Statistical Learning and Data Mining

Module 14: Model Stacking

Semester 2, 2017

Discipline of Business Analytics, The University of Sydney Business School

Model stacking

- **Model stacking** is a learning method that aims to improve predictive accuracy by combining predictions from multiple models. It is a particular case of **ensemble learning**.

- It is a highly effective strategy that is a component of several winning entries in Kaggle competitions.

- As we approach the end of the unit, it is one of the most useful and yet simple strategies that we can add to your toolbox, which is why we discuss the general idea in this short module.

Model averaging

Consider the regression setting and suppose that we estimate M models $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_M$, leading to corresponding predictions $\widehat{f}_1(\boldsymbol{x}), \ldots, \widehat{f}_M(\boldsymbol{x})$.

In **Model averaging**, we compute the prediction

$$\widehat{f}_{\text{ave}}(\boldsymbol{x}) = \sum_{m=1}^{M} w_m \, \widehat{f}_m(\boldsymbol{x}),$$

where $w_1, \ldots, w_m$ are fixed model weights.

Model averaging

How to choose the weights? One option is to simply pick the model weights, say $w_m = 1/M$ for a simple model average. Or we can choose the coefficients by empirical risk minimisation,

$$\min_{w_1,\ldots,w_m} \sum_{i=1}^{N} \left(y_i - \sum_{m=1}^{M} w_m \, \widehat{f}_m(\boldsymbol{x}_i) \right)^2 ,$$

where we may want to impose restrictions on the weights such as non-negativity and a total sum of one.

Model averaging

However, empirical risk minimisation

$$\min_{w_1,\ldots,w_m} \sum_{i=1}^{N} \left(y_i - \sum_{m=1}^{M} w_m \, \widehat{f}_m(\boldsymbol{x}_i) \right)^2 ,$$

may not work well since the regression models $\widehat{f}_1(\cdot), \ldots, \widehat{f}_M(\cdot)$ were previously estimated based on the same training data.

The minimisation does not take into account the complexity of the individual models, and will tend to overfit by putting too much weight on the most complex models (which will have low training errors).

Model stacking

Stacking overcomes this difficulty by solving the minimisation problem

$$\min_{w_1,\ldots,w_m} \sum_{i=1}^{N} \left(y_i - \sum_{m=1}^{M} w_m \, \widehat{f}_m^{(-i)}(\boldsymbol{x}_i) \right)^2,$$

where $\widehat{f}_m^{(-i)}(\boldsymbol{x}_i)$ are leave one out predictions. That is, we exclude observation i when fitting $\widehat{f}_m^{(-i)}(\cdot)$.

Sometimes, we may want to reduce the computational cost by computing K-fold predictions instead.

Model stacking

In above case, model stacking fits a linear regression model

$$\min_{w_1,\ldots,w_m} \sum_{i=1}^{N} \left(y_i - \sum_{m=1}^{M} w_m \, \widehat{f}_m^{(-i)}(\boldsymbol{x}_i) \right)^2,$$

based on constructed predictors derived from different models. We say that the linear regression model is therefore the **meta model** for the stack.

Model stacking allows us generalise this idea and use any method as a meta model. For example the lasso, nonparametric regression, or tree-based methods.

Model stacking for classification

Model averaging and stacking extends to classification. Some options are:

- Classification based on weighted majority voting across several models.
- Weighted probability averages.
- Using leave one out probabilities (or transformations thereof) as predictors in a meta model, such as logistic regression.

Model stacking

Model stacking will often lead to better predictions, at the expense of interpretability.