DS 100: Principles and Techniques of Data Science **Date: December 4, 2019**

Discussion #15

Name:

Decision Trees and Random Forests

1. (a) When creating a decision tree for classification, give two reasons why we might end up having a terminal node that has more than one class.

overlap in training set, if we specify max depth

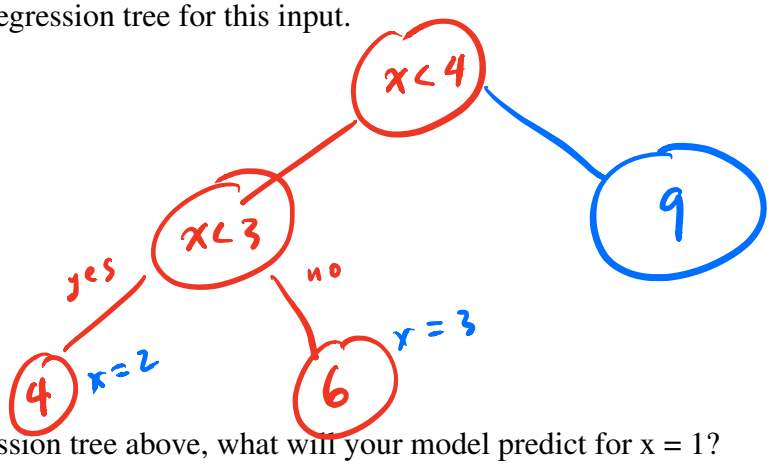
(b) Suppose we have a decision tree for classifying the iris data set. Suppose that one terminal decision tree node contains 22 setosas and 13 versicolors. If we're trying to make a prediction and our sequence of yes/no questions leads us to this node, what should we do?

- ◉ predict that the class is setosa
- ~~◯ give a probability of setosa = $\sigma(22/35)$~~
- ~~◯ refuse to make a prediction~~
- ◯ other (describe)

(c) As mentioned in lecture, we can also use decision trees for regression. Suppose we have the input table given below, where x is our 1 dimensional input value and y is our output value.

x	y
2	4
3	6
4	8
4	10

i. Draw a valid regression tree for this input.



ii. For your regression tree above, what will your model predict for x = 1?

4

iii. For your regression tree above, what prediction do you think your model should predict for x = 4?

9

(d) What techniques can we use to avoid overfitting decision trees?

covered in slides

(e) Suppose we limit the complexity of our decision tree model by setting a maximum possible node depth d, i.e. no new nodes may be created with depth greater than d. What technique should we use to pick d?

cross validation ↑ hyperparam

(f) What is the advantage of a random forest over a decision tree?

☐ lower bias ☑ lower variability ☐ lower bias and variability ☐ none of these

Clustering

unsupervised ↓

2. (a) Describe the difference between clustering and classification.

↑ supervised

(b) The process of fitting a K-means model outputs a set of K centers. We can compute the quality of the output by computing the distortion on the dataset. A Data 100 student suggests that distortion is not well-defined when evaluating the output of our agglomerative clustering algorithm because the algorithm doesn't return centers, but simply labels each point individually. Is the student correct?

no — you can compute centroids

(c) Describe qualitatively what it means for a data point to have a negative silhouette score.

closer on average to points in a different cluster than its own cluster

Broader Impact

3. Serving advertisements is a two step process. The first is *audience selection*, where advertisers select a target audience, e.g. people between 18 and 25 who live in Berkeley who have "Ziegfield Follies" as one of their interests. The second step is *ad delivery*, where proprietary algorithms automatically serve the ad to a subset of the users that match the desired attributes.

(a) During the audience targeting phase, it is illegal for advertisers to target specific groups under certain circumstances. For example, the Fair Housing Act would make it illegal to target only Asian people for an apartment listing. One way that unscrupulous advertisers might get around this is to use a ”proxy”. Give some example proxies that might allow you to effectively target a specific race, ethnicity, or gender. By proxy, we mean ”a figure that can be used to represent the value of something in a calculation.”

(b) As mentioned in lecture, a study (https://arxiv.org/abs/1904.02095) came out in Spring of this year that presented findings of race and gender stereotyping and discrimination in the second phase of the algorithm, i.e. the ad delivery phase. By testing inputs, the researchers concluded that the algorithm skews delivery according to the ”relevance” for a particular user, often due to their race or gender, especially with housing and employment opportunities. This happened ”even when advertisers set their targeting parameters to be highly inclusive”. Journalist Sam Biddle (https://theintercept.com/2019/04/03/facebook-ad-algorithm-race-gender/) summarized the issue with: ”In other words, even in the absence of bigoted landlords, the advertising platform itself appears inherently prejudiced.”.

Some have suggested that such inherent prejudice is a reflection of reality, e.g. because women click ads for secretarial jobs far more often than men, it is only sensible for an advertising platform to learn such real world patterns and serve ads appropriately.

Sam Biddle, in the Intercept, responded to such sentiments with the quote below:

”What this shallow reasoning misses is that decisions about pertinence can become self-reinforcing; it’s foolish at best to think that women are more interested in secretarial work because they keep clicking the secretary ads, rather than that they click secretarial ads because it’s all Facebook will show them.”

Many like Sam Biddle have expressed worry that large black box algorithms (e.g. Facebook’s ad deliery platform) reproduce and reinforce structural inequality, marginalization, and privilege that is already present in society. Are you personally concerned with this issue? Why or why not?

(c) What would it take to make the Facebook ad delivery algorithm fair? Who should decide what it means to be fair within an algorithm for advertising? The corporation? The government? The public?