MAS8404_Project_210431461

210431461 | 21/10/22

Introduction

At the University of Wisconsin Hospital, Dr. Wolberg (Dr. WIlliam H. Wolberg (1992)) collected breast tissue samples from women using fine needle aspiration cytology (FNAC)(Newcastle University (2022)). Histological examination of the tissue collected by this procedure allows for the physician to determine whether or not it is benign or malignant. Our objective is to build a classifier based on the cytological characteristics that determines whether a tissue sample is likely to be benign or malignant.

Data Description

Dr. Wolberg reported his data chronologically to the `mlbench` R package (Dr. WIlliam H. Wolberg (1992)). The FNAC procedure allows for the identification of nine cytological characteristics (Newcastle University (2022)): clump thickness (Cl.thickness / CT), uniformity of cell size (Cell.size / CS), uniformity of cell shape (Cell.shape / CSh), marginal adhesion (Marg.adhesion / MA), single epithelial cell size (Epith.c.size / E), bare nuclei (Bare.nuclei / BN), bland chromatin (Bl.chromatin / BC), normal nucleoli (Normal.nucleoli / NN) and mitoses (Mitoses M). These characteristics for each tissue sample are measured on a discrete scale of one to ten, where smaller numbers indicate that the sample is healthier (Newcastle University (2022)). To aid our analysis, the ordinal variables of this scale have been converted to quantitative variables.

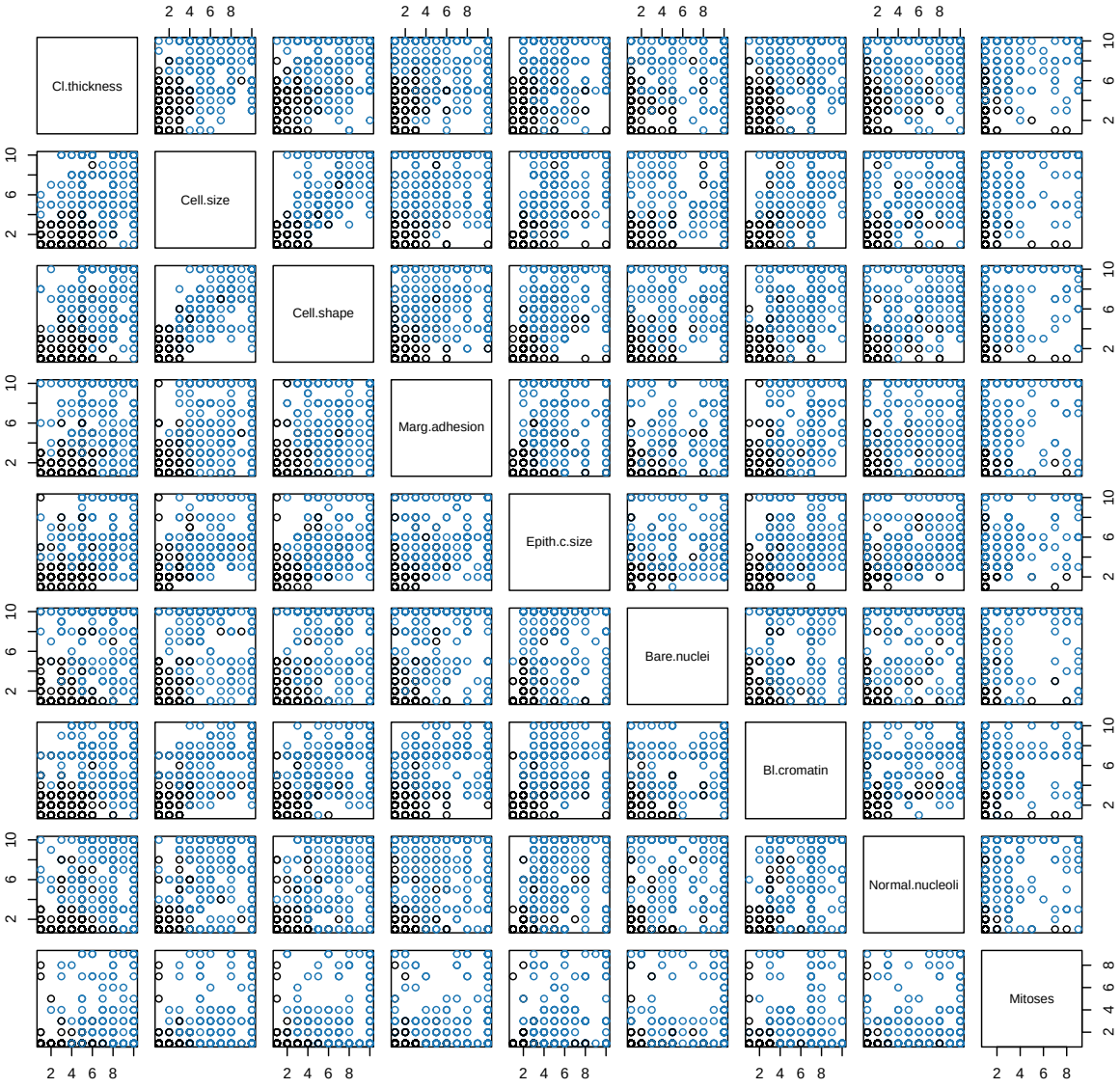
This report explores data from a sample of 699 women in the `BreastCancer` data set; please note that 16 of the 699 observations have been removed due to missing attribute values. It is assumed that this is a random sample of women experiencing symptoms of breast cancer (Newcastle University (2022)). Each woman is represented by a sample code number (Id) that reflects the chronological grouping of this data (Dr. WIlliam H. Wolberg (1992)). The data set also includes the result of further histological examination (Class), which confirms whether each woman's tissue sample was begin or malignant; this has been converted into a binary variable of 0 or 1, respectively.

Data Exploration

Of the tissues samples from 683 women, 239 are confirmed as malignant; in our sample, 53.83% of the women who are experiencing breast cancer symptoms have malignant tissue.

Whether the tissue is benign or malignant is considered the response variable, and the cytological characteristics of tissue are considered as the predictor variables. The characteristics of malignant tend to have higher numbers on the scale one to ten; this is visualized by the density distributions provided in `Appendix A`. All nine cytological characteristics of the tissue samples are presented below in a scatter plot matrix, where blue indicates the tissue is malignant.

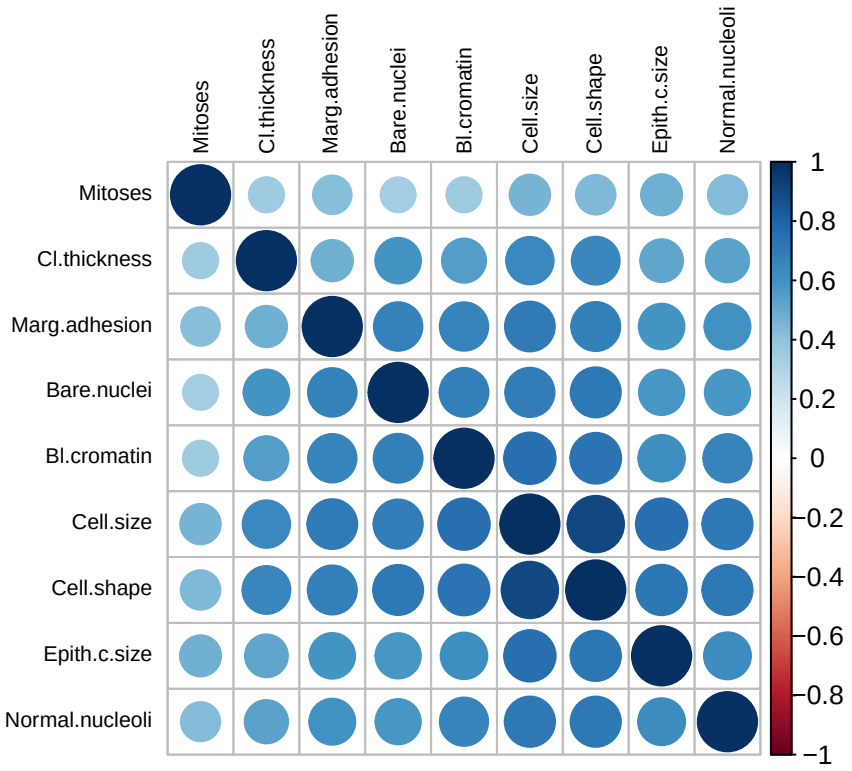
Plot 1: Scatter Plot Matrix of All Cytological Characteristics



This plot visualises the relationship between the characteristics; it suggests that there is a strong relationship between CS and CSh, indicating that larger cells have a more significant shape and are likely to be malignant. Whether this relationship is causal or correlated is worthy of further investigation.

The correlations between other characteristics are not initially obvious and require further investigation. A sample correlation matrix quantifies the strength of a linear relationship between the characteristics, and is visualised by the correlation plot below.

Plot 2: Cytological Characteristics Correlation



In this plot, the correlation is larger when the circle is both darker and larger. This plot confirms the strong linear relationship between cell shape and cell size, which suggests that any regression model is unlikely to need both. These characteristics also appear to have stronger correlations with CT, MA, E, BN, BC, and NN. M does not appear to have a strong relationship with any of the characteristics. This indicates that a selection of the predictor variables can be used to build an accurate classifier.

The two single measures of multivariate scatter help us generalise how our data varies: generalised variance and total variation. For our data, the generalised variance is 55382860.00 and the total variation is 33283.24. This tells us there is a high degree of scatter about the sample means of each variable, highlighting that it will be important to transform our data so that it is standardised, putting all the variables on a common scale.

The mean, median and standard deviation for the nine cytological characteristics of the tissue samples are also presented on the next page.

Table 1: Summary Statistics of Cytological Characteristics

Variable	CT	CS	CSh	MA	E	BN	BC	NN	M
Median	4.00	1.00	1.00	1.00	2.00	1.00	3.00	1.00	1.00
Mean	4.44	3.15	3.22	2.83	3.23	2.54	3.45	2.87	1.58
SD	2.82	3.07	2.99	2.86	2.22	3.64	2.45	3.05	1.64

Table 1 shows us that the mean is notably higher than the median for CS, CSh, BN and NN, suggesting that characteristics of malignant tissue are skewing the mean. The standard deviation is also quite large. The table below investigates these findings in more detail with separate summary statistics for benign and malignant tissue.

Table 2: Summary Statistics by Class

	Benign			Malignant		
Variable	Mean	SD	Median	Mean	SD	Median
CT	2.96	1.67	3.00	7.19	2.44	8.00
CS	1,31	0.86	1.00	6.58	2.72	6.00
CSh	1.14	0.96	1.00	6.56	2.57	6.00
MA	1.35	0.92	1.00	5.59	3.20	5.00
E	2.11	0.88	2.00	5.33	2.44	5.00
BN	1.35	1.18	1.00	7.63	3.12	10.00
BC	2.08	1.06	2.00	5.97	2.28	7.00
NN	1.26	0.95	1.00	5.86	3.35	6.00
M	1.07	0.51	1.00	2.54	2.40	1.00

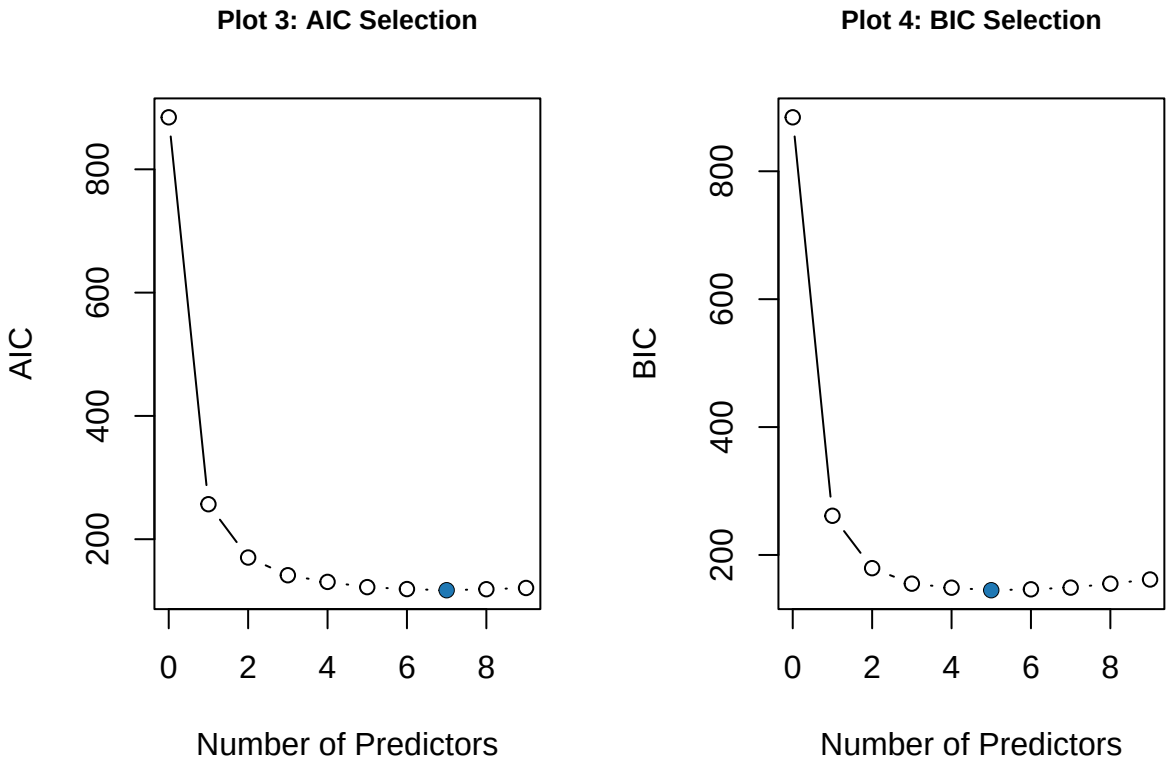
Table 2 clearly demonstrates that there is a significant difference between the characteristics of benign and malignant tissue. This is especially true for CT, BN and BC, indicating they may be the most important predictor variables.

Subset Selection

The exploratory data analysis suggests that some predictor variables are likely to be better at predicting our response variable than others. If this proves to be correct, this will allow us to learn the effects of fewer predictor variables more precisely.

Best subset selection uses least squares to fit a linear regression model to each possible subset of of the predictor variables. However, only one of two of the predictors showed an obvious relationship with `Class` (benign or malignant). To identify the best subset of predictor variables a logistic regression model for `Class` is fitted to standardised data. This highlights that CT, BN, MA and BC have a coefficient which is significantly different to zero when testing at the 5% level.

Best subset selection using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The results are presented in the plots below.



In both plots, the blue dot is number of predictor variables identified by the best subset selection approach. It appears that a model with six variables is likely to be a good compromise between five and seven. The cytological characteristics for these six variables are CT, CSh, MA, BN, BC, and NN. These align closely to the strongly correlated variables identified in our correlation matrix above.

Modelling

To increase the reliability of our approach, the data is divided into two to allow for out-of-sample validation. Good practice suggests that a ratio between 70-80% for the training data set and 20-30% for the test data set (REFERENCE). In this project, 75% of the data is randomly allocated to a training data set to construct our classifier, with the remaining 25% becoming the testing data set to compute our test errors.

Logistic Regression

Logistic regression is used to assign observations to discrete response variables. When applied to our complete data set of six predictor variables using 'in-sample validation,' the training error is 3.08%.

This is not very interesting as only the test error measures how well the method performs on previously unseen data (Newcastle University (2022)). When using our train and test data sets for out-of-sample validation,

If you have time, do it the other way round - a model conditional on maliganant and another condition not on malignant

In all three cases small values of the information criterion indicate a "better" model.

– At least one regularized form of logistic regression, i.e. with a ridge or LASSO penalty;

(One potential problem with ridge regression is that, unlike the subset selection methods from Section 4.4, the fitted model will always include all p explanatory variables. This is because the penalty term in (4.1) shrinks the coefficients towards zero but doesn't make any of them exactly equal to zero. If all we want to do is prediction using our fitted model, this is not a concern. However, when we have a large number p of predictors, including all p of them can make interpretation of the fitted model difficult. In such cases it can be convenient to set some of the regression coefficients exactly equal to zero so that we only include a subset of the predictors in our fitted model. The LASSO2 is an alternative regularisation method to ridge regression which performs subset selection in addition to shrinkage. The LASSO estimate is the point at which a contour first hits the diamond. (Chapter 4) In a big data setting when p is large relative to n, the estimators of the regression coefficients will have large variance, leading to poor predictive performance. Fortunately, all of the approaches discussed in Chapter 4 for linear regression have analogues in the context of logistic regression. For example, ridge regression and the LASSO work in exactly the same way, adding a penalty of the same form to the loss function which, in this case, is the negative of the log-likelihood function. These models can be fitted using the glmnet package in R, specifying family="binomial" (Chapter 5).) (It turns out that ridge regression is generally better than the LASSO for shrinkage, whilst only the LASSO can perform variable selection. - Chapter 4)

– At least one discriminant analyis method, i.e. the Bayes classifier for linear disciminant analysis (LDA) or quadratic discriminant analysis (QDA). (Practical 6) (As discussed in the previous section, in LDA we assume that the conditional distributions for the predictor variables X are multivariate normal, with a group-specific mean vector and a common covariance matrix. Quadratic discriminant analysis (QDA) is similar except the different groups are allowed to have different covariance matrices... The details of this estimation procedure for the mean vectors and covariance matrices are beyond the scope of this course. For the group membership probabilities k, one method of estimation is to use the sample proportions. - Chapter 5)

- For the variants of logistic regression, you should present the coefficients of the fitted model, and any other useful graphical or numerical summaries. For LDA and QDA present estimates of the group means. In each case, discuss what your results show. For example, which variables drop out of the model when you use subset selection or the LASSO? What do the parameters tell you about the relationships between the response and predictor variables?

coefficient of determination (R squred) means that x% of the variation in log PSa can be explained by regression on the eight standardised predictors

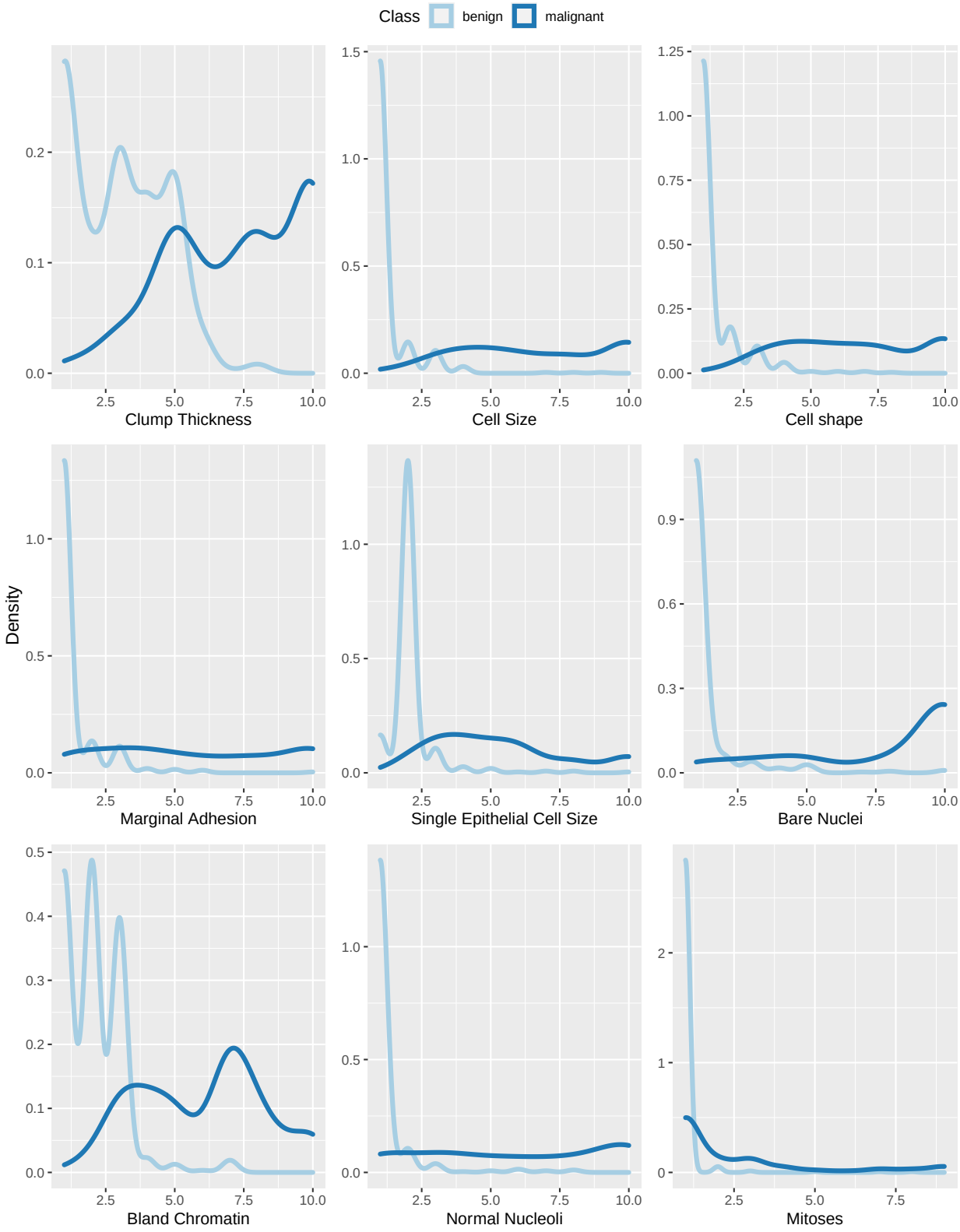
Performance %>% ggplot(aes(x = av_mem, y = av_util, color = CoV_m)) + geom_point(position = "jitter") + geom_smooth(color = "#525252") + labs(x = "Average Memory (percentage)," y = "Average Utilisation (percentage)," title = "Plot 11: Average Utilisation by Average Memory Usage per Hostname," color = "Memory Variation") # Very clear trajectory of more memory corresponds to utilising more gpu core

• Compare the performance of your models using cross-validation based on the test error. Think about how you might do this in a way that makes the comparison fair.

(We typically call the data used to construct the classifier our training data. The data used to estimate the pij , and hence "validate" the classifier, are called the validation data. If the same data are used to both train and validate the classifier, this is called in-sample validation. If the training and validation data sets are different, this is called out-of-sample validation... The K×K matrix of counts nij is often called the confusion matrix. Although the empirical method is simple and widely used, it also leads to optimistic estimates of the performance of the classification scheme due to overfitting. Again this is essentially caused by double-use of the data for constructing and testing the classifier. Often, especially when comparing classification schemes, it is useful to characterise the overall quality of the scheme using a single number summary... The training error rate ... Clearly, in the case of a perfect classifier the error rate would be 0%. We generally prefer classifiers with a low error rate. - CHapter 5) Likely to be using in-sample? Out sample mor rigerus - it is possible, but will need to change training data

• Select a final "best" classifier, justifying your choice. Does it include all the predictor variables? Why or why not? (page 120? in notes of chapter 5) assessed on thier predictie performance using techniques such as cross validation. "what is the probability that an individual has the disease if they have clinical measurements of x?"

Appendix A: Denisty Distributions of Cytological Characteristics by Class



Bibliography

Dr. WIlliam H. Wolberg, David W. Aha, Olvi Mangasarian. 1992. "R Documentation: Wisconsin Breast Cancer Database." https://www.rdocumentation.org/packages/mlbench/versions/2.1-3/topics/BreastCancer.

Newcastle University, Stefen Grunewalder on behalf of. 2022. "Mas8404: Project." https://ncl.instructure.com/courses/46421/files/6177247?wrap=1.