ISLR summary

Generalized linear models for an entire class of statistical learning methods that include both linear and logistic regression as special cases.

if there are *p* variables in our data set, then *p*(*p −* 1)*/*2 distinct scatterplots can be made,

Assessing Model Accuracy

One of the key aims of this book is to introduce the reader to a wide range

of statistical learning methods that extend far beyond the standard linear

regression approach. Why is it necessary to introduce so many different

statistical learning approaches, rather than just a single best method? There

is no free lunch in statistics: no one method dominates all others over all

possible data sets. On a particular data set, one specific method may work

best, but some other method may work better on a similar but different

data set. Hence it is an important task to decide for any given set of data

which method produces the best results. Selecting the best approach can

be one of the most challenging parts of performing statistical learning in

practice.

*Measuring the Quality of Fit*

In order to evaluate the performance of a statistical learning method on

a given data set, we need some way to measure how well its predictions

actually match the observed data. That is, we need to quantify the extent

to which the predicted response value for a given observation is close to

the true response value for that observation. In the regression setting, the

most commonly-used measure is the *mean squared error* (MSE), given by

The MSE

will be small if the predicted responses are very close to the true responses,

and will be large if for some of the observations, the predicted and true

responses differ substantially. The MSE in (2.5) is computed using the training data that was used to

fit the model, and so should more accurately be referred to as the *training*

*MSE*. But in general, we do not really care how well the method works

training

on the training data. Rather, *we are interested in the accuracy of the pre-* MSE

*dictions that we obtain when we apply our method to previously unseen*

*test data*. no guarantee that the method with the lowest training MSE will also

have the lowest test MSE. speaking, the problem is that many

statistical methods specifically estimate coefficients so as to minimize the

training set MSE. For these methods, the training set MSE can be quite

small, but the test MSE is often much larger.

*Variance* refers to the amount by which ˆ *f* would change if we estimated it using a different training data set. Since the training data

are used to fit the statistical learning method, different training data sets

will result in a different ˆ *f*. But ideally the estimate for *f* should not vary

too much between training sets. On the other hand, *bias* refers to the error that is introduced by approximating

a real-life problem, which may be extremely complicated, by a much

simpler model.

Standard errors can be used to compute *confidence intervals*. A 95%

confidence

confidence interval is defined as a range of values such that with 95% interval

probability, the range will contain the true unknown value of the parameter.

The range is defined in terms of lower and upper limits computed from the

sample of data. For linear regression, the 95% confidence interval for *β*1

approximately takes the form.

In the case of the advertising data, the 95% confidence interval for *β*0

is [6*.*130*,* 7*.*935] and the 95% confidence interval for *β*1 is [0*.*042*,* 0*.*053].

Therefore, we can conclude that in the absence of any advertising, sales will,

on average, fall somewhere between 6*,*130 and 7*,*940 units. Furthermore,

for each $1*,*000 increase in television advertising, there will be an average

increase in sales of between 42 and 53 units. We use a *confidence interval* to quantify the uncertainty surrounding

confidence

the *average* sales over a large number of cities

FTest-

The F-statistic for the multiple linear regression model obtained by regressing

sales onto radio, TV, and newspaper is shown in Table 3.6. In this

example the F-statistic is 570. Since this is far larger than 1, it provides

compelling evidence against the null hypothesis *H*0. In other words, the

large F-statistic suggests that at least one of the advertising media must

be related to sales. However, what if the F-statistic had been closer to

1? How large does the F-statistic need to be before we can reject *H*0 and

conclude that there is a relationship? It turns out that the answer depends

on the values of *n* and *p*. When *n* is large, an F-statistic that is just a

little larger than 1 might still provide evidence against *H*0. In contrast,

a larger F-statistic is needed to reject *H*0 if *n* is small. When *H*0 is true

and the errors *\_i* have a normal distribution, the F-statistic follows an

F-distribution.6 For any given value of *n* and *p*, any statistical software

package can be used to compute the p-value associated with the F-statistic

using this distribution. Based on this p-value, we can determine whether

or not to reject *H*0. For the advertising data, the p-value associated with

the F-statistic in Table 3.6 is essentially zero, so we have extremely strong

evidence that at least one of the media is associated with increased sales.

In (3.23) we are testing *H*0 that all the coefficients are zero. Sometimes

we want to test that a particular subset of *q* of the coefficients are zero.

This corresponds to a null hypothesis

Forward selection, backward selection, mixed slecetion, step sleelction

**Gradient Boosted Tree** is similar to random forest models, but the difference is that trees are built successively. With each iteration, the next tree fits the residual errors from the previous tree in order to improve the fit.

Vanishing Gradient Problem

As more layers using certain activation functions are added to neural networks, the gradients of the loss function approaches zero, making the network hard to train. Certain activation functions, like the sigmoid function, squishes a large input space into a small input space between 0 and 1. Therefore, a large change in the input of the sigmoid function will cause a small change in the output. Hence, the derivative becomes small. As an example, Image 1 is the sigmoid function and its derivative. Note how when the inputs of the sigmoid function becomes larger or smaller (when |x| becomes bigger), the derivative becomes close to zero.

Why it’s significant:

For shallow network with only a few layers that use these activations, this isn’t a big problem. However, when more layers are used, it can cause the gradient to be too small for training to work effectively.

Gradients of neural networks are found using backpropagation. Simply put, backpropagation finds the derivatives of the network by moving layer by layer from the final layer to the initial one. By the chain rule, the derivatives of each layer are multiplied down the network (from the final layer to the initial) to compute the derivatives of the initial layers.

However, when n hidden layers use an activation like the sigmoid function, n small derivatives are multiplied together. Thus, the gradient decreases exponentially as we propagate down to the initial layers.

A small gradient means that the weights and biases of the initial layers will not be updated effectively with each training session. Since these initial layers are often crucial to recognizing the core elements of the input data, it can lead to overall inaccuracy of the whole network.