Decision Trees

Most Decision Trees

Each question node:

- Asks a question about a single feature
  - Categorical variables
  - Continuous variables (with a defined threshold)
- Sorts the samples into two sets, corresponding to the ***Yes*** and ***No*** branches
  - Other perspective: cuts a part of the feature space into two pieces

Leaf nodes: assign an output (class label, probability or continuous value)

Decision Tree Construction

- Classifier: measure performance in terms of the quality of distinctions that are made (Gini Impurity / Entropy)
- Learning process: incremental / greedy
  - Add a new question that improves the performance the most
  - Which feature and (if continuous) which threshold?

Decision Tree Construction

Overfitting is a challenge here, too. Combat with regularization:

- Maximum tree depth / maximum number of leaf nodes
- Minimum number of samples in a node to be split
- Statistical tests:
  - E.g., Likelihood ratio test: does the improvement in performance (as measured by data likelihood) justify the additional parameters required to encode the new expansion?

Sir Francis Galton (1822-1911)

- Meteorology: first weather maps
- Statistics: regression
- Psychology
- Heredity
- …

Weighing a Cow

Weighing a Cow

- Individually, non-experts are generally not good at guessing the weight of a cow
- However, the distribution is ~Normal, with a mean very close to the true weight

Message: Measures from a large set of ***independent***, poor-quality predictors can give us a high-quality prediction

Mixing Many Noisy “Experts”

Ensemble-based methods:

- Create many models
- Combine the predictions of these models
  - Classifiers: voting (soft or hard)
  - Regression: some mechanism for blending the predictions (e.g., computing a mean)

- This really only works if the models provide independent assessments for the query samples

Forcing Independence

- Train each tree with a subsample of the training set:
  - Bagging: sample with replacement
  - Pasting: sample without replacement
- Sampling features:
  - Random subspaces: only use a subset of the available features for a given tree

Forcing Independence

Adding noise to tree construction. For each possible split:

- Random forest: consider only a small subset of the available features
  - Useful when there are many features possible or many possible questions
- Extra trees: consider only a subset of possible thresholds (or question parameters)

Feature Importance

Which of the features is actually important to making prediction about the data? Common approaches:

- Reduction of impurity for questions involving a specific feature
- How often does a feature occur in a tree?
- Where does a feature occur in a tree?
- Importance sampling: how does the model perform when an individual feature is corrupted

Feature Importance

Getting this right can:

- Help domain scientists focus their models
- More efficiently construct models in the future
- Refine our data collection / storage processes

Forests

So far: training of one tree is independent of another

- Natural parallelization
- Independence to varying degrees

Boosting

Alternative approach:

- Grow trees in sequence
- The current tree attempts to repair prediction errors of the prior trees
  - Each new tree is solving a new piece of the problem

The cost: lose parallelization