

Course Overview – Nonparametric Regression and Classification

STAT/BIOSTAT 527, University of Washington

Emily Fox

April 1st, 2014

©Emily Fox 2014

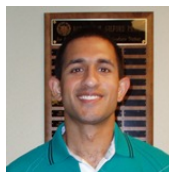
1

Course Staff

- Instructor: **Emily Fox**



- TA: **Amrit Dhar**



©Emily Fox 2014

2

Content: What is the course about?

©Emily Fox 2014

3

Course Structure

- 3 Primary Tasks:
 - Regression
 - Classification
 - Density Estimation

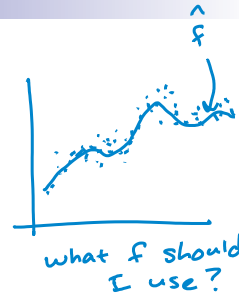
- 5 Modules:
 - Nonparametric Preliminaries
 - Splines and Kernels
 - Bayesian Nonparametrics
 - Nonparametrics for Multivariate Covariates
 - Classification

©Emily Fox 2014

4

Task 1: Regression

- Assume a sample $(x_1, Y_1), \dots, (x_n, Y_n)$
- Model: $Y_i = f(x_i) + \epsilon_i$ $E[\epsilon_i] = 0$
 \uparrow unknown



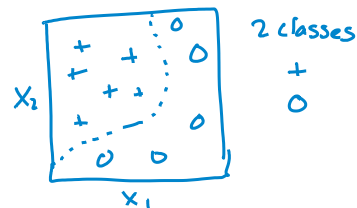
- Task involves estimating the function f
 \uparrow estimator \hat{f}
- Goals of nonparametric approach:
 - Make few assumptions about f
 - Use a large number of parameters, but constrained in some way to avoid overfitting the data
 - Complexity can grow with the sample size

©Emily Fox 2014

5

Task 2: Classification

- Assume a sample $(x_1, Y_1), \dots, (x_n, Y_n)$
 $Y_i \in \{1, \dots, K\}$
 \uparrow # classes



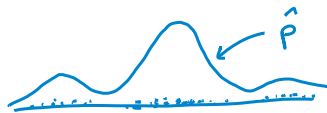
- Task involves estimating a predictive model of Y given x
- Goals of nonparametrics are as before, but now for link between x and Y with Y discrete-valued

©Emily Fox 2014

6

Task 3: Density Estimation

- Assume a sample $X_1, \dots, X_n \sim P \leftarrow \text{unknown}$



- Task involves estimating the density p
estimator \hat{p}
- Goals of nonparametric approach are as before, but applied to the estimation of p

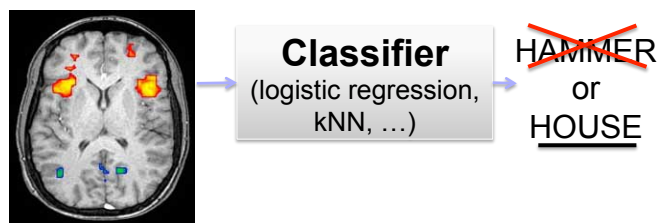
©Emily Fox 2014

7

fMRI Prediction Task *cool task involving both reg. + class*

- Goal:** Predict word stimulus from fMRI image

can we read your mind?



©Emily Fox 2014

8

fMRI



©Emily Fox 2014

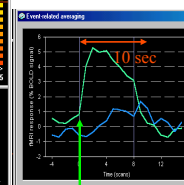
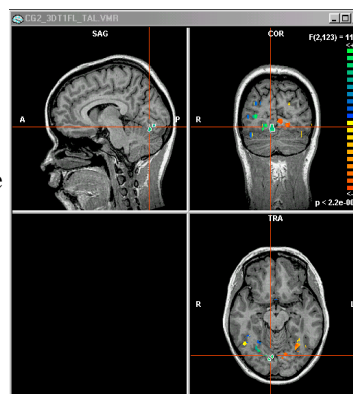
9

fMRI

very high
~1 mm resolution
pretty slow
~1 image per sec.

20,000 voxels/image
safe, non-invasive

measures Blood
Oxygen Level
Dependent (BOLD)
response

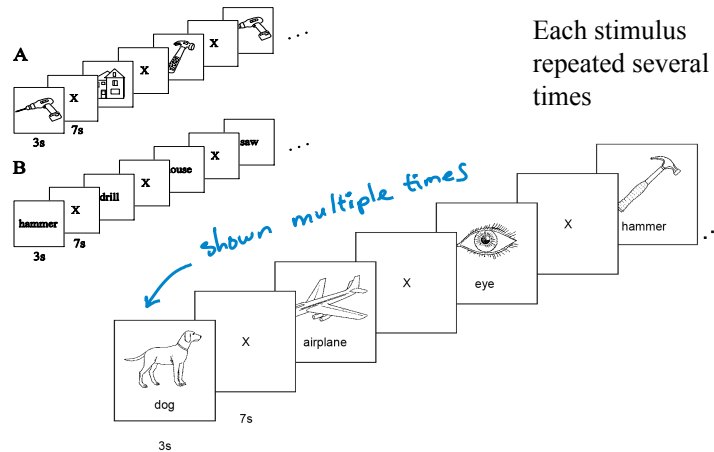


Typical fMRI
response to
impulse of
neural activity

©Emily Fox 2014

10

Typical Stimuli

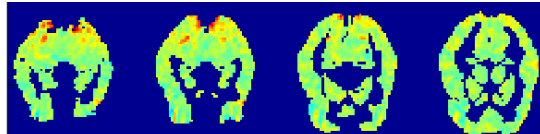


©Emily Fox 2014

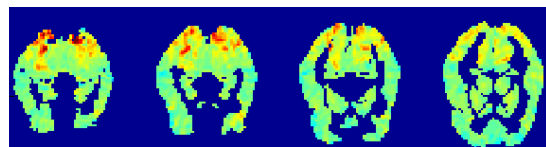
11

fMRI Activation

fMRI activation for "bottle":

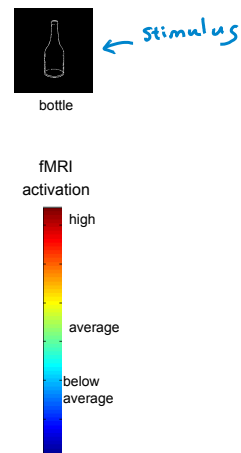
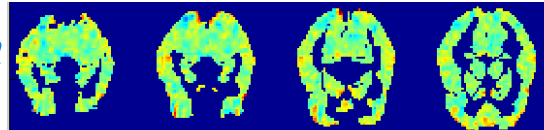


Mean activation averaged over 60 different stimuli:



"bottle" minus mean activation:

Is this enough?

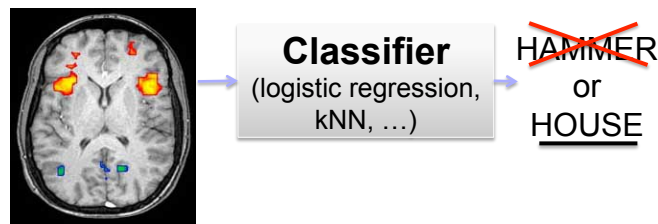


©Emily Fox 2014

12

fMRI Prediction Task

- **Goal:** Predict word stimulus from fMRI image
- **Challenges:** *# voxels*
 - $p \gg n$ (covariate dimension \gg sample size)
 - Cost of fMRI recordings is high
 - Only have a few training examples for each word



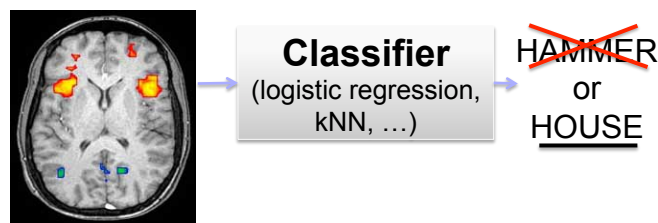
©Emily Fox 2014

13

Zero-Shot Classification

- **Goal:** Classify words not in the training set
- **Challenges:**
 - Cost of fMRI recordings is high
 - Can't get recordings for every word in the vocabulary

Never showed the word "giraffe"



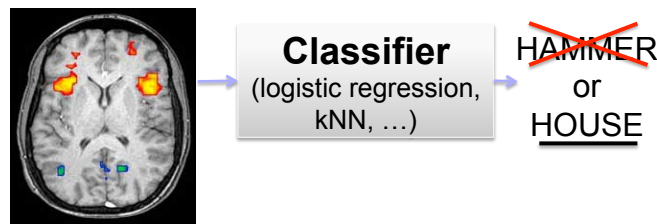
©Emily Fox 2014

14

Zero-Shot Classification

- **Goal:** Classify words not in the training set
- **Challenges:**
 - Cost of fMRI recordings is high
 - Can't get recordings for every word in the vocabulary
- We don't have many brain images, but we have a lot of info about the words and how they relate (co-occurrence, etc.)
- How do we utilize this "cheap" information?

many docs containing "giraffe" also contain "neck" "zoo" ...



©Emily Fox 2014

15

Semantic Features

Google Trillion word corpus

Semantic feature values: "celery"

0.8368, eat
0.3461, taste
0.3153, fill
0.2430, see
0.1145, clean
0.0600, open
0.0586, smell
0.0286, touch
...
...
0.0000, drive
0.0000, wear
0.0000, lift
0.0000, break
0.0000, ride

Semantic feature values: "airplane"

0.8673, ride
0.2891, see
0.2851, say
0.1689, near
0.1228, open
0.0883, hear
0.0771, run
0.0749, lift
...
...
0.0049, smell
0.0010, wear
0.0000, taste
0.0000, rub
0.0000, manipulate

©Emily Fox 2014

16

Zero-Shot Classification

- From training data, learn two mappings:

- S : input image \rightarrow semantic features
- L : semantic features \rightarrow word

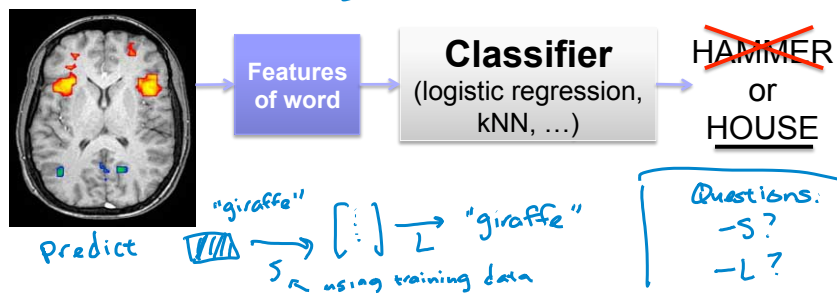
image word

$$A = \left\{ \begin{array}{c} \text{few} \\ \text{many} \end{array} \right\} \rightarrow \text{"dog"} \quad \text{semantic features}$$

$$B = \left\{ \begin{bmatrix} \vdots \end{bmatrix} \right\} \rightarrow \text{"dog"} \quad \text{semantic features}$$

- Can use "cheap" co-occurrence data to help learn L

Training $\left\{ \begin{array}{c} \text{image} \\ S \end{array} \right\} \rightarrow \left\{ \begin{bmatrix} \vdots \end{bmatrix} \right\} \rightarrow \text{"dog"} \right\}$ uses $A+B$ n examples, n small



©Emily Fox 2014

17

Assumed Background

- [Stat 502 and Stat 504] or [Biostat 514 and Biostat 515]

- Comfortable with:

- Linear algebra
- Probability
- R (or Matlab, Python, etc.)

- Computational and mathematical maturity

- Many concepts thrown at you quickly!
- Some background is not provided in above courses and requires significant dedication to keep up
- Expected to implement many methods from scratch

©Emily Fox 2014

18

Logistics: How is the course going to run?

©Emily Fox 2014

19

Website and Discussion Board

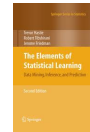
- Course website:
<http://stat.washington.edu/courses/stat527/s14>
- Catalyst:
 - ☐ Used for all discussions
 - ☐ Post all questions there (unless personal)
 - ☐ Completed assignments submitted via Catalyst dropbox
 - ☐ Homework solutions and feedback on assignments posted through Catalyst

©Emily Fox 2014

20

Reading

- Primary reference:
 - Hastie, Tibshirani, Friedman “The Elements of Statistical Learning”, Springer 2009
- Other strongly suggested textbooks (on website):
 - Wakefield, “Bayesian and Frequentist Regression Methods”, Springer 2012
 - Wasserman, “All of Nonparametric Statistics”, Springer 2005
- Papers linked on course website



©Emily Fox 2014

21

Homework

- Roughly 5 HWs total
- Assigned and due on *Thursdays*
 - Starting weekly then biweekly
- Collaboration allowed, but write-ups and coding must be done individually
- Submitted via Catalyst before start of lecture
- Allowed 2 “late days” for entire quarter

©Emily Fox 2014

22

Project

■ Options:

- ☐ Choose project from specified list
- ☐ Re-implement existing paper from specified list
- ☐ Propose own project idea

■ Individual

■ New work, but can be connected to research

■ Schedule:

- ☐ Proposal (1 page) – April 24
- ☐ Progress report (3 pages) – May 15
- ☐ Project presentation – **TBD (poster or in-class)**
- ☐ Final report (8 pages, NIPS format) – June 10

©Emily Fox 2014

23

Grading

■ HWs (60%)

- ☐ One HW treated as “midterm” and worth more

■ Final project (40%)

- ☐ Midway report (20%)
- ☐ Project presentation (20%)
- ☐ Final paper (60%)

©Emily Fox 2014

24

Support/Resources

■ Office Hours

- TA: W 12:30-2:30pm in Padelford B-302
- Emily: Th 10:30-11:30am in CSE 346

■ Recitations

- Optional tutorial/example-based sections will be held
every other week
- Very helpful for homework!
- Location TBD

©Emily Fox 2014

25

Module 1: Nonparametric Preliminaries

What to Report?,
Model Selection,
Model Assessment

STAT/BIOSTAT 527, University of Washington

Emily Fox

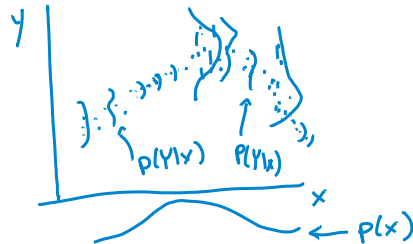
April 1st, 2014

©Emily Fox 2014

26

The Optimal Prediction

- Assume we *know* the data-generating mechanism



e.g. know $p(Y|x)$
for all $x \in \mathcal{X}$
and assume X is
random w/ known
dist. $p(x)$

- If our task is prediction, which summary of the distribution $Y|x$ should we report?

e.g. life expectancy e.g. age, income, race ...
For x , what fun $\hat{f}(x)$ should we choose to predict Y
if we can choose any $\hat{f}(\cdot)$?

©Emily Fox 2014

27

The Optimal Prediction

- Taking a decision-theoretic framework, consider the **expected loss**

predictions are penalized by $L(Y, \hat{f}(x))$

$$E_{X,Y}[L(Y, \hat{f}(X))] = E_X\{E_{Y|X}[L(Y, \hat{f}(x)) | X=x]\}$$

- $\hat{f}^*(\cdot)$ should min \rightarrow

- can min. pointwise for each x

- What are loss functions we might consider?

©Emily Fox 2014

28

Continuous Responses

- Expected loss $E_X \{E_{Y|X} [L(Y, f(x)) \mid X = x]\}$

- Example: L_2 $L(Y, \hat{f}(x)) = (Y - \hat{f}(x))^2$

Solution: $\hat{f}^*(x) = E[Y|x]$

Proofs:
HW

- Example: L_1 $L(Y, \hat{f}(x)) = |Y - \hat{f}(x)|$

Solution: $\hat{f}^*(x) = \text{median}(Y|x)$

- More generally: L_p $L(Y, \hat{f}(x)) = \left\{ \int |Y - \hat{f}(x)|^p \right\}^{1/p}$

©Emily Fox 2014

29

General Responses

- Expected loss $E_X \{E_{Y|X} [L(Y, f(x)) \mid X = x]\}$

- Example: log-likelihood $L(Y, \hat{f}(x)) = -2 \log p(Y | \hat{f}(x))$

When Gaussian: $Y|x \sim N(f(x), \sigma^2)$

$$\rightarrow L(Y, \hat{f}(x)) = \log(2\pi\sigma^2) + (Y - \hat{f}(x))^2 / \sigma^2$$

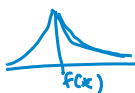
$$\rightarrow \hat{f}^*(x) = E[Y|x]$$

Gaussian model
neg. log. like. loss $= L_2$ squared error loss

When Laplace: $Y|x \sim \text{Lap}[f(x), \phi]$

$$\rightarrow \hat{f}^*(x) = \text{median}(Y|x)$$

Laplace +
NLL $= L_1$



©Emily Fox 2014

30

Incorporating Models into Prediction

- We don't actually know the data-generating mechanism
- Need an estimator $\hat{f}_n(\cdot)$ based on a random sample Y_1, \dots, Y_n , also known as **training data**

e.g. Est. $E[Y|x]$, but how? Typically, don't have multiple obs. at a given x . Maybe $\hat{f}_n(x) = \text{Avg}(y_i | x_i \in \text{Nbd}(x))$.
Can be problematic if not many obs.

- Statistical models can be used to encode knowledge about aspects of the data-generating mechanism
e.g. Assume linear form, then we know how to approx. $E[Y|x]$ s.t. linear constraint on f
- Models can provide simplifying assumptions
 - Can help cope with estimation issues due to limited data

©Emily Fox 2014

31

Incorporating Models into Prediction

- Assume some form for how the data are generated

□ E.g., $Y = f(x) + \epsilon$ $E[\epsilon] = 0$ $\text{var}(\epsilon) = \sigma^2$

$\Rightarrow f(x) = E[Y|x]$ how to est. f ?

- For non-constant variance, can consider GLMs

- Then, typically assume some form for $f(x)$

e.g. $f(x) = \beta^T x$ $\hat{f}(x) = \hat{\beta}^T x$

- Model + loss function \rightarrow some estimator

e.g. $L_2 \rightarrow \hat{\beta} = [E[XX^T]]^{-1} E[XY]$ r.v.
approx $\hat{\beta}_n = (X^T X)^{-1} X^T Y$ vector of training obs.
matrix of covariates "design matrix"

©Emily Fox 2014

32

Parametric Regression

- *Parametric* inference assumes parametric form for $f(x)$

e.g. $f(x) = \beta^T x$
↖ $f(\cdot)$ is indexed by param. β

- Advantages:

- ☐ Efficient estimation
- ☐ Concise summarization

↖ e.g. LS est. of β , $\hat{\beta}_n$,
leads to an est. \hat{f}_n of f

- What is the right parametric form for $f(x)$?

Should it change w/ sample size?

©Emily Fox 2014

33

Goals of Nonparam Regression

- Goals of *nonparametric* inference:

- ☐ Assume little prior knowledge of data-generating mechanism
- ☐ More flexibly model f (i.e., relationship between x and Y)
- ☐ Maintain “reasonable” efficiency of estimation

- Often actually assume parametric forms with large numbers of parameters

- ☐ Constrained to avoid overfitting the data

- Particularly useful when task is prediction

- ☐ Focus on accuracy of prediction rather than parameter values

- Let's discuss this idea of “complexity” more...

©Emily Fox 2014

34

Model Complexity

- How complex of a function should we choose?

- ☐ To increase flexibility, using many parameters is attractive

Reduce bias

- ☐ However, wide prediction intervals...

Fixed dataset contains a limited amt. of info

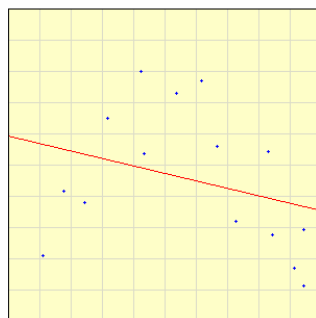
- ☐ Leads to wild predictions

©Emily Fox 2014

35

Example: Polynomial Regression

- For added flexibility, allow for high order polynomial, right?



Select points by clicking on the graph or press

Example

Degree of polynomial: 1 ☒ Fit Y to X
☐ Fit X to Y

Calculate

View Polynomial

Reset

$$y_i = \sum_{j=0}^P \beta_j x_i^j + \epsilon_i$$

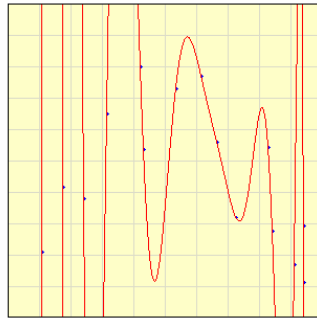
*Not always good to
add params*

©Emily Fox 2014

36

Example: Polynomial Regression

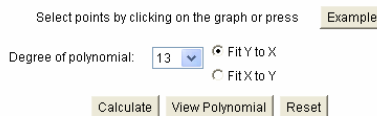
- For added flexibility, allow for high order polynomial, right?



sensitive to small changes
in data

High order = low bias, but
high var

How do we assess
an estimator \hat{f}_n ?



©Emily Fox 2014

37

Measuring Predictive Performance

- Having chosen a model, how do we assess its performance? *we'll come back to this question*
- 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

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

want small GE.
Can think of this
as a bias-var
trade off

avg. over
all possible
new obs. + cov.

fixed based on
training data

©Emily Fox 2014

38

Measuring Predictive Performance

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

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

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

©Emily Fox 2014

39

Measuring Predictive Performance

- Finally, let's average over covariates x
 - Integrated MSE**
 - Average MSE**
- Note: **avg. pred. risk** = $\sigma^2 + \text{avg. MSE}$

©Emily Fox 2014

40

Bias-Variance Tradeoff

- Minimizing risk = balancing bias and variance

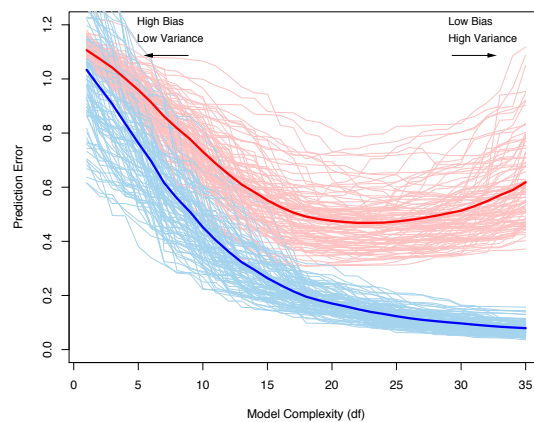
- Note: $f(x)$ is unknown, so cannot actually compute MSE

©Emily Fox 2014

41

In Practice...

- Minimizing risk = balancing bias and variance



From Hastie, Tibshirani, Friedman

©Emily Fox 2014

42

More on Nonparam Regression

- Often framed as learning functions with a complexity penalty
 - Regular behavior in small neighborhoods of the input
 - E.g., locally linear or low-order polynomial...estimator results from averaging over these local fits
- Choice of neighborhood = strength of constraint
 - Large neighborhood can lead to linear fit (very restrictive) whereas small neighborhoods can lead to interpolation (no restriction)

©Emily Fox 2014

43

More on Nonparam Regression

- Different restrictions lead to different nonparametric approaches
 - Roughness penalty → **splines**
 - Weighting data locally → **kernel methods**
 - Etc.
- Each method has associated **smoothing** or **complexity** param
 - Magnitude of penalty
 - Width of kernel (defining “local”)
 - Number of basis functions
 - ...
- Bias-variance tradeoff
- Will explore methods for choosing smoothing parameters

©Emily Fox 2014

44

Reading

- Wakefield: 10.3-10.4
- Hastie, Tibshirani, Friedman: 7.1-7.3

What you should know

- What to report when data-generating mechanism is:
 - Known (optimal prediction)
 - Unknown and constrained to a specified model + loss fcn
- Example loss functions for
 - Continuous RVs
 - General RVs
- Goals of parametric vs. nonparametric methods
- Bias-variance tradeoff
- Measures of performance of estimators