Introduction to Statistical Learning

November 20, 2020

Statistical Learning

Core concepts covered in this chapter:

- Reducible/irreducible errors
- Linear/non-linear models
- Parametric/non-parametric methods
- Supervised/unsupervised learning
- Regression/classification problems
- Quality of fit
- Bias-Variance Trade-Off

What is statistical learning?

We want to predict some output variable (also called response, dependent variable) Y based on some input variables (also called predictors, independent variables, features) X. We suppose there exists a relationship of the form $Y = f(X) + \varepsilon$. f is some unknown function of X and $\varepsilon$is a random error term, independent from X and of mean zero. f represents the *systematic information* that X provides about Y. The ε are the difference between the observations and the true underlying relationship between X and Y, which is usually unknown. Statistical learning refers to a set of approaches for estimating f.

Why estimate f?

Problems of statistical learning fall into the prediction setting, the inference setting and sometimes in both. Some models are lead to very good predictions (highly non-linear models for instance) while some other models do not yield good predictions but are highly interpretable (linear models for instance).

Prediction We can predict Y using $\hat{Y} = \hat{f}(X)$ where $\hat{f}$ represents an estimate for f. For prediction, $\hat{f}$ can be treated as a *black box.* Accuracy of the resulting prediction depends on the *reducible error* ($\hat{f}$ cannot be a perfect estimate of f) and the *irreducible error* (Y is also a function of some error term ε that cannot be predicted using X). The average of the squared distances between the observations Y and the predictions $\hat{Y}$ is thus given by: $E(Y - \hat{Y}) = \underbrace{E[f(X) - \hat{f}(X)]^2}_{Reducible} + \underbrace{Var[\varepsilon]}_{Irreducible}$. Statistical learning: estimating f and minimizing the reducible error.

Inference We are interested in understanding the relationship betwen X and Y, $\hat{f}$ cannot be treated as a *black box* here. We may wonder: what are the important predictors? what is the influence of each predictor on Y? can the relationship between the predictors and the response be summarized by a linear model?

How do we estimate f?

Linear and non-linear methods share some characteristics. In both cases, the goal is to used the training data $\{(x_i, y_i)\}$ to build some estimator $\hat{f}$ such that $Y \approx \hat{f}(X)$. These methods can be either *parametric* or *non-parametric.*

Parametric methods They involve a two-steps *model-based* approach. (1) Make modeling assumptions about the functional form of f (linear for instance). The problem is simplified, because we only have to estimate f from a subset of functions. (2) Pick procedure to fit/train the model. In linear setting for instance, we use the *ordinary least squares* method to estimate the parameters of $\hat{f}$. The issue is that these models usually do not match true relationship. One can increase the number of parameters to make the model more flexible, but there is then a risk of overfitting (following the noise).

Non-parametric methods No assumptions made about the functional form of f. They avoid the possible flaws of parametric methods but require a huge amount of observations to obtain a correct estimate of f as the problem is not restricted to a subset of functions. See *thin-plate spline* for instance.

Trade-Off Prediction Accuracy and Model Interpretability

Assessing Model Accuracy