Ensemble Methods

Chao Lan

Ensemble methods build a strong model by assembling a pool of (weaker) models. The pool is a committee, and each member is a base model. If all base models are from the same family (e.g., logistic regression), we have a homogeneous committee; otherwise, we have a heterogeneous committee. We will focus on homogeneous committee.

An ensembled model can have nonlinear decision boundary, even if each base model has a linear boundary. There are two common approaches to ensemble models: bagging and boosting.

1. Bagging

Bagging builds an ensemble model f from a set of base models $f_1, \ldots, f_m$ in the following way:

$$f(x) := \frac{1}{m} \sum_{k=1}^{m} f_k(x). \tag{1}$$

Each f_k is learned from a bootstrap sample, which is random subset (sampled with replacement) of the training set.

[Discussion] Example of bootstrap sample.

[Discussion] Example of bagging model.

The perhaps most famous bagging model is random forest, which is an ensemble of specially constructed decision trees – each tree has an additional constraint that every node split is based only on a random subset of features (no longer the entire feature set).

[Discussion] Example of random forest.

[Discussion] Why does random forest introduce that additional constraint?

Boosting

Boosting learns base models in a sequential way. A most popular algorithm is adaboost. Its basic idea is to assign higher α_k to more accurate f_k, and train each f_k using a weighted training set where previously mis-classified instances have higher weights. Detailed algorithm of adaboost is in Algorithm 1, assuming label is $\{-1, +1\}$.

Algorithm 1 AdaBoost

Input: training sample $S = \{x_1, \ldots, x_n\}$, committee size m
Initialize: weight $w_i = 1/n$ for instance x_i
for $k = 1, \ldots, m$ **do**
  1: train base model f_k on S by minimizing the following weighted loss $J(f_k) = \sum_{i=1}^{n} w_i \cdot \mathbf{1}_{f_k(x_i) \neq y_i}$.
  2: compute model weight $\alpha_k = \ln \left\{ \frac{1-\epsilon_k}{\epsilon_k} \right\}$, where $\epsilon_k = \frac{\sum_{i=1}^{n} w_i \cdot \mathbf{1}_{f_k(x_i) \neq y_i}}{\sum_{i=1}^{n} w_i}$.
  3: update instance weight $w_i = w_i \cdot \exp\{\, \alpha_k \cdot \mathbf{1}_{f_k(x_i) \neq y_i} \,\}$.
end for
Output: an ensembled model $f(x) := \sum_{k=1}^{m} \alpha_k f_k(x)$.