Practice Questions

1 Bias-Variance Trade-off

Question: Understanding Bias-Variance Trade-off in Regression

Consider a regression problem where we try to approximate the true function $f(x)$ using different models. Given a dataset with $n = 100$ training points, you fit the following models:

- **Model 1**: A simple linear regression model.
- **Model 2**: A polynomial regression model of degree 10.
- **Model 3**: A polynomial regression model of degree 3.

1. **Bias-Variance Decomposition:** Express the expected mean squared error (MSE) at a new test point x_0 in terms of bias, variance, and irreducible error.
2. **Comparing Models:**
   - Which model is likely to have the highest bias? Explain why.
   - Which model is likely to have the highest variance? Explain why.
3. **Model Selection:** Based on the bias-variance trade-off, which model is expected to generalize best on unseen data? Justify your answer.
4. **Impact of Increasing Training Data:** If the training set size increases from $n = 100$ to $n = 10,000$, how would the bias and variance change for each model?

2 Decision Tree Regression (Numerical)

2.1 Question 1: Splitting based on MSE

You are given the following dataset:

A decision tree splits based on minimizing the **mean squared error (MSE)**. If you split at $X = 3.5$, compute:

- The MSE before the split.
- The total MSE after the split.
- Compare it with splitting at $X = 5.5$. Which split is better?

X	Y
1	5
2	6
3	7
4	8
5	10
6	12
7	14

2.2 Question 2: Predicting with a Regression Tree

A decision tree is trained on the dataset:

X	Y
2	3
3	5
5	7
7	10
9	12

The tree makes a split at $X = 4$ and another at $X = 8$. The leaf nodes predict the **mean** of the points in their region.

- What would be the predicted output for $X = 6$?
- If a new point $X = 10$ with $Y = 15$ is added, how does the tree change?

2.3 Question 3: Constructing a Regression Tree

You are given the dataset:

X_1	X_2	Y
1	3	5
2	4	6
3	2	7
4	6	9
5	5	11

- Construct a regression tree with **maximum depth of 2**.
- Use **Mean Squared Error (MSE)** to determine the best splits.
- Compute the predicted value for $(X_1 = 4, X_2 = 4)$.

3 Decision Tree Classification (Numerical)

3.1 Question 4: Gini index

A dataset contains the following points:

X_1	X_2	Class
1	2	A
2	3	A
3	1	B
4	2	B
5	3	A

A decision tree considers splitting at $X_1 = 2.5$.

- Should we split on $X_1 = 2.5$ or $X_2 = 2.5$?

3.2 Question 5: Gini Impurity

A dataset has three classes: **Red, Blue, and Green**. The current node contains:

- 10 Red
- 5 Blue
- 5 Green
- Compute the **Gini impurity**.
- If the node is split into:
  - Left child: 6 Red, 2 Blue, 2 Green
  - Right child: 4 Red, 3 Blue, 3 Green

  Compute the **weighted Gini impurity** for this split.

4 Bagging and Random Forest

4.1 Question 6: Bootstrapping for Bagging

A dataset has **6 points**: **(A, B, C, D, E, F)**.

- Generate a **bootstrapped sample** of size 6.
- If 10 bootstrapped samples are taken, what is the probability that a specific data point (e.g., A) is **not included in any sample**?

4.2 Question 7: Random Forest Classification

A random forest consists of **5 decision trees**, each trained on a subset of the dataset. Given the following tree predictions for an input:

Tree 1	Tree 2	Tree 3	Tree 4	Tree 5
A	A	B	B	A

- What will the final class prediction be using **majority voting**?

4.3 Question 8: Bias-Variance Tradeoff

Consider **Decision Tree, Bagging, and Random Forest**.

- Which model has the **highest variance**?
- Which model has the **lowest bias**?