Homework 02

Brandon Hosley

University of Illinois - Springfield

September 4, 2020

Overview

1. Q1: Bias–variance tradeoff

2. Q2: Hastie and Tibshirani Summary

Bias-Variance Trade-Off

Bias Error

- Also called 'Overfitting'

Variance

Bias-Variance Trade-Off

Bias Error

- Also called 'Overfitting'

Variance

- Also called 'Underfitting'

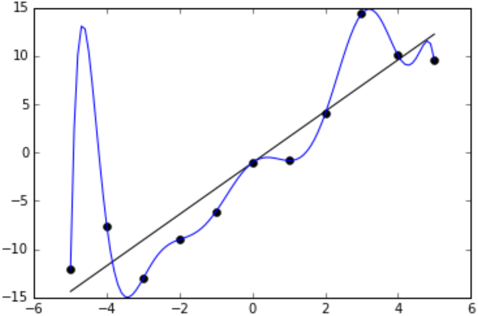
Bias-Variance Trade-Off

Bias Error

- Also called 'Overfitting'

Variance

- Also called 'Underfitting'



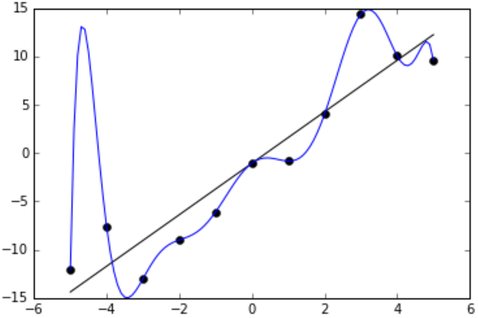
Bias-Variance Trade-Off

Bias Error

- Also called 'Overfitting'
- Predicts test data too well

Variance

- Also called 'Underfitting'



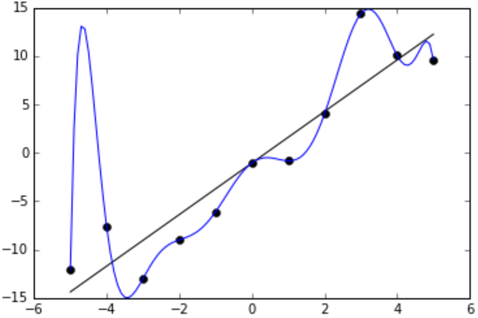
Bias-Variance Trade-Off

Bias Error

- Also called 'Overfitting'
- Predicts test data too well

Variance

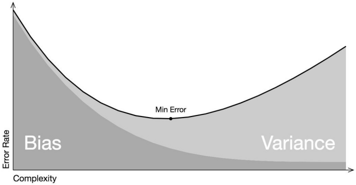
- Also called 'Underfitting'
- Generalizes too well



Bias-Variance Trade-Off

- Aiming for the lowest total error typically means finding a "middle-ground"
- A common technique for this is determining the minimum *mean squared error.*

$$
\begin{aligned}
\text{MSE} &= \left(E\left[\hat{f}(x)\right] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E\left[\hat{f}(x)\right]\right)^2\right] + \sigma_e^2 \\
&= \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}
\end{aligned}
$$



Hastie Lectures Summary

- Statistical Learning and Regression
- Dimensionality and Parametric Models
- Assessing Model Accuracy and Bias-Variance Trade-Off
- Classification Problems and K-Nearest Neighbors

Statistical Learning and Regression

- Predicting values through various *means*

Statistical Learning and Regression

- Predicting values through various *means*
- The simplest method of regression may be to use:

$$\hat{f}(x_n) = \bar{x}_n$$

Statistical Learning and Regression

- Predicting values through various *means*
- The simplest method of regression may be to use:

$$\hat{f}(x_n) = \bar{x}_n$$

- This type of model may be improved by extending the values contributing to the average to:

$$\hat{f}(x_n) = \bar{x}_{n \pm i}$$

Statistical Learning and Regression

- Predicting values through various *means*
- The simplest method of regression may be to use:

$$\hat{f}(x_n) = \bar{x}_n$$

- This type of model may be improved by extending the values contributing to the average to:

$$\hat{f}(x_n) = \bar{x}_{n\pm i}$$

- Or the range may be adjusted such that it includes the k-nearest neighbors.

Statistical Learning and Regression

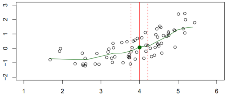
- Predicting values through various *means*
- The simplest method of regression may be to use:

$$\hat{f}(x_n) = \bar{x}_n$$

- This type of model may be improved by extending the values contributing to the average to:

$$\hat{f}(x_n) = \bar{x}_{n\pm i}$$

- Or the range may be adjusted such that it includes the k-nearest neighbors.

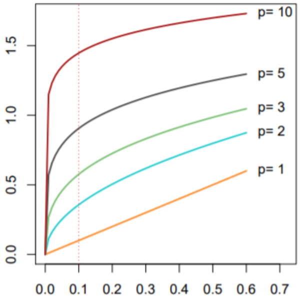


Dimensionality and Parametric Models

- Approximate proportion of a data's range needed to capture a certain fraction of the data scales very quickly with added dimensions

Dimensionality and Parametric Models

- Approximate proportion of a data's range needed to capture a certain fraction of the data scales very quickly with added dimensions



Dimensionality and Parametric Models

- Approximate proportion of a data's range needed to capture a certain fraction of the data scales very quickly with added dimensions
- Structured models may help address these scaling problems

Dimensionality and Parametric Models

- Approximate proportion of a data's range needed to capture a certain fraction of the data scales very quickly with added dimensions
- Structured models may help address these scaling problems

  The simplest structured model is a linear model:

$$f_L(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

Dimensionality and Parametric Models

- Approximate proportion of a data's range needed to capture a certain fraction of the data scales very quickly with added dimensions
- Structured models may help address these scaling problems

  The simplest structured model is a linear model:

  $$f_L(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$
- Structured models allowed allow flexible adjustment to obtain preferred level of fitting

Assessing Model Accuracy and Bias-Variance Trade-Off

- Mean-squared error is a good method for testing accuracy

$$\mathrm{MSE}_{\mathbf{Tr}} = \mathrm{AVE}_{i \in \mathbf{Tr}} \left[y_i - \hat{f}(x_i) \right]^2$$

Assessing Model Accuracy and Bias-Variance Trade-Off

- Mean-squared error is a good method for testing accuracy
- But testing against *training* data will bias toward over-fit models

$$\text{MSE}_{\textbf{Tr}} = \text{AVE}_{i \in \textbf{Tr}} \left[y_i - \hat{f}(x_i) \right]^2$$

Assessing Model Accuracy and Bias-Variance Trade-Off

- Mean-squared error is a good method for testing accuracy
- But testing against *training* data will bias toward over-fit models
- Measuring against a separate batch of *test* data will likely provide a more accurate model

$$\mathrm{MSE}_{\mathbf{Te}} = \mathrm{AVE}_{i \in \mathbf{Te}} \left[y_i - \hat{f}(x_i) \right]^2$$

Classification Problems and K-Nearest Neighbors

- Goals of classification:
  - Build classifier model $\mathcal{C}(X)$
  - Assess uncertainty of prediction
  - Understand roles of factors in prediction $X = \{X_1, X_2, \ldots, X_n\}$

Classification Problems and K-Nearest Neighbors

- Goals of classification:
  - Build classifier model $\mathcal{C}(X)$
  - Assess uncertainty of prediction
  - Understand roles of factors in prediction $X = \{X_1, X_2, \ldots, X_n\}$
- k-Nearest Neighbor works well for low dimension classifiers
- For individual classes a conditional class probability can be determined:

$$p_k = \Pr(Y = k | X = x), k = 1, 2, \ldots, K$$

Classification Problems and K-Nearest Neighbors

- Goals of classification:
  - Build classifier model $\mathcal{C}(X)$
  - Assess uncertainty of prediction
  - Understand roles of factors in prediction $X = \{X_1, X_2, \ldots, X_n\}$
- k-Nearest Neighbor works well for low dimension classifiers
- For individual classes a conditional class probability can be determined:

$$p_k = \Pr(Y = k | X = x), k = 1, 2, \ldots, K$$

- For any model the Baye's optimal classifier will be the result of combining conditional class probabilities:

$$\mathcal{C}(x) = j \text{ if } p_j(x) = \max \{p_1(x), p_2(x), \ldots, p_K(x)\}$$