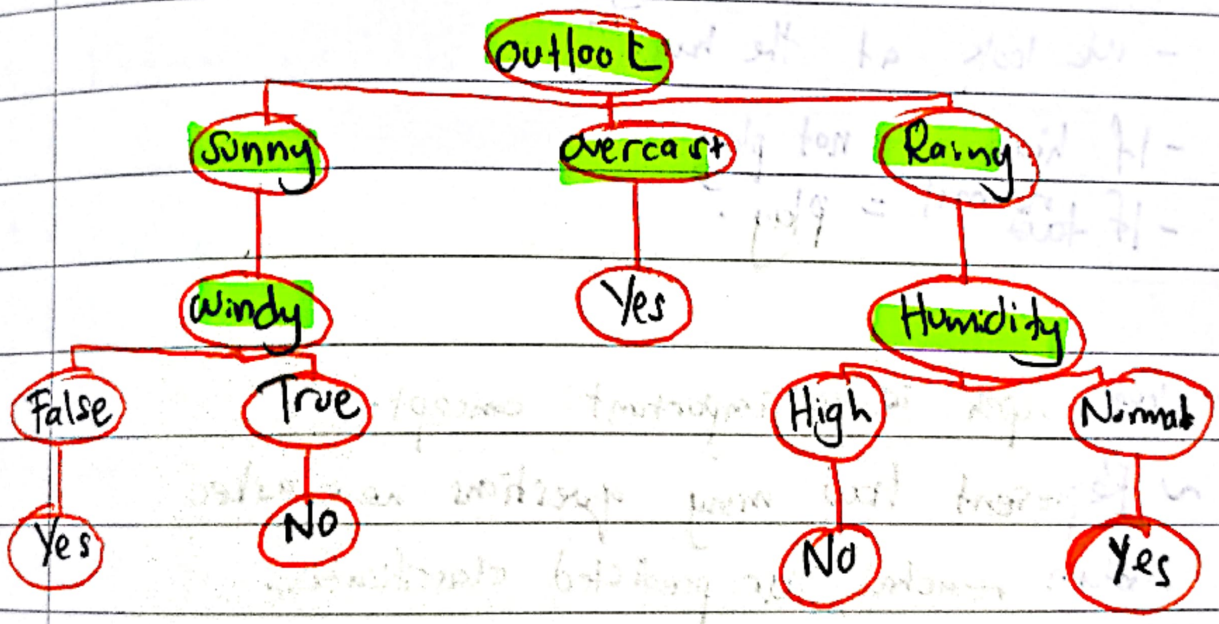
28/07/21

Machine Learning Algorithm

Decision Trees

- Trees answers sequential questions which sends a certain route of the given answer.
- It behave such as "if this than that".
- Example whether or not to play golf.



- outlook has 3 options : - sunny
  - overcast
  - Rainy

- Will it be windy ? (True / false ?)
  (if true, we choose not to play golf that day.)
  if False we play golf.

- If the outlook == overcast.
  - Play.

- If the outlook == rainy
  - we look at the humidity
  - If high = not play
  - If ~~food~~ normal = play.

- Tree depth is an important concept.
~ Represent how many questions were asked before reached our predicted classification.

- Sunny & Rainy Routes = depth of two
  Overcast = depth of One.

Advantages to using decision trees :-

1. Easy to interpret & make for straightforward visualizations.

2. Internal workings are capable of being observed & thus make it possible to reproduce work.

3. Can handle both numerical & categorial data.

4. Perform well on large datasets.

5. Are extremely fast.

Disadvantages of Decision Trees

- Requires algorithms capable of determining an optimal choice at each node.

- Hunt's algorithm is one of the popular.

- It makes most optimal decision at each step & not include the global optimum.

- Decision trees :- prone to overfitting, especially when a tree is deep.

- To overcome overfitting
  ~ set max depth
  ↓
- This will cause error due to bias.

- This will cause our model to be not a strong predictive model.

- To minimize bias and variance error we use
  ~ Random Forests.
  - Collection of decision trees, where the results are aggregated into one final result.

- Reduce variance by training on different samples of data.

(2) - Using a random subset of features.

- 'Random forest' are ~~more~~ stronger than single decision tree.
  - They aggregate many decision trees to limit overfitting as well as error due to bias & therefore yield useful results.

29/07/21

Understanding Random Forest

- Consists a large number of individual decision trees that operate as an ensemble.

  ↓ method of using multiple learning algorithm to obtain better predictive performance.

- Each individual tree in the random forest spits out a class prediction & the class with most votes becomes our model's prediction.

- low correlation between models is the key.

Prerequisites for random forest to perform well:

(1) There needs to be some actual signal in our features so that models built using those features do better than random guessing.

(2) The predictions (therefore the errors) made by the individual trees need to have low correlation with ~~eat~~ each other.

~~An Exam~~

What to do in order for our random forest to make accurate class predictions?

(1) We need features that have at least some predictive power.

(2) The trees of the forest and more importantly their predictions need to be uncorrelated.

- The algorithm tries to engineer these correlation using feature randomness.
- The feature we select & the hyper-parameters we choose will impact the ultimate correlations.