3.5 Ensemble Learning

Introduction

Multiple models (**base models**) each of them obtained by applying a learning process to a given problem, combined to make a single prediction, then combined to obtain the final prediction of the ensemble.

Intuition

Aggregation of multiple learned models with the goal of improving model quality

Ensemble learning process

1. Generation from input space
2. Pruning
3. Integration

Advantages

- **Accuracy** - majority compensates for individual errors
- **Diversity is key** - individual models specialize in different areas of the data space, but must be reasonably accurate

Disadvantages

- **Complexity** - understanding the global model, explaining decisions, computational
- **Occam's Razor** - simplicity leads to greater accuracy; identifying the best model requires identifying the proper "model complexity"

Categories of Methods

Homogeneous

- Single induction algorithm

Model Combination

Diversity comes from model combination:

- **Regression**: average, weighted average, sum, weighted sum, product, maximum, minimum, median;

- **Classification**: majority voting, weighted majority voting, borda count (based on preference ranking and voting).

Different Models

- **Data manipulation** (training set):
  - Manipulating input features;
  - Sub-sampling from the training set.
- **Modelling process manipulation**:
  - Manipulating the induction algorithm (variants of the same algorithm, otherwise heterogeneous);
  - Manipulating the parameter sets;
  - Manipulating the model (uncommon).

Heterogeneous

- Multiple induction algorithm
- Won't focus, but the same techniques are essentially applicable to heterogeneous ensemble

Popular Methods

Bagging: Bootstrap AGGregatING

- **Diagnosis analogy**: based on the majority vote of multiple doctors
- **Training**: at each iteration, training set is sampled with replacement from the original set (i.e. bootstrap), and model is learned from the training set
- **Prediction**: given an observation, make a prediction for each classifier and aggregate the predictions
- **Tasks**: classification and regression;
- **Accuracy**: often significantly better than a single classifier derived from D; robust to noise
- If classifier is unstable (a small change to the training data may lead to major decision changes): decision trees or neural networks.

Boosting

- **Training**: equal weights are assigned to each training example; learn first model and, for every following iteration, give more weight to the examples that were incorrectly predicted by the previous;
- **Prediction**: aggregation of the predictions, the weight of each classifier's vote is a function of its accuracy;
- **Task**: classification;

Boosting vs. Bagging:

- Independent sampling vs. error-dependent sampling;
- Uniform aggregation vs. weighted aggregation.
- Boosting tends to achieve greater accuracy but risks overfitting the model to misclassified data.

Random Forest

- **Training**: learn k models with changed algorithm (at each spli, randomly select a subset of the original features for tree generation);
- **Prediction**: aggregation of the predictions;
- **Task**: classification and regression;

RF vs. adaboost

Comparable in accuracy, more robust to errors and outliers;

RV vs. bagging and adaboost

Insensitive to the number of attributes at each split, faster.

Negative Correlation Learning

- **Training**: learn k models with changed algorithm (trained to minimise error function of the ensemble, i.e., it adds a penalty term with the average error of the models already trained to the error function);
- **Prediction**: aggregation of the predictions;
- **Task**: only regression, algorithms that try to minimise/maximise a given objective function (e.g. neural networks, support vector regression);
- Models negatively correlated with the averaged error of the previously generated models.

Issues

Classification

- Base classifiers should be as accurate as possible, although there is "the strength of weak classifiers
- Having diverse errors

Regression

- More amenable to theoretical analysis;
- The goal is to miminimize

- The average bias: the base learners should be as accurate (on average) as possible;
  - The average variance: the base learners should be as robust to small changes on the training data (on average) as possible;
  - The average covariance: the base learners should have negative correlation.

< Go back