Westlake University
Department of Artificial Intelligence, SOE
Tao Lin & Kaicheng YU
https://github.com/LINs-lab/course_machine_learning

Problem Set Lab 07, April 10, 2025 (Nonparametric Methods)

Goals. Goals. The goal of this week's lab is to:

- Understand the behavior of different non-parametric models.
- Learn how hyperparameters (*e.g.*, k, $depth$, *n_estimators*) affect bias-variance tradeoff.
- Gain experience using real datasets and visualizing model decisions.
- Develop skills in interpreting visual output like decision boundaries, feature importance plots, and prediction scatter plots.

Submission instructions:

- Please submit a PDF file to canvas.

1 Theory Exercises

Problem 1 (Parametric vs. Non-Parametric Models: KNN, Decision Tree, Random Forest):
In this exercise, we study the distinction between parametric and non-parametric models through three widely used learning algorithms: K-Nearest Neighbors (KNN), Decision Tree, and Random Forest.

1. Define the concepts of *parametric* and *non-parametric* models. Provide one example of each, excluding KNN, Decision Tree, and Random Forest.
2. For each of the following models: KNN, Decision Tree, and Random Forest:
   - State whether the model is parametric or non-parametric.
   - Justify your classification based on the structure or behavior of the model.
3. Explain why non-parametric models generally have higher variance compared to parametric models. How can we mitigate this issue in practice?

Problem 2 (Bias-Variance Trade-off in KNN, Decision Tree, and Random Forest):
In this exercise, we study the bias-variance trade-off in the context of three non-parametric models: K-Nearest Neighbors (KNN), Decision Tree, and Random Forest.

1. Briefly explain the bias-variance trade-off and its role in supervised learning.
2. For each model listed below, explain how its key hyperparameters affect bias and variance:
   - K-Nearest Neighbors (K)
   - Decision Tree (depth)
   - Random Forest (number of trees and tree depth)
3. Explain why bagging (bootstrap aggregating) helps reduce variance in Random Forests. How does it enhance generalization compared to a single Decision Tree?

2 KNN Classification and Random Forest Regression

Exercise 1:

Visualize the decision boundaries of k-Nearest Neighbors with different values of k.

- In this exercise, you will use the `load_iris()` dataset and reduce its dimensionality to 2 using PCA. Then, implement the function `plot_knn_decision_boundary(X_train, y_train, X_test, y_test, k)`.
- This function should:
  - Train a k-NN classifier with the specified number of neighbors k.
  - Compute and print the training and test accuracy.
  - Generate a contour plot of the decision boundary in the PCA-projected 2D space, using `matplotlib`.
- Run the function for $k = 1$, $k = 5$, and $k = 15$. Each run should produce a decision boundary plot and display the corresponding accuracies.
- Comment on the effect of k in a markdown cell: how does the decision boundary change as k increases? How does it relate to the bias-variance tradeoff?

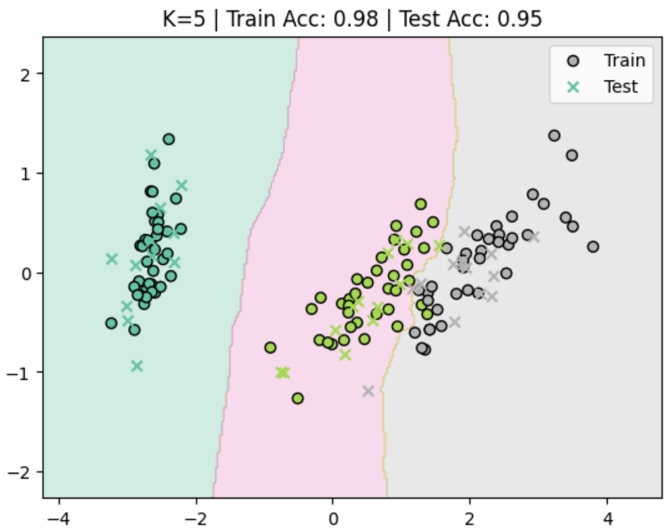

Figure 1: Decision boundary of 5-NN on the Iris dataset (after PCA).

Exercise 2:

Compare the performance of a Decision Tree and a Random Forest on a real-world regression task and visualize feature importances.

- In this exercise, you will use the `fetch_california_housing()` dataset from `sklearn.datasets`. Split the data into training and test sets using an 7: 3 ratio.
- Implement the function `compare_rf_vs_dt()` to perform the following steps:
  - Train a `DecisionTreeRegressor` with `max_depth=10` and a `RandomForestRegressor` with `n_estimators=100`.
  - Compute and print the test MSE for both models.
  - Plot two scatter plots of predicted values versus true values for both models (i.e., y_{true} vs y_{pred}).
  - Generate a horizontal bar chart that displays the feature importances from the Random Forest model.
- Comment on the generalization capability of each model based on the test MSE and the plots. Which features appear most important for predicting house prices?

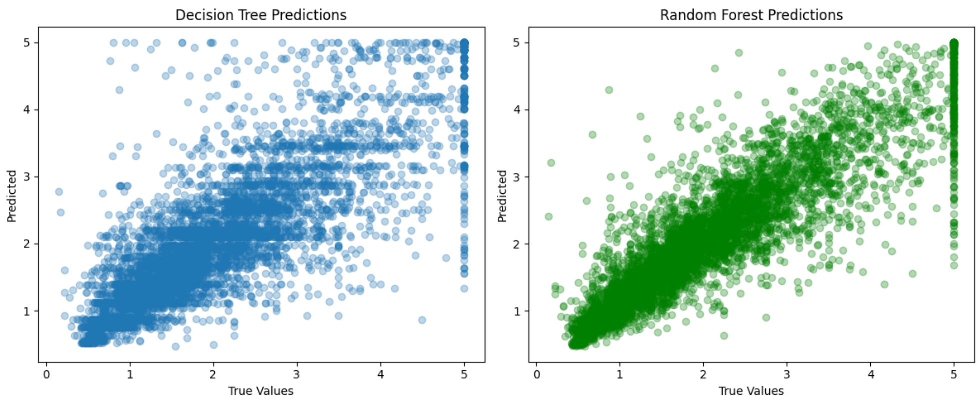

Figure 2: Scatter plots of predicted vs true values for Decision Tree and Random Forest.

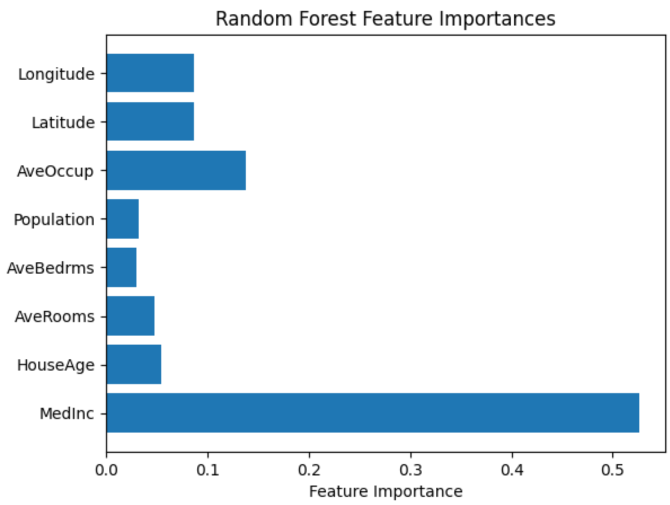

Figure 3: Feature importances from the trained Random Forest on the California Housing dataset.