \mathbb{R}^n$, we have $\lambda^\top X \sim \text{Normal}(\lambda^\top \mu, \lambda^\top \Sigma \lambda)$. We write $X \sim \text{Normal}(\mu, \Sigma)$ to denote this fact; this distribution exists for every $\mu \in \mathbb{R}^n$ and every symmetric matrix $\Sigma \in \mathbb{R}^{n \times n}$ such that $\lambda^\top \Sigma \lambda \geq 0$ for all $\lambda \in \mathbb{R}^n$ (such a matrix is called *positive semi-definite*).

Exercise

Exercise: Mean and Variance

Show that $\mu = \mathbb{E}(X)$ and $\Sigma = \mathrm{Var}(X)$ from this definition.

Exercise

Exercise: Characteristic Functions

The *characteristic function* of a random vector X is the function

$$\varphi_X(\lambda) = \mathbb{E}(e^{i\lambda^\top X})$$

where i is the imaginary unit (i.e., $i^2 = -1$). The characteristic function is similar to the moment generating function, with the important benefit that $\varphi_X(\lambda)$ is guaranteed to exist for all λ. It can be shown that, if X and Y have the same characteristic function, then X and Y have the same distribution (the proof is not difficult if you are comfortable with real analysis, but not worth our time; see this website for all the ingredients of the proof).

a. Show that the multivariate normal distribution is "well-defined" by Definition; that is, show that if X and Y both satisfy Definition then they have the same distribution. Why do we require that Σ is positive semi-definite and symmetric?

Exercise

Exercise: Density of a MVN

Suppose that Σ is full-rank and that $Y \sim \text{Normal}(\mu, \Sigma)$. Show that Y has density

$$f(y \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left\{-\frac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu)\right\}.$$

What happens if Σ is *not* full-rank? *Hint:* recall the change of variables formula, which states that if $Y = T(Z)$ where T is a smooth one-to-one function then the density of Y is

$$f_Y(y) = f_Z\{T^{-1}(y)\}|J(y)|$$

where $J(y)$ is the Jacobian matrix of T^{-1}. When $T(Z) = \mu + LZ$ where L is a full-rank matrix, this simplifies to

$$f_Y(y) = f_Z\{L^{-1}(Y-\mu)\}|L^{-1}|.$$

Exercise

Exercise: Characteristic Function of a MVN

We now derive the characteristic function of the multivariate normal distribution.

a. Show that a standard normal random variable Z has characteristic function

$$\varphi_Z(t) = \int \cos(tz)\, e^{-z^2/2}\ dz = e^{-t^2/2}.$$

This can be done in two steps: (i) because $\sin(tZ)$ is an odd function and Z is symmetric, the imaginary part disappears and (ii) by differentiating under the integral, we can establish the differential equation $\frac{d}{dt}\varphi_Z(t) = -t\varphi_Z(t)$; solving this equation with the initial condition $\varphi_Z(0) = 1$ gives the result.

b. Using this result, show that the characteristic function

of $X \sim \text{Normal}(\mu, \Sigma)$ is

MVN Properties

The multivariate normal distribution has a large number of desirable properties. First, it is *closed under marginalization*. Suppose that

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \text{Normal}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy}. \end{pmatrix}\right).$$

Here, the vectors μ_x and μ_y have the same dimension as X and Y, and Σ_{xx} and Σ_{yy} are positive semi-definite symmetric matrices with dimensions matching X and Y respectively. Because the covariance matrix of (X, Y) is symmetric, it follows that $\Sigma_{xy} = \Sigma_{yx}^{\top}$.

Exercise

Exercise: MVN Properties

We now prove some basic properties.

a. Show that $X \sim \text{Normal}(\mu_x, \Sigma_{xx})$.

b. The *covariance* of X and Y is defined to be $\text{Cov}(X, Y) = \mathbb{E}\{(X - \mu_x)(Y - \mu_y)^\top\}$. Show that (i) for any random vectors, if X and Y are independent then the covariance is equal to the zero matrix and (ii) for the multivariate normal distribution in particular the covariance matrix is Σ_{xy}.

c. Using the characteristic function, show that for the multivariate normal distribution X is independent of Y if-and-only-if Σ_{xy} is equal to zero. This is an interesting reversal – in general, the covariance being 0 does not imply independence, but it does for multivariate normal random vectors.

d. Suppose $Y \sim \text{Normal}(\mu, A)$ given μ and $\mu \sim \text{Normal}(m, B)$. Show that

$$\begin{pmatrix} Y \\ \mu \end{pmatrix} \sim \text{Normal}\left\{\begin{pmatrix} m \\ m \end{pmatrix}, \begin{pmatrix} A + B & B \\ B & B \end{pmatrix}\right\}.$$

Exercise

Exercise: The Conditionals

We will now show that the conditional distribution of X given $Y = y$ is

$$X \sim \text{Normal}(\mu_{x|y}, \Sigma_{x|y}) \tag{1}$$

where $\mu_{x|y} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)$ and $\Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$.

a. Write $X = W + (X - W)$ where $W = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(Y - \mu_y)$. Show that $\text{Cov}(Y, X - W) = 0$ so that $X - W$ is independent of Y.

b. Show that the covariance matrix of $X - W$ is $\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$ and that the mean is $\mathbf{0}$.

c. Argue that because (i) W is constant as a function of Y, (ii) $X - W$ is independent of Y, (iii) $X = W + (X - W)$, and (iv) $X - W \sim \text{Normal}(\mathbf{0}, \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})$ that we can conclude that the distribution of X given Y is given by (1).

Exercise

Exercise: GP Inference

Suppose that $Y_i \overset{\text{indep}}{\sim} \text{Normal}\{\mu(X_i), \sigma^2\}$ conditional on $\boldsymbol{X}$ for $i = 1, \ldots, N$ and μ. Let $\boldsymbol{X} = (X_1, \ldots, X_N)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$.

a. Show that the posterior distribution of μ is given by

$$[\mu \mid \boldsymbol{X}, \boldsymbol{Y}, \sigma^2] \sim \text{GP}(m^\star, K^\star)$$

where

$$m^\star(x) = m(x) + K(x, \boldsymbol{X})\{K(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \text{I}\}^{-1}\{\mathbf{Y} - m(\mathbf{X})\}$$

$$K^\star(x, x') = K(x, x') - K\left(\begin{pmatrix} x \\ x' \end{pmatrix}, \boldsymbol{X}\right)\{K(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \text{I}\}^{-1} K\left(\begin{pmatrix} x \\ x' \end{pmatrix}, \boldsymbol{X}\right)$$

b. Argue that the marginal likelihood of $\boldsymbol{Y}$ (i.e., with the random function μ integrated out) is given by

$$|2\pi(K(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \text{I})|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{Y} - m(\boldsymbol{X}))^\top (K(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \text{I})^{-1}(\boldsymbol{Y} -\right.$$

Or, equivalently, that
$[\boldsymbol{Y} \mid \boldsymbol{X}, \sigma] \sim \text{Normal}\left\{m(\boldsymbol{X}), K(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \text{I}\right\}$.

Exercise

Exercise: Squared Exponential Kernel

Let the *squared exponential* covariance function be given by

$$K(x, x') = \sigma_\mu^2 \exp\left\{ -\frac{1}{2} \sum_{j=1}^{P} \left(\frac{x_j - x'_j}{h} \right)^2 \right\} + \sigma_\delta^2 \, \delta(x, x') \, .$$

The constants $(\sigma_\mu^2, \sigma_\delta^2, h)$ are often called *hyperparameters* and $\delta(x, x')$ is the Kronecker delta function that takes the value 1 if $x = x'$ and 0 otherwise.

a. Let's start with the simple case where $\mathcal{X} = [0, 1]$, the unit interval. Write a function that simulates a mean-zero Gaussian process on $[0, 1]$ under the squared exponential covariance function. The function will accept as arguments: (1) finite set of points $x_1, \ldots, x_N$ on the unit interval; and (2) a triplet $(\sigma_\mu^2, \sigma_\delta^2, h)$. It will return the value of the random process at each point: $\mu(x_1), \ldots, \mu(x_N)$.

Curve of Dimensionality

So far: use local information around x to estimate $\mu(x)$.

Problem: In high-dimensional settings, *points are typically all equally far away from x!:

1. Large neighborhood needed around x to have nay data.
2. If neighborhood is large, the bias will be large.

Optimal MSE scales like $N^{-4/(4+P)}$.

Exercise

Exercise: Power Analysis

Supposing that the RMSE scales like $N^{-2/(4+P)}$, how large must N be in order for us to get an RMSE less than a fixed constant δ? How does this depend on the dimensionality P?

Exercise

Exercise: Geometry

In this exercise, we will demonstrate numerically the point that observations in a high-dimensional space tend to be far away from one another.

a. Write a function to generate N random data points X_i uniformly distribution within a P-dimensional hypercube.

b. For each X_i, define its *nearest neighbor* $X_{i'}$ by $i' = \min_{j \neq i}\{\|X_i - X_j\|\}$. Then, define the average nearest neighbor distance by

$$NND = \frac{1}{N} \sum_i \|X_i - X_{i'}\|.$$

Write a function that computes the average nearest neighbor distance for a given dataset.

c. Generate datasets for different dimensions P ranging from 1 to 50, keeping the number of data points N fixed. For each dataset, compute the average nearest neighbor distance divided by the average distance overall $\frac{1}{N^2} \sum_{i,j} \|X_i - X_j\|$ and store the results in a list. Plot the result as a function of P. *Explain how the result implies that all data points approach an equal amount of "closeness" to any given point as the dimension increases.*

Loopholes

Dimension reducing structures:

- Additive functions $\mu(x) = \sum_{j=1}^{P} \mu_j(x_j)$.
- Sparsity: $\mu(x)$ depends on $D \ll P$ covariates.
- Covariate structure: maybe X_i is concentrated on a "low-dimensional manifold."

Basis Functions

A basis function expansion for $\mu(x)$ refers to modeling $\mu(x)$ as

$$\mu(x) \approx \psi(x)^\top \beta$$

where $\psi(x) = \{\psi_1(x), \dots, \psi_B(x)\}$.

Options:

- **Polynomials:** $\psi_j(x) = x^j$.
- **Fourier series:** $\psi_j(x) = \cos\left(\frac{k_j \pi x}{L} + \lambda_j \pi\right)$
- **Cubic Splines:** $\psi_j(x) =$ piecewise-cubic function

Spline Basis

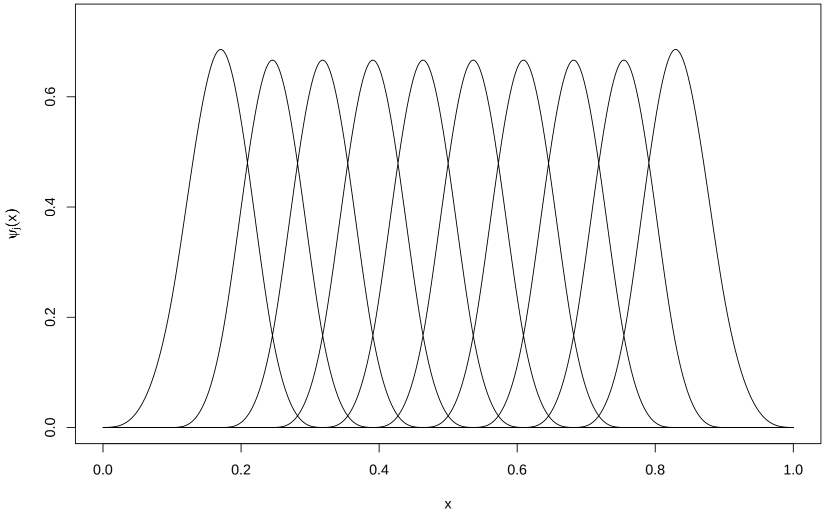
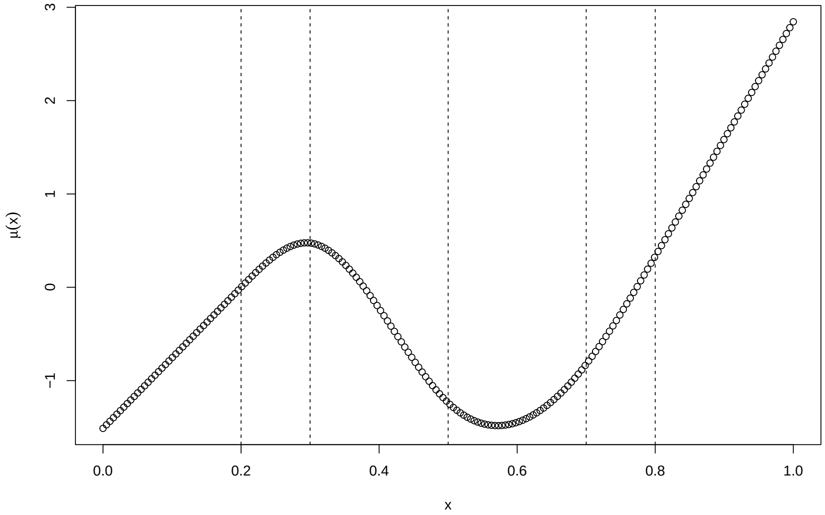


Figure 1: A set of cubic spline basis functions.

A Spline



Exercise

Exercise: Utilities Again

Consider the `utilities` dataset again.

Write a function that takes as input variables `y` and `x` and a maximal number of spline basis functions `k_max` and outputs the optimal number of basis functions to use with the `ns` function in the `splines` package according to leave-one-out cross-validation. You can do this by fitting a linear model `fit` for each `k in 1:k_max` and computing the LOOCV as `mean((fit$residuals / (1 - hatvalues(fit)))^2)`. **How many basis functions is optimal for the `utilities` dataset?**

Adaptive Basis Function Expansions

Can sometimes be advantageous to learn a basis from data.

Examples:

- Learning number/location of knots with splines.
- Wavelets.
- Decision tree boosting.

Decision Tree

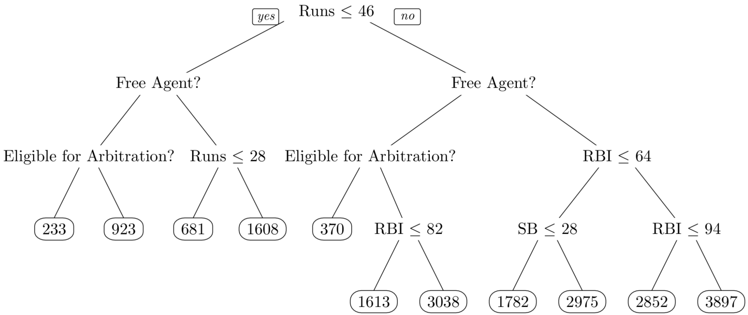


Figure 2: Salary of MLB Players

Exercise

Exercise: Basis Expansions in Decision Trees

Using tree-based basis functions as part of a basis function expansion is known as the *Bayesian additive regression trees* (BART) framework. BART typically uses a prior distribution of the form $\eta_{t\ell} \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma_\eta^2 / T)$ with the trees $\mathcal{T}_t$ sampled according to a branching process prior.

The `boston` dataset is available in the `MASS` package as `boston <- MASS::Boston`. The dataset contains information about the housing market in the Boston area in the 1970s, with 506 observations and 14 variables. The dataset is commonly used to benchmark simple machine learning methods, but the original motivation was to understand the impact o pollution (specifically, nitrous oxide) on housing prices. The outcome of interest is `medv`, the median value of a house in a given census tract. Other variables include `nox` (the predictor of interest) and confounders such as crime rate (`crim`), distance to employment centers (`dis`), and the proportion of “lower status” individuals (`lstat`).

a. Load the dataset and install the required package `BART`.

b. Perform an exploratory analysis of the marginals of the variables. What (if any) issues might there be in using the BART model, and how might you correct them?

c. Split the data into a training set consisting of 80% observations in the training set and 20% of observations in the testing set; we will be using this to compare the BART model to Gaussian process regression later.

d. Use the `wbart` function to fit a BART model with `medv` as the outcome. Does the method appear to mix well? Justify your answer. **If your approach does not mix well in the defaults, make sure to rerun the analysis with a larger number of burn-in/save iterations.**

e. A (somewhat crude) measure of the importance of a variable in a BART ensemble is the number of times a variable is used to build a decision rule; or example, in our MLB tree, the variable `Runs` is used twice, while `SB` is used once, indicating that `Runs` may be more important. BART returns the total number of uses of each variable at each iteration in the object `my_bart_it$varcount`.