Reflection Week 4

Summary

This week's lectures mainly concerns ideas of fairness and social topics in machine learning, as well as more technical concepts of decision trees and naive Bayes in the field, mainly used in classification models.

Concepts

- Six sources of bias:
  - Historical bias: bias that are inherent in our accurate data because the world we live in favors certain demographics.
  - Representation bias: bias that appear when the training data does not contain representative samples of the true population.
  - Measurement bias: bias resulted from the misinterpretation and/or wrong implementation of the measurements / features in the dataset.
  - Aggregation bias: using models that do not accurately cover every group and handle them equally can create bias.
  - Evaluation bias: bias from the benchmark dataset we used to test our models against.
  - Deployment bias: caused by the shift in actual usage of the models from their initial intentions.
- Group fairness can be defined in many different approaches, including:
  - Fairness through unawareness: prevents the model from ever looking at the protected attribute, but does not work in practice as these attributes can be unintentionally inferred from others.
  - Fairness by statistical parity: matching demographic statistics can create fairness $\Pr(\hat{Y} = +|x_1) = \Pr(\hat{Y} = +|x_2)$, which aligns with legal definitions of equity but may allow bias towards certain groups.
  - Fairness by equal opportunity: false-negative rate should be equivalent across groups, which effective controls the true outcome but unfortunately, this measure only protects this group.
  - Fairness by predictive equality: same thing as predictive equality but for false positive rate.
- 4 reasonable conditions we want in a real world ML model:
  - Statistical Parity
  - Equal Opportunity (equality across false negative rates)
  - Predictive Equality (Equality across false positive rates)
  - Good accuracy of the model across subgroups
- Fairness-accuracy tradeoff: to make a model more "fair," models often have to decrease their accuracy by some amount in order to satisfy the conditions above, as historical bias is unavoidable.
- Spaces in machine learning:
  - Construct space: true quantities of interest (unobserved)
  - Observed space: data achieved through measurement of proxies from quantities of interested that are hoped to represent these quantities
  - Decision space: decisions of the model
- Worldview assumptions to achieve individual fairness:
  - What You See Is What You Get (WYSISWYG): observed space is a good representation of the construct space, and we can confidently use the observed space to make decisions. This assumption achieves individual fairness easily, but non-discrimination methods may impair equity goals.

    - Structural Bias + "We're all Equal" (WAE): if observed space cannot be a good representation of the construct space, consider protected groups to make big impacts in observed spaces. This will distort non-discrimination measures.
- Naïve Bayes Theorem:
    - Assumption: each feature makes an independent and equal contribution to the outcome
    - Bayes theorem allows us to calculate the probability of an event A given that B happens when we already know how often B happens given that A happens, how likely A happens and B happens.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

    - For an outcome y's conditional probability given a set of features x, we can address naïve Bayes Theorem by:

$$P(y|x_1, x_2, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

      For a specific example, we can think of x as a sentence, $y = \pm 1$ as the outcome, with x_i's as the individual words in the sentence, and $P(x_i|y)$ as the probability of time we see that word in the subset of words that are in the event space of y.
- Decision trees:
    - Can be used for both numerical and categorical features
    - Anatomy: includes branch/internal nodes that split into possible values of a feature, and leaf node as final decision (class value)
    - To select the best feature, calculate classification errors and select the split with the lowest value
        - step 1: let $\hat{y}$ as the class of majority of data in the node
        - step 2: calculate the classification error of predicting $\hat{y}$ for this data:

$$Error = \frac{\#mistakes}{\#data\ points}$$

    - Greedy algorithm:
        - Start at root node and calculate error: if classification for data at this node is perfect, then stop
        - Else, repeat split selection with the next stump
    - Threshold split for numerical data:
        - sort values of a feature $h_j(x)$ and let $\{v_1, v_2, \ldots, v_N\}$ denote sorted values
        - for $i = 1$ to $N - 1$:
            1. calculate the middle point between v_i and v_{i+1}
            2. compute classification error for the threshold split using this point $h_j(x) > t_i$
        - Choose the value t^* that has the lowest classification error

Concerns

- For the Paneto Frontier, do we have an algorithm on how to determine the frontier without having to find all possible models?
- How to prevent evaluation bias?