Random Forest

Oliver Zhao

1 Description

Random Forest classifiers are generated largely through the same methodology of the Bagging classifier. Consequently, refer to the document titled "Bagging" for more information.

The central difference is that when generating Decision Trees, only a subset of features is considered for each node split. In contrast, generated Decision Trees in Bagging considers all features for each node split. In short, Bagging subsamples the training data, while Random forest subsamples the feature space.

The main advantage of Random Forests is that variance is decreased by reducing correlation between Decision Trees, due to the random selection of a feature subset for each node split.

2 Algorithm

Description: Consider a training set T with k features $X_1, \ldots, X_k$.

---

1. Using random sampling with replacement, generate n feature subsets $F_1, \ldots, F_n$ from the feature space $\mathbf{X} = \{X_1, \ldots, X_n\}$.
2. For each feature subset F_i, generate a classifier C_i. See the algorithm for Decision Trees for how each classifier is generated.
3. Find the predictions of each classifier $C_1, \ldots, C_n$.
4. The master classifier determines which class received the most votes, to determine the final output.
