ORIE 4741: Learning with Big Messy Data

Review: through Linear Models

Professor Udell

Operations Research and Information Engineering
Cornell

October 16, 2021

Review

- learning
- big
- messy
- data

Review: data

- first of all: look at it!
- are there missing values?
- decide what you want to learn or predict
- input space $\mathcal{X}$, output space $\mathcal{Y}$
  - real, boolean, nominal, ordinal, text, ...

Review: messy

- probabilistic model: $(x, y) \sim P(x, y)$
- deterministic model: $y = f(x)$
- additive noisy model: $y = f(x) + \varepsilon$
  - additive noise model makes no sense for non-real data types (boolean, ordinal, nominal)
- feature engineering
  - can convert other data to real valued features
  - enables easy fitting of complex nonlinear models

Review: learning

- view data as samples from $P(x, y)$
- goal is to learn $f : \mathcal{X} \to \mathcal{Y}$
- how?
  - using an iterative procedure, like the **perceptron** method
  - by minimizing some **loss function**, like **least squares**
- complex models fit both data and noise better
- underdetermined problems give uninterpretable results
- generalization: how do we know if we're overfitting?
  - bootstrap: how big are the error bars?
  - crossvalidate: how big are the out-of-sample errors?
  - compute error on test set + use Hoeffding bound
  - posit a probabilistic model + use bias variance tradeoff
  - improve generalization with regularization

Review: big

- algorithms for big data should be **linear** in the number of samples n
- three big data algorithms for least squares:
  - gradient descent ($O(nd)$ per iteration)
  - QR ($O(nd^2)$)
  - SVD ($O(nd^2)$) (mostly used as analysis tool)

Studying for the exam

go through your notes (or the lecture slides).

for each technique we've learned,

- why would you use it?
- when would you use it?
- how would you use it?

Studying for the exam

go through your notes (or the lecture slides).

for each technique we've learned,

- why would you use it?
- when would you use it?
- how would you use it?

- look at the sample questions (released tonight)
- go to a review session Friday or Monday