Tutorial: Bias-Variance Decomposition

Cheat Sheet and Practice Problems

ES335 - Machine Learning
IIT Gandhinagar

July 23, 2025

1 Summary from Slides

1.1 The Three Sources of Error

Any prediction made by a machine learning model is affected by three sources of error:

- **Noise**: Irreducible error inherent in the data
- **Bias**: Error due to overly simplistic assumptions
- **Variance**: Error due to sensitivity to small changes in training data

Mathematical Formulation: For a true function $f_{\theta_{\text{true}}}(x)$ with noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, the observed data follows:

$$y_t = f_{\theta_{\text{true}}}(x_t) + \epsilon_t$$

1.2 Bias

Definition: Bias measures how well a model can capture the true underlying relationship. It represents the error introduced by approximating a complex real-world problem with a simplified model.

Mathematical Definition:

$$\text{Bias}(x_t) = f_{\theta_{\text{true}}}(x_t) - f_{\bar{\theta}}(x_t)$$

where $f_{\bar{\theta}}(x_t) = E_{\text{train}}[f_{\hat{\theta}}(x_t)]$ is the average fit over all possible training sets.

Key Property: As model complexity increases, bias decreases because the model becomes more flexible and can better approximate the true function.

1.3 Variance

Definition: Variance measures how much the predictions change when trained on different training sets. High variance indicates that small changes in training data lead to very different models.

Mathematical Definition:

$$\text{Variance}(f_{\hat{\theta}}(x_t)) = E_{\text{train}}[(f_{\hat{\theta}}(x_t) - f_{\bar{\theta}}(x_t))^2]$$

Key Property: As model complexity increases, variance increases because complex models are more sensitive to small changes in training data.

1.4 The Bias-Variance Trade-off

Fundamental Relationship: There exists an inherent trade-off between bias and variance:

- Simple models: High bias, low variance
- Complex models: Low bias, high variance
- Optimal models: Balance between bias and variance

Expected Prediction Error Decomposition:

$$E_{\text{train},y_t}[(y_t - f_{\hat{\theta}}(x_t))^2] = \sigma^2 + [\text{Bias}(f_{\hat{\theta}}(x_t))]^2 + \text{Variance}(f_{\hat{\theta}}(x_t))$$

1.5 Underfitting and Overfitting

Underfitting (High Bias):

- Model is too simple to capture underlying patterns
- Poor performance on both training and test data
- Example: Linear model for non-linear data

Overfitting (High Variance):

- Model learns noise and specific patterns in training data
- Good training performance, poor test performance
- Example: Deep decision tree memorizing training data

Good Fit:

- Balanced bias and variance
- Generalizes well to unseen data
- Found through techniques like cross-validation

2 Practice Problems

Problem : Basic Bias-Variance Concepts

Consider three different models trained to predict house prices:

- Model A: Always predicts the average price (constant function)
- Model B: Linear regression
- Model C: Decision tree with maximum depth

a) Which model likely has the highest bias? Explain why.
b) Which model likely has the highest variance? Explain why.
c) If the true relationship is non-linear, how would this affect each model's bias?
d) Rank the models from lowest to highest variance, justifying your answer.

Problem : Mathematical Understanding

Given the bias-variance decomposition:

$$\text{Expected Error} = \sigma^2 + \text{Bias}^2 + \text{Variance}$$

Assume for a particular point:

- Irreducible error: $\sigma^2 = 4$
- Bias = 3
- Variance = 5

a) Calculate the total expected error
b) If we could reduce bias to 1, what would be the new expected error?
c) If we could reduce variance to 2 (keeping original bias), what would be the expected error?
d) Which reduction (bias or variance) provides greater improvement?

Problem : Decision Tree Depth Analysis

Consider a decision tree classifier with varying depths on a dataset:

Depth	Training Accuracy	Test Accuracy
1	0.65	0.63
3	0.78	0.75
5	0.89	0.79
10	0.95	0.71
15	1.00	0.68

a) Identify the underfitting, optimal, and overfitting regions
b) At what depth does overfitting begin?
c) Explain the bias-variance trade-off observed in this data
d) What depth would you choose for deployment? Why?

Problem : Ensemble Methods and Bias-Variance

Consider the following ensemble methods:

- Bagging: Average of 100 decision trees trained on bootstrap samples
- Single deep tree: One decision tree with maximum depth
- Random Forest: 100 trees with random feature selection

a) Compare the bias of these three approaches
b) Compare the variance of these three approaches
c) Why does bagging reduce variance?
d) How does random feature selection in Random Forest affect bias and variance?

Problem : Polynomial Regression Analysis

You fit polynomial models of different degrees to a dataset:

- Degree 1: $\hat{y} = \theta_0 + \theta_1 x$
- Degree 5: $\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + ... + \theta_5 x^5$
- Degree 15: High-degree polynomial

Assume the true function is quadratic: $y = 2 + 3x + 0.5x^2 + \epsilon$
a) Which model has the highest bias? Explain.
b) Which model has the highest variance? Explain.
c) As training set size increases, how does bias change for each model?
d) As training set size increases, how does variance change for each model?

Problem : Regularization Effects

Consider three linear regression models:

- Model A: Standard linear regression
- Model B: Ridge regression with $\lambda = 0.1$
- Model C: Ridge regression with $\lambda = 10.0$

a) Order these models from highest to lowest bias
b) Order these models from highest to lowest variance
c) How does increasing regularization strength affect the bias-variance trade-off?
d) In what scenarios would you prefer Model C over Model A?

Problem : Cross-Validation and Model Selection

You perform 5-fold cross-validation on different models:

Model	Mean CV Score	Standard Deviation
Linear	0.75	0.02
Polynomial (degree 3)	0.82	0.05
Polynomial (degree 10)	0.78	0.15

a) Which model shows the highest variance across folds?
b) Which model likely has the best bias-variance trade-off?
c) Why is the standard deviation a good indicator of model variance?
d) How would you use this information for model selection?

Problem : Learning Curves Analysis

You observe the following learning curves (training set size vs. error):
For a simple model:

- Training error: starts low, increases slowly, plateaus at high value
- Validation error: starts high, decreases slowly, plateaus at high value
- Final gap between training and validation: small

For a complex model:

- Training error: stays very low throughout
- Validation error: starts very high, decreases but remains high
- Final gap between training and validation: large

a) Which model suffers from high bias?
b) Which model suffers from high variance?
c) What would adding more training data likely achieve for each model?
d) Suggest remedies for each model's problems.

Problem : K-Nearest Neighbors Analysis

Consider K-NN with different values of K:

- K = 1: Use only the nearest neighbor
- K = 5: Use 5 nearest neighbors
- K = 100: Use 100 nearest neighbors

a) Order these models from highest to lowest bias
b) Order these models from highest to lowest variance
c) How does the choice of K affect the decision boundary complexity?
d) In a dataset with 1000 samples, what problems might K = 100 cause?

Problem : Noise Impact Analysis

Two datasets with the same underlying function but different noise levels:

- Dataset A: Low noise ($\sigma^2 = 0.1$)
- Dataset B: High noise ($\sigma^2 = 5.0$)

You train the same model architecture on both datasets.
a) How does noise level affect the optimal model complexity?
b) Would you expect the same optimal regularization strength for both datasets?
c) How does noise affect the bias-variance trade-off curve?
d) Which dataset would benefit more from ensemble methods? Why?

Problem : Feature Selection Impact

Consider three scenarios:

- Scenario A: Use all 100 features
- Scenario B: Use top 20 features selected by correlation
- Scenario C: Use top 5 features selected by mutual information

a) How does feature selection affect model bias?
b) How does feature selection affect model variance?
c) Compare the bias-variance trade-off across the three scenarios
d) When might Scenario A be preferred despite higher variance?

Problem : Boosting vs Bagging

Compare two ensemble approaches:

- AdaBoost: Sequential ensemble that focuses on difficult examples
- Random Forest: Parallel ensemble with bootstrap sampling

Both use decision trees as base learners.
a) Which method primarily reduces bias?
b) Which method primarily reduces variance?
c) How does the sequential nature of boosting affect the bias-variance trade-off?
d) In what scenarios might boosting lead to overfitting?

Problem : Derivation Challenge

Prove that for the squared loss function, the expected prediction error can be decomposed as:

$$E[(y - \hat{f}(x))^2] = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

Given:

- $y = f(x) + \epsilon$ where $E[\epsilon] = 0$ and $\text{Var}(\epsilon) = \sigma^2$
- $\hat{f}(x)$ is our prediction
- $\bar{f}(x) = E[\hat{f}(x)]$ is the expected prediction

a) Start with $E[(y - \hat{f}(x))^2]$ and add/subtract $\bar{f}(x)$ and $f(x)$
b) Expand the squared terms
c) Show that cross terms vanish due to independence assumptions
d) Identify each component in the final decomposition

Problem : Advanced Application

You're building a medical diagnosis system with the following constraints:

- Limited training data (1000 samples)
- High-dimensional features (500 features)
- False negatives are more costly than false positives
- Model interpretability is important for regulatory approval

a) Would you prefer a high-bias or high-variance model? Justify.
b) How would the cost asymmetry affect your bias-variance considerations?
c) Suggest three techniques to manage the bias-variance trade-off in this scenario
d) How would you validate that your chosen model has the right bias-variance balance?
e) If you could collect more data, how would this change your approach?