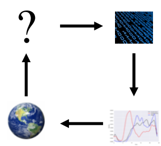
The Bias Variance Tradeoff and Regularization

Slides by:

Joseph E. Gonzalez jegonzal@cs.berkeley.edu

Spring'18 updates:

Fernando Perez fernando.perez@berkeley.edu



Quick announcements

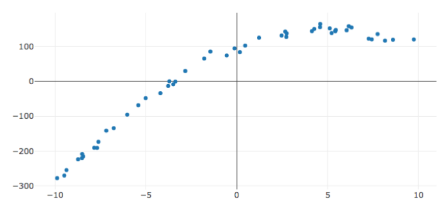
- Please **be respectful** on Piazza
  - Both of your fellow students and of your teaching staff.
  - The teaching team monitors Piazza, but you can report any incidents directly to Profs. Gonzalez and/or Perez.
- Our infrastructure isn't perfect
  - We're working hard on improving it.
  - We're building the plane while we fly it, full of passengers.
- We have a textbook: textbook.ds100.org
  - It's a **work in progress**!

Linear models for non-linear relationships

Advice for people who are dealing with non-linear relationship issues but would really prefer the simplicity of a linear relationship.

Is this data Linear?

What does it mean to be linear?



What does it mean to be a linear model?

$$f_\theta\left(\phi(x)\right) = \phi(x)^T\theta = \sum_{j=1}^{k} \phi(x)_j \theta_j$$

In what sense is the above **model linear**?

Are linear models linear in the

1. the features?
2. the parameters?

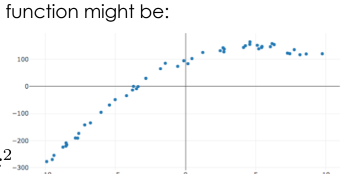
Introducing Non-linear Feature Functions

- One reasonable feature function might be:

$$\phi(x) = [1, x, x^2]$$

- That is:

$$f_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

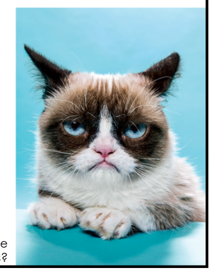


- This is **still a linear model**, in the parameters θ

What are the fundamental challenges in learning?

Fundamental Challenges in Learning?

- ***Fit* the Data**
  - Provide an explanation for what we observe
- ***Generalize* to the World**
  - Predict the future
  - Explain the unobserved



Is this cat grumpy or are we overfitting to human faces?

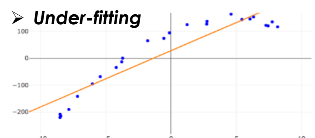
Fundamental Challenges in Learning?

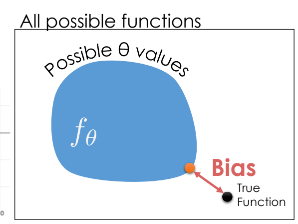
- **Bias:** the expected deviation between the predicted value and the true value
- **Variance:** two sources
  - **Observation Variance:** the variability of the random noise in the process we are trying to model.
  - **Estimated Model Variance:** the variability in the predicted value across different training datasets.

Bias

The expected deviation between the predicted value and the true value

- Depends on both the:
  - **choice** of f
  - **learning procedure**
- ***Under-fitting***



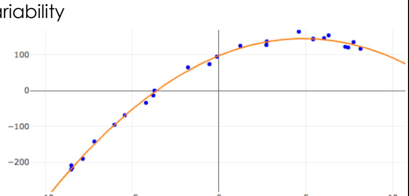


Observation Variance

the variability of the random noise in the process we are trying to model

- measurement variability
- stochasticity
- missing informati

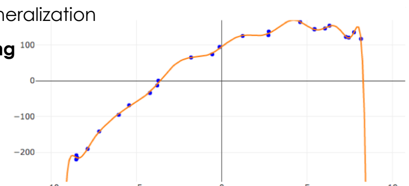
Beyond our control (usually)



Estimated Model Variance

variability in the predicted value across different training datasets

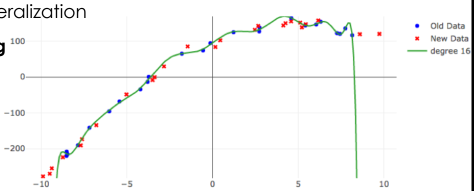
- Sensitivity to variation in the training data
- Poor generalization
- **Overfitting**



Estimated Model Variance

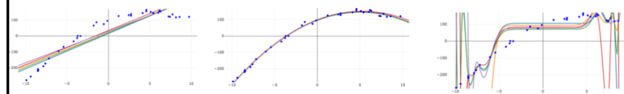
variability in the predicted value across different training datasets

- Sensitivity to variation in the training data
- Poor generalization
- **Overfitting**



The Bias-Variance Tradeoff

Estimated Model Variance →



← Bias

Demo

Analysis of the Bias-Variance Trade-off

Analysis of Squared Error

- For the test point $\boldsymbol{x}$ the expected error:
  - *Random variables are* ***red***

Noise term:
$\mathbf{E}[\epsilon] = 0$
$\mathbf{Var}[\epsilon] = \sigma^2$

True Function

Assume noisy observations → y is a random variable

$y = h(x) + \epsilon$

$$\mathbf{E}\left[\left(y - f_{\hat{\theta}}(x)\right)^2\right]$$

Assume ***training data*** is random → θ is a random variable

Analysis of Squared Error

Goal:

$$\mathbf{E}\left[\left(y - f_{\hat{\theta}}(x)\right)^2\right] =$$

Obs. Var. + **(Bias)**^{2} + **Mod. Var.**

Other terminology:

"Noise" + **(Bias)**^{2} + **Variance**

$$\mathbf{E}\left[\left(y - f_{\hat\theta}(x)\right)^2\right] = \mathbf{E}\left[\left(\underbrace{y - h(x) + h(x)}_{\text{Subtracting and adding } h(x)} - f_{\hat\theta}(x)\right)^2\right]$$

Useful Eqns:
$y = h(x) + \epsilon$
$\mathbf{E}[\epsilon] = 0$
$\mathbf{Var}[\epsilon] = \sigma^2$

$$\mathbf{E}\left[\left(y - f_{\hat\theta}(x)\right)^2\right] = \mathbf{E}\left[\left(\underbrace{y - h(x)}_{a} + \underbrace{h(x) - f_{\hat\theta}(x)}_{b}\right)^2\right]$$

Expanding in terms of *a* and *b*: $(a+b)^2 = a^2 + b^2 + 2ab$

$$= \overbrace{\mathbf{E}\left[\left(y - h(x)\right)^2\right]}^{a^2} + \overbrace{\mathbf{E}\left[\left(h(x) - f_{\hat\theta}(x)\right)^2\right]}^{b^2}$$

$$+2\mathbf{E}\left[\left(y - h(x)\right)\left(h(x) - f_{\hat\theta}(x)\right)\right] \quad 2ab$$

$y = h(x) + \epsilon$ ⟹ ϵ

$$+2\mathbf{E}\left[\epsilon\left(h(x) - f_{\hat\theta}(x)\right)\right]$$

Useful Eqns:
$y = h(x) + \epsilon$
$\mathbf{E}[\epsilon] = 0$
$\mathbf{Var}[\epsilon] = \sigma^2$

$$\mathbf{E}\left[\left(y - f_{\hat\theta}(x)\right)^2\right] = \mathbf{E}\left[\left(y - h(x) + h(x) - f_{\hat\theta}(x)\right)^2\right]$$

Expanding in terms of *a* and *b*:

$$= \mathbf{E}\left[\left(y - h(x)\right)^2\right] + \mathbf{E}\left[\left(h(x) - f_{\hat\theta}(x)\right)^2\right]$$

$$+2\mathbf{E}\left[\epsilon\left(h(x) - f_{\hat\theta}(x)\right)\right]$$

Independence of ϵ and θ

$$+2\underbrace{\mathbf{E}[\epsilon]}_{0}\mathbf{E}\left[\left(h(x) - f_{\hat\theta}(x)\right)\right]$$

Useful Eqns:
$y = h(x) + \epsilon$
$\mathbf{E}[\epsilon] = 0$
$\mathbf{Var}[\epsilon] = \sigma^2$

$$\mathbf{E}\left[\left(y - f_{\theta}(x)\right)^2\right] =$$

$$\mathbf{E}\left[\left(y - h(x)\right)^2\right] + \quad = \sigma^2$$

Obs. Value — True Value

Obs. Variance
"Noise" *Term*

$$\mathbf{E}\left[\left(h(x) - f_{\hat\theta}(x)\right)^2\right]$$

True Value — Pred. Value

Model Estimation Error

Useful Eqns:
$y = h(x) + \epsilon$
$\mathbf{E}[\epsilon] = 0$
$\mathbf{Var}[\epsilon] = \sigma^2$

$$\mathbf{E}\left[\left(h(x) - f_{\hat\theta}(x)\right)^2\right] = \text{Next we will show....}$$

$$\left(h(x) - \mathbf{E}\left[f_{\hat\theta}(x)\right]\right)^2 + \mathbf{E}\left[\left(\mathbf{E}\left[f_{\hat\theta}(x)\right] - f_{\hat\theta}(x)\right)^2\right]$$

(Bias)² **Model Variance**

- How?
- Adding and Subtracting what?

$$\mathbf{E}\left[\left(h(x) - f_{\hat\theta}(x)\right)^2\right] =$$

$$\mathbf{E}\left[\left(\underbrace{h(x) - \mathbf{E}\left[f_{\hat\theta}(x)\right]}_{a} + \underbrace{\mathbf{E}\left[f_{\hat\theta}(x)\right] - f_{\hat\theta}(x)}_{b}\right)^2\right]$$

Expanding in terms of *a* and *b*: $(a+b)^2 = a^2 + b^2 + 2ab$

$$\overbrace{\mathbf{E}\left[\left(h(x) - \mathbf{E}\left[f_{\hat\theta}(x)\right]\right)^2\right]}^{a^2} + \overbrace{\mathbf{E}\left[\left(\mathbf{E}\left[f_{\hat\theta}(x)\right] - f_{\hat\theta}(x)\right)^2\right]}^{b^2}$$

$$\underbrace{+2\mathbf{E}\left[\left(h(x) - \mathbf{E}\left[f_{\hat\theta}(x)\right]\right)\left(\mathbf{E}\left[f_{\hat\theta}(x)\right] - f_{\hat\theta}(x)\right)\right]}_{2ab}$$

$$\mathbf{E}\left[\left(h(x) - f_{\hat{\theta}}(x)\right)^2\right] =$$

$$\mathbf{E}\left[\left(h(x) - \mathbf{E}\left[f_{\hat{\theta}}(x)\right]\right)^2\right] + \mathbf{E}\left[\left(\mathbf{E}\left[f_{\hat{\theta}}(x)\right] - f_{\hat{\theta}}(x)\right)^2\right]$$

$$+2\mathbf{E}\left[\underbrace{\left(h(x) - \mathbf{E}\left[f_{\hat{\theta}}(x)\right]\right)}_{\text{Constant}}\left(\mathbf{E}\left[f_{\hat{\theta}}(x)\right] - f_{\hat{\theta}}(x)\right)\right]$$

$$+2\left(h(x) - \mathbf{E}\left[f_{\hat{\theta}}(x)\right]\right)\mathbf{E}\left[\left(\mathbf{E}\left[f_{\hat{\theta}}(x)\right] - f_{\hat{\theta}}(x)\right)\right]$$

$$\mathbf{E}\left[\left(h(x) - f_{\hat{\theta}}(x)\right)^2\right] =$$

$$\mathbf{E}\left[\left(h(x) - \mathbf{E}\left[f_{\hat{\theta}}(x)\right]\right)^2\right] + \mathbf{E}\left[\left(\mathbf{E}\left[f_{\hat{\theta}}(x)\right] - f_{\hat{\theta}}(x)\right)^2\right]$$

$$+2\left(h(x) - \mathbf{E}\left[f_{\hat{\theta}}(x)\right]\right)\mathbf{E}\left[\left(\underbrace{\mathbf{E}\left[f_{\hat{\theta}}(x)\right]}_{\text{Constant}} - f_{\hat{\theta}}(x)\right)\right]$$

$$+2\left(h(x) - \mathbf{E}\left[f_{\hat{\theta}}(x)\right]\right)\underbrace{\left(\mathbf{E}\left[f_{\hat{\theta}}(x)\right] - \mathbf{E}\left[f_{\hat{\theta}}(x)\right]\right)}_{0}$$

$$\mathbf{E}\left[\left(h(x) - f_{\hat{\theta}}(x)\right)^2\right] =$$

$$\underbrace{\mathbf{E}\left[\left(h(x) - \mathbf{E}\left[f_{\hat{\theta}}(x)\right]\right)^2\right]}_{\text{Constant}} + \mathbf{E}\left[\left(\mathbf{E}\left[f_{\hat{\theta}}(x)\right] - f_{\hat{\theta}}(x)\right)^2\right]$$

$$\left(h(x) - \mathbf{E}\left[f_{\hat{\theta}}(x)\right]\right)^2 +$$

$$\mathbf{E}\left[\left(h(x) - f_{\hat{\theta}}(x)\right)^2\right] =$$

$$\left(h(x) - \mathbf{E}\left[f_{\hat{\theta}}(x)\right]\right)^2 + \mathbf{E}\left[\left(\mathbf{E}\left[f_{\hat{\theta}}(x)\right] - f_{\hat{\theta}}(x)\right)^2\right]$$

(Bias)^2 **Model Variance**

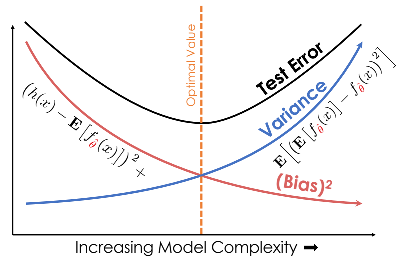
$$\mathbf{E}\left[\left(y - f_{\theta}(x)\right)^2\right] =$$

$$\mathbf{E}\left[\left(y - h(x)\right)^2\right] + \quad = \sigma^2$$ **Obs. Variance** "Noise"

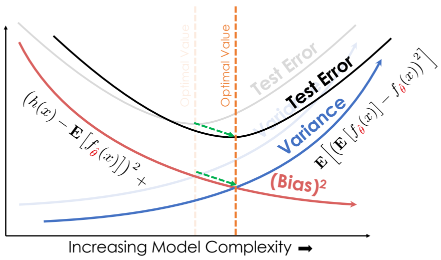
$$\left(h(x) - \mathbf{E}\left[f_{\hat{\theta}}(x)\right]\right)^2 +$$ **(Bias)^2**

$$\mathbf{E}\left[\left(\mathbf{E}\left[f_{\hat{\theta}}(x)\right] - f_{\hat{\theta}}(x)\right)^2\right]$$ **Model Variance**

Bias Variance Plot

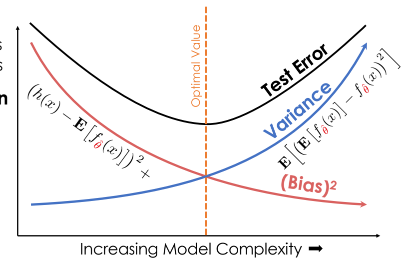


Bias Variance **Increasing Data**



How do we control **model complexity**?

- So far:
  - Number of features
  - Choices of features
- **Next: Regularization**



Bias Variance Derivation Quiz

- Match each of the following:

http://bit.ly/ds100-sp18-bvt

(1) $\mathbf{E}\left[y\right]$

(2) $\mathbf{E}\left[\epsilon^2\right]$

(3) $\mathbf{E}\left[\left(h(x) - \mathbf{E}\left[f_{\hat{\theta}}(x)\right]\right)^2\right]$

(4) $\mathbf{E}\left[\epsilon\left(h(x) - f_{\hat{\theta}}(x)\right)\right]$

A. 0
B. Bias²
C. Model Variance
D. Obs. Variance
E. h(x)
F. h(x) + ϵ

Bias Variance Derivation Quiz

- Match each of the following:

http://bit.ly/ds100-sp18-bvt

(1) $\mathbf{E}\left[y\right]$

(2) $\mathbf{E}\left[\epsilon^2\right]$

(3) $\mathbf{E}\left[\left(h(x) - \mathbf{E}\left[f_{\hat{\theta}}(x)\right]\right)^2\right]$

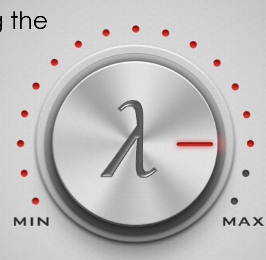
(4) $\mathbf{E}\left[\epsilon\left(h(x) - f_{\hat{\theta}}(x)\right)\right]$

A. 0
B. Bias²
C. Model Variance
D. Obs. Variance
E. h(x)
F. h(x) + ϵ

Regularization

Parametrically Controlling the *Model Complexity*

- Tradeoff:
  - **Increase bias**
  - **Decrease variance**



Basic Idea of Regularization

Fit the Data

Penalize Complex Models

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \mathbf{Loss}\left(y_i, f_\theta(x_i)\right) + \lambda \mathbf{R}(\theta)$$

Regularization Parameter

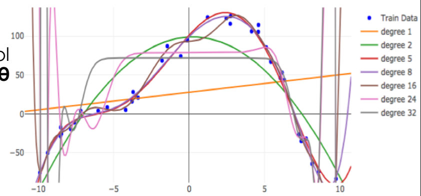
- How should we define **R(θ)**?
- How do we determine λ?

The Regularization Function **R(θ)**

Goal: *Penalize model complexity*

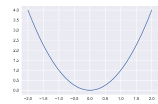
Recall earlier: $\phi(x) = [x, x^2, x^3, \ldots, x^p]$

- More features → overfitting …
- How can we control overfitting through **θ**
- **Proposal:** set weights = 0 to remove features



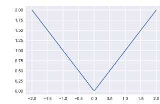
Common Regularization Functions

Ridge Regression (L2-Reg) $R_{\text{Ridge}}(\theta) = \sum_{i=1}^{d} \theta_i^2$



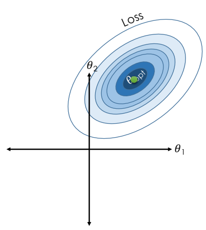
- Distributes weight across related features (robust)
- Analytic solution (easy to compute)
- Does not encourage sparsity → small but non-zero weights.

LASSO (L1-Reg) $R_{\text{Lasso}}(\theta) = \sum_{i=1}^{d} |\theta_i|$

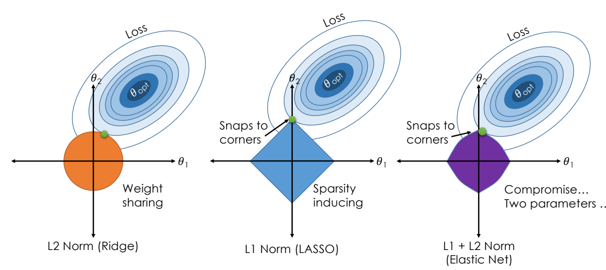


- **Encourages sparsity** by setting weights = 0
  - Used to select informative features
- Does not have an analytic solution → numerical methods

Regularization and Norm Balls



Regularization and Norm Balls



Python Demo!

The shapes of the norm balls.

Maybe show reg. effects on actual models.

Determining the Optimal λ

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathbf{Loss}\left(y_i, f_\theta(x_i)\right) + \lambda \mathbf{R}(\theta)$$

- Value of λ determines bias-variance tradeoff
  - Larger values → more regularization → more bias → less variance

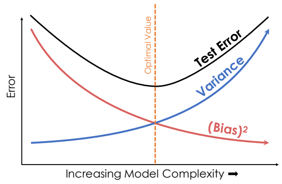
Summary

$$\mathbf{E}\left[(y - f_{\theta}(x))^2\right] =$$

$$\mathbf{E}\left[(y - h(x))^2\right] +$$

$$\left(h(x) - \mathbf{E}\left[f_{\hat{\theta}}(x)\right]\right)^2 +$$

$$\mathbf{E}\left[\left(\mathbf{E}\left[f_{\hat{\theta}}(x)\right] - f_{\hat{\theta}}(x)\right)^2\right]$$



Regularization

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \mathbf{Loss}\left(y_i, f_{\theta}(x_i)\right) + \lambda \mathbf{R}(\theta)$$