View on GitHub

stats-learning-notes

Notes from Introduction to Statistical Learning

Chapter 2 - Statistical Learning

Inputs, also known as predictors, independent variables, features, or more generally, variables.

Outputs, also known as response or dependent variable.

Suppose an observed quantitative response Y and p different predictors $x_1, x_2, \dots, x_p$. The assumed relationship between Y and $X = (x_1, x_2, \dots, x_p)$ can be generalized as:

$$Y = f(X) + \epsilon$$

where f is some fixed, but unknown function of X and ϵ is a random error term that is independent of X and has a mean of zero. In such a scenario, f represents the systematic information that X provides about Y.

In general, an estimation of f, denoted by $\hat{f}$, will not be perfect and will introduce error.

The error introduced by the discrepancy between f and $\hat{f}$ is known as irreducible error because it can never be reduced regardless of the accuracy $\hat{f}$.

The irreducible error will be larger than zero because ϵ may contain unmeasured variables needed to predict Y or ϵ may contain unmeasured variation. The irreducible error always enforces an upper bound on the accuracy of predicting Y. In practice, this bound is almost always unknown.

Estimating f

Parametric Methods

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

1. First, make an assumption about the functional nature, or shape, of f. For example, assume that f is linear, yielding a linear model.
2. Once a model has been selected, use training data to fit, or train, the model. In the case of a linear model of the form

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p,$$

the training procedure should yield estimates for the parameters $\beta_0, \ \beta_1, \ \ldots, \ \beta_p$ such that

$$Y \approx f(X) \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p.$$

A model-based approach like that outlined above is referred to as parametric because it simplifies the problem of estimating f down to estimating a set of parameters.

In general, it is much simpler to estimate a set of parameters than it is to estimate an entirely arbitrary function f. A disadvantage of this approach is that the specified model won't usually match the true form of f.

Using more flexible models is one means to attempt to combat inaccuracies in the chosen model. However, more flexible models have the disadvantage of requiring a greater number of parameters to be estimated and they are also more susceptible to overfitting.

Overfitting is a phenomenon where a model closely matches the training data such that it captures too much of the noise or error in the data. This results in a model that fits the training data very well, but doesn't make good predictions under test or in general.

Non-Parametric Methods

Non-parametric methods don't make explicit assumptions about f and instead seek to estimate f by getting as close to the data points as possible without being too coarse or granular, preferring smoothness instead.

Non-parametric approaches can fit a wider range of possible shapes for f since essentially no assumptions about the form of f are made. However, since non-parametric approaches don't simplify the problem of estimating f, they tend to require a very large number of observations to accurately estimate f.

A thin-plate spline is one example of a non-parametric method.

Though less flexible, more restrictive models are more limited in the shapes they can estimate, they are easier to interpret because the relation of the predictors to the output is more easily understood.

Supervised Learning vs. Unsupervised Learning

Supervised learning refers to those scenarios in which for each observation of the predictor measurements X_i there is an associated response measurement Y_i. In such a scenario, it is often desirable to generate a model that relates the predictors to the response with the goal of accurately predicting future observations or of better inferring the relationship between the predictors and the response.

Unsupervised learning refers to those scenarios in which for each observation of the predictor measurements X_i, there is no associated response Y_i. This is referred to as unsupervised because there is no response variable that can supervise the analysis that goes into generating a model.

Cluster analysis, a process by which observations are arranged into relatively distinct groups, is one form of unsupervised learning.

Regression Problems vs. Classification Problems

Quantitative values, whether a variable or response, take on numerical values. Problems with a quantitative response are often referred to as regression problems.

Qualitative values, whether a variable or response, take on values in one of K different classes or categories. Problems with a qualitative response are often referred to as classification problems.

Which statistical learning method is best suited to a problem tends to depend on whether the response is qualitative or quantitative.

Measuring Quality of Fit

To evaluate the performance of a model it is necessary to evaluate how well the model's predictions match the observed data, to quantify to what extent the predicted response is close to the observed response data.

Mean squared error is one common measure in the regression setting.

Mean squared error is defined as

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

The mean squared error will be small when the predicted responses are close to the true responses and large if there's a substantial difference between the predicted response and the observed response for some observations.

Mean squared error is applied both when training a model and when testing a model.

Though it may be tempting to optimize the training mean squared error, the reality is that the model is judged by the accuracy of its predictions against unseen test data. As such, **the model that yields the best test mean squared error is preferable** to the model that yields the best training mean squared error.

The Bias-Variance Trade-Off

The number of degrees of freedom quantifies the number of values in the model that are free to vary. The degrees of freedom is a quality that summarizes the flexibility of a curve.

As a model's flexibility increases, the training mean squared error will decrease, but the test mean squared error may not. When the training mean squared error is small but the test mean squared error is large, the model is described as overfitting the data. That said, the training mean squared error will almost always be less than the test mean squared error because most methods either directly or indirectly aim to minimize the training mean squared error.

Overfitting refers specifically to scenarios in which a less flexible model would have yielded a smaller test mean squared error.

The expected test mean squared error for a given value x_0 can be decomposed into the sum of three quantities: The variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$, and the variance of the error term, ϵ. Formally,

$$\mathrm{E}(y_i - \hat{f}(x_i))^2 = \mathrm{Var}(\hat{f}(x_i)) + [\mathrm{Bias}(\hat{f}(x_i))]^2 + \mathrm{Var}(\epsilon).$$

To minimize the expected test error, it's necessary to choose a method that achieves both low variance and low bias. It can be seen that the expected test mean squared error can never be less than $\mathrm{Var}(\epsilon)$, the irreducible error.

Variance refers to the amount by which $\hat{f}$ would change if it were estimated using a different training data set. In general, more flexible methods have higher variance.

Bias refers to the error that is introduced by approximating a potentially complex function using a simple model. More flexible models tend to have less bias.

In general, the more flexible the statistical learning method, the more variance will increase and bias decrease.

The relationship between bias, variance, and test set mean squared error is referred to as the bias-variance trade-off. It is called a trade-off because it is a challenge to find a model that has both a low variance and a low squared bias.

Assessing Classification Accuracy

In classification scenarios, the most common means of quantifying the accuracy of $\hat{f}$ is the training error rate. The training error rate is the proportion of errors that are made when applying $\hat{f}$ to the training observations. Formally stated as,

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(y_i \neq \hat{y})$$

where I is an indicator variable that equals 0 when $y = \hat{y}$ and equals 1 when $y \neq \hat{y}$.

In simple terms, the error rate is the ratio of incorrect classifications to the observation count.

As in the regression scenario, a good classifier is one for which the test error rate is smallest.

The Bayes Classifier

It is possible to show that the test error rate is minimized on average by a very simple classifier that assigns each observation to the most likely class given its predictor variables.

In Bayesian terms, a test observation should be classified for the predictor vector x_0 to the class j for which

$$\Pr(Y = j|X = x_0)$$

is largest. That is, the class for which the conditional probability that $Y = j$, given the observed predictor vector x_0, is largest. This classifier is called the Bayes classifier.

In a two-class scenario, this can be restated as $\Pr(Y = 1|X = x_0) > 0.5$ matching class A when true and class B when false.

The threshold where the classification probability is exactly 50% is known as the Bayes decision boundary.

The Bayes classifier yields the lowest possible test error rate since it will always choose the class with the highest probability. The Bayes error rate can be stated formally as

$$1 - \mathrm{E}(\max_j \Pr(Y = j|X)).$$

The Bayes error rate can also be described as the ratio of observations that lie on the “wrong” side of the decision boundary.

Unfortunately, the conditional distribution of Y given X is often unknown, so the Bayes classifier is most often unattainable.

K-Nearest Neighbors

Many modeling techniques try to compute the conditional distribution of Y given X and then provide estimated classifications based on the highest estimated probability. The K-nearest neighbors classifier is one such method.

The K-nearest neighbors classifier takes a positive integer K and first identifies the K points that are nearest to x_0, represented by N_0. It next estimates the conditional probability for class j based on the fraction of points in N_0 who have a response equal to j. Formally, the estimated conditional probability can be stated as

$$\Pr(Y = j|X = x_0) = \frac{1}{k}\sum_{i \in N_0}\mathrm{I}(y_i = j)$$

The K-Nearest Neighbor classifier then applies Bayes theorem and yields the classification with the highest probability.

Despite its simplicity, the K-Nearest Neighbor classifier often yields results that are surprisingly close to the optimal Bayes classifier.

The choice of K can have a drastic effect on the yielded classifier. Too low a K yields a classifier that is too flexible, has too high a variance, and low bias.

Conversely, as K increases, the yielded classifier becomes less flexible, with a low variance, but high bias.

In both regression and classification scenarios, choosing the correct level of flexibility is critical to the success of the model.

Next: Chapter 3 - Linear Regression

stats-learning-notes maintained by tdg5

Published with GitHub Pages