S&DS 265 / 565

Introductory Machine Learning

Stochastic Gradient Descent and Bias-Variance Tradeoffs

September 27

Yale

Goings on

- Assignment 1 due Thursday
- Assignment 2 posted Thursday
- Quiz 2 scores out: Average 84%, great!
- Quiz 3 posted next week
- Midterm: October 18 (in class)

Readings

					classification	
4	Sept 20, 22	Stochastic gradient descent	SGD examples	Sept 20: Classification (continued) Sept 22: Stochastic gradient descent	ISL Section 6.2.2 ISL Section 10.7.2	Thu: Quiz 2
5	Sept 27, 29	Bias and variance, cross-validation	Bias-variance tradeoff Covid trends (revisited) California housing	Sept 27: Bias and variance Sept 29: Cross-validation	ISL Section 2.2 ISL Section 5.1	Thu: Assn 1 in Assn2 out
6	Oct 4, 6	Tree-based methods	Trees and forests Visualizing trees	Oct 4: Trees Oct 6: Forests	ISL Sections 8.1, 8.2	Thu: Quiz 3

Readings

Winner of the 2014 Eric Ziegel award from Technometrics.

As the scale and scope of data collection continue to increase across virtually all fields, statistical learning has become a critical toolkit for anyone who wishes to understand data. *An Introduction to Statistical Learning* provides a broad and less technical treatment of key topics in statistical learning. Each chapter includes an R lab. This book is appropriate for anyone who wishes to use contemporary tools for data analysis.

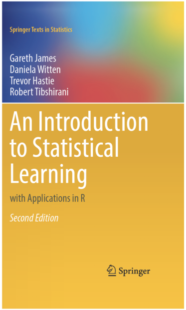
The book has been translated into Chinese, Italian, Japanese, Korean, Mongolian, Russian and Vietnamese.

The First Edition topics include:

- Sparse methods for classification and regression
- Decision trees
- Boosting
- Support vector machines
- Clustering

The Second Edition adds:

- Deep learning
- Survival analysis
- Multiple testing
- Naive Bayes and generalized linear models
- Bayesian additive regression trees
- Matrix completion



statlearning.com

Outline for today

- Stochastic gradient descent (redux)
- Regularization
- Jupyter notebook example
- Bias-variance tradeoffs

SGD idea

- For each parameter β_j, see what happens to the loss if that parameter is increased a little bit.
- If the loss goes down (up), then increase (decrease) β_j proportionately
- Do this simultaneously for all of the parameters
- Rinse and repeat

Stochastic gradient descent

Initialize all parameters to zero: $\beta_j = 0$, $j = 1, \ldots, p$.

Read through the data one record at a time, and update the model.

1. Read data item x
2. Make a prediction $\widehat{y}(x)$
3. Observe the true response/label y
4. Update the parameters β so $\widehat{y}$ is closer to y

Stochastic gradient descent

Suppose we are doing *linear regression*. We initialize all parameters to zero: $\beta_j = 0$, $j = 1, \dots, p$.

We read through the data one record at a time, and update the model.

1. Read data item x
2. Make a prediction $\widehat{y}(x) = \sum_{j=1}^{p} \beta_j x_j$
3. Observe the true response/label y
4. Update the parameters β so $\widehat{y}$ is closer to y

SGD: Computation for squared error

Change β_j by a little bit:

$$\beta_j \to \beta_j + \varepsilon$$

SGD: Computation for squared error

Change β_j by a little bit:

$$\beta_j \to \beta_j + \varepsilon$$

What happens to the squared error?

$$\begin{aligned}(y - \widehat{y})^2 &\to (y - \widehat{y} - \varepsilon x_j)^2 \\ &\approx (y - \widehat{y})^2 + \underbrace{-2(y - \widehat{y})x_j}_{\textit{derivative of loss}} \varepsilon\end{aligned}$$

SGD: Computation for squared error

Change β_j by a little bit:

$$\beta_j \to \beta_j + \varepsilon$$

What happens to the squared error?

$$\begin{aligned}(y - \widehat{y})^2 &\to (y - \widehat{y} - \varepsilon x_j)^2 \\ &\approx (y - \widehat{y})^2 + \underbrace{-2(y - \widehat{y})x_j}_{\textit{derivative of loss}} \varepsilon\end{aligned}$$

Use adjustment

$$\begin{aligned}\beta_j &\to \beta_j \overbrace{-\eta \cdot \text{derivative of loss}}^{\varepsilon} \\ &= \beta_j + \eta \cdot 2(y - \widehat{y})x_j\end{aligned}$$

SGD: Computation for squared error

Change β_j by a little bit:

$$\beta_j \to \beta_j + \varepsilon$$

What happens to the squared error?

$$\begin{aligned}(y - \widehat{y})^2 &\to (y - \widehat{y} - \varepsilon x_j)^2 \\ &\approx (y - \widehat{y})^2 + \underbrace{-2(y - \widehat{y})x_j}_{\textit{derivative of loss}} \varepsilon\end{aligned}$$

Use adjustment

$$\begin{aligned}\beta_j &\to \beta_j \overbrace{-\eta \cdot \text{derivative of loss}}^{\varepsilon} \\ &= \beta_j + \eta \cdot 2(y - \widehat{y})x_j\end{aligned}$$

Squared error then decreases:

$$(y - \widehat{y})^2 \approx (y - \widehat{y})^2 - \eta \cdot \text{derivative of loss } \underline{\text{squared}}$$

SGD for general loss

Suppose $L(y, \beta^T x)$ is the loss for an input (x, y), e.g., $(y - \beta^T x)^2$

SGD update, for a small step size $\eta > 0$:

$$\beta_j \longleftarrow \beta_j - \eta \frac{\partial L(y, \boldsymbol{\beta}^T x)}{\partial \beta_j}$$

$$\boldsymbol{\beta} \longleftarrow \boldsymbol{\beta} - \eta \nabla_{\beta} L(y, \boldsymbol{\beta}^T x) \quad \text{(vector notation)}$$

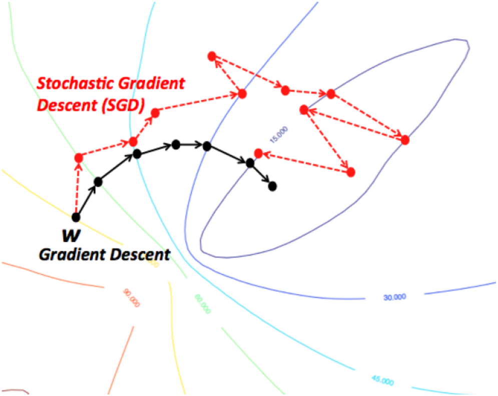
"Batch" gradient descent uses the entire training set in each step of gradient descent.

Stochastic gradient descent computes a quick approximation to this gradient, using only a single or a small "mini-batch" of data points

Batch vs. stochastic gradient descent

- The average derivative over a mini-batch can be thought of as a noisy version of the average derivative over the entire data set
- (Which can in turn be thought of as a sample estimate of a population)
- The stochastic gradient is computed more cheaply, and updating the parameters makes progress more quickly

Batch vs. stochastic gradient descent



https://wikidocs.net/3413

SGD for logistic regression

SGD Update:

$$\beta_j \longleftarrow \beta_j + \eta(y - p(x))x_j$$

$$\beta_j x_j \longleftarrow \beta_j x_j + \eta(y - p(x))x_j^2$$

$$p(x) = \frac{1}{1 + \exp(-\beta^T x)}$$

Case checking:

- Suppose $y = 1$ and probability $p(x)$ is high?

SGD for logistic regression

SGD Update:

$$\beta_j \longleftarrow \beta_j + \eta(y - p(x))x_j$$

$$\beta_j x_j \longleftarrow \beta_j x_j + \eta(y - p(x))x_j^2$$

$$p(x) = \frac{1}{1 + \exp(-\beta^T x)}$$

Case checking:

- Suppose $y = 1$ and probability $p(x)$ is high? *small change*
- Suppose $y = 1$ and probability $p(x)$ is small?

SGD for logistic regression

SGD Update:

$$\beta_j \longleftarrow \beta_j + \eta(y - p(x))x_j$$

$$\beta_j x_j \longleftarrow \beta_j x_j + \eta(y - p(x))x_j^2$$

$$p(x) = \frac{1}{1 + \exp(-\beta^T x)}$$

Case checking:

- Suppose $y = 1$ and probability $p(x)$ is high? *small change*
- Suppose $y = 1$ and probability $p(x)$ is small? *big change* $\uparrow$
- Suppose $y = 0$ and probability $p(x)$ is small?

SGD for logistic regression

SGD Update:

$$\beta_j \longleftarrow \beta_j + \eta(y - p(x))x_j$$

$$\beta_j x_j \longleftarrow \beta_j x_j + \eta(y - p(x))x_j^2$$

$$p(x) = \frac{1}{1 + \exp(-\boldsymbol{\beta}^T x)}$$

Case checking:

- Suppose $y = 1$ and probability $p(x)$ is high? *small change*
- Suppose $y = 1$ and probability $p(x)$ is small? *big change* $\uparrow$
- Suppose $y = 0$ and probability $p(x)$ is small? *small change*
- Suppose $y = 0$ and probability $p(x)$ is big?

SGD for logistic regression

SGD Update:

$$\beta_j \longleftarrow \beta_j + \eta(y - p(x))x_j$$

$$\beta_j x_j \longleftarrow \beta_j x_j + \eta(y - p(x))x_j^2$$

$$p(x) = \frac{1}{1 + \exp(-\boldsymbol{\beta}^T x)}$$

Case checking:

- Suppose $y = 1$ and probability $p(x)$ is high? *small change*
- Suppose $y = 1$ and probability $p(x)$ is small? *big change* $\uparrow$
- Suppose $y = 0$ and probability $p(x)$ is small? *small change*
- Suppose $y = 0$ and probability $p(x)$ is big? *big change* $\downarrow$

SGD: choice of learning rate

A conservative choice of learning rate is

$$\eta_t = \frac{1}{t}$$

A more agressive choice is

$$\eta_t = \frac{1}{\sqrt{t}}$$

In practice: Try learning rates $C/\sqrt{t}$ for different choices of C, and monitor the error

$$\frac{1}{T}\sum_{t=1}^{T}(Y_t - \widehat{Y}_t)^2$$

SGD: Regularization

A “ridge” penalty $\frac{1}{2}\lambda\sum_{j=1}^{d}\beta_j^2$ is easily handled.

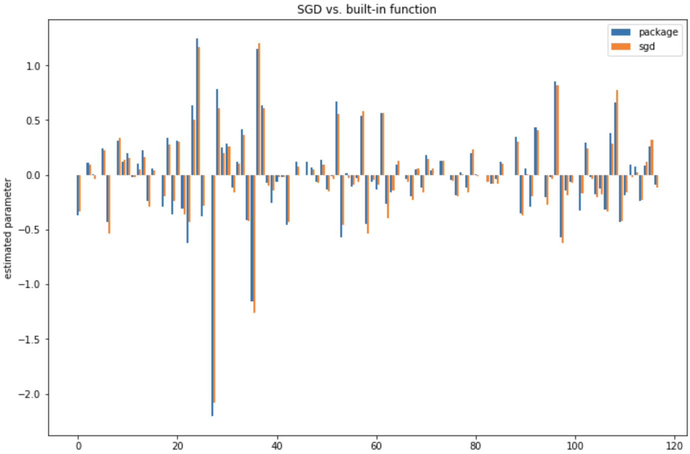
Gradient changes by an additive term $2\lambda\beta_j$. Update becomes

$$\begin{aligned}\beta_j &\longleftarrow \beta_j + \eta\{(y - p(x))x_j - \lambda\beta_j\} \\ &= (1-\eta\lambda)\beta_j + \eta(y - p(x))x_j\end{aligned}$$

Check that this “does the right thing” whether β_j wants to be large positive or negative.

- *The penalty shrinks β_j toward zero*

Recall from demo



Each bar indicates an estimated parameter $\widehat{\beta}_j$. The estimates from SGD are very similar to those obtained using the package.

Bias and variance

Bias: How much are we off—on average?

Variance: How variable are we—on average?

Bias and variance

Bias: $\theta - \mathbb{E}\widehat{\theta}$

Variance: $\mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2$

Bias and variance

Examples of $\theta, \widehat{\theta}$:

Estimating height, population, election outcome, ad click rate...

Bias and variance

Bias: $\theta - \mathbb{E}\widehat{\theta}$

Variance: $\mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2$

Bias and variance

$$\text{Bias: } \theta - \mathbb{E}\widehat{\theta}$$

$$\text{Variance: } \mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2$$

- $\widehat{\theta}$ is an estimate from a sample
- $\mathbb{E}$ is the expectation (average) with respect to the sample
- So $\mathbb{E}\widehat{\theta}$ is the average estimate
- We can only directly compute $\widehat{\theta}$ for the sample we have
- *We don't know θ*

Bias and variance

In machine learning, bias and variance are two sides of a coin:
As squared bias goes up, variance goes down (and vice-versa)

Bias-variance tradeoff

$$\text{Risk} = \text{Bias}^2 + \text{Variance}$$

Bias-variance tradeoff

$$\mathbb{E}(\theta - \widehat{\theta})^2 = \text{Bias}(\widehat{\theta})^2 + \text{Variance}(\theta)$$

Bias-variance tradeoff

$$\mathbb{E}(\theta - \widehat{\theta})^2 = (\theta - \mathbb{E}\widehat{\theta})^2 + \mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2$$

Bias-variance tradeoff

Proof:

$$\mathbb{E}(\theta - \widehat{\theta})^2 = \mathbb{E}(\theta - \mathbb{E}\widehat{\theta} + \mathbb{E}\widehat{\theta} - \widehat{\theta})^2$$

Bias-variance tradeoff

Proof:

$$
\begin{aligned}
\mathbb{E}(\theta - \widehat{\theta})^2 &= \mathbb{E}(\theta - \mathbb{E}\widehat{\theta} + \mathbb{E}\widehat{\theta} - \widehat{\theta})^2 \\
&= \mathbb{E}(\theta - \mathbb{E}\widehat{\theta})^2 - 2\mathbb{E}\left\{(\theta - \mathbb{E}\widehat{\theta})(\widehat{\theta} - \mathbb{E}\widehat{\theta})\right\} + \mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2
\end{aligned}
$$

Bias-variance tradeoff

Proof:

$$
\begin{aligned}
\mathbb{E}(\theta - \widehat{\theta})^2 &= \mathbb{E}(\theta - \mathbb{E}\widehat{\theta} + \mathbb{E}\widehat{\theta} - \widehat{\theta})^2 \\
&= \mathbb{E}(\theta - \mathbb{E}\widehat{\theta})^2 - 2\mathbb{E}\left\{(\theta - \mathbb{E}\widehat{\theta})(\widehat{\theta} - \mathbb{E}\widehat{\theta})\right\} + \mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2 \\
&= \mathbb{E}(\theta - \mathbb{E}\widehat{\theta})^2 - 2(\theta - \mathbb{E}\widehat{\theta})\mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta}) + \mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2
\end{aligned}
$$

Bias-variance tradeoff

Proof:

$$\begin{aligned}
\mathbb{E}(\theta - \widehat{\theta})^2 &= \mathbb{E}(\theta - \mathbb{E}\widehat{\theta} + \mathbb{E}\widehat{\theta} - \widehat{\theta})^2 \\
&= \mathbb{E}(\theta - \mathbb{E}\widehat{\theta})^2 - 2\mathbb{E}\left\{(\theta - \mathbb{E}\widehat{\theta})(\widehat{\theta} - \mathbb{E}\widehat{\theta})\right\} + \mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2 \\
&= \mathbb{E}(\theta - \mathbb{E}\widehat{\theta})^2 - 2(\theta - \mathbb{E}\widehat{\theta})\mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta}) + \mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2 \\
&= \mathbb{E}(\theta - \mathbb{E}\widehat{\theta})^2 + \mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2
\end{aligned}$$

Bias-variance tradeoff

Proof:

$$\begin{aligned}
\mathbb{E}(\theta - \widehat{\theta})^2 &= \mathbb{E}(\theta - \mathbb{E}\widehat{\theta} + \mathbb{E}\widehat{\theta} - \widehat{\theta})^2 \\
&= \mathbb{E}(\theta - \mathbb{E}\widehat{\theta})^2 - 2\mathbb{E}\left\{(\theta - \mathbb{E}\widehat{\theta})(\widehat{\theta} - \mathbb{E}\widehat{\theta})\right\} + \mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2 \\
&= \mathbb{E}(\theta - \mathbb{E}\widehat{\theta})^2 - 2(\theta - \mathbb{E}\widehat{\theta})\mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta}) + \mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2 \\
&= \mathbb{E}(\theta - \mathbb{E}\widehat{\theta})^2 + \mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2 \\
&= \text{Bias}(\widehat{\theta})^2 + \text{Variance}(\widehat{\theta})
\end{aligned}$$

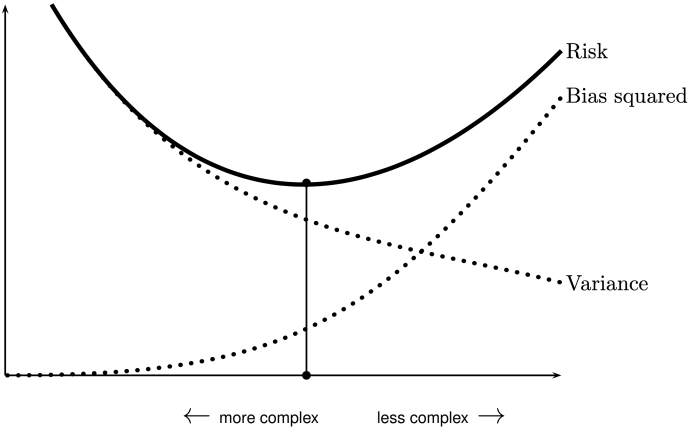
Bias-variance tradeoff

If $Y = \theta +$ noise, with $\mathbb{E}(\text{noise}) = 0$ and $Var(\text{noise}) = \sigma^2$,

$$\text{Risk} = \mathbb{E}[(Y - \widehat{\theta})^2] = \text{Bias}^2 + \text{Variance} + \sigma^2$$

Please work through the proof of this!

Bias-variance tradeoff



Example: Regularization

Suppose that $\mathbb{E}(Y) = \theta^*$ and we estimate

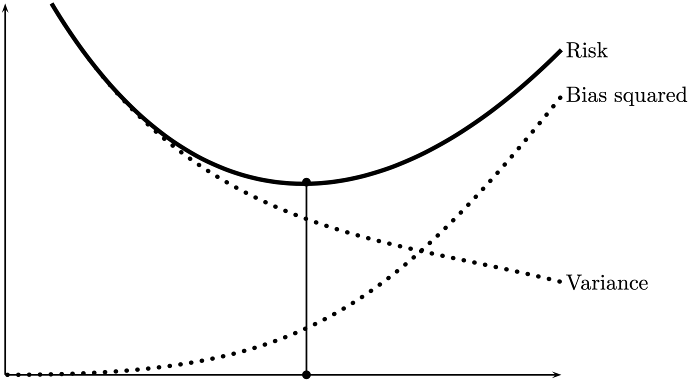
$$\widehat{\theta} = \arg\min_{\theta} (Y - \theta)^2 + \lambda\theta^2$$

Then $\widehat{\theta} = \frac{Y}{1+\lambda}$. What are the squared bias and variance?

$$\text{Bias}^2 = \theta^{*2} \left(\frac{\lambda}{1+\lambda}\right)^2$$

$$\text{Variance} = \left(\frac{1}{1+\lambda}\right)^2 \text{Variance}(Y)$$

Bias-variance tradeoff

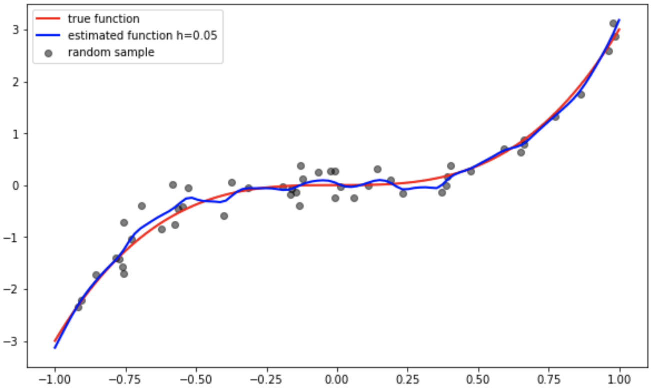


← less regularization more regularization →

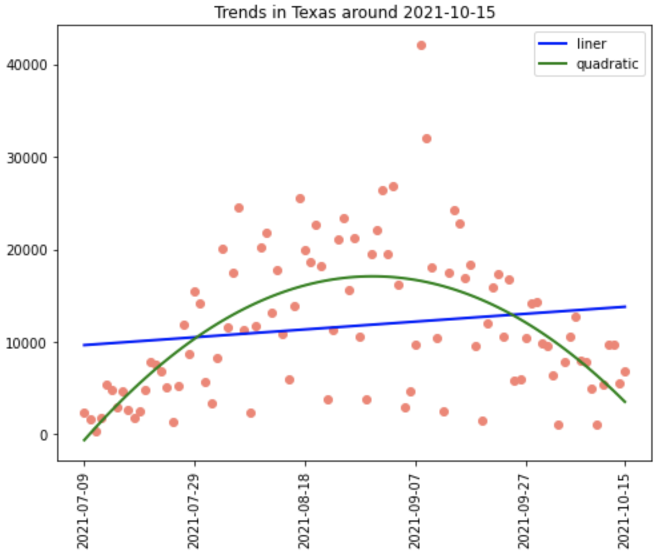
← less shrinkage more shrinkage →

← smaller penalty bigger penalty →

Let's go to the first notebook



Let’s go to the second notebook



What did we learn today?

- In SGD, a parameter is updated according to how much the loss changes when that parameter is changed by a little bit
- Mean squared error splits into squared bias plus variance
- As model complexity increases, squared bias decreases while variance increases