Statistical Learning: Notes

Mo D Jabeen

May 10, 2023

Contents

1 Sources

The main source for the information and figures below is the book *An Introduction to Statistical Learning* [1]. Other elements were also taken from the book *Applied Predictive Modelling* [2]. Markov models was taken from the book *Introduction to machine learning* [3].

2 General

Statistical is learning is based on making predictions or inferences on data inputs. Via approximating f(x), where y = f(x) + error.

2.1 What are the types of statistical problems?

- Regression: Determine outcome variable based on predictors, continuos problem ie Range of values
- Classification: Discrete choice of answers (normally qualitative)
- Clustering: Determine similar groups of data (no natural output variable)

2.2 Notation

$$x_{ij}, i : 1, 2, \cdots n, j : 1, 2, \cdots p \tag{2.1}$$

$$x_i = (xi1, x_{i2}, x_{i3} \cdots, x_{ip}) \tag{2.2}$$

i is the observation and j is the predictor.

2.3 What is parametric and non parametric?

Parametric: Assume form of the desired function and calculate parameters based on the assumption.

Non Parametric: Do not make any explicit assumptions, instead fit function best to the data given.

Choosing an non parametric avoids the problem of having a function form very different to reality however opens up the possibility of overfitting (following noise too closely) and needs more data for an accurate form.

Furthermore, if the main goal is inference, restrictive (parametric) are much easier to interpret.

2.4 What is unsupervised learning?

There is no response/outcome variable, ie clustering.

Semi supervised learning: some response variables.

2.5 What is RSS?

Residual sum of squares is minimised in regression to calculate the parameters.

$$RSS = (y_1 - \hat{y_1})^2 + \cdots + (y_n - \hat{y_n})^2 \tag{2.3}$$

3 Quality of fit

Measure how well the model fits the the true model.

3.1 How is Mean Squared Error used?

Regression compares the predicted outcome to the true value and measures the MSE. However many statistical methods minimise the **training** MSE, therefore a test MSE should be used!

If there is a small training MSE and a large test MSE this can indicate overfitting.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2 \tag{3.1}$$

3.2 What is Bias Variance Trade off?

MSE can be deconstructed to give variance of the estimate function, squared bias of the estimate function and variance of error.

$$E(y_o - \hat{f}(x_o))^2 = Var(\hat{f}(x_o)) + (Bias(\hat{f}(x_o)))^2 + Var(\epsilon) \tag{3.2}$$

Aim is to minimise variance and bias as error can not be removed.
The rate of change between var and bias determines optimal flexibility of a model.

3.2.1 What is variance?

The change in outcome/model if the data set is changed. Higher flexibility of model often increases this.

3.2.2 What is bias?

Error from approximations.

3.3 Error Rate

Classification accuracy is measured by error rate, the frequency of values that predicted the wrong class. In this case test error rate is also preferred.

$$Error\ rate = \frac{1}{n} \sum I(y_i \neq \hat{y}_i), I : y_i \neq \hat{y}_i?1:0 \tag{3.3}$$

3.4 What is Bayes Classification?

Choosing the max probability an observation is a class will minimise the error rate.
Ie if there are only two classes, choose the class that fits:

$$P(Y = 1|X = x_o) > 0.5 \tag{3.4}$$

3.5 What is Bayes error rate?

$$Bayes\ Error\ Rate = 1 - E(maxP(Y = j|X = x_o)) \tag{3.5}$$

E is the average for all values of X. ie if the probability of a 2 class, setup where one class is 0.7 the error rate will be 0.3. Not possible to actually use Bayes as the probability is unknown, however, this is the gold standard.

3.6 What is residual standard error?

Variance is often unknown (population) and therefore RSE is used:

$$RSE = \sqrt{\frac{RSS}{n-2}} \tag{3.6}$$

Multiple predictors:

$$RSE = \sqrt{\frac{RSS}{(n-p-1)}} \tag{3.7}$$

3.7 Hypothesis tests for multiple variables?

The F stat:

$$F = \frac{(TSS-RSS)/p}{RSS/(n-p-1)} \tag{3.8}$$

TSS: Total sum of squares.

Methods that account for all predictors such as the F stat should be used instead of individual stats for each variable inferred from collectively as this will be less accurate.

3.8 What is the interaction effect?

When predictors have links or synergies between them and therefore an additive method using a single to represent each predictor individually is not accurate.

Should then use combined parameters:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \tag{3.9}$$

β_3 is the effect on x1 for one unit change of x2 or vice versa.

3.8.1 What is the hierarchy principal?

If an interaction is included, the individuals should also be.

3.9 Which assumptions can be problematic?

3.9.1 Uncorrelated error terms

If the error terms are correlated, stats will be less accurate. RSE will increase (more variation in results) and confidence intervals will be lower.

Checking for correlation can be done visually by plotting the residual terms, if there tracking in the residuals (adjacent values are similar) this is a sign of correlation.

3.9.2 Variance of error terms

Can check for heteroscedasticity by plotting residuals and observing the spread. Can reduce the spread via transformation ie log or sqrt. Can also use weighted methods to counteract the changing var.

3.9.3 Outliers

Outliers are unexpected values in the response, given an observation within the normal range. These have less of an effect on the models fit but do effect the stats used for checking model accuracy ie RSE or R^2.
Check for outliers by plotting studentised residuals:

$$SR = \frac{e}{std} \tag{3.10}$$

If the SR is > 3 possible the value is an outlier.

3.10 High leverage points

If the observation is unusual it is considered a high leverage point, and can have significantly bigger impact on the model than outliers.

With multiple predictors it is possible to have each individual predictor in the standard range, but the overall set of predictors to be unusual. This is hard to spot, can use leverage stat.

For linear regression:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \tag{3.11}$$

3.10.1 Collinearity

When predictors are closely related to each other the accuracy of the coefficients will decrease increasing the standard error, therefore effecting hypothesis tests.

For 2 predictors can check using a correlation matrix, above that when there is multicollinearity use Variance Inflation Factor (VIF).

$$VIF = \frac{1}{1 - R^2_{X_j|X_{-j}}} \tag{3.12}$$

$R^2_{X_j|X_{-j}}$ is the R squared if Xj was the response and the other predictors are used to predict.

Possible fix it to drop linear terms or combine them.

4 Classification

4.1 What is Naive Bayes?

Using bayes theorem and the idea that all predictors are independent from each other. Can simply think of the problem as the frequency of class y in the data set to determine the probability. The distribution of P(X|Y) is the differing forms of naive bayes.

Very efficient and can be effective in some world scenarios when there is independent predictors.

4.2 What is K nearest neighbours?

KNN identifies the K points closest to training point x, shown as η_o. The conditional probability is then the fraction of the points in η_o that equal class j.

$$P(Y = j | X = X_o) = \frac{1}{K} \sum_{i \epsilon \eta_o} I(y_i = j) \tag{4.1}$$

Then classify point x as the class with the highest probability. K =1 has low bias but high variance.

4.3 What if there is high dimensions?

KNN performs worse if p « n due to the lack of neighbours. Generally parametric methods are preferred if few observations per predictor.

4.4 Why logistic regression instead of linear regression?

To create a regression problem from a classification, you would be forced to create some type of ordering of the qualitative variables if >2. This ordering may not be true for the data set. However with a binary classification least squares is possible, the issue is that least squares does not have the required boundaries for a binary decision ie 0-1 so you would create useless areas.

4.5 What is logistic regression?

To bound p(X) between 0 and 1, exponential equation is used:

$$p(X) = P(Y = 1 | X) \tag{4.2}$$

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \tag{4.3}$$

This will produce an S curve, and if the log is taken shows that 1 unit increase in X gives a change of log odds by β_1.

$$log\ odds : ln(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 X \tag{4.4}$$

This also shows that if β_1 is positive, increasing X will increase p(X), if negative the increasing X will decrease p(X).

4.6 How is logistic regression used?

For binary class scenarios you are trying to maximise the likelihood function to estimate the beta parameters.
Likelihood function:

$$(\beta_o, \beta_1) = \Pi_{i:Y_i=1} p(x_i) + \Pi_{i:Y_i=0}(1 - p(x_i)) \tag{4.5}$$

4.7 What is standard error?

$$SE(\alpha^2) = x \tag{4.6}$$

4.6 shows for each sample expect α to very by x on average.
The below is the standard error of the mean ie std/n:

$$SE(\mu)^2 = \frac{\sigma^2}{n} \tag{4.7}$$

The below is based on how the Betas are calculated (this is for linear regression):

$$SE(\hat{\beta}_0)^2 = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}) \tag{4.8}$$

$$SE(\beta_1)^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \tag{4.9}$$

4.8 What ways can you validate the parameters?

Use confidence intervals, z/t tests and hypothesis tests.

The null hypothesis is that there is no relationship between predictor and outcome.

4.9 What are dummy variables?

If there is qualitative category for the observational data, this can be used a dummy value which is used to represent the observations. This then requires new parameters to be calculated.

I.e. set x_i as 1 if male, 0 if female. If more than one category use multiple pairs as predictors, ie $x_{i1} = X == Asian?1:0, x_{i2} = X == Jamaican?1:0$.

These parameters from this can be validated in the same way.

4.10 How do you include multiple predictors in logistic regression?

Match the number of predictors to the number of parameters:

$$\ln \frac{p(x)}{1 - p(x)} = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip} \tag{4.10}$$

Multiple predictors can show how the interweaving between each other and the outcome. Not shown when not included if not included in the equation.

Multi-class logistic regression exists but is not really used.

4.11 WHat is the equivocal zone?

Adjusting the boundary of the required class boundary before it is accepted ie a min prob for one class maybe adjusted to 0.6, in a binary scenario 0.6 for one class is equivalent to 0.4 for the other.

This threshold is varied when creating an ROC graph.

4.12 What is linear discriminant analysis?

Model X for each class Y as a distribution, then use Bayes theorem and compare all dist to choose the max P(Y=k|X=x). **Assumes normal distribution for each class.**

4.12.1 LDA Equation

Π_k : Probability of the randomly chosen observation X being in class K. (Fraction of observations that belong to the kth class)

$f_k(X)$: Density function for X belonging to class K.

$$P(Y = k|X = x) = \frac{\Pi_k f_k(x)}{\sum_{l=1}^{K} \Pi_l f_l(X)} \tag{4.11}$$

$$P(Y = k|X = x) = \frac{Overall\ Prob * density\ of\ x\ for\ k}{\sum_{all\ classes}(Overall\ Prob * density\ of\ x\ for\ k)} \tag{4.12}$$

The focus is estimating f(x) to fit Bayes Classifier. A point on the density function should be maximised as shown by: **Assuming the var is common for all classes and a normal dist**

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} \tag{4.13}$$

Calculating the above may not be possible as the population is not available so estimates of mean and std are used:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i \tag{4.14}$$

$$\hat{\sigma^2} = \frac{1}{n - K} \sum_{1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k) \tag{4.15}$$

This being the average sample var for each class (on the premise they are common).

These are then plugged in to give the boundary points as:

$$\hat{\delta_k}(x) = x\frac{\hat{\mu}_k}{\hat{\sigma^2}} - \frac{\hat{\mu}_k^{\,2}}{2\hat{\sigma^2}} \tag{4.16}$$

If possible compare to Bayes error rate to determine classifier performance.

4.12.2 What if the number of predictors > 1?

Multi-variate normal dist is then used. From which a class specific mean vector is used and a common covariance matrix.

Each predictor follows a one dimensional normal dist with some correlation between predictors.
Think of one dimensional as collapsing into the left side or down. If the var is equal and correlation=0 then the image on the left in 4.1 with a circular bottom is produced, if otherwise skewed, as shown on the right.

$$X \sim N(\mu, \textstyle\sum); \mu : mean\ vector, \textstyle\sum : covariance\ matrix \tag{4.17}$$
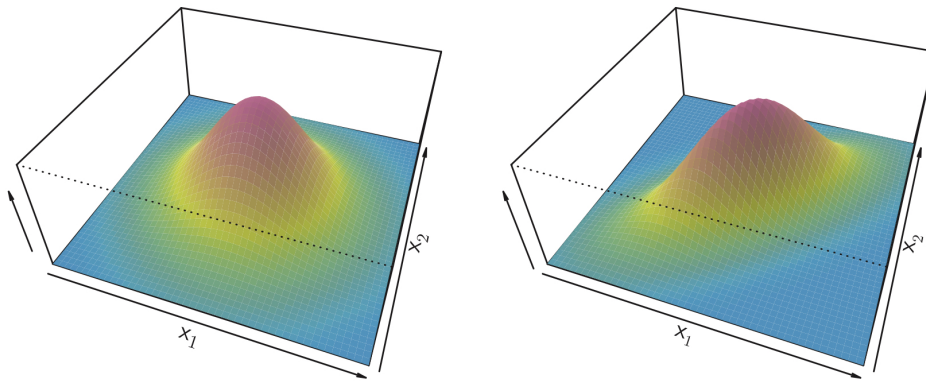
FIGURE 4.5. *Two multivariate Gaussian density functions are shown, with* $p = 2$. Left: *The two predictors are uncorrelated.* Right: *The two variables have a correlation of* 0.7.

Figure 4.1: Multivariate Normal Dist

4.12.3 How is multi preditor LDA used?

Using the multi variate normal dist values can create the vector/matrix version of the Bayes classifier boundary. If there are >2 classes the boundary will be shown as pair of classes.

4.12.4 Issues to be aware of from multi LDA

Small distribution between classes in the data set can result in good training error rates, but poor test error rates.

4.13 What is a confusion matrix?

Determine which type of error is being made in terms of the classes, comparing the error rate for each class.

Accuracy: Correctly predicted/Total

Sensitivity (Recall): Sensitivity is the percentage of true positives (divided by total of values actually positive) (e.g. 90% sensitivity = 90% of people who have the target disease will test positive)

Precision: Percentage of predicted positive and are (divided by total of values predicted positive)

Specificity: Specificity is the percentage of true negatives (e.g. 90% specificity = 90% of people who do not have the target disease will test negative).

False positive rate = 1 - specificity

F1 Score: Harmonic mean of precision and recall (trade off).

4.14 What is Kappa?

Useful for multi class or imbalanced problems. Accuracy is misleading in imbalanced class problems, precision and recall are preferred. Keppa is also weighted based on the freq of each class and therefore more representative.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{4.18}$$

p_o: Observed agreement
p_e: Expected agreement.

Essentially measures how much better the classifier is working vs a random guesser dependant on a class freq.

4.15 How to accommodate non equal classes?

If the bias between classes is not equal the boundary can be altered to instead of being the max, match the criteria of the problem.

4.16 What is ROC curve?

Shows the error rate using different thresholds of class boundary, the overall performance of the model is the area under the curve.

Can also be considered as showing the trade off between specificity and sensitivity.

If area under the curve is <=0.5 its assumed the performance is no better than chance.

The plot is true postive rate (freq of correctly finding pos) vs false postive rate (freq of incorrectly labelled as pos ie is actually neg).
Want a 1 of TPR (all correctly labelled) and 0 of FPR (none wrongly labelled as positive).

4.17 Different classification measurements

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

Figure 4.2: Types of classification measurements

4.18 Quadratic Discriminant Analysis

The main difference to LDA is that there is not an assumption of common variance between distributions, this leads to a quadratic boundary function.

4.18.1 What are the pros/cons of QDA?

More flexible, accommodates differing class distributions and therefore has lower bias.

It is more compute intensive and has higher variance.

4.18.2 How do you choose between LDA and QDA?

Compare the correlation between class and observations for each class, if they are all similar shows the variance is likely similar. And therefore LDA is a good choice otherwise QDA is.

4.19 How do you choose between LDA and Logistic for binary scenarios?

Main differential is that LDA assumes normal distribution of class, logistic does not.

4.20 Differences with KNN?

Decision boundaries cant be weighted towards classes, however it is the most flexible choice.

4.21 Methods to improve non linear performance?

Can improve non linear performance by adding transformed predictors, ie X_1^2, X_2^2 etc.

5 Resampling methods

Completely separate data for the test set may not make sense to use as that could have been used as valuable data for training. So instead you can repeatedly draw samples from the data set you have and refit the model and compare to other samples as test data to gain info about the models setup.

Model assessment: Evaluating a models performance.
Model selection: Determining a models flexibility.

Side note: Can determine if a transformation is valid using hypothesis tests.

5.1 Cross Validation

5.1.1 What is the validation set approach?

Randomly divide the data into a training and test set. The issue is there will be less data used for training the model with each split, will most likely overestimate the test error.

5.1.2 What is leave one out cross valiation (LOOCV)?

Use a single observation as the validation set, the rest as training and loop through all data points.
Use the average MSE from all models and validations point:

$$CV_n = \frac{1}{n} \sum MSE \tag{5.1}$$

The benefits to this are:

1. less bias in model as the training set is much bigger

2. No randomness in method so consistent results produced

This is a very flexible method, however, does require large computation.

5.1.3 What is K fold cross validation?

Separate into K sets, 1 set is the validation and the rest are used for training. Loop through switching the validation set to each Kth set, taking the average across n sets as the error. The main benefit is the computational costs.
Generally the result produced for using K as a reasonable number (>5), produces a result similar to LOOCV (K=n).

The mean of many highly correlated samples has a higher variance than if uncorrelated. And therefore kCV will produce a low variance than LOOCV but higher bias (less training data used).

A good var-bias balance is found around k=5 and k=10, and cleaner to use a multiple of the n as k. K fold is the main method to compare different models looking for the the bottom of the error over the different models.

5.2 Bootstrap

Bootstrapping involves altering the original data set to make estimates on the accuracy of the data. An example of bootstrapping is randomly duplicating and replacing observations throughout the data set to get new data to train and validate on. Other types include "Smooth sampling" zero centred (mean) adding noise to normal dist data.

This can be applied to a wide range of statistical methods, especially useful for models whose variance is difficult to measure.

The bootstrap approach can be used to assess the variability of the coefficient estimates and predictions from a statistical learning method.

5.2.1 Types of sampling

- **Basic Sampling:** Random resampling with replacement (BasicSampling)
- **Antithetic Sampling:** Antithetic resampling, introducing negative correlation between samples
- **Balanced Sampling:** Balanced random resampling, reducing bias
- **Exact Sampling:** Exact resampling, iterating through all unique resamples, deterministic bootstrap, suited for small samples sizes
- **Residual Sampling, Wild Sampling:** Resampling of residuals in generalized linear models
- **Maximum Entropy Sampling:** Maximum Entropy bootstrapping for dependent and non-stationary datasets

6 Regularisation

If the number of predictors is greater than the number of observations, least square method cannot be used, it will produce incorrect results.
Irrelevant predictors will lead to unwanted complexity in the model and wasteful computation, the following are methods to reduce the number of predictors.

Methods below normally assume linear regression is being used and therefore RSS is being minimised.

1. Subset: Choose a subset of predictors.
2. Shrinkage (Normalisation): Set some parameters to 0.
3. Dimension reduction: Project p onto M space.

6.1 How do you use subsets?

Compare all possible combination of predictors (2^p). Choose the best in terms of RSS or R^2 for each set number of predictors. Choose the best overall with AIC, BIC or adjusted R^2.

6.1.1 How does it work with logistic regression?

Use deviance instead of RSS, which is the -2* maximum log likelihood.

$$D_M = -2\log(L_M) \tag{6.1}$$

L_M : Max achievable likelihood.

6.2 How do you perform step wise subset ?

Add one predictor on every iteration, each time trying all possible predictor combination for that set number of predictors. Then based on the previous best add another predictor.

When comparing two models with the same number of predictors can use R^2 or RSS, however differing predictor number AIC,BIC or adjusted R^2 should be used.

6.2.1 What is the pros/cons to this method?

Reduce the iterations to p(1+p)/2. The con is possibly missing different combinations that together produce a better model than working in the best approach method.

To combat this backward stepwise can be used as well as some hybrid methods that allow better scoping of possible combinations.

6.3 How do you compare test error for different models?

Two methods: Adjust training data to remove bias or use cross validation.

How to determine the error parameter to compare:

- C_p
- Alkaline Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- R^2

$$C_p = \frac{1}{n}(RSS + 2d\sigma^2); \; d : predictors \tag{6.2}$$

This adds a penalty of 2d var to the RSS error. For least squares AIC is proportional to Cp.

$$AIC = \frac{1}{n\sigma^2}(RSS + 2d\sigma^2) \tag{6.3}$$

$$BIC = \frac{1}{n}(RSS + log(n)d\sigma^2) \tag{6.4}$$

BIC provides a larger penalty for more predictors as when n>7 log(n) >2.

$$Adjusted\ R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)} \tag{6.5}$$

$$TSS = \sum(y_i - \bar{y})^2 \tag{6.6}$$

Balances RSS by increasing the denominator in terms of the number of predictors, ensuring the number of predictors provides enough benefit in RSS to make it worth it.

6.4 How does cross validation compare?

Cross validation does not assume any model form and so is preferred in comparing models.

6.5 What is the one standard error rule?

Choose the lowest predictor error score in a cross validation which is within one standard error (of the MSE for each model) of the lowest CV error.

6.6 How are t stats used to compare predictors?

The mean output of a model can be compared using different predictors, via the t stat or if mean does not follow a normal dist use Wicoxon ranksum test.

In this case it is better to compare in a binary predictor format.

6.7 What is the Relief algo?

For a binary case compare random observations predictions to the predictions of closest to it with the one from each class (hit (same class) miss (diff class)). Depending on the distance, hits should be close misses far the average score is calculated. The higher the score the more important the predictor.

6.8 What is MIC?

Creates many grids in 2D space on the graph showing the relationship between predictor and outcome. Determining the grid with the largest association value which is the score assigned to the predictor. Similar to tree splitting.

6.9 What are two ways feature selection can be thought of?

Wrapper: Compare models to determine predictor of choice.
Filter: Compare predictors before models.

7 Shrinkage

Reduce the parameters of the model, two methods: Ridge regression and LASSO.

7.1 Ridge regression

Adds a penalty to the standard RSS, to push during the minimisation the parameters towards 0.

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{7.1}$$

$\lambda = 0$: no effect
$\lambda - > \infty$: Impact grows
$\beta_0 = \bar{y}, x_i = 0$

7.1.1 How is ridge regression effected by scaling?

As scaling will directly effect the penalty value (unlike standard linear regression), therefore predictors should be standardised to the same unit before applying ridge regression.

7.1.2 How does the penalty effect bias variance trade off?

As λ increases the bias will also increase and variance will decrease. However, there is a balancing point up to which there is not much impact on bias.

7.1.3 In non linear settings how are vectors compared?

l_2 (ell-2) is used to compare vectors and is a measured of vector from 0.

$$||\beta||_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2} \tag{7.2}$$

Can be used to compare ridge parameters vs standard LSE parameters.

7.2 LASSO

Ridge regression does not set any of the values to 0, instead they tend toward 0. LASSO uses a slightly different penalty that allows for parameters to become 0, allowing variable selection.

7.2.1 What is the penalty?

$$RSS + \sum |\beta_j| \tag{7.3}$$

7.2.2 How are the regularisation methods shown in a constraint form?

LASSO:

$$minimise_\beta(RSS) \quad subject \ to \ \sum_{j=1}^{p} |\beta_j| \le s \tag{7.4}$$

Ridge Regression:

$$minimise_\beta(RSS) \quad subject \ to \ \sum_{j=1}^{p} \beta_j^2 \le s \tag{7.5}$$

Subset:

$$minimise_\beta(RSS) \quad subject \ to \ \sum_{j=1}^{p} I(\beta_j \ne 0) \le s \tag{7.6}$$
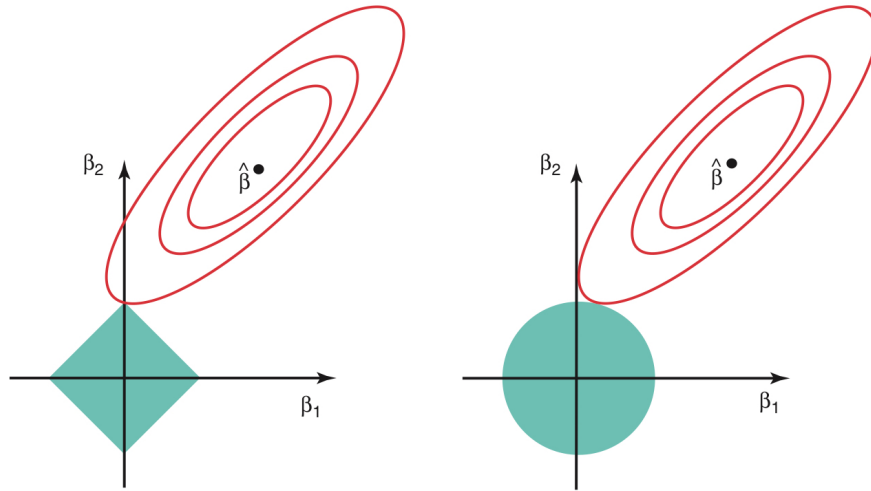
FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Figure 7.1: ridge vs lasso; p=2, ellipses are constant RSS vals

7.2.3 How does graphing shows the interaction of the constraints?

7.2.4 How does ridge compare to lasso?

Ridge should be used if all predictors truly relate to the outcome, otherwise lasso is the best bet.

Cross validation should be used to calculate the best lambda and compare ridge and lasso solutions.

7.3 Dimension reduction?

Projection of p predictors to M where M<p:

Below are the vectors Z which is X predictors linearly combined to create M predictors.

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \tag{7.7}$$

Theta combined with phi for M results in the beta parameters. Giving a smaller M parameters, due to this mapping all M parameters are combined with a specific phi set of values to give the original parameter.

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} \tag{7.8}$$

$$\beta_j = \sum_{m=1}^{M} \phi_{jm} \theta_m \tag{7.9}$$

Reducing the dimensions will reduce the variance, and can avoid overfitting as it is a concentration of the data.

7.3.1 What is principal component analysis?

The first principle component is the line which if the points were projected onto would give the most variance (ie distance from points to line). To confine this we use the constraint $\phi_{11}^2 + \phi_{21}^2 = 1; (\phi_{jm})$.

The second principal has max variance but is also uncorrelated to the first pc, and therefore will be perpendicular to it.

7.3.2 What is principal components regression approach?

Determine the first M Z principal components and uses them as predictors for a regression line. The idea is that most of the relationship with the response is in the first few principal components.

Or the direction in which the observations show the most variation is directions associated with the response, which is what the principle components are.

If this is true dimension reduction will produce a less overfit model which performs better. Recommended to standardise the variables onto the same scale, as it will effect the variance between predictors.

7.4 What to do in high dimensional problems (p≫n)?

Things to be aware of:

- Dimensional model comparison isn't effective (AIC,BIC,R^2)
- Careful of collinearity, where removing predictors is difficult due to their linear relationship.

8 Non linear regression

5 main methods:

1. **Polynomial regression:** Add extra predictors by raising each by a power
2. **Step functions:** Split the range of a variable into sections making it a classification problem
3. **Regression splines:** Divide into sections, fit a polynomial that is constrained to be smooth at the boundary.
4. **Smoothing splines:** Use residual sum of squares subject to a smoothness penalty similar to regression splines.
5. **Local regression:** Allows splines to overlap at the boundaries in a smooth way.
6. **Generalised additive models:** Extend methods to deal with more predictors.

8.1 Generalised additive Models

Create additive non linear model for p predictors.

For regression:

$$y = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 \tag{8.1}$$

Can be transformed to:

$$y = \beta_0 + f(x_{i1}) + f(x_{i2})$$ (8.2)

In which each function is its own model for that predictor.

For classification:

$$log\frac{p(x)}{1-p(x)} = \beta_0 + f(x_{i1})... + f(x_{ip})$$ (8.3)

8.1.1 What is backfitting?

For some methods its not as simple as adding the models, in which case backfitting is used. Combines models by updating the fit for each predictor in turn keeping the others fixed.

8.1.2 What is the con of GAMs?

As it is additive the interaction between predictors can be lost, but can mitigate this by adding combined function predictors ie f(x_{i1}, x_{i2}).

9 Tree based methods

The idea behind tree based methods is to segment the observations, with rules that can be summarised as a tree.

Simple tree methods don't compare well in terms of accuracy to other modelling methods, however, are easy to visualise and use for inference.

Trees are also less sensitive to data processing than for linear regression.
To improve predictive performance can use multiple tree together; bagging, random forests and boosting are examples of this.

9.1 How do trees work?

Splitting the observations into sections based on the predictor values, this can be done by stepping through each predictor.
The mean in each section is often used as the default response, if the observation lies in that section. It can also be used for inference.

9.2 How to determine the regions?

A possible method is minimising the RSS from the mean value of each region:

$$\sum_{j=1}^{J} \sum_{i \epsilon R_j} (y_i - \hat{y}_{Rj})^2$$ (9.1)

J: is the region.
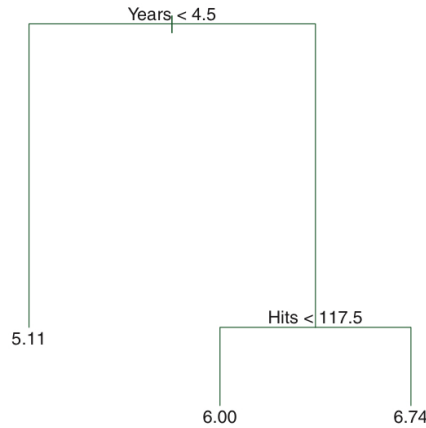$\hat{y}_{R_j}$: is the mean of the region.

FIGURE 8.1. *For the* `Hitters` *data, a regression tree for predicting the log salary of a baseball player, based on the number of years that he has played in the major leagues and the number of hits that he made in the previous year. At a given internal node, the label (of the form $X_j < t_k$) indicates the left-hand branch emanating from that split, and the right-hand branch corresponds to $X_j \geq t_k$. For instance, the split at the top of the tree results in two large branches. The left-hand branch corresponds to* `Years<4.5`, *and the right-hand branch corresponds to* `Years>=4.5`. *The tree has two internal nodes and three terminal nodes, or leaves. The number in each leaf is the mean of the response for the observations that fall there.*

Figure 9.1: Tree split example

9.3 What is top down greedy approach recursive binary splitting?

Working from the top of the tree splitting into two for each predictor, greedy as it does not look ahead only works step by step. Until a stopping criterion is reached ie each region will have n>5. This could lead to overfitting of the tree.

9.4 When should the splitting stop?

An alternative is to split until the a chosen RSS value. But a better method is to let the full tree be created and then work through the tree subsets comparing each other in terms of error rate.

9.5 How can you limit the number of subsets compared?

Pruning allows a selection of the subset via penalties added to the minimisation:

$$\sum_{m=1}^{|T|} \sum_{i \epsilon R_m} (y_i - \hat{y}_{Rm})^2 + \alpha |T| \tag{9.2}$$

T: Number of terminal nodes on the tree.

Similar to LASSO this will allowing reduce the number of terminal nodes, alpha is chosen via cross validation.

9.6 How is it used with Classification?

Use the most commonly occurring class in the region as the default response. And use error rate instead of RSS to choose regions.

$$E = 1 - max_k(\hat{p}_{mk}) \tag{9.3}$$

$\hat{p}_{mk}$: Portion of training observations in the mth region that from the kth class.

9.6.1 Alternatives to using error rate?

Error rate can be not sensitive enough so Gini Index is used:

$$G = \sum_{1}^{k} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{9.4}$$

Gini index will take a small value if the p value is close to 0 or 1 and therefore kind of shows node purity in that the when the proportion of the class is very low or high Gini is low.
Can also use cross entropy function which is similar.

9.7 What are pros/cons?

Pros:

- Easy to explain
- Easy to display
- Easily handles qualitative (easy dummy values as a split)

Cons:

- Poor predictive accuracy

If the relationship is highly non linear and complex trees maybe the preferred model.

9.8 What is bagging?

Bootstrapping or bagging reduces the variance of a model, combining the averages of observations sets will give a lower variance than each set.
So using bootstrapped models and aggregating them will achieve a lower variance.

9.8.1 What is out of bag?

If during the bootstrap a model uses 2/3 of the observations, the other 1/3 is OOB. This is used as the test set and if any other model has the same OOB the error is aggregated.
This can be used with subset selection methods to get better trees to combine via bootstrapping.

9.9 What are Random forests?

The trees are decorrelated and combined just as in bagging. Ensuring the trees when combined have a mixed structure is done by randomly selecting a sample of m predictors to build the tree. m is often the square root of p.

9.10 What is Boosting?

Similar to bagging but sequential. Calculate tree, add factor of tree to overall function, take away response from that model from the residuals (shown below):
Fraction addition of model, where r the residuals are used as the response:

$$\hat{f}(x) += \lambda \hat{f}^b(x) \tag{9.5}$$

Update the residuals:

$$r_i -= \lambda \hat{f}^b(x_i) \tag{9.6}$$

Final model:

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x) \tag{9.7}$$

λ: Shrinkage parameter, controls the speed of learning.
B: number of tree models created.

Due to the learning process increasing B does not cause overfitting.

10 Support Vector Machines

One of the best out of the box classifiers, intended for binary class cases.

10.1 What is the hyperplane?

In p dimensional space, plane is p-1 subspace. Ie in 3 dimensional space the plane is flat.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p = 0 \tag{10.1}$$

$$X = (X_1, X_2, \cdots X_p)^T \tag{10.2}$$

A point that satisfies the above two equation lies on the hyperplane. The hyperplane cuts the space in half about 0, can determine on which side of the hyperplane the point is by using the equation.

Also shows the certainty of the response based on the distance from the hyperplane boundary.

10.2 What is the maximal margin classifier?

The minimum perpendicular distance from all observations to the plane (the margin), is maximised. Support vectors are the points that the hyperplane relies on and used to determine the margin (the closest points).

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip}) \geq M \tag{10.3}$$

M: Instead of being 0, adds a cushion for the decision boundary, this is also the margin.

The margin (perp distance to hyperplane) is given by:

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip}) \tag{10.4}$$

10.3 What if there is no perfect hyperplane?

In many cases cannot maximise M>0, as there is no hyperplane perfectly separating the two. In this case a soft margin is used, which almost completely separates the observations, this is called the **support vector classifier**.

Furthermore, in many cases even if there exists a perfect solution the variance will be very high as a single observation could change the hyperplane.

10.4 What is the support vector classifier?

Maximise M:

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip}) \geq M(1 - \epsilon_i) \tag{10.5}$$

$$\epsilon_i \geq 0, \sum \epsilon_i \leq C \tag{10.6}$$

C: Non negative tuning parameter
ϵ: Slack variable

This allows the support vectors to have a distance less than the margin, depending on epsilon and C. If $\epsilon > 0$ the point is on the wrong side of the margin if $\epsilon > 1$ its on the wrong side of the hyperplane. C determines the severity of the violations, controlling the bias-var trade off and is chosen via cross validation.

Support vector classifiers are not effected by outliers!

10.5 How do you make the plane non linear?

Could use added predictor polynomials to make it non linear but instead the general method is to use kernels in the support vector machine.

10.6 What is the support vector machine?

Focusing on only the support vectors (S) and using inner products (which is used to solve the support vector classifier) gives the classifier as:

$$f(x) = \beta_0 + \sum_{i \epsilon S} \alpha_i \sum_{j=1}^{p} x_{ij} x_{i'j} \tag{10.7}$$

Can then generalise the inner product of points to be function (kernel):

$$K\{x_i, x_{i'}\} \tag{10.8}$$

Linear kernels version:
$$K : \sum x_i x_{i'} \tag{10.9}$$

Polynomial kernel version:

$$K : (1 + \sum x_i x_{i'})^d; \ d : Dimensions \tag{10.10}$$

Other kernel options exist ie, radial kernel.
Use ROC curve to compare SVM and LDA.

10.7 What if the number of classes is >2?

Use one versus one classification, create pairs of classes for all possible combinations and run all SVMs choosing the highest shown class.

One versus all classification, one class vs the rest to check with certainty if the observation belongs to the chosen class.

10.8 Show SVC in loss penalty form

Fitting the support vector classifier $f(x_i) = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip})$:

$$minimise(\sum_{i=1}^{n} max(0, 1 - y_i f(x_i)) + \lambda \sum_{j=1}^{p} \beta_j^2) \qquad (10.11)$$

λ acts like C as the penalty term. Can directly compare this form to other methods, which are made of loss functions and penalty terms, known as hinge loss.

10.9 How does logistic and SVM compare?

If the classifiers are clearly separated SVM is preferred, in overlapping regions logistic regression is good.

Can also use kernel methods in other classifiers.

10.10 What about Support vector regression?

The margin is setup so that absolute values greater than a constant contribute to the loss to determine the parameters.

11 Markov Models

11.1 Markov Process

System has S states, the state at time t is shown as q_t.

$$q_t = S_i \qquad (11.1)$$

The probability of a state is only dependant on the current state, nothing before:

$$P(q_{t+1} = S_j | q_t = S_i) \qquad (11.2)$$

Transition probability $(S_i -> S_j)$ is independent of time.

A sequence of states is the sum of the transition probabilities and the initial probability.

11.2 Hidden Markov Model

In the hidden model states are not observable, but depending on the state of the system the model will produce a probability function for an observation. I.e. in a known model the box 1 contains blue balls, in a hidden model the blue ball may come from any box.

The HMM model is used to find the maximum likelihood of an observation given the current state, or to deduce the maximum likelihood of the current state given a sequence of observations.

HMM models are specialised for known sequenced observations that are only dependant on the present, ie speech synthesis. Can also be layered to combine probability functions.

HMM models allow inference on hidden generative processes via indirect observations correlated to these process (Determine the state of a system via observations correlated to the state).Can be considered within reinforcement learning.

Given a set of data the state of a HMM is estimated by bayesian estimation (expectation maximisation).

11.2.1 Prob function

The prob function for an observation sequence or state could be in the form of Multivariate gaussian:

$$p(x_t|z_t = k, \theta) \qquad (11.3)$$

11.2.2 State space modelling

Three methods to use these models:

- Predicition
- Filtering: Estimate current values
- Smoothing: Estimate past values

12 Bayesian Estimation

Treats the value to be determined as a random variable not an unknown constant and therefore each response has a probability distribution and related stats methods.

References

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning – with Applications in R* (Springer Texts in Statistics). New York: Springer, 2013, vol. 103, ISBN: 978-1-4614-7137-0. DOI: 10.1007/DOI.

[2] M. Kuhn and K. Johnson, *Applied predictive modeling.* New York, NY: Springer, 2013, ISBN: 9781461468493 1461468493 1461468485 9781461468486. [Online]. Available: http://www.amazon.com/Applied-Predictive-Modeling-Max-Kuhn/dp/1461468485/.

[3] E. Alpaydin, *Introduction to Machine Learning* (Adaptive Computation and Machine Learning), 3rd ed. Cambridge, MA: MIT Press, 2014, ISBN: 978-0-262-02818-9.