Big Data and Security

Jeffrey Borowitz, PhD

Lecturer

Sam Nunn School of International Affairs

Local Regression

k Nearest Neighbors

- One way to avoid making an assumption while predicting:
  - For each unit you want to predict, find the most similar unit in your data
  - Use the outcome for that unit
- Often this is generalized to "*k*-nearest neighbor", so you pick the nearest few and average
- This gives you a function $\hat{y}(x)$, even if it doesn't have a nice functional form.

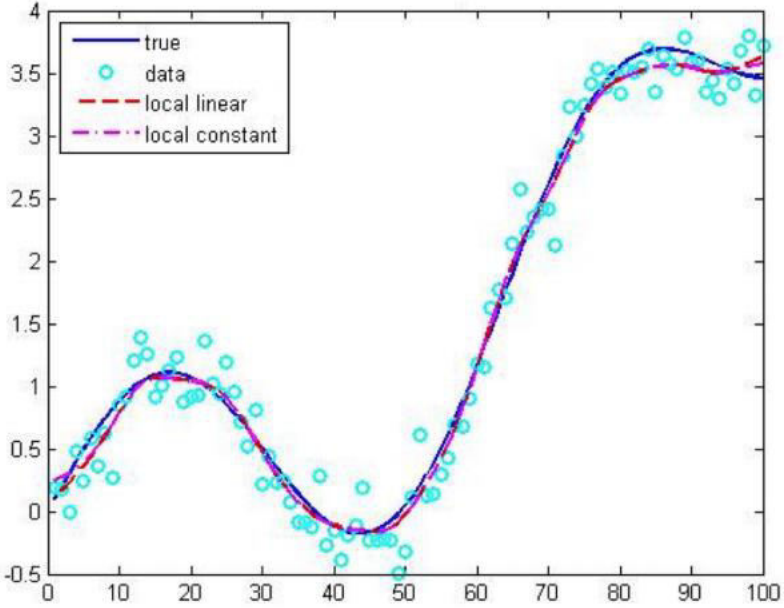
$$\hat{y} = \frac{1}{k}\sum_{1}^{k} y_i$$

Georgia Tech

A Simple Generalization: Local Regression

- Instead of just taking *k* neighbors, weight all points by their closeness
    - This is called local linear regression
    - Instead of just averaging, you can do a linear regression, hence the name

Local Regression Examples



Assumptions and Local Regression

- This is great!
  - We can fit curvy shapes
  - And we don't have to assume the function is linear, or the residuals are normal or anything
- What are some downsides?
  - Compared to using a parametric model, you often have less precision for predictions and results
  - Often this takes more data than we have, due to the curse of dimensionality
  - If you have a specific theory, a local regression might not be as easily interpretable
- What assumptions do we make?
  - Comparing data points is an assumption
  - Choosing X variables is an assumption

An Aside on the Great Recession

- Subprime mortgages were repackaged (in a generally reasonable way)
- Ratings agencies (Moodys, S&P) use statistical models to determine how likely groups of mortgages are to default.
- Where do they get data?
  - Historical mortgage payment rates, for borrowers of a particular quality
- The key assumption: tomorrow's mortgage payments will be drawn from the same distribution as the historical rates.
  - This assumption turned out to be very wrong! historical payment rates were not useful guides to future rates because house prices only increased in historical data sets.
  - This is the assumption of choosing particular data points to compare - you can never get rid of this one!

Georgia Tech

The Curse of Dimensionality

- Example: try to predict what type of computer an individual will buy, based on site history (i.e. what computer bought in the past), and survey data
- Let’s say you have a survey where people say “strongly disagree”, “disagree”, “neutral”, “agree”, “strongly agree”
  - Let’s say you have “all the data” on these questions (one from every person in the world, or 6 billion)
  - Now let’s say these data are independent
  - If you had 14 questions, you would have on average **1 person with each possible combination of responses**! (5^{14})
- Generally, more dimensions use more data exponentially and you can never have enough to do fully non-parametric stuff with lots of dimensions
  - This is the Curse of Dimensionality

Big Data and Nonparametric Statistics

- But what about “big data”? We have **all** the data! Can’t that help?
- All the data might not be enough. . .
  - Every person in the world can answer your 14 questions and it won’t help
- The Curse of Dimensionality trumps “Big Data”
- Remember, we’re trying to learn about a **random variable**, so our population isn’t necessarily represented by our data

Nonparametric Flavored Statistics

- There are a bunch of things that can be done that are nonparametric as you get more data
- The idea:
  - Math!
  - Any function can be approximated everywhere with an infinite series.
  - The intuition: if you get more terms, you get more ability to wiggle your line around
  - So as you get more data, use more terms
  - But just don’t use more terms than the data can support
  - In principle, as you get infinite data points, you would have infinite flexibility, but for now you have as much flexibility as you can have.
- Since you can’t actually have infinite data, your estimates might be wrong
- These are called sieves

Georgia Tech

How To Think About Nonparametric Statistics

- One element of this is, are there parameters, α or β, or something like this, which go into a model of the data?
- The broader conceptual way to think about it:
  - We don't know what's going on
  - So let's let the data tell us what's happening
- This means the data should be able to give us any answer, right?
  - In the case of least squares, we were only ever going to get straight lines
  - In the case of local linear regression or nearest neighbors, we could have any possible shape
  - In the case of sieves, we would eventually have any possible shape

Georgia Tech

Lesson Summary

- “*k*-nearest neighbors” and local linear regression help avoid making an assumption about our data
    - Instead, you can use local linear regression
    - Local linear regression weighs all points by their closeness
- The Curse of Dimensionality is when what you need to estimate grows faster than your data size. It means we can never really have "enough" data.