Week 8 solutions

1 Question 1

Part a

An example of a loss function for linear regression is $L(\beta, x, y) = (y - x^T\beta)^2$. And example of a loss function for classification is $L(h(\cdot), y, x) = I(h(x) \neq y)$.

Part b

It would be ok to **generalize** a machine learning model fitted on students in STA314 to predict things about statistics majors at the University of Toronto only if the students in STA314 were representative of the general stats major population. Otherwise, the model may over-fit to idiosyncratic features of the STA314 cohort that are not present in the larger group. This is because machine learning methods do not extrapolate.

2 Question 2

Part a

Single Linkage: All calculations are based on nearest neighbours.

(i) First Merge: From inspection of the squared distance matrix, there are two pairs of nearest neighbours: 51 and 53 and 52 and 101. So the first merge results in the clusters,

$$(51, 53), (52, 101), (102), (103).$$

(ii) Second Merge: The next smallest squared distance is 0.05 between 51 and 103 and also 53 and 103, and resulting in 103 being merged with (51,53). The clusters are now

$$(51, 53, 103), (52, 101), (102).$$

(iii) Third Merge: The next smallest distance is 0.26 between 52 and 53 ans so (51,53,103) are merged with (52,101).The clusters are now

$$(51, 52.53.101, 103), (102).$$

(iv) Fourth Merge: The shortest squared distance between 102 and the rest is 0.61,(both 52 and 101) and so all observations are merged into one cluster at this height.

Complete Linkage: We have slightly more complicated calculations as distances between clusters are measured by furthest neighbours.

(i) First Merge: This is identical to that for single linkage above as all clusters start as individual observations. So the clusters are

$$(51, 53), (52, 101), (102), (103).$$

(ii) Second Merge: We now need to find the distances between furthest neighbours in all pairs of the above clusters. Extracting the relevant information we find,

	51	53
52	0.36	0.26
101	**0.5**	0.4

	51	53
102	**1.69**	1.37

	51	53
103	**0.05**	**0.05**

	52	101
102	**0.61**	**0.61**

	52	101
103	0.53	**0.73**

	102
103	**1.78**

The distances between furthest neighbours are in bold and we see that the smallest of these is 0.05 between (51,53) and (103) and so we merge these to obtain the clusters

$$(51, 53, 103), (52, 101), (102).$$

(iii) Third Merge: Repeating the above procedure on our current clusters,

	51	53	103
52	0.36	0.26	0.53
101	0.5	0.4	**0.73**

	51	53	103
102	1.69	1.37	**1.78**

	52	101
102	**0.61**	**0.61**

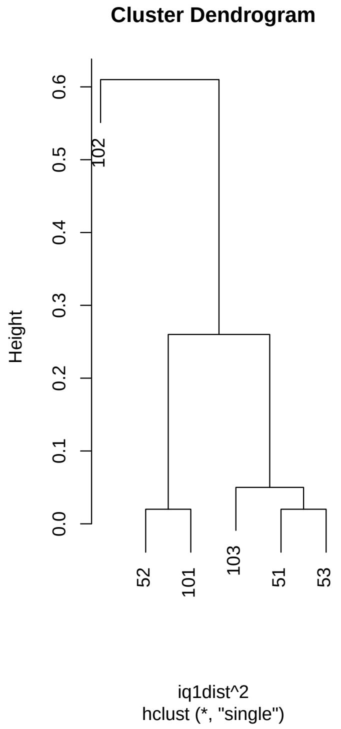
The shortest squared distance between furthest neighbours is 0.61 between both 52 and 102 and also 101 and 102 and so our next clustering is

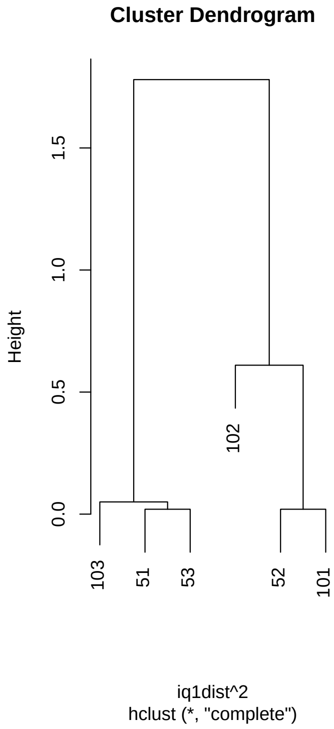
$$(51, 53.103), (52, 101, 102).$$

(iv) Fourth Merge: The two remaining clusters are merged at the height of their furthest neighbour which we see from the table below is 1.78.

	51	53	103
52	0.36	0.26	0.53
101	0.5	0.4	0.73
102	1.69	1.37	**1.78**

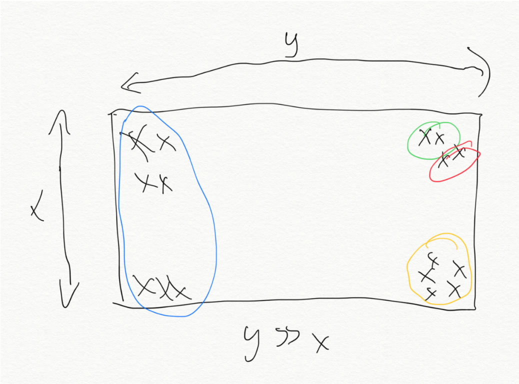
The resulting dendrograms should look something like:





From these we can see that at height 0.5, single linkage leads to clusters {51,52,53,101,103} and {102} and complete linkage leads to clusters {51,53,103}, {52,101} and {102}.

Part b



The above figure is an example where the data has four natural cluster that have been badly classified using k-means. The problem is that the three of the initial cluster centres were on the right hand side of the plot, which means that all of the points on the left were forced into the same cluster. The k-means++ algorithm tries to prevent this happening by attempting to choose initial cluster centres that are as far apart from each other as possible. This would make it unlikely for three initial cluster centres to end up on the same side of the rectangle in the above figure.

3 Question 3

Part a

- Linearity. Plot the residuals and look for nonlinearities
- Independent error. Plot the residuals and look for evidence of serial correlation
- Identically distributed errors. Look at the plot of the residuals and check for changes in variances (heteroskedasticity)
- Normality of errors. A histogram, and empirical CDF or a QQ-Plot.

Part b

Let $A = uv^T$. Then the eigenvalue equation $Ax = \lambda x$ becomes

$$\begin{aligned} Ax &= \lambda x \\ uv^T x &= \lambda x \\ (v^T x)u &= \lambda x. \end{aligned} \tag{1}$$

Now, $v^T x$ and λ are both scalars, so the only way the final vector equality can hold is if $x = \alpha u$, for some $\alpha \in \mathbb{R}$. By convention, eigenvectors have unit norm, so the eigenvector is

$$x = \frac{u}{\|u\|}.$$

Substituting that into (1), we get

$$\frac{v^T u}{\|u\|} u = \frac{\lambda}{\|u\|} u,$$

which implies $\lambda = v^T u$. To show that this is the largest eigenvector (in absolute value), we note that if we choose x to be any of the $n-1$ vectors w that satisfy $v^T w = 0$, then $Ax = \lambda x$ simplifies to $0 \cdot w = \lambda w$, which implies that all of the other eigenvalues of A are zero.

4 Question 4

Part a

We want to find the direction that explains the **third** largest amount of information. Given that we already know the first two PCs, we don't need to explain any more information in those particular directions, so we're only interested in directions that are **orthogonal** to v_1 and v_2. From the lectures, we know that the variance in a direction u is given by

$$(n-1)^{-1}\|Xu\|$$

, so we need to solve the following optimization problem

$$v_2 4 = \arg \max_{\substack{u^T v_1 = 0 \\ u^T v_2 = 0 \\ \|u\|=1}} u^T X^T X u = \arg \max_{\substack{u^T v_1 = 0 \\ u^T v_2 = 0}} \frac{u^T X^T X u}{u^T u}.$$

Following the same reasoning as the lectures, we compute the eigendecomposition $X^T X = V \Lambda V^T$ and set $z = V^T u$. Because $u \perp v_1, v_2$, the first two components of z will be zero. Then it follows that

$$u^T X^T X u = z^T \Lambda z = \sum_{i=1}^{p} \lambda_i z_i^2 = \sum_{i=3}^{p} \lambda_i z_i^2 \leq \lambda_3 \sum_{i=3}^{p} z_i^2 = \lambda_3 u^T V V^T u = \lambda_3 u^T u.$$

As in class, that inequality is an equality if $z_4 = z_5 = \ldots = z_p = 0$, which occurs when $u = v_3$. As we have shown that the Rayleigh quotient cannot be larger than λ_3 when computed on vectors that are orthogonal to v_1 and v_2 **and** that value is attained when $u = v_3$, it follows that v_3 maximizes the constrained Rayleigh quotient and hence the second principal component is the eigenvector of $X^T X$ that corresponds to the second largest eigenvalue.

Part b

In forward variable selection you start with a model that only has an intercept. At each step you add the feature that maximizes the prediction error on the test set. Continue until all features have been added.

In backward variable selection, you start with a model that has all the features in it. At each step you remove the feature that reduces the prediction error the least on the test set. Continue until there are no features left.

Forwards selection is preferable if there are more features than observations.