Statistical Methods for Machine Learning

Derek Li

Contents

1 Introduction

1.1 Terminology

$X = (X_1, \cdots, X_p)$ are p predictor variables and suppose there are n observations, $X_p = (X_{p1}, \cdots, X_{pn})$. 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, but Xs could be quantitative or qualitative. The goal is to predict a new observation not in the training data set - what is the expected value of y_0, given x_0.

In the classification problem, Y takes values in a finite, unordered set (e.g., binary or dichotomous, categorical), but Xs could be quantitative or qualitative.The goal is to predict a new observation not in the training data set - what is the class probability of y_0, given x_0.

1.2.2 Unsupervised Learning

$X = (X_1, \cdots, X_p)$ are measured variables on n observations and usually p is large. Note that there is no response (Y) to supervise the algorithm.

Two common methods for unsupervised learning:

- Clustering: Understand the relationships between the ***observations.*** Cluster the observations on the basis of the variables measured and

identify the group to which each observation belongs (number of groups is unknown).

- Principal components analysis (PCA): Understand the relationships between the ***variables*** and reduce the number of variables to a smaller number.

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. 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.

1.3.1 Prediction

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,

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

□

1.3.2 Inference

Inference is the goal of classical statistical methods: we want to understanding the relationship between Y and Xs and how strong is the relationship, etc.

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 Flexibility and Interpretability

Restrictive models (e.g., linear regression) are more interpretable, but flexible models (e.g., splines, boosting) are less interpretable and they can have complicated estimates of f.

Generally, as flexibility increases, interpretability decreases. Also, as flexibility increases, prediction accuracy may decrease due to over-fitting.

1.6 Assessing Model Accuracy

1.6.1 Measuring the Quality of Fit

In the regression setting, the most commonly-used measure is the mean squared error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2,$$

where $\widehat{f}(x_i)$ is the prediction that $\widehat{f}$ gives for the ith observation.

Since the MSE above is computed using the training data that was used to fit the model, so we call it training MSE. Nevertheless, we are interested in the accuracy of the predictions that we obtain when we apply our model to previously unseen test data not used to train the model.

We want to choose the model that gives the lowest test MSE

$$\text{Ave}(y_0 - \widehat{f}(x_0))^2,$$

as opposed to the lowest training MSE.

Recall that $\text{Var}[\varepsilon]$ is the irreducible error, and it corresponds to the lowest achievable test MSE among all possible methods.

Here is a fundamental property of statistical learning that holds regardless of the data set and the model being used.

Property 1.1. As the flexibility of the model increases, there is a monotone decrease in the training MSE and a U-shape in the test MSE.

Definition 1.1 (Overfitting)**.** When a given model yields a small training MSE but a large test MSE, we are said to be overfitting the data.

Overfitting happens because the model may pick up some patterns that are just caused by random chance rather than by true properties of the unknown function f. The test MSE will be large because the patterns do not exist in the test data.

Note that we almost always expect the training MSE to be smaller than the test MSE since most models seek to minimize the training MSE.

In practice, test observations usually are not available, we still cannot select a model that minimizes the training MSE, because there is no guarantee that the method with the lowest training MSE will also have the lowest test MSE. To estimate test MSE, we may use cross-validation.

1.6.2 The Bias-Variance Trade-Off

Suppose we have fit a model $\widehat{f}(x)$ to some training data and let (x_0, y_0) be a test observation drawn from the population. Suppose the true model is $Y = f(X) + \varepsilon$ and $f(x) = \mathbb{E}[Y|X = x]$, then for given x_0, test MSE is

$$\mathbb{E}\left[\left(y_0 - \widehat{f}(x_0)\right)^2\right] = \mathrm{Var}\left[\widehat{f}(x_0)\right] + \left[\mathrm{Bias}\left(\widehat{f}(x_0)\right)\right]^2 + \mathrm{Var}[\varepsilon].$$

Proof. We have

$$\begin{aligned}
\mathbb{E}\left[\left(y_0 - \widehat{f}(x_0)\right)^2\right] &= \mathbb{E}\left[\left(y_0 - \mathbb{E}\left[\widehat{f}(x_0)\right] + \mathbb{E}\left[\widehat{f}(x_0)\right] - \widehat{f}(x_0)\right)^2\right] \\
&= \mathbb{E}\left[\left(\mathbb{E}\left[\widehat{f}(x_0)\right] - y_0\right)^2\right] + \mathbb{E}\left[\left(\widehat{f}(x_0) - \mathbb{E}\left[\widehat{f}(x_0)\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(\mathbb{E}\left[\widehat{f}(x_0)\right] - f(x_0) - \varepsilon\right)^2\right] + \mathrm{Var}\left[\widehat{f}(x_0)\right] \\
&= \mathbb{E}\left[\left(\mathbb{E}\left[\widehat{f}(x_0)\right] - f(x_0)\right)^2\right] + \mathbb{E}[\varepsilon^2] - (\mathbb{E}[\varepsilon])^2 + \mathrm{Var}\left[\widehat{f}(x_0)\right] \\
&= \left[\mathrm{Bias}\left(\widehat{f}(x_0)\right)\right]^2 + \mathrm{Var}[\varepsilon] + \mathrm{Var}\left[\widehat{f}(x_0)\right].
\end{aligned}$$

□

Here are some comments:

- Expected test MSE can never lie below $\mathrm{Var}(\varepsilon)$, the irreducible error.
- To minimize the expected test error, we need to select a model that simultaneously achieves low variance and low bias.
- Variance refers to the amount by which $\widehat{f}$ would change if we estimated it using a different training data set. Ideally, the estimate for f should not vary too much between training sets. More flexible models have higher variance.
- Bias refers to the error introduced by approximating a real life problem, which may be extremely complicated, by a much simpler model. Generally, more flexible models result in less bias.

- As flexibility increases, the variance will increase and the bias will decrease. The relative rate of change in variance and bias determines whether the test MSE increases or decreases.
- As we increase the flexibility, the bias tends to initially decrease faster than the variance increases so that the expected test MSE declines. At some point increasing flexibility has little impact on the bias but starts to significantly increase the variance so that the test MSE increases.

2 Linear Regression

Linear regression is a simple model to supervised learning. It assumes that the dependence of Y on $X_1, \cdots, X_p$ is linear.

2.1 Simple Linear Regression

We assume a model

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where β_0 and β_1 are two unknown parameters represents the intercept and slope respectively, and ε is the error term.

ε is to catch what we miss with the model. Assume ε is independent of X and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Given some estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for the model coefficients, we can predict Y given X by

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x,$$

where $\widehat{y}$ indicates a prediction of Y given $X = x$.

2.1.1 Estimation of the Parameters by Least Squares

Let $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ and $e_i = y_i - \widehat{y}_i$, where $i = 1, \cdots, n$, and n is the sample size and e_i is the ith residual of observation i.

We define residual sum of squares

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2,$$

and by the least squares method, we choose $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to minimize RSS:

$$\begin{aligned} \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x}, \\ \widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}. \end{aligned}$$

2.1.2 Assessing the Accuracy of the Coefficient Estimates

The standard error of an estimator reflects how it varies under repeated sampling:

$$\mathrm{SE}(\widehat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right], \mathrm{SE}(\widehat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

where $\sigma^2 = \mathrm{Var}[\varepsilon]$.

SE can be used to compute confidence intervals. A 95% confidence interval is defined as a range of values s.t. with 95% probability, the range will contain the true unknown value of the parameter. For linear regression, the 95% confidence interval for $\beta_i, i = 0$ or 1, is

$$\left[\widehat{\beta}_i - 1.96 \cdot \mathrm{SE}(\widehat{\beta}_i), \widehat{\beta}_i + 1.96 \cdot \mathrm{SE}(\widehat{\beta}_i) \right],$$

i.e., there is approximately a 95% chance that the interval will contain the true value of β_i.

SE can be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing if there is some relationship between X and Y:

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0.$$

To test the null hypothesis, we compute a t-statistic given by

$$t = \frac{\widehat{\beta}_1 - 0}{\mathrm{SE}(\widehat{\beta}_1)},$$

which will have a t distribution with $n - 2$ degrees of freedom, assuming $\beta_1 = 0$. We can compute the probability of observing any value equal to $|t|$ or larger under H_0 and we call the probability the p-value.

2.1.3 Assessing the Accuracy of the Model

The residual standard error (RSE) is an estimate of the standard deviation of ε, given by

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}}.$$

The R^2 statistic provides an alternative measure of fit:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}},$$

where $\text{TSS} = \sum_{i=1}^{n}(y_i - \overline{y})^2$ is the total sum of squares.

In simple linear regression, $R^2 = r^2$, where r is the correlation between X and Y :

$$r = \text{Cor}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$