Unit 1 Regression Modeling

Variance-Bias Trade off

October 2, 2019

Introduction

predictive model : It is important to understand prediction errors (bias and variance)
Gaining a proper understanding of these errors would help us not only to build accurate models but also to avoid the mistake of overfitting and underfitting.
start w/ the basics and see how they make difference to models.

What is bias?

the difference btw the average prediction of model and the correct value. which we are trying to predict.

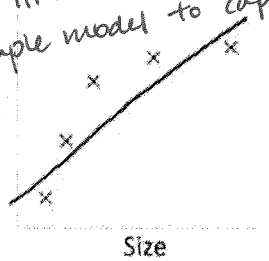
$$I(y - \hat{y}).$$

model w/ high bias → under fitting.

What is variance?

variability of model prediction for a given data point. (spread of data)
models w/ high variance. perform very well on trainig data but has high error rates on test data.

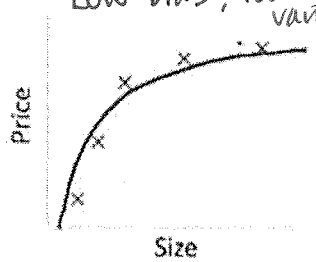
happens when a model unable to capture the underlying pattern of data.
⇒ very less amount of data to build the model.
⇒ try quit linear model w/ non-linear data.
⇒ fit simple model to capture the complex patterns.



$\theta_0 + \theta_1 x$

High bias (underfit)

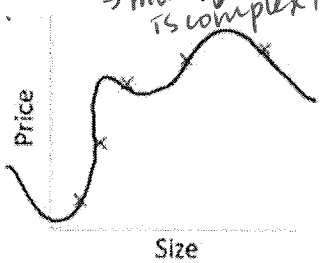
Low bias, low variance.



$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"

happens when our model captures the noise along w/ the underlying pattern in data.
→ training set includes lots of noise.
→ model is complex!
Hard to generalize the model



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High variance (overfit)

Error = Bias^2 + variance + Irreducible Error.
↑ "data" under- or over-fitted model
↑ model selection.
↑ cannot be reduced by creating good models no matter how good model we make (noise of data).

There is a trade off between a model's ability to minimize bias and variance.

Why is Bias Variance Trade off?

too simple model (has very few parameters) ⇒ High bias and low variance.

lager # of parameters ⇒ High variance and low bias.

⇒ Need to find the right / good balance w/o overfitting and underfitting the data.

This tradeoff in complexity is why there is a tradeoff btw bias and variance.

Total Error = Bias^2 + variance + Irreducible error

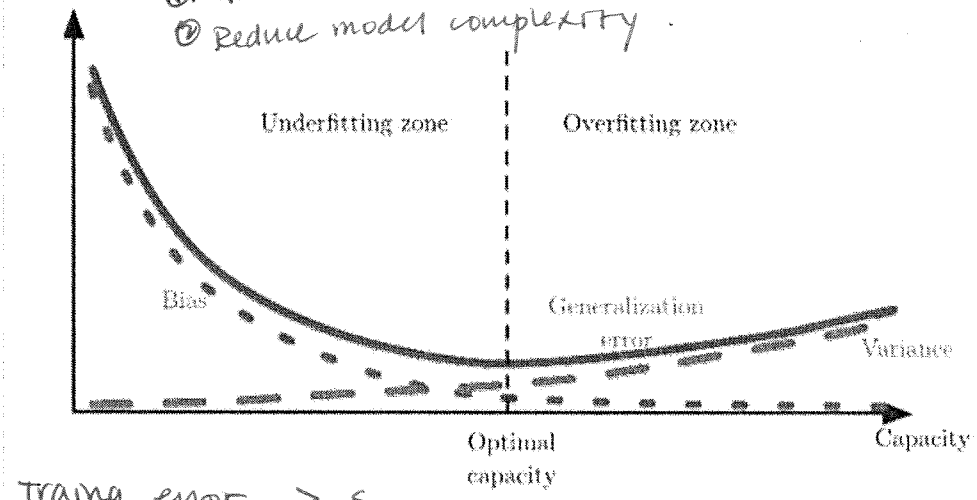
↑ ↓ fixed.

↓ ↑

symptoms

High variance: ① Training error is much lower than test error.
② Training error is lower than ε.
③ Test error is above ε.

solution ⇓
① Add more training data
② Reduce model complexity.



High Bias : Traing error > ε.

① complex model (ex) non-linear model).
② Add features.
③ Boosting.