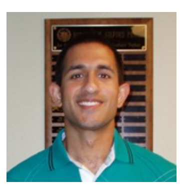
Course Overview – Nonparametric Regression and Classification

STAT/BIOSTAT 527, University of Washington
Emily Fox
April 1st, 2014

Course Staff

- Instructor: **Emily Fox**

- TA: **Amrit Dhar**



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
- Model:

- Task involves estimating the function *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

Task 2: Classification

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

- 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

Task 3: Density Estimation

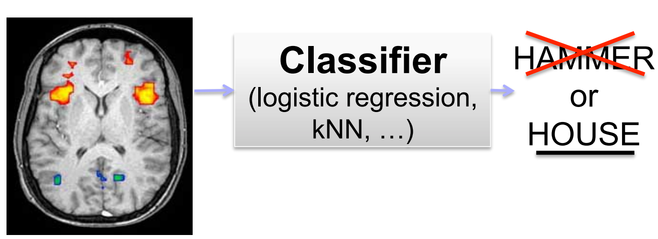
- Assume a sample

- Task involves estimating the density *p*

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

fMRI Prediction Task

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



fMRI

fMRI

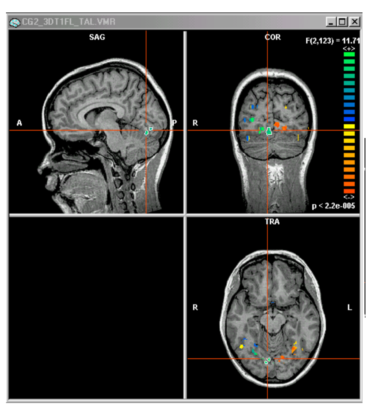
~1 mm resolution

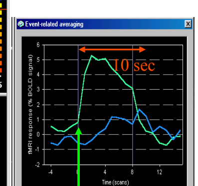
~1 image per sec.

20,000 voxels/image

safe, non-invasive

measures Blood Oxygen Level Dependent (BOLD) response

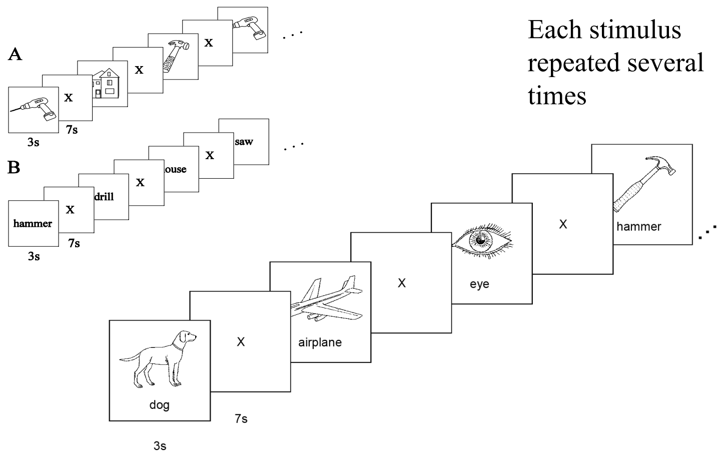




Typical fMRI response to impulse of neural activity

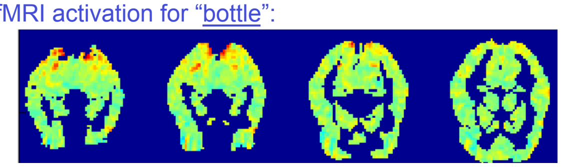
Typical Stimuli

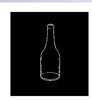
Each stimulus repeated several times



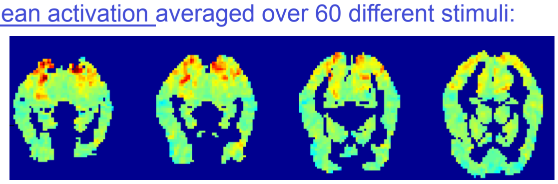
fMRI Activation

fMRI activation for "bottle":



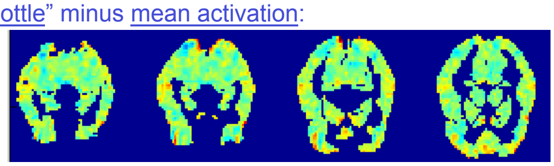


Mean activation averaged over 60 different stimuli:



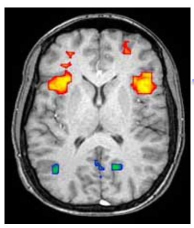
fMRI activation

"bottle" minus mean activation:



fMRI Prediction Task

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

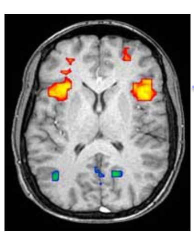


Classifier
(logistic regression, kNN, …)

~~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

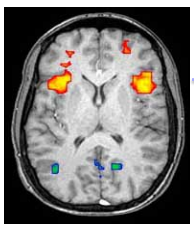


Classifier
(logistic regression, kNN, …)

~~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
- 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

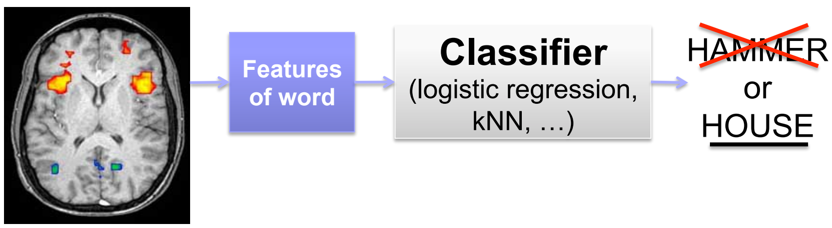
Semantic Features

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:
  - S: input image → semantic features
  - L: semantic features → word
- Can use “cheap” co-occurrence data to help learn L



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

Logistics: How is the course going to run?

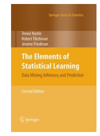
Website and Discussion Board

- Course website:
  http://stat.washington.edu/courses/stat527/s14

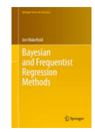
- Catalyst:
  - Used for all discussions
  - Post all questions there (unless personal)
  - Completed assignments submitted via Catalyst dropbox
  - Homework solutions and feedback on assignments posted through Catalyst

Reading

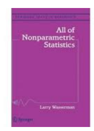
- Primary reference:
  - Hastie, Tibshirani, Friedman “The Elements of Statistical Learning”, Springer 2009



- 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

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

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

Grading

- HWs (60%)
  - One HW treated as "midterm" and worth more
- Final project (40%)
  - Midway report (20%)
  - Project presentation (20%)
  - Final paper (60%)

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

Module 1: Nonparametric Preliminaries

What to Report?, Model Selection, Model Assessment

STAT/BIOSTAT 527, University of Washington

Emily Fox

April 1st, 2014

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?

The Optimal Prediction

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

- What are loss functions we might consider?

Continuous Responses

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

  Solution:
- Example: L_1

  Solution:
- More generally: L_p

General Responses

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

  When Gaussian:

  When Laplace:

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***
- Statistical models can be used to encode knowledge about aspects of the data-generating mechanism
- 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., $Y = f(x) + \epsilon \qquad E[\epsilon] = 0 \quad \mathrm{var}(\epsilon) = \sigma^2$
  - For non-constant variance, can consider GLMs
- Then, typically assume some form for *f(x)*
- Model + loss function → some estimator

Parametric Regression

- *Parametric* inference assumes parametric form for $f(x)$
- Advantages:
  - Efficient estimation
  - Concise summarization
- What is the right parametric form for $f(x)$?

Goals of Nonparam Regression

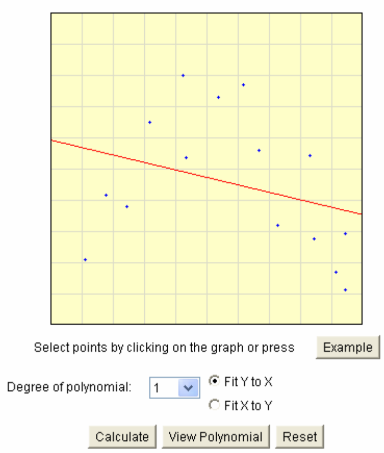
- Goals of *nonparametric* inference:
  - Assume little prior knowledge of data-generating mechanism
  - More flexibly model *f* (i.e., relationship between *x* and *Y*)
  - Maintain "reasonable" efficiency of estimation
- Often actually assume parametric forms with large numbers of parameters
  - Constrained 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
  - However, wide prediction intervals…
  - Leads to wild predictions

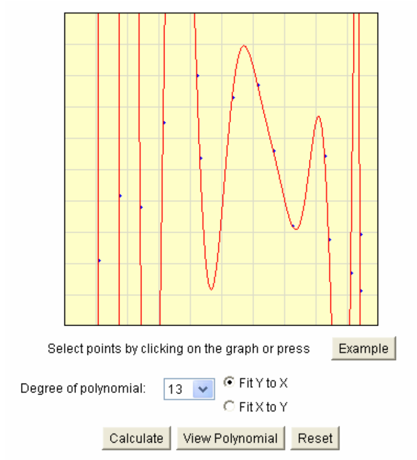
Example: Polynomial Regression

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



Example: Polynomial Regression

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



Measuring Predictive Performance

- Having chosen a model, how do we assess its performance?
- Assume estimate $\hat{f}_n(\cdot)$ based on training data $y_1, \ldots, y_n$
- The ***generalization error*** provides a measure of predictive performance

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

Measuring Predictive Performance

- Assume L_2 loss
- Averaging over repeat training sets $\mathbf{Y}_n = Y_1, \ldots, Y_n$ we get the ***predictive risk*** at x^*

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

- 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***
  - ***Average MSE***
- Note: ***avg. pred. risk =*** σ^2 + ***avg. MSE***

Bias-Variance Tradeoff

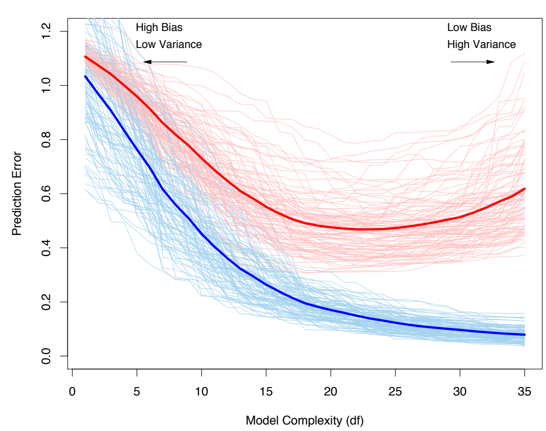
- Minimizing risk = balancing bias and variance

- 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)

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

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