Business Analytics

Session 4a. Classification and Regression Trees

Renyu (Philip) Zhang

New York University Shanghai

Spring 2019

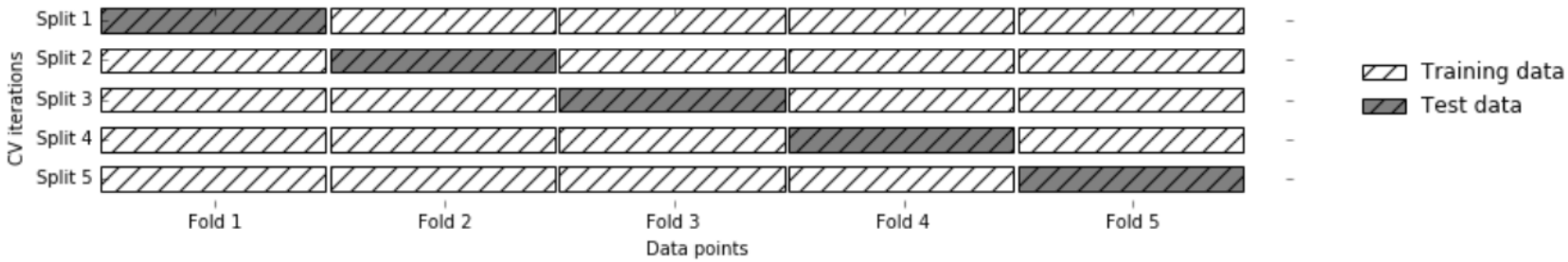
Warm-up Exercises

- In logistic regression, what will happen if we increase the classification threshold t?
  - FN error rate will increase and FP rate will decrease.
- For a classification problem,

	Predicted=0	Predicted=1
Actual=0	24	6
Actual=1	8	12

  What is the FN rate? What is the FP rate?
  - FN Rate=40%, FP Rate=20%.
- What is the underlying assumption for the k−nearest neighbors approach?
  - Distance between covariates reliably represents their similarities.
- If the model complexity increases, what will happen to its bias and variance in its generalization error?
  - The bias will decrease whereas the variance will increase.

K−fold Cross Validation



1. Split the training set into k pieces.

2. For each candidate model, use $k-1$ folds to fit/train the model, and test the model on the remaining fold (the test set).

3. Pick up the model with the smallest average error of the k-folds. Then, re-train the model on the whole training set.

Healthcare Cost Management for D2Hawkeye

Healthcare Cost Management

- D2Hawkeye tries to predict the healthcare costs of patients.
- Medical costs often related to severity of health problems, an issue for both patients and physicians.
- Goal: Improve the quality of cost predictions.

Pre-Analytics Approaches

- Human judgment.
- Limited data sets.
- Costly and inefficient.
- Question: How can we use analytics to improve the prediction?

Data

- Claims data
  - Requests for reimbursement submitted to insurance companies or state-provide insurances.
  - Rich and structured
  - Very high-dimensional
  - Does not capture all the relevant information (many things have to be inferred)
  - Does not reveal the results of a test, only that a test was administered
- Eligibility information

- Demographic information

Variables

- We are most interested in which cost bucket this patient is in?
  - $Y_i = 1$ (low): $<$\$3000
  - $Y_i = 2$ (emerging): \$3000-\$8000
  - $Y_i = 3$ (moderate): \$8000-\$19000
  - $Y_i = 4$ (high): \$19000-\$55000
  - $Y_i = 5$ (very high): $>$\$55000
- Covariates:
  - Whether the patient has certain diseases
  - Age
  - Reimbursement information in 2008 and 2009
  - Cost bucket information in 2008

Error Measures

- Typically, we use overall accuracy.
- In the case of D2Hawkeye, failing to classify a high-cost patient correctly is worse than failing to classify a low-cost patient.

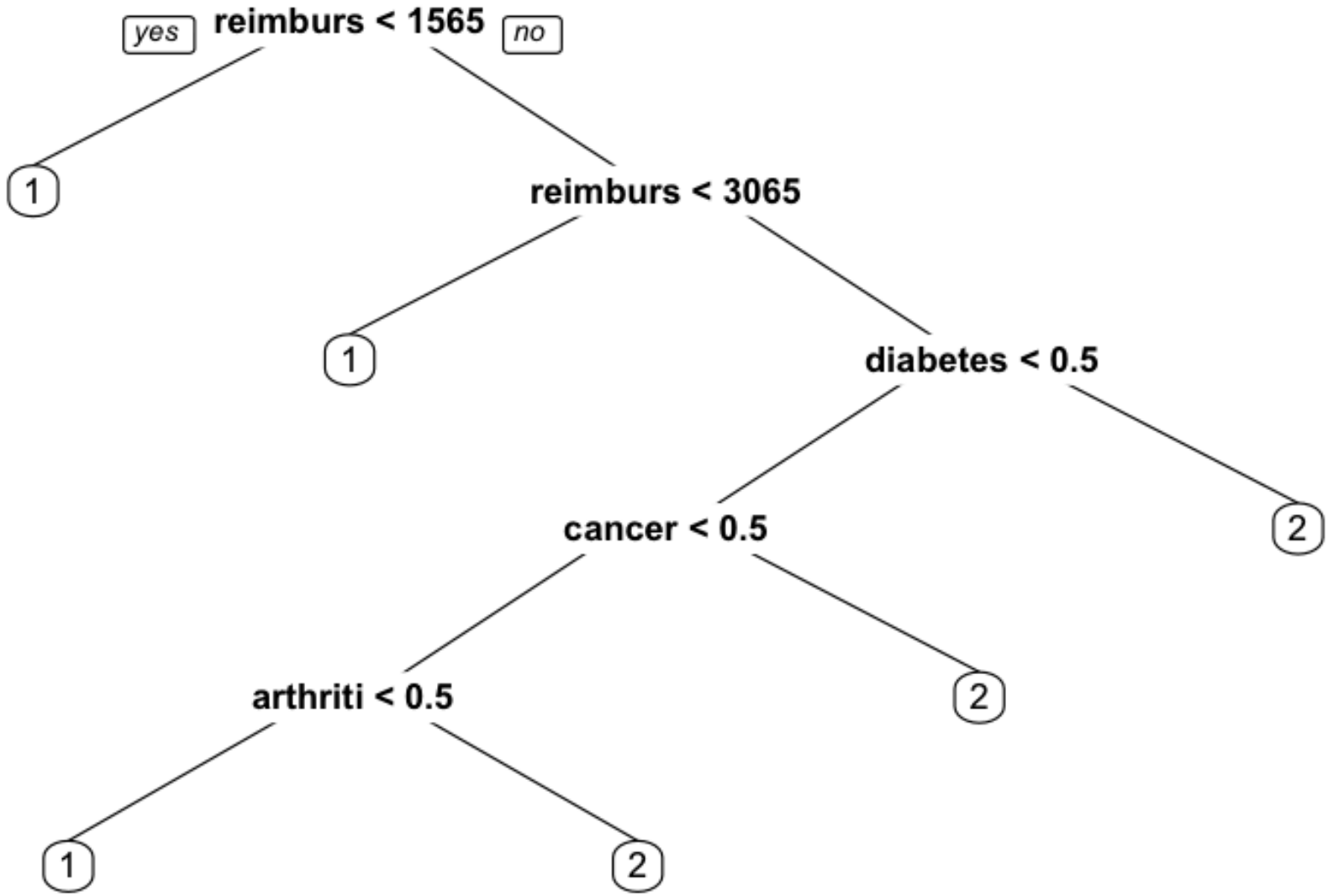
Penalty Matrix (the cost of making wrong predictions):

	Forecast=1	Forecast=2	Forecast=3	Forecast=4	Forecast=5
Actual=1	0	1	2	3	4
Actual=2	2	0	1	2	3
Actual=3	4	2	0	1	2
Actual=4	6	4	2	0	1
Actual=5	8	6	4	2	0

Baseline Forecast

- Essentially the naïve forecast: Y_i =Cost bucket of patient i in 2008.
- Overall Accuracy=68.4%.
- Penalty Error = 0.739

Classification and Regression Tree (CART)



Interpreting CART

- Nodes close to the root: Most important factors.
- Nodes near the leaves: Secondary factors.
- Examples of predicting $Y_i = 1$:
  - reimbursement$2008 < 1565$
  - reimbursement2008 between 1565 and 3065
  - reimbursement$2008 \geq 3065$ and no diabetes, no cancer, no arthritis
- Overall accuracy: 71.0% (68.4%), Penalty error: 0.774 (0.739).
  - Physicians were able to improve the model by identifying new variables and refining existing variables.
  - Very imbalanced data. How to address it? (We will talk about it later.)

Parameter Tuning

- Limit the tree size to address the over-fitting issue.
  - max_depth
  - max_leaf_nodes
  - min_samples_split
  - min_impurity_decrease
- Use k-fold cross validation to tune the parameter.
- Post-pruning: Create a pure tree first and pune some leaves the minimize the validation errors.

CART Trees

Classification Method Revisited

- Logistic regression: Essentially a linear classifier.
- k−nearest neighbors: Too slow when the training data set is large.
- CART: Exploiting a tree structure to implement the idea of k−nearest neighbors.

Building a CART

- Goal: Build a tree that is
  - Maximally compact;
  - With pure leaves only.
  - Bad news: This is NP-hard.
  - Good news: We have a good approximation strategy.

- p_k : The fraction of data points with $Y_i = k$ ($k \in \{1, 2, \cdots, C\}$).

$$\text{Gini Index of a set S: } G(S) = \sum_{k=1}^{C} p_k(1 - p_k)$$

$$\text{Gini Index of a tree T: } G^T(S) = \frac{|S_L|}{|S|} G^T(S_L) + \frac{|S_R|}{|S|} G^T(S_R),$$

  where S_L is the left subtree, S_R is the right subtree.

- Gradually grow the tree until the stopping criterion is met.
- Intuition: To increase the "purity" of each leaf as much as possible.

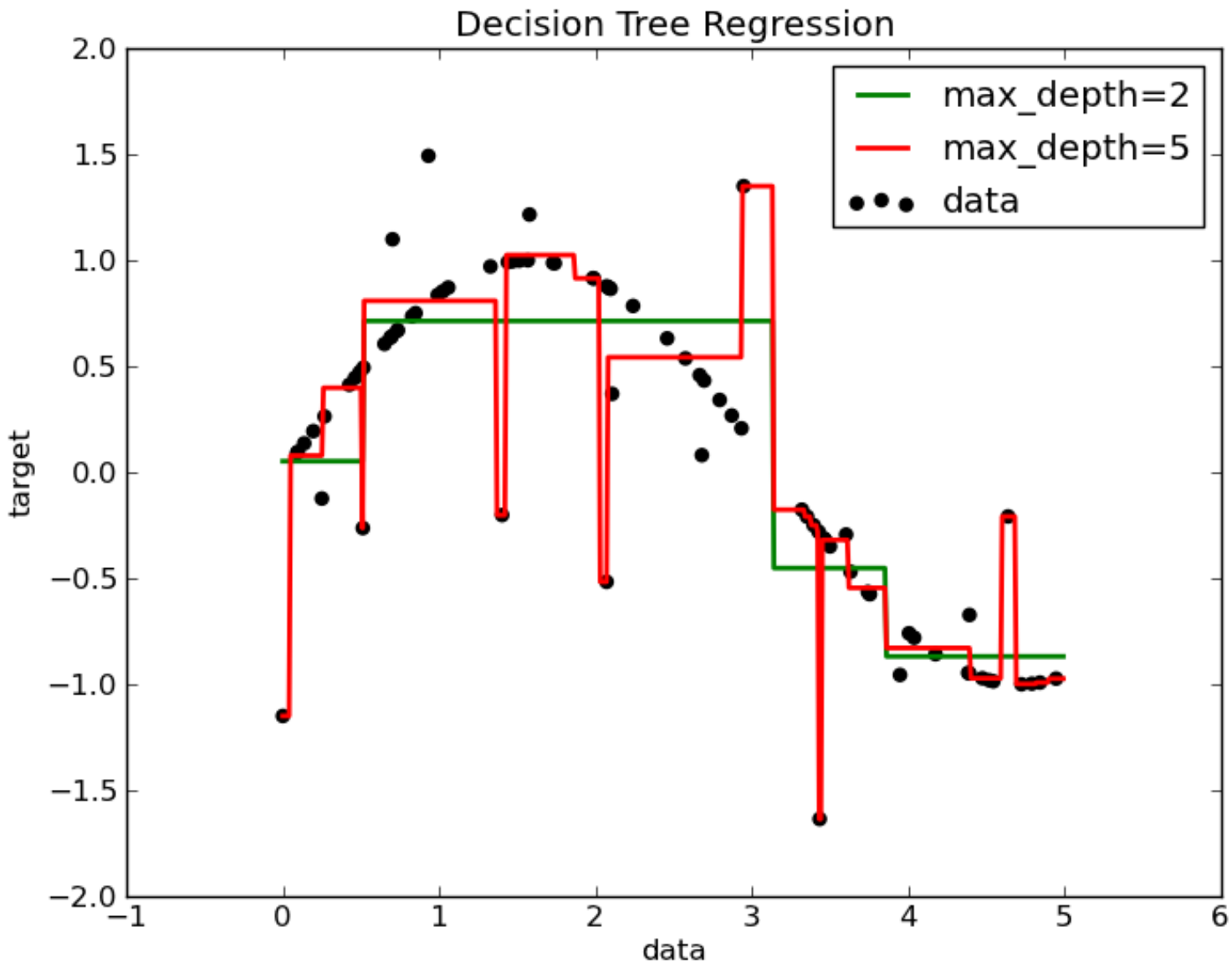
Regression Tree

- The outcome Y_i is a continuous variable.
- Minimize the impurity: Squared losses.

$$\mathcal{L}(S) = \frac{1}{|S|} \sum_{(X,Y)\in S} (Y - \bar{Y}_S)^2, \text{ where } \bar{Y}_S = \frac{1}{|S|} \sum_{(X,Y)\in S} Y$$

- Gradually grow the tree until some stopping criterion is met (e.g., each leaf has fewer than some minimum number of observations).
  - Some times need to post-prune the tree to reduce over-fitting

Regression Tree Illustration



Random Forests

Random Forests

- Build a large number of CARTs based on the training data set.
- Each tree is built from a "bootstrapped" sample of the data.
  - Select random samples from the training data set **with replacement**.
  - Build a CART on each random sample; each tree can split on only a random subset of the variables.
- Each tree makes a prediction for a new observation. The RF model picks the outcome receiving the majority of the votes.
  - The RF model is less interpretable than CART.
  - Reduces the variance of the predicted value.