Variance- Bias Trade off

Agenda

- Bias & Variance
- Trade-off
- K fold cross validation
- Regularization methods for regression

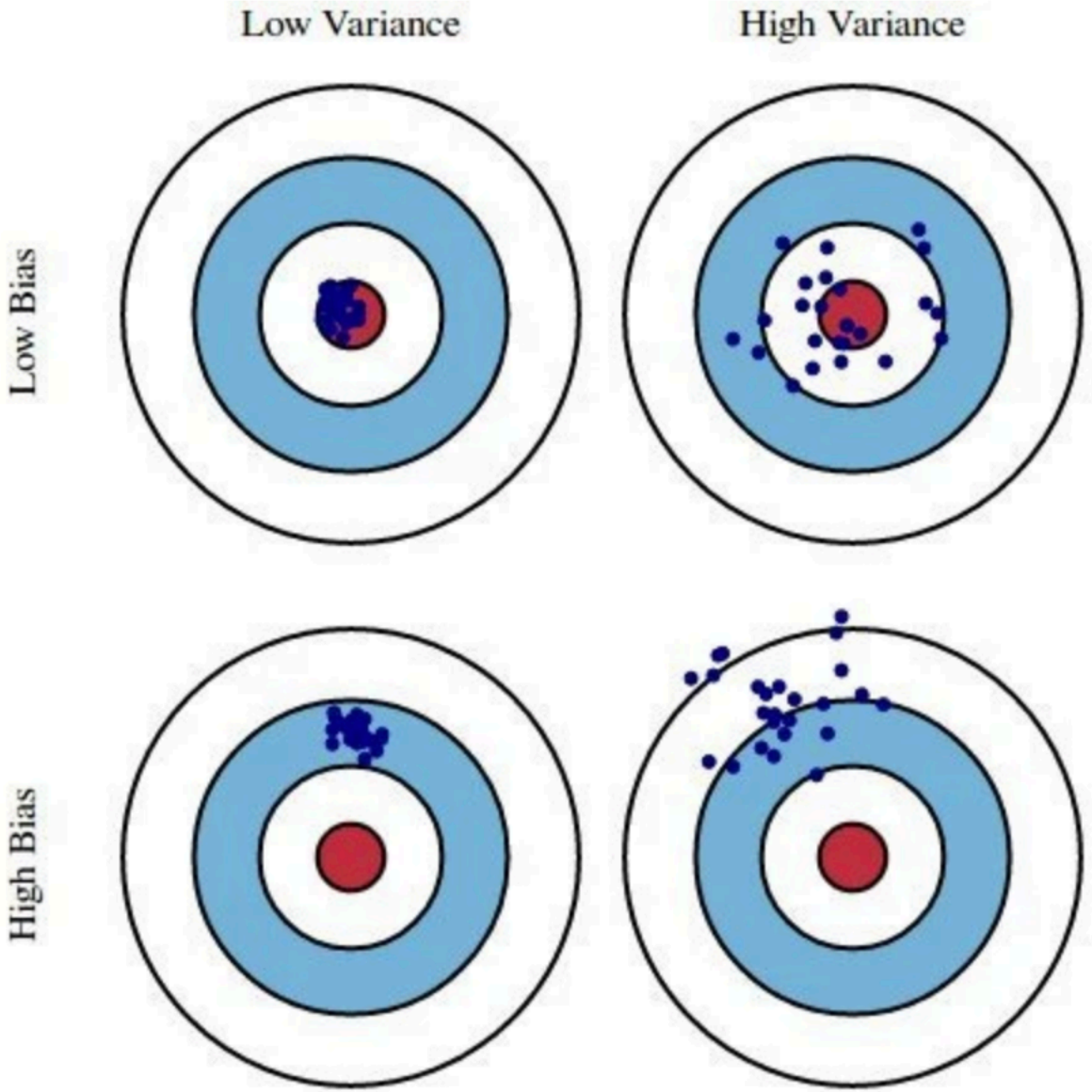
Bias and Variance

- "Model Bias" is the expected prediction error from your expected trained model
- "Model Variance" is the expected variation in predictions, relative to your expected trained model

- What do we mean by high bias and low bias?
- What do we mean by high variance and low Variance?

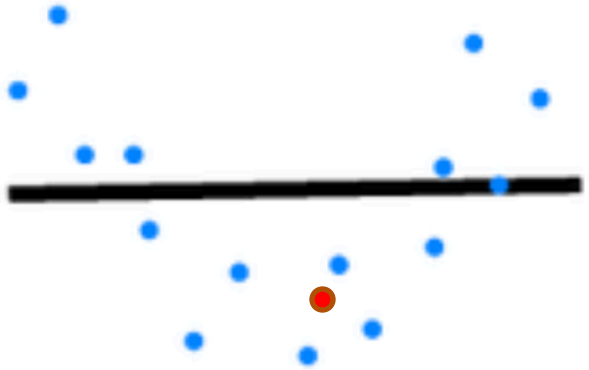
Archery target example:

Which should be an ideal model?

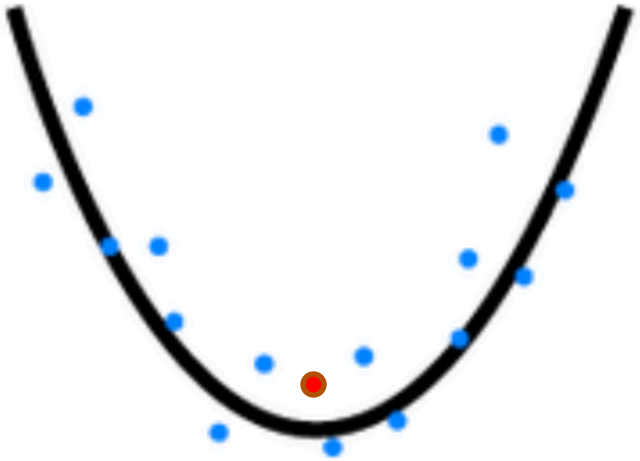


Which model is best? Why?

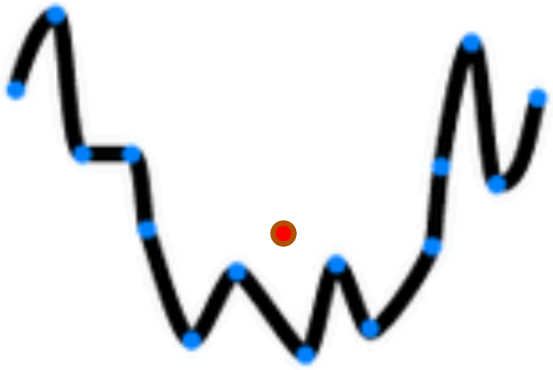
Write an explanation for why the quadratic model is “best”, using the terms bias, variance, and error. Share it with a neighbor, then synthesize your explanations.



Linear Model



Quadratic Model



High-Order Polynomial Model


Y
X
Underfitting
Y
X
Just right!
Y
X
overfitting

Overfitting

- Model that fits our training data well but fails to estimate the real relationship among variables beyond the training set. Therefore our model performs poorly on the test data. This problem is called as **over-fitting**.

It occurs :

- when a statistical model or machine learning algorithm captures the noise/outliers of the data.
- It occurs when the model or the algorithm fits the data too well.

Overfitting a model result in good accuracy for training data set but poor results on new data sets.

Underfit Model

- A statistical model or machine learning algorithm cannot capture the underlying trend of the data. This condition is known as underfitting.
- Occurs when the model or the algorithm does not fit the data well enough.
- Underfitting is often a result of an excessively simple model.

Question

- Based on your knowledge of bias and variance how would you define overfitting?

High Variance

Symptoms:

- Training error is much lower than test error

Remedies:

- Add more training data
- Reduce model complexity -- complex models are prone to high variance
- Bagging (will be covered later in the course)
- Regularization

High Bias

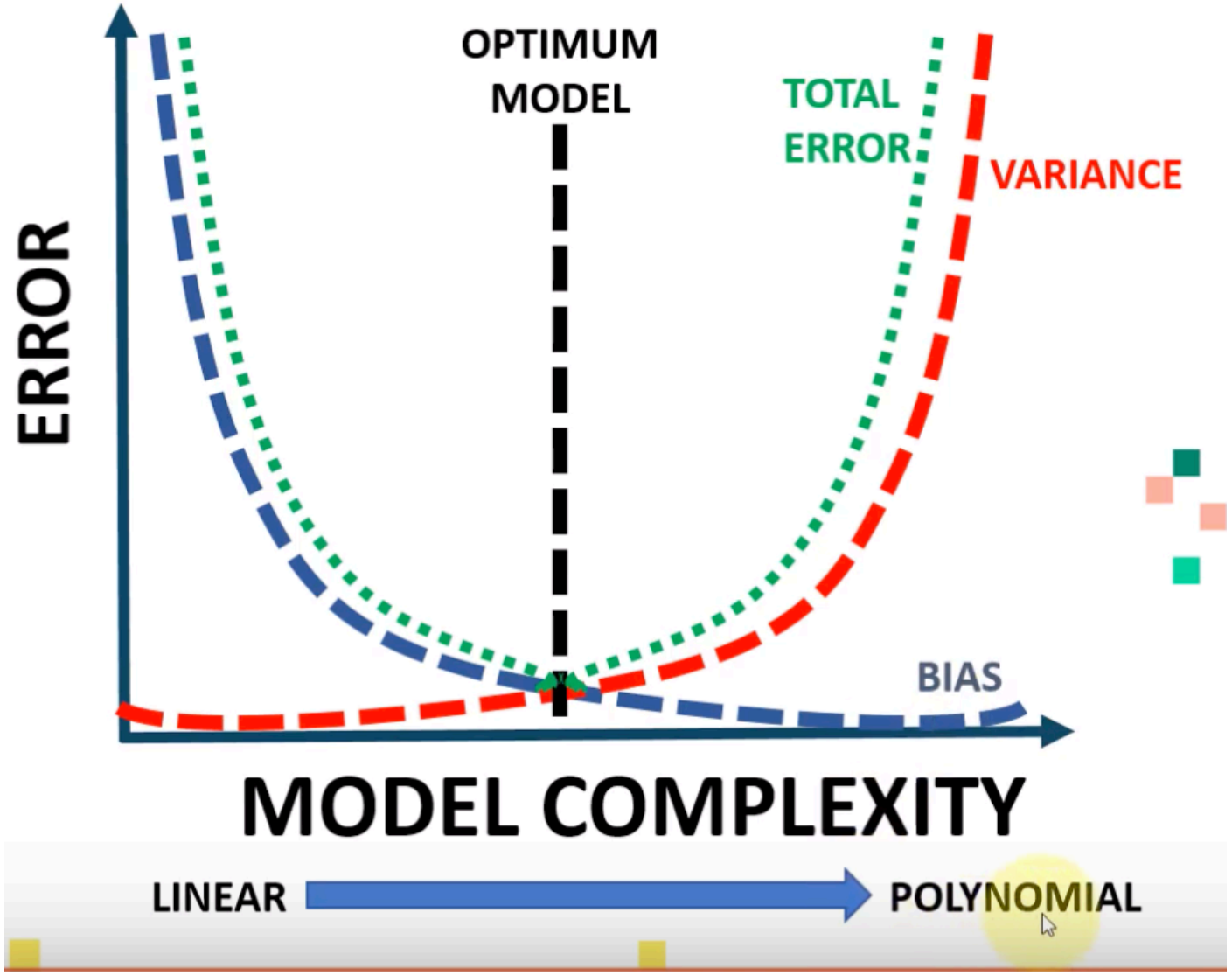
Symptoms:

- Training error is higher than ϵ

Remedies:

- Use more complex model (e.g. kernelize, use non-linear models)
- Add features
- Boosting (will be covered later in the course)

Bias Variance trade off



What are the methods to avoid overfitting

- Regularization
- Boosting/ Bagging
- Reduce the model complexity
- Cross Validation

Cross-validation

- **Cross-validation** is a technique that is used for the assessment of how the results of statistical analysis generalize to an independent data set.

Why?

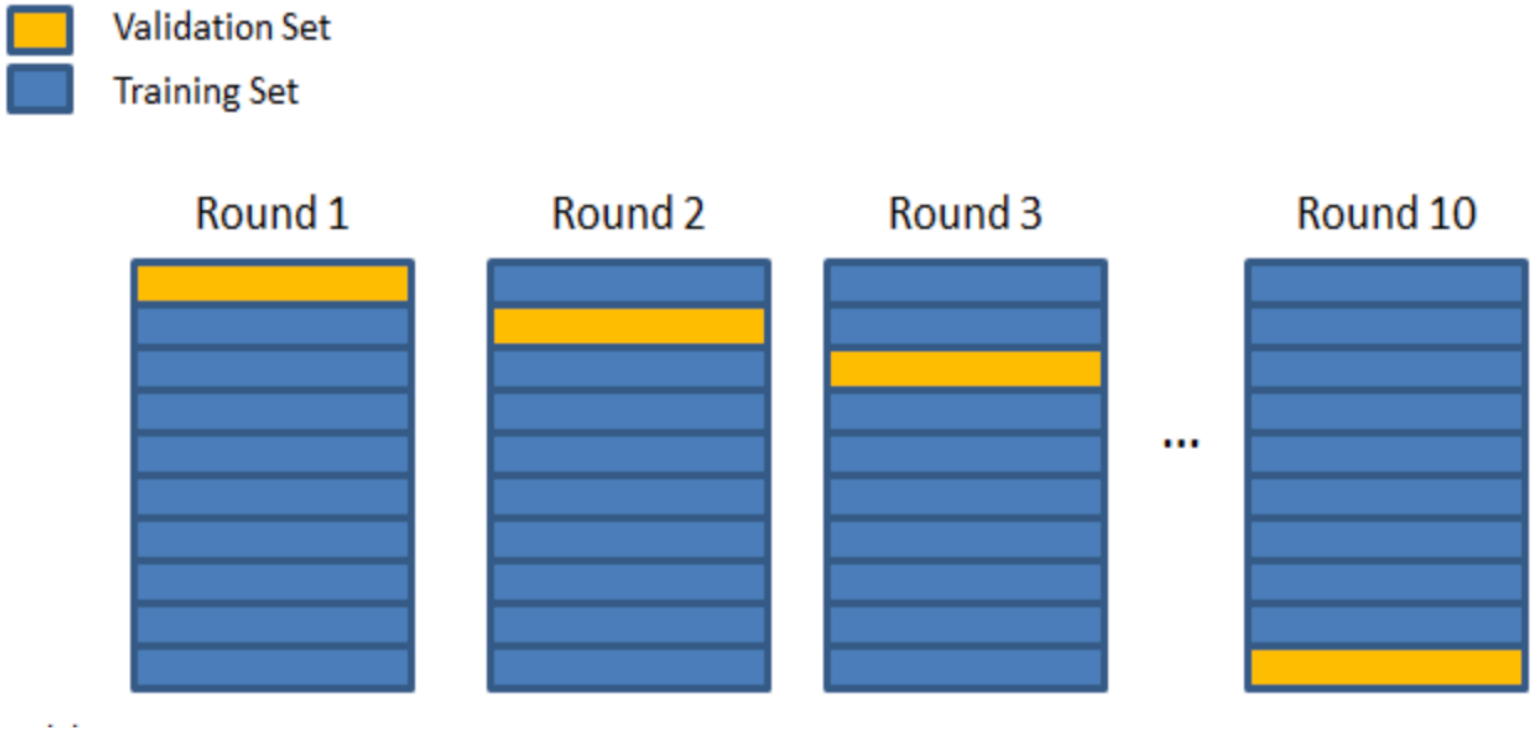
- Evaluate our model
- Avoid overfitting

Methods:

- K fold cross validation
- Leave one out cross validation

K fold cross validation

- To partition the data into a number of subsets
- Hold out a set at a time and train the model on remaining set
- Test model on hold out set



Regularization

Types: Lasso & Ridge

Why?

- Reduces complexity
- Reduce the chance of overfitting in regression model
- Reduces model's variance at the expense of introducing small bias
- Increases model's interpretability
- Help to deal with collinearity

Regularization

How we do it?

- It works by reducing the variance at the cost of adding some bias to the model
- We normally we keep the same number of features, but reduce the magnitude of the coefficients.

Lasso

- **Lasso Regression** (Least Absolute Shrinkage and Selection Operator) adds "*absolute value of magnitude*" of coefficient as penalty term to the loss function.
- The magnitude of the coefficients decreases, where the values may reach 1 or zero removing the feature altogether
- It uses L1 regularization technique (will be discussed later in this article)
- It is generally used when we have more number of features, because it automatically does feature selection.

Ridge

- **Ridge regression** adds "*squared magnitude*" of coefficient as penalty term to the loss function.
- The magnitude of the coefficients decreases, where the values reaches to zero but not absolute zero.
- It shrinks the parameters, therefore it is mostly used to prevent multicollinearity.
- It reduces the model complexity by coefficient shrinkage.
- It uses L2 regularization technique.

Difference between Lasso and Ridge

Lasso shrinks the less important feature's coefficient to zero thus, removing some feature altogether.

So, this works well for **feature selection** in case we have a huge number of features.

Q. When we know that Ridge and Lasso is better than vanilla linear regression?

- A. High variation in your model --> Collinearity and too many variables.

Q. How do we know whether we should choose Lasso or Ridge?

- A. Most of the time they perform very similar but Lasso has the feature selection property, ridge doesn't have this.

Thank you!