**Jeopardy Week 6 Questions:**

* **with regards to the bias/variance trade off: With bagging what are we trying to minimize? What about with boosting?**

-bagging trying to reduce variance

- boosting trying to reduce bias

* **Describe the difference between bagged decision trees and an extra trees model**

A bag of decision trees will have all of the features available to it at each split. An Extra Trees model will have a random subset of features available at each split, and will have randomly chosen decision boundaries



**What is the difference between a parametric and non-parametric model?**

With a parametric model like linear regression we have parameters: An intercept, and a coefficient per each variable. We use the state of the data to calculate these parameters and there are assumptions around how a parameteric model like linear regression works well. In order to make predictions, we simply need these parameters to make new predictions, not data.

For a non parametric model, predicting on new data is based on not just the parameters but also in the current state of data that has been observed. With decision trees, data is partitioned until it can accurately describe the target output for prediction and largely doesn't follow any distribution assumptions (ie: doesn't matter if data is normally distributed or linear relationships exist).

* **What does CART stand for?**

Classification and Regression Trees

* **Give 3 benefits of decision trees:**

They are not influenced by scale (Don’t need to scale data!)

They don’t care about the distribution of the data

Interpretable

Don’t require special encoding (no dummies encoding needed)

* **What is the main downside of Decision Trees?**
  + Decision trees are a high variance model and tend to overfit.

**Describe the relationship between gini, entropy, information gain and purity**.

We use **gini** or **entropy** to measure this idea of "purity" within a partion (bacon vs vegetable in this example). Another way to think about **purity**is how "unbalanced" the proportion of classes are. The more *unbalanced* a set classes are within a given partition, the more **pure** the partition is.

**What are 3 major hyperparameters you might want to tune over (with Gridsearch) when using a Random Forest Classifier.**

* + N\_estimators, criterion, max\_features, max\_depth, min\_sample\_split

**Write out the steps involved as information flows down a single decision tree:**

1. All data enters the root node.

-

2. We test all possible conditions for the best possible split, given the amount of information gain.[¶](http://localhost:8909/notebooks/6.01-trees-CARTs-lesson/1-intro-to-CART.ipynb#2.-We-test-all-possible-conditions-for-the-best-possible-split,-given-the-amount-of-information-gain.)

3. gini or entropy is used to calculate information gain.

4. A split is determined based on information gain. Data is partitioned.

5. Nodes that are still impure, can be "split" into new partions until purity is achieved.

**when we have a 50/50 split in our data – what is the value of our gini index? What is the value of our entropy index?**

* + - Gini= 0.5
    - Entropy = 1.0

**name all the regression and classification models we have learned so far.**

* + Regression:
    - Linear regression, knn regressor, decision tree regressor, random forest regressor, extra trees regressor,
  + Classification:
    - Logistic regression, knn, decision tree classifier, random forest classifier, extra trees classifier, SVC

**Name 2 things we can do if we have skewed target variable (more samples in one class than the other)**

Shift the decision boundry threshold, take the log of the target variable to make it have a more gaussian distribution

**What does bootstrapping mean?**

* + Taking random sub samples from the full data set with replacement

**In train test split, what does it mean to stratify?**

* The stratify parameter makes a split so that the proportion of values in the sample produced will be the same as the proportion of values provided to parameter stratify.
* For example, if variable y is a binary categorical variable with values 0 and 1 and there are 25% of zeros and 75% of ones, stratify=y will make sure that your random split has 25% of 0's and 75% of 1's.

**Can you determine how specific features play into your model when using an ensemble method, and if so, how?**

Yep, model.feature\_importances\_ - will give a ranking of how important each feature is in your model

**Give 2 Pros of boosting:**

**Pros**

* Achieves higher performance than bagging when the hyperparameters are properly tuned.
* Works equally well for classification and regression.
* Can use "robust" loss functions that make the model resistant to outliers.

**2 cons to boosting** **Cons**

* Difficult and time consuming to properly tune hyperparameters.
* Cannot be parallelized like bagging (bad scalability when there are huge amounts of data).
* Higher risk of overfitting compared to bagging.

**Generally, What does AdaBoost Do?**

The core principle of AdaBoost is to **fit a sequence of weak learners** (i.e., models that are only slightly better than random guessing, such as a single-split tree) **on repeatedly modified versions of the data**. After each fit, the importance weights on each observation need to be updated.

**Generally, what does Gradient Boosting do?:**

starts with a weak learner, second tree is trying to predict the residuals of the first tree. # then each subsequent tree is trying to predict the residuals. # aggregate all those together, then that's your prediction'

**Name 2 kernels you can use in SVM:**

Linear, rbf, polynomial, sigmoid, custom

**Define Linear Seperability:**

The data above were **linearly separable**. This simply means that they can be separated by a line.

**What does the hyperpameter C do in support vector classifer?**

C, as described above, is our "budget" that controls how imperfect our classifications are.

**What does C do when it’s large and small?**

* **If C is small**: We get a less flexible boundary between our classes, leading to a less perfect classification of our training data.
* **If C is large**: We get a more flexible boundary between our classes, leading to a more perfect classification of our training data.

**What is the role of alpha in adaboost?**

Alpha is the contribution weight of the weak learner All training examples start with equal importance weighting. When we finish training a classifier, we update the importance weighting of the classifier, represented by alpha.

**How can we determine if a model is overffit? Give 2 examples.**

Compare the training and testing scores

For regression, compare the RMSE

For classification, compare F1 scores

**Do we need to use cross val score when using gridsearch?**

Nope – cv is included in gridsearch