Course Overview – Nonparametric Regression and Classification

STAT/BIOSTAT 527, University of Washington

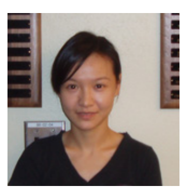
Emily Fox

April 2nd, 2013

Course Staff

- Instructor: **Emily Fox**

- TA: **Shirley You Ren**



Content: What is the course about?

Course Structure

- 3 Primary Tasks:
  - **Regression**
  - **Classification**
  - **Density Estimation**
- 5 Modules:
  - **Nonparametric Preliminaries**
  - **Splines and Kernels**
  - **Bayesian Nonparametrics**
  - **Nonparametrics for Multivariate Covariates**
  - **Classification**

Task 1: Regression

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

  f ← unknown

- Task involves estimating the function *f*

What f should I use?

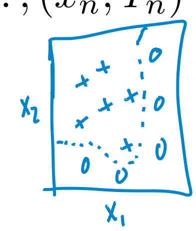
- Goals of nonparametric approach:
  - Make few assumptions about *f*
  - Use a large number of parameters, but constrained in some way to avoid overfitting the data
  - Complexity can grow with the sample size

Task 2: Classification

- Assume a sample $(x_1, Y_1), \dots, (x_n, Y_n)$

  $Y_i \in \{1, \dots, K\}$

  K ← # classes



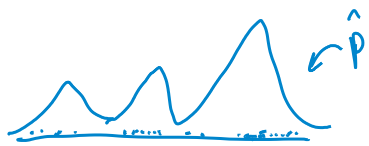
2 classes
+
o

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

Task 3: Density Estimation

- Assume a sample $X_1, \dots, X_n \sim p$

unknown



- Task involves estimating the density *p*

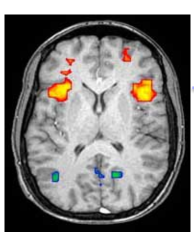
- Goals of nonparametric approach are as before, but applied to the estimation of *p*

fMRI Prediction Task

cool task involving both reg. + class.

- **Goal:** Predict word stimulus from fMRI image

Can we read your brain?



→ **Classifier** (logistic regression, kNN, …) → ~~HAMMER~~ or HOUSE

fMRI

fMRI

very high

~1 mm resolution

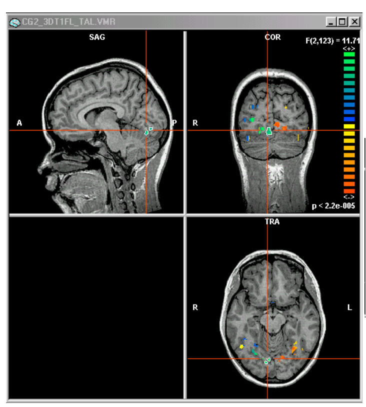
pretty slow

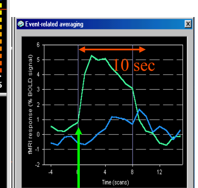
~1 image per sec.

20,000 voxels/image

safe, non-invasive

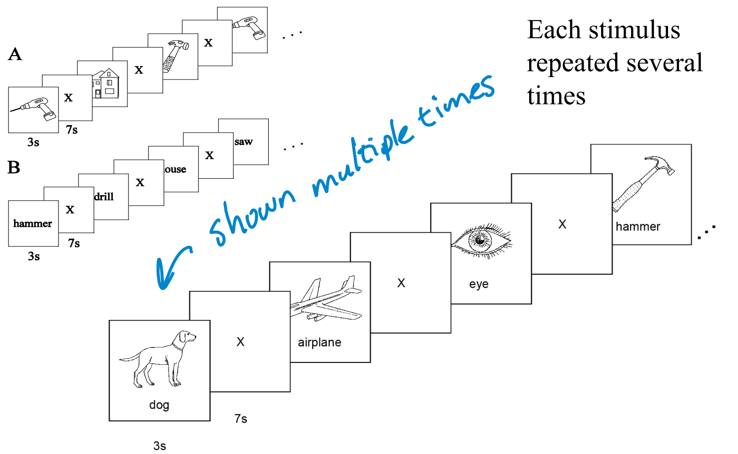
measures Blood Oxygen Level Dependent (BOLD) response





Typical fMRI response to impulse of neural activity

Typical Stimuli

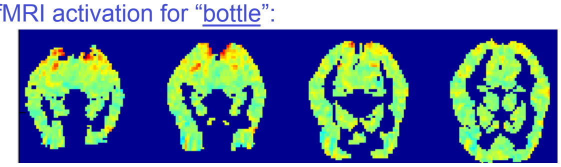


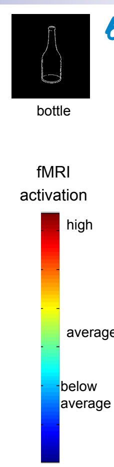
shown multiple times

Each stimulus repeated several times

fMRI Activation

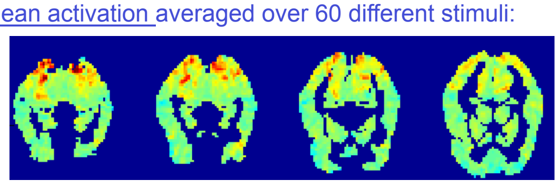
fMRI activation for "bottle":





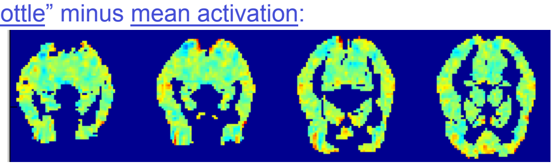
← stimulus

Mean activation averaged over 60 different stimuli:



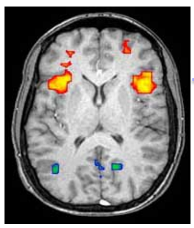
"bottle" minus mean activation:

Is this enough?



fMRI Prediction Task

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



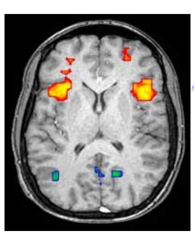
Classifier
(logistic regression, kNN, …)

~~HAMMER~~
or
HOUSE

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"



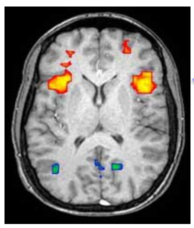
Classifier
(logistic regression, kNN, …)

~~HAMMER~~
or
HOUSE

Zero-Shot Classification

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

- **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?



→ **Classifier** (logistic regression, kNN, …) → ~~HAMMER~~ or HOUSE

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

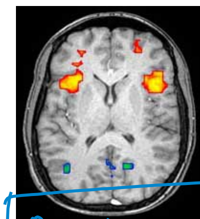
Zero-Shot Classification

- From training data, learn two mappings: $A = \{ \text{[image]} \rightarrow \text{"dog"} \}$ few
  - S: input image → semantic features
  - L: semantic features → word

  $B = \{ [\vdots] \rightarrow \text{"dog"} \}$ many

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

Training $\{ \text{[image]} \rightarrow [\vdots] \rightarrow \text{"dog"} \}$ uses A+B   n examples, n small



Features of word → Classifier (logistic regression, kNN, …) → ~~HAMMER~~ or HOUSE

Questions: - S? - L?

Predict "giraffe" $\text{[image]} \xrightarrow{S} [\vdots] \xrightarrow{L} \text{"giraffe"}$

↑ using training data

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**

Logistics: How is the course going to run?

©Emily Fox 2013

19

Website and Discussion Board

- Course website:
  http://stat.washington.edu/courses/stat527/s13

- Catalyst:
  - Used for all discussions
  - Post all questions there (unless personal)
  - See website for sign-up details

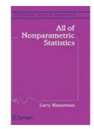
©Emily Fox 2013

20

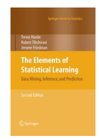
Reading

- No required textbook
- Three suggested textbooks (on website):
  - Wakefield, “Bayesian and Frequentist Regression Methods", Springer 2012

  - Wasserman, “All of Nonparametric Statistics”, Springer 2005



  - Hastie, Tibshirani, Friedman “The Elements of Statistical Learning”, Springer 2009



- Papers linked on course website

Homework

- 7 HWs total
- Assigned and due weekly on *Thursdays*
- Collaboration allowed, but write-ups and coding must be done individually
- Submitted at beginning of class
- Allowed 2 “late days” for entire quarter

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 25
  - Progress report (3 pages) – May 16
  - Poster presentation – June 6
  - Final report (8 pages, NIPS format) – June 11

Grading

- HWs 1, 2, 4, 5, 6 (10% each)
- HWs 3, 7 (5%) – short, due dates coincide with project due dates
- Final project (40%)

Support/Resources

- Office Hours
  - TA: W 2-4pm in Padelford A-316
  - Emily: Th 11am-12pm in Padelford B-305

- Recitations
  - Optional tutorial/example-based sections will be held *every other* week
  - Choose from:
    - Monday, 2-3pm
    - Monday, 5-6pm
    - Tuesday, 4-5pm
  - Location TBD

Module 1: Nonparametric Preliminaries

Intro,
What to Report?

STAT/BIOSTAT 527, University of Washington
Emily Fox
April 2nd, 2013

The Optimal Prediction

- Assume we *know* the data-generating mechanism
- If our task is prediction, which summary of the distribution $Y \mid x$ should we report?

For x, what fcn $f(x)$ should we choose to predict Y
if we can choose any $f(\cdot)$

- Taking a decision-theoretic framework, consider the ***expected loss*** predictions are penalized by $L(\cdot,\cdot)$

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

- $\hat{f}(\cdot)$ should min
- can min. pointwise

Continuous Responses

- Expected loss $E_X\left\{E_{Y|X}\left[L(Y,f(x)) \mid X=x\right]\right\}$
- Example: L_2 $\quad L(Y,f(x)) = (Y-f(x))^2$

  Solution: $\hat{f}(x) = E[Y|x]$
- Example: L_1 $\quad L(Y,f(x)) = |Y-f(x)|$

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

Proofs: HW

General Responses

- Expected loss $E_X \left\{ E_{Y|X} \left[L(Y, f(x)) \mid X = x \right] \right\}$
- Example: log-likelihood $L(Y, f(x)) = -2 \log p(Y \mid f(x))$

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

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

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

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

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

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 $Y_1, \ldots, Y_n$, also known as ***training data***

e.g. Est. $E[Y|x]$, but how? Typically don't have mult. obs. at a given x. Maybe $\hat{f}_n(x) = \text{Avg}(Y_i \mid X_i \in \text{Nbhd}(x))$

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. a linear const. on f

- Models can provide simplifying assumptions
  - Can help cope with estimation issues due to limited data

Incorporating Models into Prediction

- Assume some form for how the data are generated
  - E.g., $Y = f(x) + \epsilon \qquad E[\epsilon] = 0 \quad \mathrm{var}(\epsilon) = \sigma^2$

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

  - For non-constant variance, can consider GLMs
- Then, typically assume some form for *f(x)*

e.g., $f(x) = \beta^T x$

- Model + loss function → some estimator r.v.

e.g. $L_2 \rightarrow \hat{\beta} = (E(XX^T))^{-1} E(XY)$

approx. $\hat{\beta}_n = (X^T X)^{-1} X^T y$ ← obs. vec.; matrix of covariates

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?

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…

Model Complexity

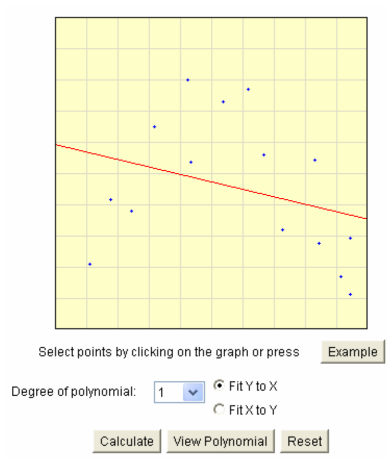
- How complex of a function should we choose?
  - To increase flexibility, using many parameters is attractive

    Reduce bias
  - However, wide prediction intervals…

    Fixed dataset contains a limited amt of info
  - Leads to wild predictions

Example: Polynomial Regression

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

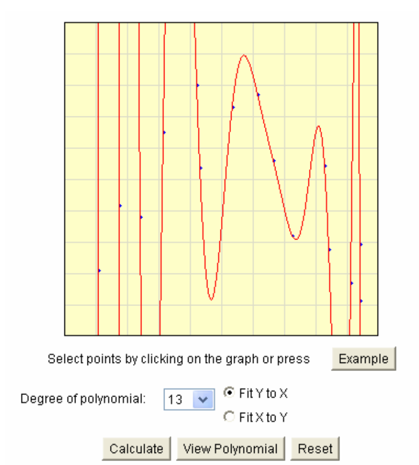


$$y_i = \sum_{j=0}^{p} \beta_j x_i^j + \epsilon_i$$

Not always good to add params

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 fn ?

Measuring Predictive Performance

- Assume estimate $\hat{f}_n(\cdot)$ based on training data $y_1, \ldots, y_n$

fixed

- The ***generalization error*** provides a measure of predictive performance

$$GE(\hat{f}_n) = E_{Y,X}\left[L(Y, \underbrace{\hat{f}_n}_{\text{fixed}}(X))\right]$$

Measuring Predictive Performance

- Assume L_2 loss

$Y = f(X) + \epsilon$ ☆ $E[\epsilon] = 0$ $var(\epsilon) = \sigma^2$

- Averaging over repeat training sets $\mathbf{Y}_n = Y_1, \ldots, Y_n$ we get the ***predictive risk*** at x^*

$$E_{Y^*,\mathbf{Y}_n}\left[(Y^* - \hat{f}_n(x^*))^2\right] = E_{Y^*,Y_n}\left[\left(Y^* - f(x^*) + f(x^*) - \hat{f}_n(x^*)\right)^2\right]$$

test, training

fcn of training data Y_n

$$= E_{Y^*}\left[(Y^* - f(x^*))^2\right] + \underbrace{E_{Y_n}\left[(\hat{f}_n(x^*) - f(x^*))^2\right]}_{MSE(\hat{f}_n(x^*))} + 2\,\cancel{E_{Y^*}[Y^* - f(x^*)]}^{\,0}\, E_{Y_n}[\hat{f}_n(x^*) - f(x^*)]$$

$$= \sigma^2 + MSE(\hat{f}_n(x^*))$$ ✓

"irreducible error" "risk"

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

Measuring Predictive Performance

- Finally, let's average over covariates *x*
  - ***Integrated MSE*** $\int MSE(\hat{f}_n(x))\, p(x)\, dx$
    summary over all inputs
  - ***Average MSE*** $\frac{1}{n}\sum_{i=1}^{n} MSE(\hat{f}_n(x_i))$
    Monte Carlo est. $x_i \sim p$
- Note: ***avg. pred. risk =*** σ^2 + ***avg. MSE***

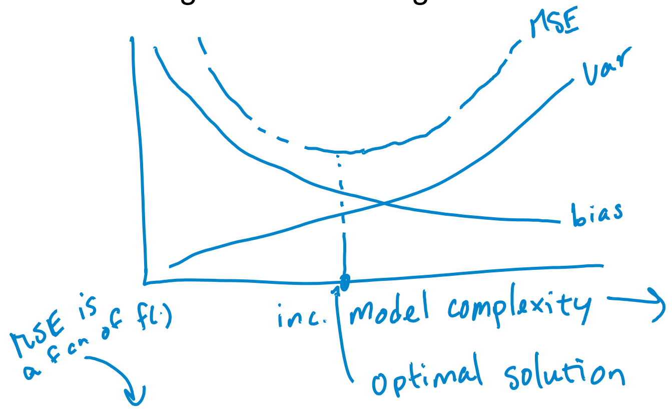
$$\frac{1}{n}\sum_{i=1}^{n} E_{Y_n, Y_n^*}\left[(Y_i^* - \hat{f}(x_i))^2\right]$$

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

Bias-Variance Tradeoff

recall polynomial reg. example

- Minimizing risk = balancing bias and variance



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

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)

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