Data Mining
W4240 Section 001

Giovanni Motta

Columbia University, Department of Statistics

September 30, 2015

Dimensionality Reduction

Summary of last week:

- ▶ have high dimensional continuous data
- ▶ want to find low dimensional representation
- ▶ Principal Components Analysis (PCA) finds best K-dimensional rotation of data
- ▶ gives a parsimonious description of the data

...but preprocessing **for what?**

Data Questions

Some questions are purely about prediction:

- ▶ Can I predict the yield of a crop next year (based on historicals, weather, ...) ?
- ▶ Can I predict a patient outcome (based on condition, behavior, ...) ?

Note that we naturally say things like "as a function of historicals, weather, ... "

We can wait and see what actually happens and then measure how well we did.

Data Questions

Some questions are about relationships between inputs and outputs:

- are average crop yields increasing over time?
- is there a negative health implication of doing z?

We need a model that tells us about the relationship between these values and explains the data well.

These patterns can be useful for directing our actions.

Data Questions

Some questions are about summarizing the data:

- ▶ what are the basic patterns that make up patients?
- ▶ which patient types are most similar?

We need a model that summarizes the data well.

These patterns can be useful for directing further investigation.

Data Questions

The questions fall into two categories: **supervised learning** and *unsupervised learning*

Supervised learning:

- ▶ predicting an output
- ▶ understanding the relationship between an input and an output

Unsupervised learning:

- ▶ summarizing the data
- ▶ understanding underlying (hidden) factors

PCA is unsupervised learning. For most of the rest of the class, we will be studying supervised learning.

Supervised Learning

Today we are learning about concepts that will be used throughout this course:

- Types of variables
- Estimating a function
- Prediction and inference
- Parametric and nonparametric estimation
- Interpretability

Supervised Learning

What is supervised learning? Let's start with an example.

- Let's look at the Advertising data set (not in the course data package!!!)
- In the data:
  - `sales` for each of 200 products
  - advertising budgets for each product in `TV`, `radio` and `newspaper`

If you have a new product, how should you spend your advertising money to generate the most sales? How much? In which media?

How can we use data to answer these questions?

Supervised Learning: Advertising

If you have a new product, how should you spend your advertising money to generate the most sales?

Idea: predict the level of sales from TV, radio and newspaper

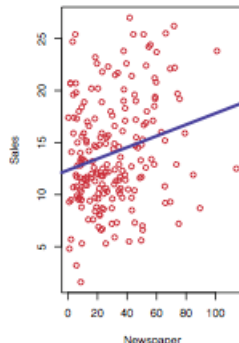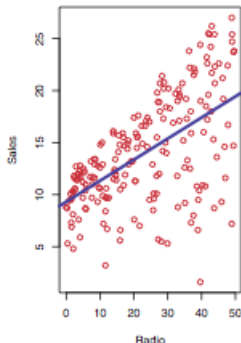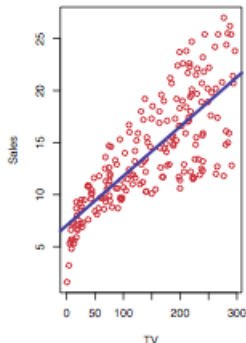Input variables (covariates, independent variables, predictors, features):

- TV
- radio
- newpaper

Output variable (response, dependent variable):

- sales

Supervised Learning: Advertising

If you have a new product, how should you spend your advertising money to generate the most sales?

Supervised Learning: Advertising

Formula:

$$\texttt{sales} = f\left(\texttt{TV}, \texttt{radio}, \texttt{newspaper}\right) + noise$$

We want to find f!

In general,

$$
\begin{aligned}
X &= (X_1, X_2, \ldots, X_p)^T & &\text{inputs} \\
Y & & &\text{output} \\
Y &= f(X) + \epsilon & &\underline{\text{relationship}}
\end{aligned}
$$

Supervised Learning: Regression and Classification

$$X = (X_1, X_2, \ldots, X_p)^T \qquad \text{inputs}$$
$$Y \qquad \text{output}$$
$$Y = f(X) + \epsilon \qquad \text{relationship}$$

We are interested in studying f in two settings: regression and classification.

Regression: Y has continuous values, like \$81,200 or 72.

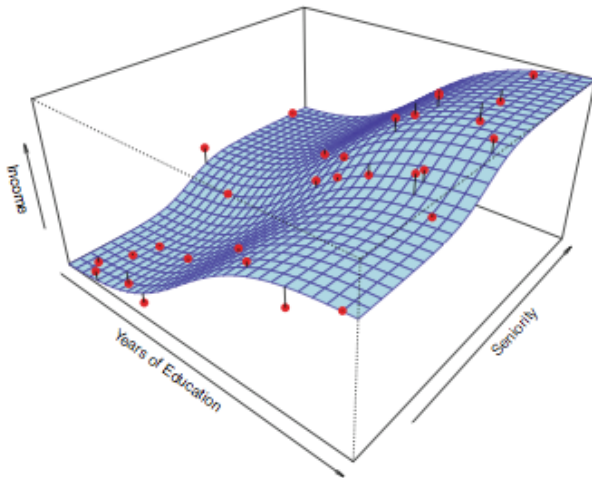**Classification:** Y has categorical values, like low/medium/high or red/green/blue

Supervised Learning: Income

Another example: we want to predict the annual income (income) of an individual based on years of education (years of education).
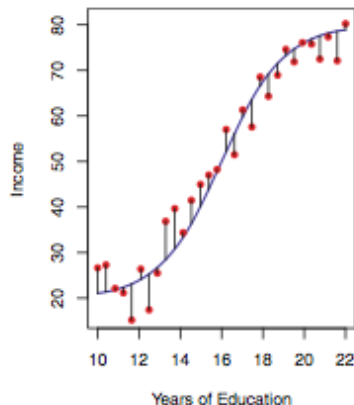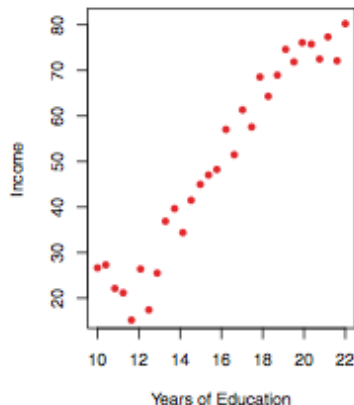
Supervised Learning: Income

Let's include seniority (`seniority`) as well.

Prediction and Inference

Why estimate f?

- prediction
- inference

Prediction:

- We have a new product with a set advertising budget TV, radio and newspaper). What will its sales be?
- Alice has 16 years of education and 0 years of seniority. What will her income be?

Goal: accurately estimate output for new inputs.

Prediction

In general, prediction is a two-step process:

1. Use data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to estimate f with $\hat{f}$
2. Feed new input X through $\hat{f}$ to get estimated output:

$$\hat{Y} = \hat{f}(X)$$

How accurate can we make $\hat{Y}$?

$$Y - \hat{Y} = \epsilon + f(X) - \hat{f}(X)$$

- irreducible error: $\underline{\epsilon}$
- reducible error: $\underline{f(X) - \hat{f}(X)}$

Prediction

Generally, we measure our success by the expected **mean squared error** (MSE):

$$\underline{\mathbb{E}(Y - \hat{Y})^2}$$

Fix both X and $\hat{f}$. What are the reducible and irreducible errors with the MSE?

$$\begin{aligned}
\mathbb{E}(Y - \hat{Y})^2 &= \mathbb{E}\left[f(X) + \epsilon - \hat{f}(X)\right]^2 \\
&= \mathbb{E}\left[(f(X) - \hat{f}(X))^2 + 2\epsilon(f(X) - \hat{f}(X)) + \epsilon^2\right] \\
&= (f(X) - \hat{f}(X))^2 + \mathrm{Var}(\epsilon)
\end{aligned}$$

Inference

If you have a new product, how should you spend your advertising money to generate the most sales?
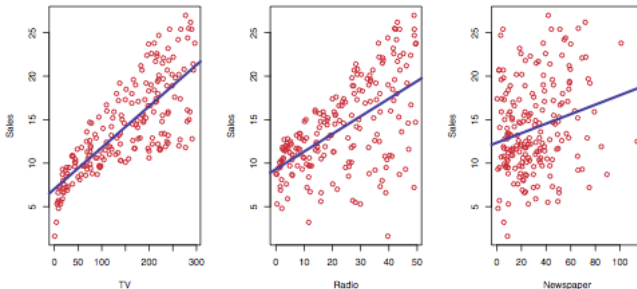
Which media contribute to sales? Which gives the biggest boost? If I spend more on advertising, how much should sales increase?

Inference

In **inference**, we want to learn about relationships between inputs
and outputs:

- ▶ how will increasing one input affect the output?
- ▶ is a specific combination of inputs associated with an increase
  in the output?

Let's load some approval data.

Can you think of some inference questions? Prediction questions?
What is the difference between the two?
(wartime approval? is the world getting angry?)
(which party will win the next election? what was the approval in
1985?)

Fitting f

Suppose we have n observations, $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]^T$.

OK, I now have:

- a question that I want to answer (prediction or inference)
- data

To answer my question, I need to estimate the relationship

$$Y = f(X) + \epsilon.$$

How do I find $\hat{f}$ using $(x_1, y_1), \ldots, (x_n, y_n)$?

How do I find $\hat{f}$ using $(x_1, y_1), \ldots, (x_n, y_n)$?

1. select a statistical model
2. select the model parameters using the data

What types of statistical models are there?

- parametric: described by a finite number of parameters
- non-parametric: not described by a finite number of parameters

Parametric Models

A **parametric model** is a statistical model described by a finite number of parameters. Examples include:

- a Gaussian distribution (parameters are μ and σ^2)
- a Bernoulli distribution (parameter is π)
- a *linear model*

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{years of education} + \beta_2 \times \texttt{seniority}$$

What are the parameters of a linear model?

$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{years of education} + \beta_2 \times \texttt{seniority}$$

What does this model say about the structure of f?

Parametric Models

$$\text{income} \approx \beta_0 + \beta_1 \times \text{years of education} + \beta_2 \times \text{seniority}$$
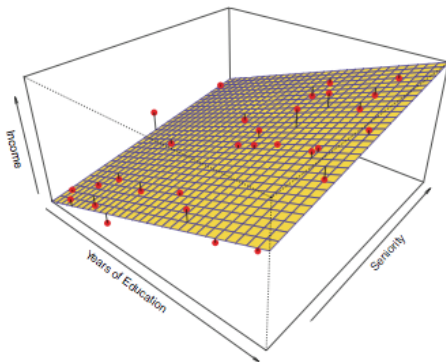
Is this model good for <u>prediction</u>? What can it tell us for inference?

Parametric Models

Parametric models:

- few parameters (when is this good?)
- well-described interactions between inputs, parameters and output (when is this useful?)
- limited flexibility (desirable or undesirable?)

Nonparametric Models

Nonparametric models are not described by a finite number of parameters.

So, what does that mean?

Let's learn about an example: $k-$nearest neighbors (kNN)

$k-$Nearest Neighbors

Idea: average the values of the k closest observations

$$\hat{Y} := \hat{f}(\boldsymbol{x}) = \frac{1}{k} \sum_{\boldsymbol{x}_i \in N_k(\boldsymbol{x})} y_i$$

where $N_k(x)$ is the set of observations with the k smallest distances to the query point x

- *Classification:* pick the majority label
- *Regression:* average the values

Why is this a nonparametric model? Does it have parameters?

Is It a Good Day to Go for a Run?

I have data on my past running. I recorded the temperature and whether the run was fun:

- ▶ Temperature (degrees F)
- ▶ Wind Speed (mph)
- ▶ Fun (yes, no)

It is now 65 degrees and the wind is 9 mph. Will my run be fun?

k-Nearest Neighbors: Classification
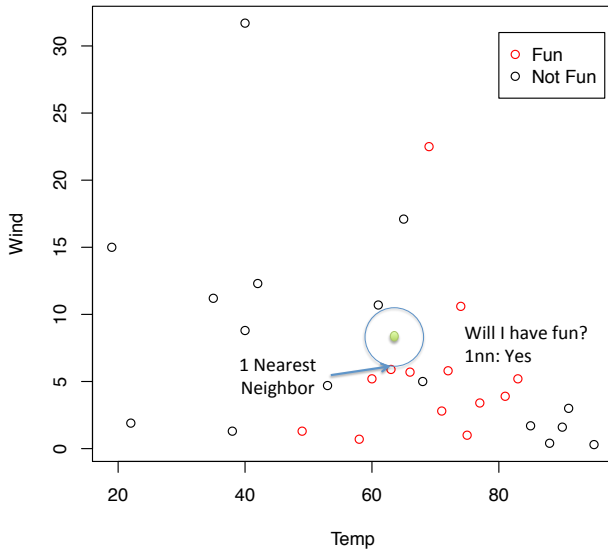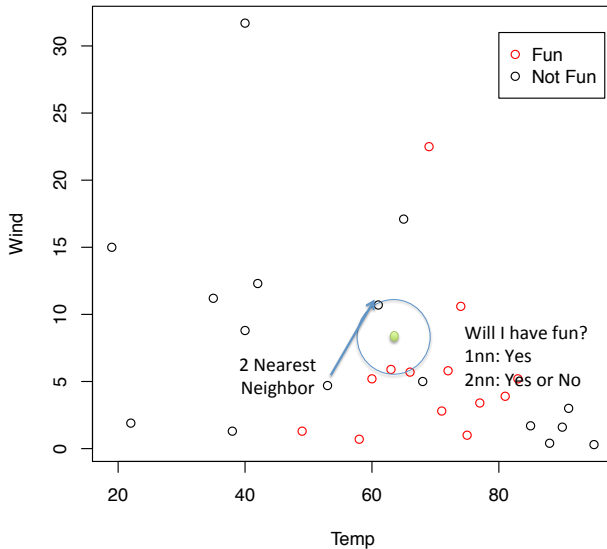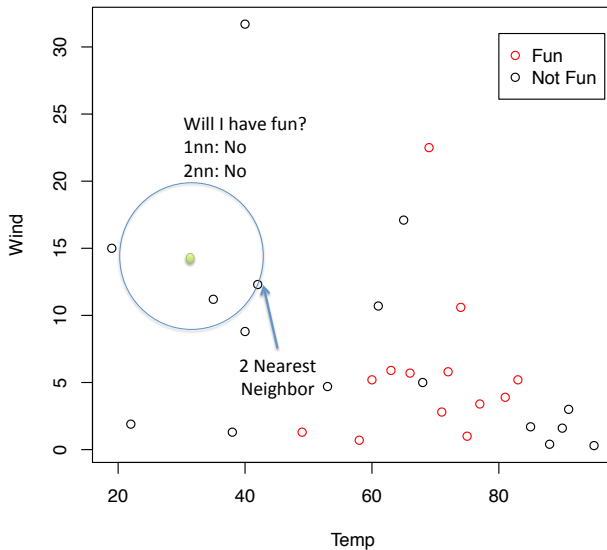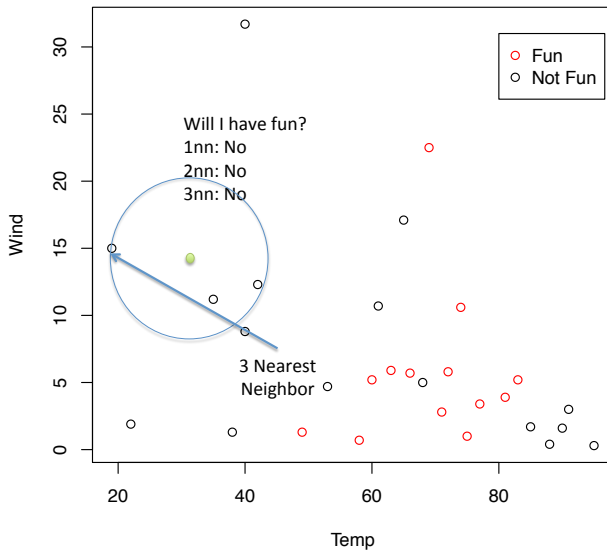
k-Nearest Neighbors: Classification

k-Nearest Neighbors: Classification
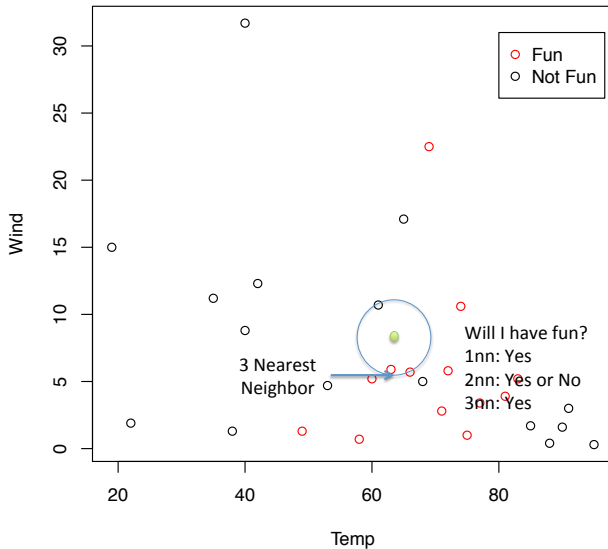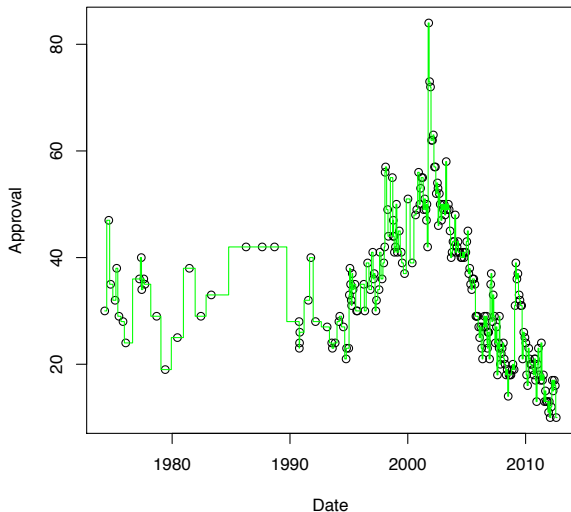
k-Nearest Neighbors: Classification
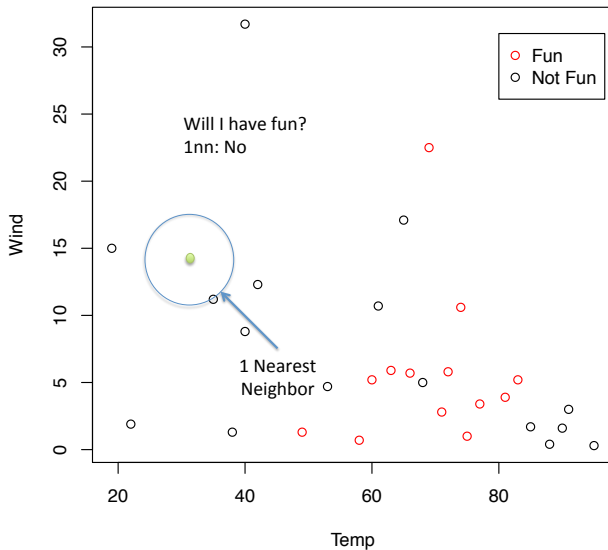
k-Nearest Neighbors: Classification

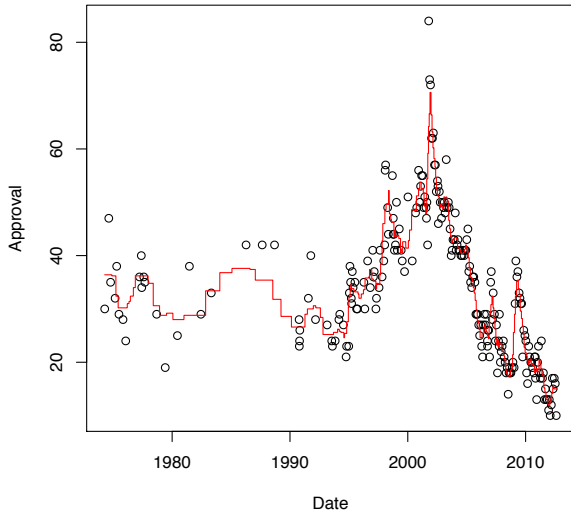# k-Nearest Neighbors: Classification

k-Nearest Neighbors: Regression

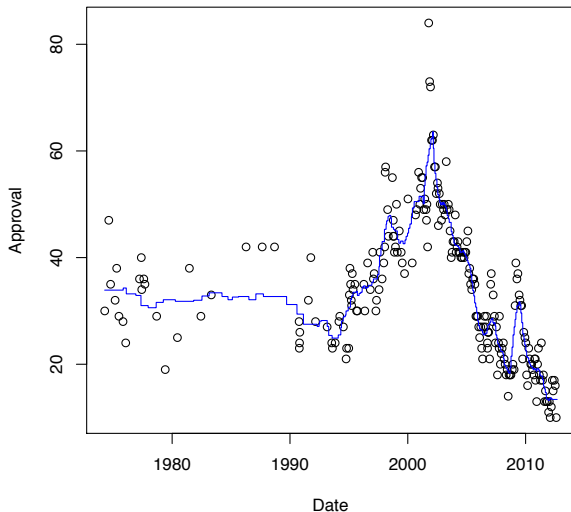1-Nearest Neighbor: Congressional Approval

k-Nearest Neighbors: Regression

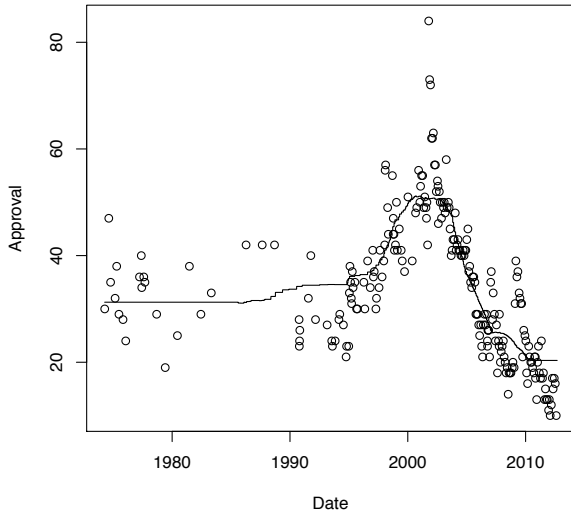5-Nearest Neighbors: Congressional Approval

k-Nearest Neighbors: Regression
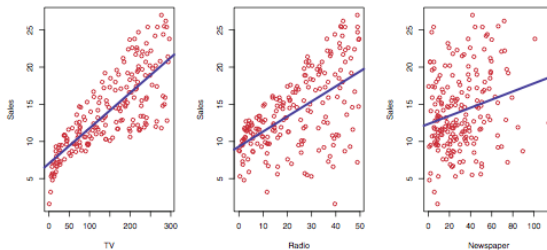
10-Nearest Neighbors: Congressional Approval

k-Nearest Neighbors: Regression

50-Nearest Neighbors: Congressional Approval

Nonparametric vs Parametric Models

Let's go back to `Advertising`. Suppose that we fit it with a linear model and kNN.

- ▶ which model will produce a more accurate prediction? with a lot of data? with a little data?
- ▶ which model will tell us about the which media will produce the best return?
- ▶ which model will be faster to evaluate when we have a lot of data (n is large)?

Interpretability

Nonparametric models are generally more *flexible* than parametric models. Why would we ever want a more restrictive model?

- prevent overfitting
- data compression
- **interpretability:** model parameters often mean something

Often, more interpretable models are less flexible and vice versa.

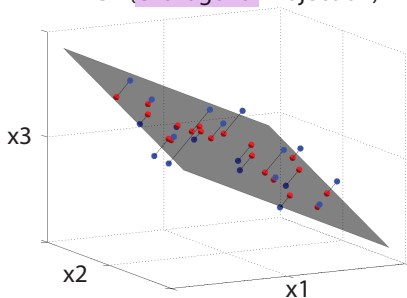# Big Picture Concepts

- Inference vs. Prediction

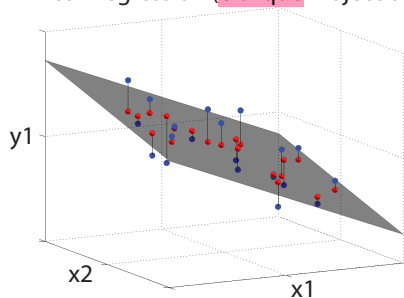- Parametric vs. Nonparametric

- Supervised vs. Unsupervised

Big Picture: Supervised vs Unsupervised

- Unsupervised learning seeks explanatory factors
- Supervised learning asserts explanatory factors

PCA (Orthogonal Projection)

Linear Regression (Oblique Projection)

Next Time

We will talk more about:

- classification
- decomposing predictive error
- evaluating predictive error