Intro To Statistical Learning: Notes

Mo D Jabeen

December 6, 2022

1 General

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

1.1 What are the types of statistical problems?

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

1.2 Notation

$$x_{ij}, i : 1, 2, ...n, j : 1, 2, ...p \tag{1.1}$$

$$x_i = (xi1, x_{i2}, x_{i3}..., x_{ip}) \tag{1.2}$$

i is the observation and j is the predictor.

1.3 What is paramtric 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 differnt 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.

1.4 What is unsupervised learning?

There is no response/outcome variable, ie clustering.

Semi supervised learning: some response variables.

2 Quality of fit

Measure how well the model fits the the true model.

2.1 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.

2.2 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.

$$E(y_o - \hat{f}(x_o))^2 = Var(\hat{f}(x_o)) + (Bias(\hat{f}(x_o)))^2 + Var(\epsilon) \tag{2.1}$$

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

2.2.1 What is variance?

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

2.2.2 What is bias?

Error from approximations.

2.3 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 preffered.

$$Error\ rate = \frac{1}{n}\sum I(y_i \neq \hat{y}_i), I : y_i \neq \hat{y}_i ? 1 : 0 \tag{2.2}$$

2.4 What is Bayes Classification?

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

$$P(Y = 1|X = x_o) > 0.5 \tag{2.3}$$

2.5 What is Bayes error rate?

$$Bayes\ Error\ Rate = 1 - E(maxP(Y = j|X = x_o)) \tag{2.4}$$

E is the average for all values of X. ie if the probablilty 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 probablilty is unknown, however, this is the gold standard.

3 Classification

3.1 What is K nearest neighbours?

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

$$P(Y = j|X = X_o) = \frac{1}{K}\sum_{i\epsilon\eta_o} I(y_i = j) \tag{3.1}$$

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

3.2 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 decison ie 0-1 so you would create a useless areas.

3.3 What is logistic regression?

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

$$p(X) = P(Y = 1|X) \tag{3.2}$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{3.3}$$

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.

$$log\ odds : ln(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X \tag{3.4}$$

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

3.4 How is logistic regression used?

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

$$(\beta_o, \beta_1) = \Pi_{i:Y_i=1} p(x_i) + \Pi_{i:Y_i=0}(1-p(x_i)) \tag{3.5}$$

3.5 What is standard error?

$$SE(\mu)^2 = \frac{\sigma^2}{n} \tag{3.6}$$

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

$$SE(\hat{\beta_0})^2 = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}) \tag{3.7}$$

$$SE(\beta_1)^2 = \frac{\sigma^2}{\sum(x_i-\bar{x})^2} \tag{3.8}$$

3.6 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.

3.7 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 preidctors, ie $x_{i1} = X == Asian?1:0, x_{i2} = X == Jamican?1:0$.

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

3.8 How do you include multiple predictors in logistic regression?

Match the number of preidctors to the number of parameters:

$$ln\frac{p(x)}{1-p(x)} = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_p x_{ip} \tag{3.9}$$

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.

3.9 What is linear discriminant analysis?

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

3.9.1 LDA Equation

Π_k : Probablilty 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.

$$P(Y=k|X=x) = \frac{\Pi_k f_k(x)}{\sum_{l=1}^{K} \Pi_l f_l(X)} \tag{3.10}$$

$$P(Y=k|X=x) = \frac{Overall\ Prob * density\ of\ x\ for\ k}{\sum_{all\ classes}(Overall\ Prob * density\ of\ x\ for\ k)} \tag{3.11}$$

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**

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} \tag{3.12}$$

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

$$\hat{\mu_k} = \frac{1}{n_k} \sum_{i:y_i=k} x_i \tag{3.13}$$

$$\hat{\sigma^2} = \frac{1}{n-K} \sum_{1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu_k}) \tag{3.14}$$

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:

$$\hat{\delta_k(x)} = x\frac{\hat{\mu_k}}{\hat{\sigma^2}} - \frac{\hat{\mu_k}^2}{2\hat{\sigma^2}} \tag{3.15}$$

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

3.9.2 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.

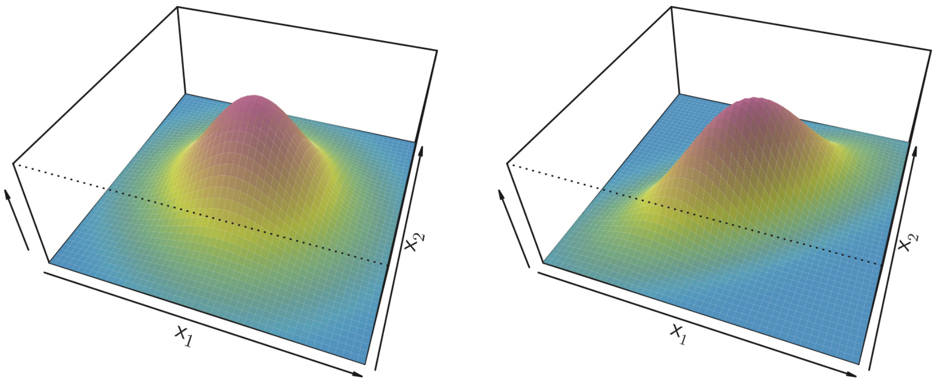

FIGURE 4.5. *Two multivariate Gaussian density functions are shown, with* $p = 2$. Left: *The two predictors are uncorrelated.* Right: *The two variables have a correlation of* 0.7.

Figure 3.1: Multivariate Normal Dist

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 3.1 with a circular bottom is produced, if otherwise skewed, as shown on the right.

$$X \sim N(\mu, \sum); \mu: mean\ vector, \sum: covariance\ matrix \tag{3.16}$$

3.9.3 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.

3.9.4 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.

3.10 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.

Sensitivity: Sensitivity is the percentage of true positives (e.g. 90% sensitivity = 90% of people who have the target disease will test 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).

3.11 How to accomdate 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.

3.12 What is ROC curve?

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

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

3.13 Different classification measurments

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 3.2: Types of classification measurments

3.14 Quadratic Discrimanant Analysis