Title: Ensemble averaging (machine learning)

URL: https://en.wikipedia.org/wiki/Ensemble_averaging_(machine_learning)

PageID: 27669989

Categories: Category:Artificial intelligence engineering

In machine learning , ensemble averaging is the process of creating multiple models (typically artificial neural networks) and combining them to produce a desired output, as opposed to creating just one model. Ensembles of models often outperform individual models, as the various errors of the ensemble constituents "average out". [citation needed]

Overview

Ensemble averaging is one of the simplest types of committee machines . Along with boosting , it is one of the two major types of static committee machines. In contrast to standard neural network design, in which many networks are generated but only one is kept, ensemble averaging keeps the less satisfactory networks, but with less weight assigned to their outputs. The theory of ensemble averaging relies on two properties of artificial neural networks:

In any network, the bias can be reduced at the cost of increased variance

In a group of networks, the variance can be reduced at no cost to the bias.

This is known as the bias–variance tradeoff . Ensemble averaging creates a group of networks, each with low bias and high variance, and combines them to form a new network which should theoretically exhibit low bias and low variance. Hence, this can be thought of as a resolution of the bias–variance tradeoff. The idea of combining experts can be traced back to Pierre-Simon Laplace .

Method

The theory mentioned above gives an obvious strategy: create a set of experts with low bias and high variance, and average them. Generally, what this means is to create a set of experts with varying parameters; frequently, these are the initial synaptic weights of a neural network, although other factors (such as learning rate, momentum, etc.) may also be varied. Some authors recommend against varying weight decay and early stopping. The steps are therefore:

Generate N experts, each with their own initial parameters (these values are usually sampled randomly from a distribution)

Train each expert separately

Combine the experts and average their values.

Alternatively, domain knowledge may be used to generate several classes of experts. An expert from each class is trained, and then combined.

A more complex version of ensemble average views the final result not as a mere average of all the experts, but rather as a weighted sum. If each expert is y i {\displaystyle y_{i}} , then the overall result y ~ {\displaystyle {\tilde {y}}} can be defined as:

where α {\displaystyle \mathbf {\alpha } } is a set of weights. The optimization problem of finding alpha is readily solved through neural networks, hence a "meta-network" where each "neuron" is in fact an entire neural network can be trained, and the synaptic weights of the final network is the weight applied to each expert. This is known as a linear combination of experts .

It can be seen that most forms of neural network are some subset of a linear combination: the standard neural net (where only one expert is used) is simply a linear combination with all α j = 0 {\displaystyle \alpha _{j}=0} and one α k = 1 {\displaystyle \alpha _{k}=1} . A raw average is where all α j {\displaystyle \alpha _{j}} are equal to some constant value, namely one over the total number of experts.

A more recent ensemble averaging method is negative correlation learning, proposed by Y. Liu and X. Yao. This method has been widely used in evolutionary computing .

Benefits

The resulting committee is almost always less complex than a single network that would achieve the same level of performance

The resulting committee can be trained more easily on smaller datasets

The resulting committee often has improved performance over any single model

The risk of overfitting is lessened, as there are fewer parameters (e.g. neural network weights) which need to be set.

See also

Ensemble learning

References

Further reading

Perrone, M. P. (1993), Improving regression estimation: Averaging methods for variance reduction with extensions to general convex measure optimization

Wolpert, D. H. (1992), "Stacked generalization", Neural Networks , 5 (2): 241– 259, CiteSeerX 10.1.1.133.8090 , doi : 10.1016/S0893-6080(05)80023-1

Hashem, S. (1997), "Optimal linear combinations of neural networks", Neural Networks , 10 (4): 599– 614, doi : 10.1016/S0893-6080(96)00098-6 , PMID 12662858

Hashem, S. and B. Schmeiser (1993), "Approximating a function and its derivatives using MSE-optimal linear combinations of trained feedforward neural networks", Proceedings of the Joint Conference on Neural Networks , 87 : 617– 620