

Module 1: Nonparametric Preliminaries

Selecting Smoothing Parameters

STAT/BIOSTAT 527, University of Washington

Emily Fox

April 11th, 2013

©Emily Fox 2013

1

Smoothing Parameter

- In both ridge and lasso regression, we saw that the parameter λ controlled the solution
 - Often, can straightforwardly equate with effective degrees of freedom
- Which λ (\rightarrow estimator) should we choose???

©Emily Fox 2013

2

Two Goals

- **Model Selection:** estimating the performance of models in order to select the best one
 - E.g., choosing λ
- **Model Assessment:** having chosen a final model, estimate its prediction error (generalization error) on new data
- Ideally, divide data into 3 parts



©Emily Fox 2013

3

Focus on Model Selection

- Which estimator/smoothing parameter should we choose?



- Recall metrics for assessing the performance of an estimator...

©Emily Fox 2013

4

Measuring Predictive Performance

- Assume estimate $\hat{f}_n(\cdot)$ based on training data y_1, \dots, y_n
↑ fixed
- The **generalization error** provides a measure of predictive performance

$$GE(\hat{f}_n) = E_{Y, X} [L(Y, \hat{f}_n(X))]$$

⇨
↑ fixed

©Emily Fox 2013

5

Measuring Predictive Performance

- Assume L_2 loss $Y = f(X) + \epsilon$ ★ $E[\epsilon] = 0$ $\text{var}(\epsilon) = \sigma^2$
- Averaging over repeat training sets $\mathbf{Y}_n = Y_1, \dots, Y_n$ we get the **predictive risk** at x^*

$$\begin{aligned}
 E_{Y^*, \mathbf{Y}_n} [(Y^* - \hat{f}_n(x^*))^2] &= E_{Y^*, \mathbf{Y}_n} [(Y^* - f(x^*) + f(x^*) - \hat{f}_n(x^*))^2] \\
 &= E_{Y^*} [(Y^* - f(x^*))^2] + E_{\mathbf{Y}_n} [(\hat{f}_n(x^*) - f(x^*))^2] + 2 E_{Y^*, \mathbf{Y}_n} [(Y^* - f(x^*))(\hat{f}_n(x^*) - f(x^*))] \\
 &= \sigma^2 + \text{MSE}(\hat{f}_n(x^*)) \quad \checkmark \\
 &\quad \uparrow \text{"irreducible error"} \quad \uparrow \text{"risk"}
 \end{aligned}$$

test ↑ training ↑ ↑ fcn of training data ↑ 0
↑ ↑ ↑

- Recall $MSE[\hat{f}_n(x)] = \text{bias}(\hat{f}_n(x))^2 + \text{var}(\hat{f}_n(x))$

©Emily Fox 2013

6

Measuring Predictive Performance

- Finally, let's average over covariates x

- Integrated MSE** $\int \text{MSE}(\hat{f}_n(x)) p(x) dx$
summary over all inputs

- Average MSE** $\frac{1}{n} \sum_{i=1}^n \text{MSE}(\hat{f}_n(x_i))$ Monte Carlo est.
 $x_i \sim p$

- Note: **avg. pred. risk** = $\sigma^2 + \text{avg. MSE}$

$$\frac{1}{n} \sum_{i=1}^n E_{Y_n, Y_n^*} [(Y_i^* - \hat{f}(x_i))^2]$$

training new obs. $Y_n^* = Y_1^*, \dots, Y_n^*$

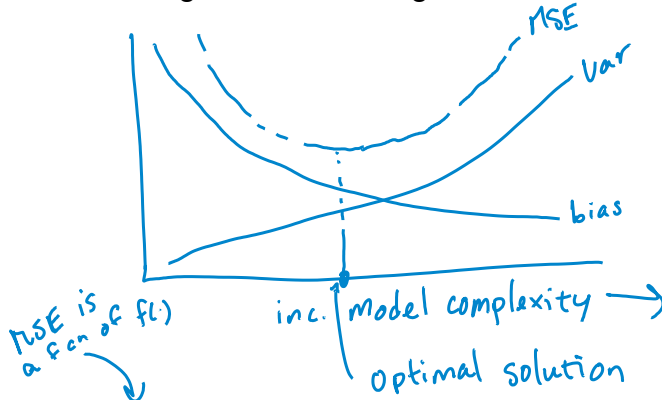
©Emily Fox 2013

7

Bias-Variance Tradeoff

recall polynomial reg. example

- Minimizing risk = balancing bias and variance



- Note: $f(x)$ is unknown, so cannot actually compute MSE

©Emily Fox 2013

8

Focus on Model Selection

- Which estimator/smoothing parameter should we choose?



- We saw that minimizing (average) prediction error can be equated with minimizing (average) MSE
- With a validation set, we can estimate the prediction error

©Emily Fox 2013

9

Data Scarce Approximations

- Often, we do not have enough data to form suitably sized training and validation sets
 - What is a good training/test split? Sensitivity?
 - Typically want to use as much data for training as possible
- Rely on other approximations

©Emily Fox 2013

10

Approx 1: Training Data Only

- **Goal:** Minimize average MSE

$$\min_{\lambda} E \left[\frac{1}{n} \sum_{i=1}^n (f(x_i) - \hat{f}_n^{\lambda}(x_i))^2 \right]$$

- **Solution:** Use training error

Approx 2: Cross Validation

- **Goal:** Minimize average MSE

$$\min_{\lambda} E \left[\frac{1}{n} \sum_{i=1}^n (f(x_i) - \hat{f}_n^{\lambda}(x_i))^2 \right]$$

- **Solution:** Mimic heldout data using *training* data
- Leave-one-out (LOO) cross validation (CV) algorithm:
 - Estimate fit using all but i^{th} data point
 - Predict i^{th} observation
 - Repeat for all i
- Repeat for all values of λ

Approx 2: Cross Validation

- Reasoning
- For linear smoothers
- Warning: Curves can be very flat...Don't just choose and use without thinking. Some rules of thumb (see Elements of Statistical Learning)

©Emily Fox 2013

13

Approx 2: Cross Validation

- K-fold cross validation



- Algorithm

1. Fit model using data with k^{th} fraction removed
2. Using fitted model, compute

$$CV_k = \frac{1}{n_k} \sum_{i \in J(k)} (y_i - \hat{f}_{-k}^\lambda(x_i))$$

3. Store

$$CV = \frac{1}{K} \sum_{k=1}^K CV_k$$

4. Repeat for each value of λ using same split of the data

©Emily Fox 2013

14

Approx 3: Generalized CV

- Recall LOO ordinary CV for linear smoothers

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{f}_n^\lambda(x_i)}{1 - L_{ii}} \right)^2$$

- Instead of L_{ii} , use $\frac{1}{n} \sum_{i=1}^n L_{ii}$

- Often very close to OCV solution

©Emily Fox 2013

15

Approx 3: Generalized CV

$$GCV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{f}_n^\lambda(x_i)}{1 - \frac{\nu_\lambda}{n}} \right)^2$$

- One motivation: Invariance to orthonormal transformations

©Emily Fox 2013

16

Approx 3: Generalized CV

$$GCV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{f}_n^\lambda(x_i)}{1 - \frac{\nu_\lambda}{n}} \right)^2$$

- Using $(1 - x)^{-2} \approx 1 + 2x$

Approx 4: Mallows C_p Statistic

- **Goal:** Minimize average MSE

$$\min_{\lambda} E \left[\frac{1}{n} \sum_{i=1}^n (f(x_i) - \hat{f}_n^\lambda(x_i))^2 \right]$$

- **Solution:** Approximate directly

$$\text{avg. MSE} = \frac{1}{n} E \left[(f - \hat{f}_n^\lambda)^T (f - \hat{f}_n^\lambda) \right]$$

Approx 4: Mallows C_p Statistic

$$\text{avg. MSE} = \frac{1}{n} E [(Y - L^\lambda Y)^T (Y - L^\lambda Y)] - \sigma^2 + \frac{2}{n} \nu_\lambda \sigma^2$$

- Estimate as

- Note: Arises from considering L_2 loss. Log-likelihood loss leads to AIC. For BIC, consider Bayesian model selection

©Emily Fox 2013

19

Bayesian Model Selection

- Assume some M possible models
 - Model M_m $m=1, \dots, M$ has parameters θ_m and prior $p(\theta_m | M_m)$
 - Prior over models $p(M_m)$

- Model posterior

$$\begin{aligned} p(M_m | Z) &\propto p(M_m) p(Z | M_m) \\ &\propto p(M_m) \int p(Z | \theta_m, M_m) p(\theta_m | M_m) d\theta_m \end{aligned}$$

- Compare models:

$$\frac{p(M_m | Z)}{p(M_\ell | Z)} = \frac{p(M_m) p(Z | M_m)}{p(M_\ell) p(Z | M_\ell)} \gtrless 1$$

©Emily Fox 2013

20

Bayesian Model Selection

- For Bayes factor, approximate
$$\log p(Z | M_m) \approx \log p(Z | \hat{\theta}_m, M_m) - \frac{\nu_m}{2} \log n + O(1)$$
- If loss is $-2 \log p(Z | \hat{\theta}_m, M_m)$, then equivalent to BIC
 - Minimizing BIC = maximizing approximated posterior
- However, in addition to being able to select the best model, in Bayesian framework we also get the relative merit of each

$$\approx \frac{e^{-\frac{1}{2} \text{BIC}_m}}{\sum_{\ell=1}^M e^{-\frac{1}{2} \text{BIC}_\ell}}$$

- BIC is asymptotically consistent, but AIC is not
- For finite samples, BIC tends to choose too simple models

©Emily Fox 2013

21

Module 2: Splines and Kernel Methods

Spline Model Overview, Regression Splines, Smoothing Splines

STAT/BIOSTAT 527, University of Washington

Emily Fox

April 11th, 2013

©Emily Fox 2013

22

Moving Beyond Linearity

- So far we have assumed standard linear models
- In the case of many predictors relative to number of observations, we considered penalized regression to avoid overfitting
- Often a convenient form, and necessary to assume simple structure to avoid overfitting in data-scarce regimes, but linear assumption rarely holds in practice

©Emily Fox 2013

23

Moving Beyond Linearity

- Consider generic functional forms (univariate x for now)
 - If constrained to linear forms \rightarrow
 - If arbitrary \rightarrow
- As before, penalize complexity. Here, in terms of roughness.
 - If $\lambda \rightarrow 0$,
 - If $\lambda \rightarrow \infty$,
- Remarkable result: Explicit, finite-dimensional minimizer

©Emily Fox 2013

24

Backtrack a bit...

- Instead of just considering input variables x (potentially mult.), augment/replace with transformations = “input features”

- **Linear basis expansions** maintain linear form in terms of these transformations

$$f(x) = \sum_{m=1}^M \beta_m h_m(x)$$

- What transformations should we use?

- $h_m(x) = x_m \rightarrow$
- $h_m(x) = x_j^2, \quad h_m(x) = x_j x_k \rightarrow$
- $h_m(x) = I(L_m \leq x_k \leq U_m) \rightarrow$
- ...

©Emily Fox 2013

25

Piecewise Polynomial Fits

- Again, assume x univariate

- Polynomial fits are often good locally, but not globally

- Adjusting coefficients to fit one region can make the function go wild in other regions

- Consider **piecewise polynomial** fits

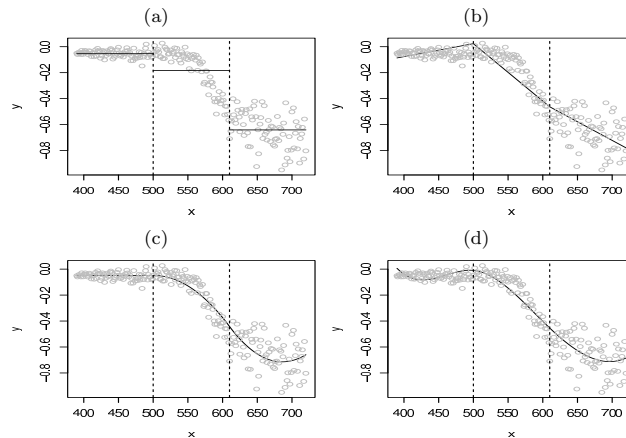
- Local behavior can often be well approximated by low-order polynomials

©Emily Fox 2013

26

Piecewise Polynomial Fits

LIDAR Data Example



From Wakefield book

©Emily Fox 2013

27

Piecewise Constant/Linear Fits

■ Example 1: Piecewise constant, with 3 basis functions

$$h_1(x) =$$

$$h_2(x) =$$

$$h_3(x) =$$

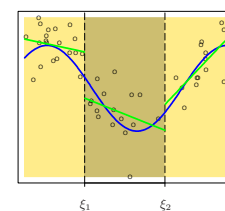
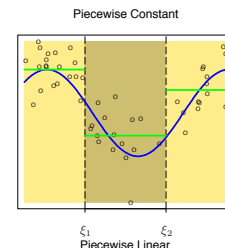
■ Resulting model: $f(x) = \sum_{m=1}^3 \beta_m h_m(x)$

■ Fit: Take mean of data in each region

■ Example 2: Piecewise linear

■ Add three basis functions:

$$h_{m+3} = h_m(x)x$$



From Hastie, Tibshirani, Friedman book

©Emily Fox 2013

28

Regression Splines – Linear

- Resulting piecewise linear model:

$$f(x) = I(x < \xi_1)(\beta_1 + \beta_4 x) + I(\xi_1 \leq x < \xi_2)(\beta_2 + \beta_5 x) + I(\xi_2 \leq x)(\beta_3 + \beta_6 x)$$

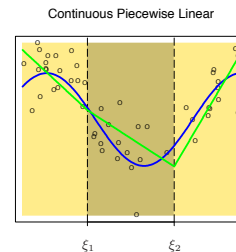
- # of params?

- Typically prefer continuity...

- Enforce

- Which implies

- # params?



From Hastie, Tibshirani, Friedman book

Regression Splines – Linear

- More directly, we can use the **truncated power basis**

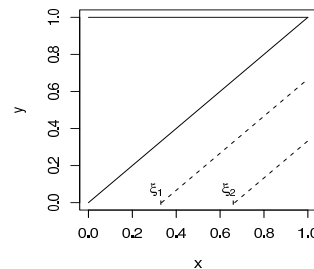
$$h_1(x) = 1$$

$$h_2(x) = x$$

$$h_3(x) = (x - \xi_1)_+$$

$$h_4(x) = (x - \xi_2)_+$$

- Resulting model:



From Wakefield book

- Continuous at the knots because all prior basis functions are contributing to the fit up to any single x

Regression Splines – Cubic

- Naively, extend as

$$f(x) = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3(x - \xi_1)_+ + \beta_4(x - \xi_1)_+^2 + \beta_5(x - \xi_2)_+ + \beta_6(x - \xi_2)_+^2$$

- But, 1st derivate is discontinuous (check this)
- Drop the truncated linear basis:
- Has continuous 1st derivative (check), but not 2nd

- Popular to consider **cubic spline**:

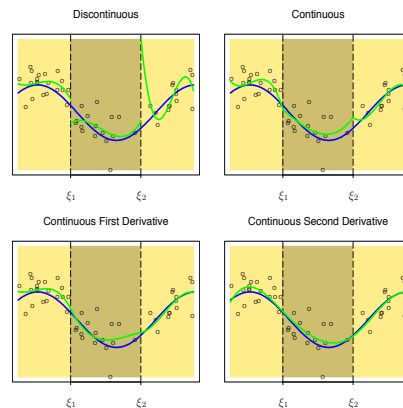
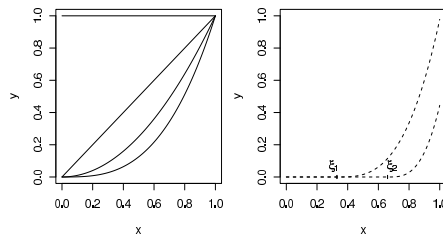
$$f(x) = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + b_1(x - \xi_2)_+^3 + b_2(x - \xi_2)_+^3$$

- Has continuous 1st and 2nd derivatives
- Typically people stop here

©Emily Fox 2013

31

Cubic Spline Basis and Fit



©Emily Fox 2013

32

Cubic Splines as Linear Smoothers

- Cubic spline function with K knots:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^K b_k (x - \xi_k)_+^3$$

- Simply a linear model

- Estimator:

- Linear smoother:

©Emily Fox 2013

33

Natural Cubic Splines

- For polynomial regression, fit near boundaries is erratic.
 - Problem is worse for splines: each is fit locally so no global constraint

- **Natural cubic splines** enforce linearity beyond boundary knots

- Starting from a cubic spline basis, the natural cubic spline basis is

$$N_1(x) = 1 \quad N_2(x) = x \quad N_{k+2}(x) = d_k(x) - d_{K-1}(x)$$

$$d_k(x) = \frac{(x - \xi_k)_+^3 - (x - \xi_K)_+^3}{\xi_K - \xi_k}$$

- Derivation

©Emily Fox 2013

34

Regression Splines – Summary

- Definition:

*An **order- M spline** with knots $\xi_1 < \xi_2 < \dots < \xi_K$ is a piecewise $M-1$ degree polynomial with $M-2$ continuous derivatives as the knots*

*A spline that is linear beyond the boundary knots is called a **natural spline***

- Choices:

- Order of the spline
- Number of knots
- Placement of knots

©Emily Fox 2013

35

Return to Smoothing Splines

- Objective:

$$\min_f \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx$$

- Solution:

- Natural cubic spline**
- Place knots at every observation location x_i

- Proof: See Green and Silverman (1994, Chapter 2) or Wakefield textbook

- Notes:

- Would seem to overfit, but penalty term shrinks spline coefficients toward linear fit
- Will not typically interpolate data, and smoothness is determined by λ

©Emily Fox 2013

36

Smoothing Splines

- Model is of the form: $f(x) = \sum_{j=1}^n N_j(x)\beta_j$

- Rewrite objective:

$$(y - N\beta)^T(y - N\beta) + \lambda\beta^T\Omega_N\beta$$

- Solution:

- Linear smoother:

©Emily Fox 2013

37

Splines – Summary

- **Regression splines:**

Fewer number of knots and no regularization

- **Smoothing splines:**

Knots at every observation and regularization (smoothness penalty) to avoid interpolators

©Emily Fox 2013

38