Summary:

Statistical learning is the process of learning about relationships and predictions using variables. This is useful for predicting how marketing spend across different channels will affect sales.

One use of statistical learning is finding the relationship of one variable Y with many variables (x1,x2,...,xp). The main questions asked about the relationship of the variables are:

1) Are there certain features which have the greatest relationship with the target variable?

2) What is the effect of the target variable as we change certain features or combinations of features?

3) Is the relationship between the features and the target variable linear or complicated?

Another reason to use statistical learning is to create a black box where you input the variables (x1,x2,...,xp) and output a prediction for Y. This is useful when you want to find out if a commodity is priced accurately based on it's properties.

The mathematical functions are:

Y = f(X) + epsilon

Y-hat = f-hat(X) where epsilon averages to zero

{(x1,y1),(x2,y2),...,(xn,yn)} where xi = (xi1,xi2,...,xip)-Transpose

Y ~= B0 + B1\*X1 + B2\*X2 ... linear model

f-hat is estimated by either parametric or non-parametric methods. The parametric method defines the structure of the function f-hat; it is simpler, requires less data for accuracy, is intuitive, but is prone to under-fit the correct model f. The non-parametric method will try to fit the training set will without using too many curves; it fits the data well without error introduced by using an inappropriate structure for f-hat, but it requires a lot of data, and can overfit the data. Overfitting with when the model follows the errors, or noise, too closely.

Main Question: How do I tell if my model f-hat is too biased or overfitting the data?

Answer:

We minimize the test set(s). This appears to be good at giving a low variance of f-hat and bias squared of f-hat.

Summary:

This section describes the trade-off between a complicated (flexible) model and a simple (interpretable) model. A simple model is good for understanding by humans so it's good for inference problems. A complicated model is good at predicting accurately, but steps have to be taken to make sure it's not fitting a complicated and inaccurate model to a simple model.

Supervised learning needs target variable (yi) values with each observation. Unsupervised learning doesn't have a target variable and one method is clustering. Clustering is hard to visualize with many variables, so it's important to automate the clustering job.

Classification is used in supervised learning when the target variable (yi) is qualitative. Regression is used in supervised learning when a target variable (yi) is quantitative.

Our goal is to have a low variance of f-hat and low bias square of f-hat. Minimizing the test MSE correlates with a lower variance of f-hat and lower bias squared of f-hat.

variance of f-hat is a measure of how much f-hat will change with a different training set. bias squared of f-hat is a measure of how our f-hat is not complicated enough or fitting the true f well enough.

Formulas:

MSE = 1/n \* summation(yi - f-hat(xi)) squared

E(y0 - f-hat(x0)) squared = var(f-hat(x0)) + [bias(f-hat(x0)] squared + var(epsilon)

bias squared of f-hat and variance of f-hat are both non-negative.

Likely main topics:

Theory behind some classification solutions, Bayes and K neighbors. Also basic functions and graphics.

Main Questions:

1)What are the formulas related to classification basics?

2)What is Bayes?

3)What is K neighbors?

4)What are some good R functions I don't know about?

Short Answers:

1)Take the average of correct predictions to find your error rate. The Bayes classifier formula P(Y=j|X=x) will choose the best class given predictor values.

The Bayes error rate formula is the rate at which we incorrectly classify

The K nearest neighbors formula estimates the Bayes classifier at a given point by taking the closest K nearest neighbors and finding the rate it was correct. Repeat this to get it for more points, or to search for the 50% areas, or all points.

2) Bayes classifier is used to find the best class at a given set of predictor variables.

3) K neighbors uses the Bayes classifier but first needs to estimate the Bayes classifier. To do this it takes the K nearest neighbors and takes the rate at which the neighbors were class j. This rate is now the probability that an observation with those predictors will be class j.

4) cor(), var(), cov(), rnorm(), c(), seq(), plot(), contour(), image(), persp()

pdf(), dev.off(), rm(list=ls())

Summary:

K nearest neighbors has a U-shaped test error curve, so it has the trade-off of bias vs variance. Choosing K=10 is good while K=1 and K=100 weren't as good. Contour() was hard to understand but interesting and image was the same.

Main Questions:

1)How do we know a relationship exists between X and Y?

2)How do we know if the coefficients for a simple linear regression is good?

3)How do we know if our model is good enough?

Answers:

1)Simple linear regression has the form Y = f(x) + epsilon; Y = B0 + B1X1 + epsilon. We know a relationship exists if the parameter B1 is not equal to zero. We can estimate the parameters using a formula or machine learning method. The standard error is used when using a sample to infer characteristics about a population. The standard error is sigma squared divided by n. We can estimate the standard error of B1 by the formulas given. We then can use the null hypothesis B1=0 and find the t-statistic of how many standard errors B1 hat our B1 hat estimation is away from zero. We check the t-calculator with n-2 degrees of freedom to see the probability of B1 equaling zero and deviating by |t|. If the p-value is very low (usually .01 or .05) we can reject the null hypothesis and accept the alternative hypothesis that B1 != 0 and there is a relationship between X and Y.

2) The SE(B1 hat) and SE(B0 hat) are values that tell you how good your coefficients are. This is the variability of the confidence that the estimate of the parameter is correct. The formulas to calculate these are included in the chapter.

3) Two methods of knowing if the goodness of fit of your model are the methods RSE and R squared. The RSE = SQRT(RSS/(n-2)). RSS = summation of squared (yi - (yi hat)). The value of the RSE can vary because it describes the average variation of Y in terms of Y. The RSE could be 3000 when our estimate is 14000. So it can vary by 3000 which is about 23% variation (3000/14000).

The R squared method gives a number between 0 and 1. The value 1 means it has a good goodness of fit. The R squared formula is: R squared = 1 - RSS/TSS. RSS is the variation after regression, TSS is the variation before regression (imagine a flat horizontal line at the average y value and we measure the variability of the data from this line). TSS - RSS equals the amount of variation that we explain away using the regression. R squared is always between 0 and 1 so it can be easier to understand what the value means right away compared to RSE. Sometimes we aim for R squared values that are high, and other times when we are okay with R squared values of .1 when we have a very rough model to the data. R squared is equal to r (r is correlation) squared.

Summary:

The residual squared error, RSE, (written in formulas as sigma) is equal to the square root of (RSS / (n-2))

t statistic is calculated for parameters like B1 as the following with null hypothesis B1=0:

t = (B1 hat - 0) / SE(B1 hat). The is just the # of SE(B1) away from the null hypothesis the estimate B1 is. This will help get the p-value

confidence intervals can be created for each parameter such as B0 and B1. It is approximately 2 SE(B0 hat) or 2 SE(B1 hat) from the approximate B0 hat or B1 hat values found for 95% confidence.

RSE is an estimate of how far away actual values will fall from the regression line/plane. RSE = 3.24 means on average, actual values will deviate by 3.24 from the regression line/plane.

Questions:

1) What does the text say about the F statistic?

2) How do you tell if a relationship exists between one of your variables?

3) How do you know which variables to keep and which to throw out?

Answers:

1) The F-statistic is necessary to test whether a relationship exists between at least one predictor and the response variable. The F-statistic formula is:

F-statistic = (TSS-RSS)/p divided by RSS/(n-p-1)

The denominator of the F-statistic is approximately sigma squared when linear regression is used. The numerator of the F-statistic is approximately sigma squared when no relationship exists and greater than sigma squared when there is a relationship. So a F-statistic of approximately 1 will tell you that no relationship exists, and a larger F-statistic tends to be when there is a relationship.

Our null hypothesis is that no relationship exists between the predictor variables and the response variables. To disprove this null hypothesis and accept the alternative hypothesis we find the F-statistic. The point at which we disprove the null hypothesis with the F-statistic depends on the values of n and p. When n is large then the F-statistic only needs to be a bit higher than 1 to disprove the null hypothesis. When n is small then the F-statistic needs to be a bit larger to disprove the null hypothesis.

We cannot just use the t statistic of each parameter because even 5% of the time the p-value will be below 0.05 due just to chance. The F-statistic is necessary to see if it is likely that a relationship exists between the Y and at least one response variable.

2) A relationship exists if the F-statistic is large enough given values n and p.

3) You can use a selection method. The forward selection method and mixed selection method can be the go to methods. They work even if p > n because they start with no predictors in the model, and only add more if it is statistically significant. The mixed selection will remove variables that have been added if they are no longer statistically significant. The backwards selection method starts with all predictors and removes the least statistically significant predictors. This approach stops usually when all predictors left are under a threshold p-value such as 5%.

Summary:

Learned about F-statistic for seeing if a relationship exists between at least one of your predictor choices using multi-variate linear regression. A large F-statistic can mean we can disprove the null hypothesis. Also learned about variable selection methods such as forward selection method and mixed selection method.

Summary:

1) How do you know if there is a good model fit?

2) How do you find the confidence interval of your model prediction?

Answers:

1) You can use RSE and R squared statistic.

R squared statistic (between 0 and 1) is considered a good fit if the R squared value is close to 1. R squared is equvalent to Cor(Y,Y-hat) squared. Adding variables will always increase the R squared value because the model will fit the training data better, but not necessarily the test set. If the R squared increases by very little when a variable is added to the model, that variable might not end up being a good predictor for the model.

The formula for RSE is sqrt(RSS/(n-2)) for simple linear regression. Now it is generally RSE = sqrt(RSS/(n-p-1)) for multi-variate regression. RSE is best when it is a low value. Adding variables can sometimes increase this value giving a worse RSE value for the model.

2) Prediction is straight forward you simply use the formula Y = B0 + B1X1 + ...

What is important is to look at the uncertainty. The three forms of uncertainty are your estimations of parameters B0,B1,...,Bp, the bias squared error of using a linear model to a complicated model (often the case in real life), and the epsilon. To understand our degree of uncertainty we create 99% or 95% confidence intervals around the predicted value.

Questions:

1)What are some variables in the credit data set that are qualitative?

Answer:

1) gender, student (status), status (marital), and ethnicity.

Columns such as age, cards, education (years) are quantitative predictors.

Summary:

Just intros about qualitative data in an example data set.

Questions:

1) Do we need to treat qualitative variables differently; if so, how for 2 level?

2) Do we need to treat qualitative variables differently; if so, how for multiple level?

3) How can we update the additive model?

Answers:

1) Assign dummy variables 1 or 0. Don't treat it differently than normal variables, it just takes more work to understand what it means.

2) Create different parameters for 1 or 0 based on each level. Use all of them or none of them; use an F-test for this. You only need 2 new dummy variables for a 3 level variable; so z levels requires z-1 dummy variables of 0 or 1. The one without a dummy variable is known as the baseline.

3) Multiply two features together and use this as a new variable. Y=B0+B1X1+B2X2+B3X1X2. We must keep the B1 and B2 if we keep B3 due to hierarchal principle. Simple, if we use the interaction of feature 1 and 2, we have to keep the feature 1 and 2 in the model where they are separate from each other.

Summary:

Qualitative variables need the dummy values 1 and 0. For multiple levels (z levels) make z-1 dummy variables of 1 or 0 values. The level without a 0 or 1 value is called the baseline variable. Test whether you should keep variables with the t test, measuring the p-value. However, if you use a multiple level qualitative variable and made many dummy variables, you need to do an F-test on all the dummy variables. Your null hypothesis is that B1=B2=0 where B1 and B2 are parameters next to the dummy variables. If you can reject this null hypothesis then you want to keep these dummy variables created for the multi-level qualitative variable.

The first limitation of the linear model is the additive assumption. To overcome this we add interaction variables. We test if we should keep them using the t-test statistics and p-values from the t-tests. Just create a new variable like B3\*X1\*X2 to your model. It is useful to use interaction variables especially with qualitative variables because it allows the slow to change based on whether the value is 0 or 1, otherwise it only changes the intercept a flat amount.

Summary:

Identify non-linear fit by plotting. x-axis is the fitted model, the y-axis is the residuals. We expect it to random with no pattern if it is a good fit. To use polynomial regression just transform one of the features with itself squared or square root, keeping also the original feature as a term elsewhere. Two other problems with regression is correlated error terms. This is often due to time, such as time series; other times it's due to sampling failing to be random, and there may be other causes. Another error is heterschedasticity which is when the variance increases at different fitted values. Measure the fitted model vs residuals again and if you see a funnel this indicates that this is a problem.

Questions:

1)What can you do about outliers?

2) What are high leverage points and what can you do to fix this problem.

Summary:

A lot of the problems are solved with residual plots.

1)Outliers can be identified with residual plots.

2) High leverage points are strange values of predictor variables that are well outside the range of most observations. For multi-variate linear regression you must consider if a value is high leverage by the leverage statistic because we can't graphically see if the mixture of variables is normal. The leverage statistic is always on average (p+1)/n. Each value is always between 1/n and 1, the higher the value the more likely that the observation is a high leverage point.

You can't separate out highly correlated predictors, either drop one or combine them. To identify this find the correlation between predictors.

Correlation matrix is a simple way to find these, also there is a formula.

VIF formula is the best way; a value greater than 5 or 10 suggests at least one problem term.

Summary:

Create additive terms, and polynomial terms based on R squared and RSE and forward/mixed/reverse method of adding features; including the use of t statistic and F statistic for parameters. Then to prepare against the 6 largest problems for non-linear regression:

1) non-linearity, 2) correlated epsilons, 3) epsilons that aren't constant over predictor values, 4) outliers, 5) high leverage poinst, 6) collinearity.

To fix this, use residual graphs.

1) look to fit a straight line to the residual graph. 2) Make sure it's a randomized sample and if time series take precautions on this one. 3) Look for a funnel in the residual graphs. 4) remove outliers identified by residual graphs to improve prediction, but removing them can make your model inaccurate. Definately remove if it was likely an error in measuring. 5) Remove high leverage points by calculating the observations' leverage statistic, a value close to 1 is significantly a leverage point and the average is (p+1)/n. Graphing can't help always. 6) Graphing can't help always here as well. Collinearity can sometimes be seen in a correlation matrix, but the formula is best. A value of 5 or 10 indicates a trouble variable that should be combined with another or removed. The p-value will be large for collinear variables when the other variable is present. The formula might be automated in programs.

You can't separate out highly correlated predictors, either drop one or combine them. To identify this find the correlation between predictors.

Correlation matrix is a simple way to find these, also there is a formula.

VIF formula is the best way; a value greater than 5 or 10 suggests at least one problem term.

Summary:

Create additive terms, and polynomial terms based on R squared and RSE and forward/mixed/reverse method of adding features; including the use of t statistic and F statistic for parameters. Then to prepare against the 6 largest problems for non-linear regression:

1) non-linearity, 2) correlated epsilons, 3) epsilons that aren't constant over predictor values, 4) outliers, 5) high leverage poinst, 6) collinearity.

To fix this, use residual graphs.

1) look to fit a straight line to the residual graph. 2) Make sure it's a randomized sample and if time series take precautions on this one. 3) Look for a funnel in the residual graphs. 4) remove outliers identified by residual graphs to improve prediction, but removing them can make your model inaccurate. Definately remove if it was likely an error in measuring. 5) Remove high leverage points by calculating the observations' leverage statistic, a value close to 1 is significantly a leverage point and the average is (p+1)/n. Graphing can't help always. 6) Graphing can't help always here as well. Collinearity can sometimes be seen in a correlation matrix, but the formula is best. A value of 5 or 10 indicates a trouble variable that should be combined with another or removed. The p-value will be large for collinear variables when the other variable is present. The formula might be automated in programs.

CLASSIFICATION

Questions:

1) why not linear regression for classification?

2) What is logistic regression?

3) How do we get coefficients for logistic regression?

4) How do we predict for logistic regression?

Answers:

1) We would have to assume ordering which wouldn't work well with multi-level classification. Linear regression can estimate probability approximations outside the range of 0 to 1, so it's not very practical to use.

2) The p(X) values will range from 0 to 1 that you're estimating. It uses maximum likelihood for coefficient estimation. It works well with multi-variate regression.

3) maximum likelihood, then check the table for the p-values of each statistic.

4) Plug in your estimated and tested (by p-value) coefficients to the logistic formula p(X) = ...

SUMMARY:

I think the four most important take aways from these 10 pages are these:

1) The logistic formula

2) Odds and the log-odds formula

3) Maximum Likelihood and the formula for it

4) Test Coefficients

---

1) This is p(X) = e / (1 + e) where e is raised to the power B0 + B1X1 + ... + BpXp

2) This is log(p(X)/(1-p(X))) = B0 + B1X1 + ... + BpXp

3) Used for both simple and multi-variate logistic regression to estimate coefficients. The formula is just multiplying all the probabilities estimated given your B values each observation where Y=1, and multiplying all the (1- probabilities) estimated given your B values for each observation where Y=0. We want the best estimates that match up with the observations.

4) Take a look at the estimated values, and standard deviations to find the p-values. You should usually ignore the intercept term. (B-hat / SE(B-hat))

Questions:

1) How does linear discriminant analysis work for p=1?

2) What formulas and concepts are most important for LDA?

Answers:

1) You estimate u-hat(k), variance-hat, and pi-hat(k). The u-hat and pi-hat must be found for each class K. Plug it into the formula for delta-hat(k)(x) and you choose the k with the highest value at a given x value.

2) estimation of u-hat(k)s and pi(k)s and variance then plugging into the formula.

Summary:

They only really talked about the decision boundary when pi(k) equaled a half. Then you just take the middle of their two means for an x value.

The idea of f(k)(X) and pi(k) were what led to the formula for delta(k)(x). It's also important to remember the f(k)(x) \* pi(k) divided by the summation of each of these values. That is the provability of the k value at that point.

Sections:

1) LDA for p > 1

2) quadratic LDA

3) Comparing methods

Summary:

KNN, logistic regression, LDA, QDA; none of these are best in all situations. LDA and logistic regression are best when the decision boundary is linear. LDA makes the assumption of a Gaussian distribution. QDA can work better for non-linear decision boundaries. KNN can work best for very complicated decision boundaries.

CROSS VALIDATION

Questions:

1) Why are there multiple MSE in the plot for the validation set topic?

2) How do you do leave one out? How do you use the formula if applicable to speed it up?

Answers:

1) It is run 10 times; 50% of the data randomly chosen 10 times.

2) Leave one out, find the test error. Do this n times. Average the test error. The formula for a computational shortcut is yi-yi-hat squared divided by 1- leverage value; summation of all that divided by n.

Summary:

1) The simple method of estimating your test error is to use a validation set of half the data. You train using half the data and find the test error using the other half of the data. It is simple and easy to use, but has some drawbacks. First it has variance introduced due to the randomness of which 50% is put in each set. Second it only uses half the data and statistical learning models tend to work best with large training sets, this causes overestimation of the test error (we couldn't train the model as well as our data would allow us with all n observations).

2) An attempt to fix these drawbacks can be found in LOOCV (leave one out cross validation). This allows the training set to have n-1 observations. There is no randomness in putting data into sets. There is a formula that can be used with linear and polynomial regression so you only have to run it once, but generally you have to run the model fit n times to go across all the data.

1 validation set and k-1 training sets is a very good choice for bias-variance trade off. Using more data leads to less bias error but using LOOCV causes greater variance. For LOOCV all model fits are very similar just different by one observation. Overall k-Fold k=5 or k=10 are great options.

Bootstrap

Model Selection, stepwise or subset

R-squared adjusted, AIC, BIC, or Cp.