ORIE 4741: Learning with Big Messy Data

Review: through Bootstrap and Bias Variance Tradeoff

Professor Udell

Operations Research and Information Engineering
Cornell

October 16, 2021

Announcements

- midterm Tuesday in class
- hw 4 out, due in two weeks (10-31-17)
- OH schedule changed; look at website
- I will not hold OH Friday or Tuesday

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
  - can encourage smoothness or locality

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?
  - cross validate: how big are the out-of-sample errors?
  - bootstrap: what is the variance of the estimate?
  - compute error on test set + use Hoeffding bound
  - posit a probabilistic model + use bias variance tradeoff
  - improve generalization with regularization

Generalization and Overfitting

- goal of model is **not** to predict well on $\mathcal{D}$
- goal of model is to predict well **on new data**

if the model has ____ training set error and ____ test set error, we say the model:

	low test set error	high test set error
low training set error	generalizes	overfits
high training set error	?!?!	underfits

Generalization and Overfitting

- goal of model is **not** to predict well on $\mathcal{D}$
- goal of model is to predict well **on new data**

if the model has ____ training set error and ____ test set error, we say the model:

	low test set error	high test set error
low training set error	generalizes	overfits
high training set error	?!?!	underfits

Q: How to fix underfitting?

Generalization and Overfitting

- goal of model is **not** to predict well on $\mathcal{D}$
- goal of model is to predict well **on new data**

if the model has ____ training set error and ____ test set error, we say the model:

	low test set error	high test set error
low training set error	generalizes	overfits
high training set error	?!?!	underfits

Q: How to fix underfitting?

A: Use more complex model, or add new features.

Generalization and Overfitting

- goal of model is **not** to predict well on $\mathcal{D}$
- goal of model is to predict well **on new data**

if the model has ____ training set error and ____ test set error, we say the model:

	low test set error	high test set error
low training set error	generalizes	overfits
high training set error	?!?!	underfits

Q: How to fix underfitting?

A: Use more complex model, or add new features.

Q: How to fix overfitting?

Generalization and Overfitting

- goal of model is **not** to predict well on $\mathcal{D}$
- goal of model is to predict well **on new data**

if the model has ____ training set error and ____ test set error, we say the model:

	low test set error	high test set error
low training set error	generalizes	overfits
high training set error	?!?!	underfits

Q: How to fix underfitting?

A: Use more complex model, or add new features.

Q: How to fix overfitting?

A: Use less complex model, remove features, regularize, or find more data.

Review: big

algorithms for big data should be **linear** in the number of samples n

	GD	SGM	Gram GD	Parallel GD	QR or SVD
initial	0	0	nd^2	nd^2/P	nd^2
per iter	nd	$\lvert S\rvert d$	d^2	d^2	0

(numbers in flops, omitting constants)

- gradient descent (most flexible, $O(nd)$ flops per iteration)
- QR factorization (most efficient exact solution method, $O(nd^2)$ flops)
- SVD factorization (exact solution method, works for underdetermined problems, $O(nd^2)$ flops)

Review: big

algorithms for big data should be **linear** in the number of samples n

	GD	SGM	Gram GD	Parallel GD	QR or SVD
initial	0	0	nd^2	nd^2/P	nd^2
per iter	nd	$\lvert S\rvert d$	d^2	d^2	0

(numbers in flops, omitting constants)

- gradient descent (most flexible, $O(nd)$ flops per iteration)
- QR factorization (most efficient exact solution method, $O(nd^2)$ flops)
- SVD factorization (exact solution method, works for underdetermined problems, $O(nd^2)$ flops)

how to compute flop counts:

- count the number of inner products

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

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
- we'll release solutions to the practice exam on Monday
- go to section on Monday

Questions?

Questions?

- I think we should go over the math behind bootstrapping again.

Questions?

- I think we should go over the math behind bootstrapping again.
- What are some reasons that you would accept some bias in favor of smaller variance.

Questions?

- I think we should go over the math behind bootstrapping again.
- What are some reasons that you would accept some bias in favor of smaller variance.
- Besides linear regression, what are some other unbiased estimators?

Questions?

- I think we should go over the math behind bootstrapping again.
- What are some reasons that you would accept some bias in favor of smaller variance.
- Besides linear regression, what are some other unbiased estimators?
- Examples of reducing bias / variance