8

COMBINING METHODS AND ENSEMBLE LEARNING

We are each of us angels with only one wing,
and we can only fly by embracing one another.
Lucretius

Most of the learning methods presented in earlier chapters assume that a set of admissible models, or approximating functions, $f(\mathbf{x}, \omega)$ is specified a priori. For example, this is a linear combination of sigmoid functions (of projections) for MLP neural networks, whereas for SVM methods it is a linear combination of kernel functions. Such learning methods are defined by:

- a flexible model parameterization $f(\mathbf{x}, \omega)$,
- a loss function $L\left(y, f(\mathbf{x}, \omega)\right)$,
- an optimization method for minimizing average training error using this loss function.

Then a predictive model $f(\mathbf{x}, \omega^*)$ is estimated via single application of a learning method to available training data $(\mathbf{x}_i, y_i)$ $i = 1,2,..n$.

Methods discussed in this chapter follow a different approach, where many component (or 'base') models are evaluated first, and then they are

Predictive Learning
By Vladimir Cherkassky
Copyright © 2013 Vladimir Cherkassky

combined to produce the final predictive model. Aggregating or averaging several models may have important benefits, such as avoiding overfitting, improving flexibility and producing more stable models. We have already presented one such approach called Bayesian averaging in Chapter 3, where it was related to the philosophical Principle of Multiple Explanations. Collective decision making is common in our daily lives. Examples include:

- Politburo, the principal decision-making committee of a Communist country;
- jury trial in the American legal system;
- seeking multiple medical expert opinions for a major medical procedure.

There are two main motivations for this approach. First, several expert opinions may be better than one. Second, collective opinion of many non-expert individuals is better than the judgment of a single person. This idea underlies the notion of Western democracy. Of course, collective decision making does not always guarantee the best outcome. For example, all revolutionary scientific discoveries are anti-democratic, because they contradict currently prevailing scientific wisdom. This has been brilliantly noted by Galileo who said:

> *In questions of science, the authority of a thousand is not worth the humble reasoning of a single individual.*

Similar ideas have been also explored in machine learning and statistics. This chapter presents several learning methods collectively known as *combining methods*, or *ensemble learning*. These methods have been developed in different fields, so our goal is to describe representative combining methods under a coherent framework. Hence, we focus on the underlying assumptions necessary for understanding relative strengths and limitations of these techniques. All the combining methods presented in this chapter follow a standard inductive learning setting, where the goal is to estimate a predictive model from the training data. Under this setting, there are two general strategies leading to combining methods:

1. *Apply different learning methods to the same training data* and then combine individual predictors (estimated by each method). The rationale is that no single method can consistently yield the best model for every data set, so combining individual predictors may improve generalization. This leads to the Committee of Networks approach (developed in neural networks), or the similar statistical technique called Stacking. These methods are presented in Section 8.1.

2. *Apply the same learning method to several modified versions of the training data* and then combine individual models. In this case, the learning method is often called the 'base learning method', and the component models are known as base models. This strategy may pursue two different objectives:
   - Apply a learning method to many *statistically identical* realizations of the training data, and average the results, in order to reduce model variability due to random variations of training samples. This idea is implemented in the bagging methods described in Section 8.2.
   - Another objective is to achieve flexible data fitting by applying a simple *base learning method* (called 'weak learner') to many *statistically different* realizations of the training data. In this case, the local features of the final model can be estimated (by a weak learner) by selectively assigning larger weights to the training samples responsible for the local variation. This approach is implemented in boosting methods discussed in Section 8.3.

The combining techniques presented in this chapter can be developed for classification or regression settings. For regression, the final predictive model is a (weighted) average of the component models:

$$F(\mathbf{x}) = \sum_{k=1}^{N} w_k f_k(\mathbf{x}) \tag{8.1a}$$

For classification problems, the final predictive model is defined via (weighted) majority voting of the component predictors:

$$F(\mathbf{x}) = sign\left(\sum_{k=1}^{N} w_k f_k(\mathbf{x}) \right) \tag{8.1b}$$

Different combining methods use different strategies for estimating individual models $f_k(\mathbf{x})$ and for calculating the combining weights w_k in (8.1). For some combining methods, these weights are positive and sum up to one. This leads to a probabilistic (Bayesian) interpretation, where the final model can be described by several potentially true component models. Then the combined model (8.1) is an average with weights reflecting the level of confidence in each component model. See more discussion on Bayesian averaging in Section 3.8.

Combining methods are algorithmic procedures that are conceptually quite different from the methods discussed in earlier chapters. Such earlier methods yield a *single analytic model*, such as linear regression, decision tree, or MLP network. This single model can be interpreted, in principle, even though some methods, such as decision trees, offer an easier interpretation than others. Combining methods make many passes over training data, and the final model $F(\mathbf{x})$ depends on the aggregated results from all passes. So interpretation becomes very difficult, even for low-dimensional models. Also, it is important to keep in mind that combining methods have been originally introduced using informal arguments. The initial motivation was to improve certain performance aspects of 'conventional' inductive learning algorithms, such as overfitting, stability, etc. As a result, there is no complete theoretical framework for combining methods, and no good theoretical understanding of the factors controlling their generalization. Discussion of these theoretical and methodological issues is provided in Section 8.4.

Finally, we point out important differences between ensemble methods in machine learning and collective decision making in social systems. That is, the methods described in this chapter aim to achieve improved prediction accuracy, under well-defined statistical assumptions (e.g., inductive learning setting). In contrast, the typical goals of such methods in social systems are: achieving consensus opinion and/or winning the final vote. As a result, individual voting decisions are not 'statistically independent'. Moreover, the voters' preferences often change according to the reflexive property of social systems (see Chapter 3), so they cannot be modeled by stationary distributions. This herd-like behavior is often created and actively managed in the course of political campaign using mass media and, more recently, digital social networks.