Prediction Part II

Classification, Tree, Ensemble Learning

Classification v.s. Regression

From previous lectures:

Problem type	Output Space	Example Tasks
Regression	Numeric (continuous or *discrete*)	House prices, number of accidents
Classification	Categorical	Spam/non-spam, disease/no disease

A (Sometimes) Blurred Boundary

Consider the following outcomes. Are they classification or regression?

1. Hand-written digit recognition (0 to 9).
2. Housing price.
3. COVID-19 PCR test.
4. Weather forecast: probability of precipitation.

Decision tree

Questions

- What is a decision tree? **Tree (graph theory)**
- When and why would you use a decision tree (instead of other models)?

CART

- CART stands for **C**lassification **A**nd **R**egression **T**ree introduced by Breiman, Friedman, Olshen & Stone (1984).
- CART is an *algorithm* for growing a decision tree (perhaps the most popular algorithm)
- **More discussion in jupyter notebook**

*Further reading: An insightful note by Leo Breiman on "**two cultures in the use of statistical modeling to reach conclusions from data**".*

Ensemble Learning

Ensemble learning: methods that aggregate multiple models to improve learning performance.
“Many weak models, when combined cleverly, beat a single strong model.” --- The ensemble learning mantra by *ChatGPT 3o*

Tree-based Ensemble Learning

A single decision tree is

- ✅ fast and interpretable
- ❌ Sensitive and prone to overfitting

Overcome these flaws using ensembel learning:

Method	Main idea	What it mainly reduces
Boosting	Build trees **sequentially**, each fixing the last model's errors.	Bias
Bagging	Build trees **in parallel** on bootstrapped data; vote/average.	Variance
Random Forest	Bagging **plus** feature-level randomness at each split.	Variance (even more)