CS456: Machine Learning

Bias-Variance tradeoff

Jakramate Bootkrajang

Department of Computer Science
Chiang Mai University

March 16, 2020

Objective

- Understand tradeoff between model bias and variance

Outlines

- Expected Error
- Decomposition of expected error
- Bias Variance
- Practical guideline

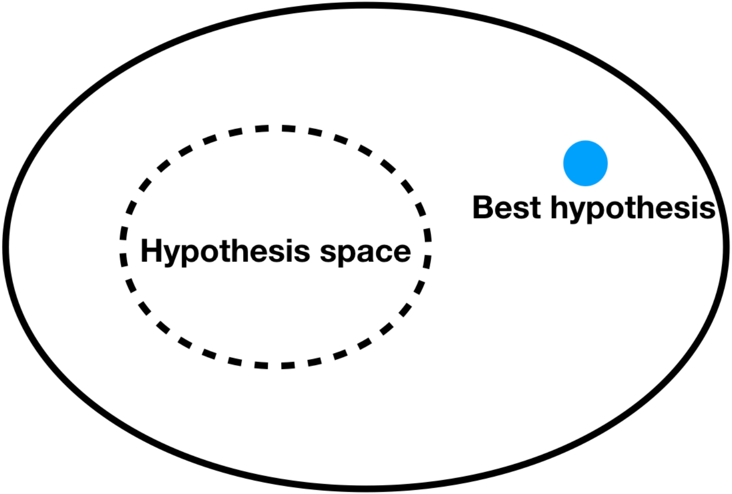
Learning as searching

- Recall, for example in case of LR, how we train the model for the best weight vector
- Learning a LR can be thought of as searching for best model (weights) among all possible model in model space
- The space that contains all models is named hypothesis space, and one particular model is called a hypothesis

(Learner) hypothesis space

- In practice, we need to make a choice as to what kind of model will be used to learn from data
- The bias in selecting model types is called inductive bias
- If we are lucky, the best hypothesis might be in the hypothesis space

Hypothesis space



Optimal model

- The *size* of hypothesis space characterises the resulting model
- Size is measured by
  - Counting: only for finite hypothesis space
  - Vapnik–Chervonenkis (VC) dimension
- This is known also known as model's
  - complexity, capacity, richness, expressive power

Complex vs simple model

- Complex model (complex hypothesis space)
  - fits data well (low bias)
  - can overfit data if training data is of poor quality
  - performance varies with training data (high variance)
- Simple model
  - might not fit data well (high bias)
  - performance varies less with training data (low variance)

Error analysis: Notation

Let

- D be a training data
- $h_D()$ be a model trained using D
- $(\mathbf{x}, y)$ be any test instance and its correct answer (unknown to classifier)
- Assuming regression task, expected error is given by

$$E_{\mathbf{x},y,D}[(h_D(\mathbf{x}) - y)^2] \tag{1}$$

Error decomposition

- Error in Eq.(1) can be decomposed into 3 components

$$
\begin{aligned}
& E_{\mathbf{x},y,D}\left[[h_D(\mathbf{x}) - y]^2\right] \\
& = E_{\mathbf{x},y,D}\left[\left[\left(h_D(\mathbf{x}) - \bar{h}(\mathbf{x})\right) + \left(\bar{h}(\mathbf{x}) - y\right)\right]^2\right] \\
& = E_{\mathbf{x},D}\left[(\bar{h}_D(\mathbf{x}) - \bar{h}(\mathbf{x}))^2\right] + \\
& \quad 2\ E_{\mathbf{x},y,D}\left[\left(h_D(\mathbf{x}) - \bar{h}(\mathbf{x})\right)\left(\bar{h}(\mathbf{x}) - y\right)\right] \\
& \quad + E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x}) - y\right)^2\right]
\end{aligned} \tag{2}
$$

The middle term

Let $\bar{h}(\mathbf{x})$ be the expected hypothesis (over training data)

$$\begin{aligned}
E_{\mathbf{x},y,D}\left[\left(h_D(\mathbf{x}) - \bar{h}(\mathbf{x})\right)\left(\bar{h}(\mathbf{x}) - y\right)\right] &= E_{\mathbf{x},y}\left[E_D\left[h_D(\mathbf{x}) - \bar{h}(\mathbf{x})\right]\left(\bar{h}(\mathbf{x}) - y\right)\right] \\
&= E_{\mathbf{x},y}\left[\left(E_D\left[h_D(\mathbf{x})\right] - \bar{h}(\mathbf{x})\right)\left(\bar{h}(\mathbf{x}) - y\right)\right] \\
&= E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x}) - \bar{h}(\mathbf{x})\right)\left(\bar{h}(\mathbf{x}) - y\right)\right] \\
&= E_{\mathbf{x},y}\left[0\right] \\
&= 0
\end{aligned}$$

The error reduced to

$$E_{\mathbf{x},y,D}\left[\left(h_D(\mathbf{x})-y\right)^2\right] = \underbrace{E_{\mathbf{x},D}\left[\left(h_D(\mathbf{x})-\bar{h}(\mathbf{x})\right)^2\right]}_{\text{Variance}} + E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x})-y\right)^2\right] \tag{3}$$

The last term

$$
\begin{aligned}
E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x}) - y\right)^2\right] &= E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x})) + (\bar{y}(\mathbf{x}) - y\right)^2\right] && (4) \\
&= \underbrace{E_{\mathbf{x},y}\left[(\bar{y}(\mathbf{x}) - y)^2\right]}_{\text{Noise}} + \underbrace{E_{\mathbf{x}}\left[\left(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x})\right)^2\right]}_{\text{Bias}^2} \\
&\quad + 2\ E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x})\right)(\bar{y}(\mathbf{x}) - y)\right] && (5)
\end{aligned}
$$

The last term of Eq.5

$$
\begin{aligned}
E_{\mathbf{x},y}\left[\left(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x})\right)\left(\bar{y}(\mathbf{x}) - y\right)\right] &= E_{\mathbf{x}}\left[E_{y|\mathbf{x}}\left[\bar{y}(\mathbf{x}) - y\right]\left(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x})\right)\right] \\
&= E_{\mathbf{x}}\left[E_{y|\mathbf{x}}\left[\bar{y}(\mathbf{x}) - y\right]\left(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x})\right)\right] \\
&= E_{\mathbf{x}}\left[\left(\bar{y}(\mathbf{x}) - E_{y|\mathbf{x}}\left[y\right]\right)\left(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x})\right)\right] \\
&= E_{\mathbf{x}}\left[\left(\bar{y}(\mathbf{x}) - \bar{y}(\mathbf{x})\right)\left(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x})\right)\right] \\
&= E_{\mathbf{x}}\left[0\right] \\
&= 0
\end{aligned}
$$

Finally

$$\underbrace{E_{\mathbf{x},y,D}\left[(h_D(\mathbf{x}) - y)^2\right]}_{\text{Expected Test Error}} = \underbrace{E_{\mathbf{x},D}\left[\left(h_D(\mathbf{x}) - \bar{h}(\mathbf{x})\right)^2\right]}_{\text{Variance}} + \underbrace{E_{\mathbf{x},y}\left[(\bar{y}(\mathbf{x}) - y)^2\right]}_{\text{Noise}} + \underbrace{E_{\mathbf{x}}\left[\left(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x})\right)^2\right]}_{\text{Bias}^2}$$

- Variance: how much classifier changes if you train on a different training set
- Bias: how much classifier being "biased" to a particular kind of hypothesis (e.g. linear classifier).
- Noise: measures ambiguity due to your data distribution and feature representation.

Bias-Variance

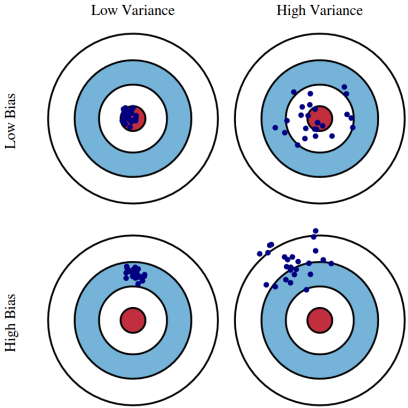


Figure: `http://scott.fortmann-roe.com/docs/BiasVariance.html`

Bias-Variance Tradeoff

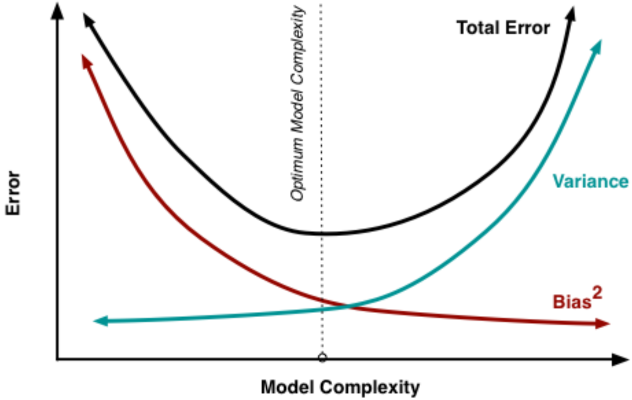


Figure: http://scott.fortmann-roe.com/docs/BiasVariance.html

Practical guideline (1/2)

High variance

- Symptoms:
  - Training error is much lower than test error
  - Training error is lower than
  - Test error is above
- Remedies:
  - Add more training data
  - Reduce model complexity – complex models are prone to high variance

Practical guideline (1/2)

High bias

- Symptoms:
  - Training error is higher than
- Remedies:
  - Use more complex model (e.g. kernelize, use non-linear models)
  - Add features

Objective: revisited

- Understand tradeoff between model bias and variance

Reference

- https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote12.html