Statistical Methods for Machine Learning

Derek Li

Contents

1 Introduction

1.1 Terminology

Input variables typically denoted by $X_1, \cdots, X_p$. Output varible typically denoted by Y.

1.2 Supervised and Unsupervised

Supervised statistical learning involves building a statistical model for predicting or estimating an output based on one or more inputs. For unsupervised statistical learning, there are inputs but no supervising output.

Alternatively we could say for supervised learning, all data is labeled and the algorithms learn to predict the output from the input data; for unsupervised learning, all data is unlabeled and the algorithms learn to inherent structure from the input data.

1.2.1 Supervised Learning

In the regression problem, Y is quantitative; in the classification problem, Y takes values in a finite, unordered set.

1.2.2 Unsupervised Learning

There is no outcome variable, and only a set of predictors (features) measured on a set of samples. The objective is more fuzzy: we could find groups of samples or features that behave similarly, or we could find linear combinations of features with the most variation. It is difficult to know how well you are doing.

Unsupervised learning is different from supervised learning but can be useful as a pre-processing step for supervised learning.

1.3 Simple Model

Suppose that we observe a quantitative response Y and p different predictors, $X_1, \cdots, X_p$. Assume a general form of a relationship by a model

$$Y = f(X) + \varepsilon,$$

where f is fixed and unknown and ε is a random error term, which is independent of X and has mean zero. ε captures measurement errors and other discrepancies such as omitted predictors.

There are two main reasons that we estimate f : prediction and inference.

1.3.1 Prediction

With a more accurate f we can make predictions of Y at new points $X = x$, understand which components of $X = (X_1, X_2, \cdots, X_p)$ are important in explaining Y, and depending on the complexity of f, understand how each component X_i of X affects Y.

We can predict Y using $\widehat{Y} = \widehat{f}(X)$. The accuracy of $\widehat{Y}$ depends on:

- Reducible error: an error introduced by because of $\widehat{f}$ not being the perfect estimate of f but by using the most appropriate algorithm, this error can be reduced.
- Irreducible error: Y is a function of ε that cannot be predicted using X, and we cannot reduce the error introduced by ε.

Theorem 1.1. Consider a given estimate $\widehat{f}$ and a set of predictors X that yields the prediction $\widehat{Y} = \widehat{f}(X)$. Assume $\widehat{f}$ and X are fixed. Then

$$\mathbb{E}\left[\left(Y - \widehat{Y}\right)^2\right] = \underbrace{\left[f(X) - \widehat{f}(X)\right]^2}_{\text{Reducible}} + \underbrace{\text{Var}[\varepsilon]}_{\text{Irreducible}} .$$

Proof. We have

$$\begin{aligned}\mathbb{E}\left[\left(Y - \widehat{Y}\right)^2\right] &= \mathbb{E}\left[\left(f(X) + \varepsilon - \widehat{f}(X)\right)^2\right] \\ &= \mathbb{E}\left[\left(f(X) - \widehat{f}(X)\right)^2\right] + \mathbb{E}[\varepsilon^2] + \mathbb{E}\left[2\varepsilon\left(f(X) - \widehat{f}(X)\right)\right].\end{aligned}$$

Since $\mathbb{E}[\varepsilon] = 0$, we have

$$\mathbb{E}\left[2\varepsilon\left(f(X) - \widehat{f}(X)\right)\right] = 0.$$

Besides,

$$\mathbb{E}[\varepsilon^2] = \mathbb{E}[\varepsilon^2] - (\mathbb{E}[\varepsilon])^2 = \mathrm{Var}[\varepsilon].$$

Therefore,

$$\mathbb{E}\left[\left(Y - \widehat{Y}\right)^2\right] = \left[f(X) - \widehat{f}(X)\right]^2 + \mathrm{Var}[\varepsilon].$$

□

1.4 Estimation Methods

We can use linear and non-linear approaches to estimate f. For these methods, we observe a set of n different data points (training data) and train the specific method to find an estimate $\widehat{f}$ for f.

1.4.1 Parametric Methods

Parametric methods involve a two-step model-based approach.

- Assume the functional form of f.
- Use the training data to train the model.

The problem of estimating f reduces to one of estimating a set of parameters without fitting an entirely arbitrary function f so that we call this method parametric.

There are disadvantages such as chosen model will not match the true unknown form of f.

1.4.2 Non-Parametric Methods

Non-parametric methods do not make explicit assumptions about the functional form of f, but seek an estimate of f that gets as close to the data points as possible.

These methods can fit a wider range of possible shapes for f but problem of estimating f is complex since the methods cannot reduce the problem to estimating a small number of parameters as in parametric case. Also, these methods need a very large number of observations to obtain an accurate estimate for f.

1.5 Prediction Accuracy and Model Interpretability

There is a trade-off between accuracy and interpretability so that choosing an appropriate model depending on your goal.