Summary of Statistical Learning Methods

Topic	Advantages	Disadvantages/drawbacks	Additional Notes
Simple Linear Regression	• Simple to apply and interpret • Has many explicit results • Doesn't need a lot of data to fit • LOOCV result is easy to calculate	• One of the most restrictive models • Usually has low predicting power	• High bias, low variance
Multiple Linear Regression	• Easy to explain and to fit • Easy for inference tasks • LOOCV result is easy to calculate	• Variable selection can be a challenging task • Low predicting power	• Categorical variables • Interaction effect • High bias, low variance • Outliers, high leverage points • Collinearity
Best Subset Selection	• Can find the best subset of predictors based on the given criterion	• Can be computationally expensive when *p* is large	• It is better than stepwise selections if *p* is small
Stepwise Selection	• More computationally efficient	• Not working properly for $n \leq p$ • Not guarantee the best selection of variables	•
Ridge Regression	• Can reduce the variance in fitted model for a carefully chosen tuning parameter λ • Computationally efficient for selected λ values	• Can't be used as a feature selection approach • Finding the optimal tuning parameter can be a challenging task	• Cross validation is useful in selecting the tuning parameter λ • λ increases, higher bias, lower variance • works well when most of the predictors are related to the response variable
Lasso Approach	• Can reduce the variance in fitted model for a carefully chosen tuning parameter λ • Can achieve a feature selection goal	• Finding the optimal tuning parameter can be a challenging task	• Cross validation is useful in selecting the tuning parameter λ • λ increases, higher bias, lower variance

			• works well when very few of the predictors are related to the response variable
Principal Component Analysis (PCA)	• Can help to reduce the dimensions in modelling.	• The number of PCs to include needs to be determined. • It is not an approach to select features.	• The standardisation of predictors is usually needed. • CV is useful when determining the number of PCs.
Partial Least Squares	• Can help to reduce the dimensions in modelling. • The response variable is involved in transforming the predictors.	• The number of transformed variables to include in the model needs to be determined. • It is not an approach to select features.	• CV is useful when determining the number of PCs. •
Validation Set Approach	• Can estimate the test MSE or test error rate when test data are not available • It is a general approach	• The results can vary significantly due to randomness in data split	• Has the highest bias in the CV approaches
LOOCV Approach	• It is a widely applicable approach • Can estimate the test MSE/test error rate when test data are not available • The results are stable as no randomness in data split • Have shortcuts in certain cases, eg MLR, polynomial regression, smoothing splines	• Can be computationally expensive • The estimated test MSE or test error rate can be inaccurate	• Has the lowest bias in the CV approaches
***K*-fold CV Approach**	• It is a widely applicable approach • Can estimate the test MSE/test error rate when test data are not available • The results are more stable than the validation set approach	• The estimated test MSE or test error rate can be inaccurate	• Has lower variance than LOOCV method • Usually *K* = 5 or 10

	• More efficient than the LOOCV approach		
Bootstrap Approach	• It is a widely applicable approach • Can estimate the test MSE/test error rate when test data are not available	• The estimated test MSE or test error rate can be inaccurate	• The original data quality and number of bootstrap samples to drawn do matter.
Polynomial Regression	• Easy to fit using the least squares method and the result is smooth • Can produce extremely non-linear fit • LOOCV result is easy to calculate	• The model can become overly flexible for if *d* is large	• Usually *d* is 3 or 4.
Regression Splines	• Combining low-degree polynomials at the chosen knots with constraints gives a smooth and non-linear fit • Variations of results in certain ranges can be lowered down through additional constraints (natural spline) • Can produce more stable results than high-degree polynomials • The flexibility of results can be adjusted through number of knots	• The number and location of knots can be hard to optimise.	
Smoothing Splines	• The results are just natural cubic splines • The tuning parameter λ controls the flexibility of results • LOOCV result is easy to calculate	• The number of knots involved in modelling can be very large (the number of unique predictor values) • The tuning parameter λ needs to be optimised	• The effective degrees of freedom play an important role in model performance.
Local Regression	• Adapts well to bias problems at boundaries and in regions of high curvature.	• Need to find optimal span/bandwidth values. • May need to optimise the parametric form for the local models.	• Quadratic or cubic functions are common choices for the local models.

	• Easy to understand and interpret. • Methods have been developed that provide fast computation for one or more independent variables. • Because of its simplicity, can be tailored to work for many different distributional assumptions. • Having a local model enables derivation of response adaptive methods for span value and polynomial order selection in a straightforward manner.	• Need to define a suitable weight function.	• The tricube weight function is a commonly used one.
GAMs	• Allows us to fit a non-linear model to each predictor. • Retains a nice additive structure within the model. • Can deal with different distributional assumptions. • The individual effect of each predictor on the response variable can be assessed. • The smoothness of each building block can be summarized via degrees of freedom.	• The method is restricted to be additive. • Interaction effects are hard to incorporate (not impossible). • The fitted models often don't have explicit expressions.	• It can combine parametric and non-parametric methods.
Logistic regression	• Provides a proper model to study the conditional probability P(Y=1|X). • It retains the linear framework for predictors.	• Logistic regression works better with bivariate response variables. • When the classes are well separated, the logistic regression results can be unstable.	• It results in a linear decision boundary.

LDA	• It works well when n is small and the predictors are approximately normal in each class. • It can handle multi-class response variable well.	• The normal assumption is key to this method. • The predictors are assumed to have same variances among different classes.	• It results in a linear decision boundary. • When the true decision boundary is linear or close to linear, LR and LDA tend to work better, in particular for small data set. If normality is absent, then LR outperforms LDA.
QDA	• It works well when the predictors are approximately normal in each class when the covariance matrix vary among classes.	• It works well when the true decision boundary is close to quadratic. • It wouldn't work well if n is too small.	• It results in a quadratic decision boundary. • Though not as flexible as KNN, QDA can perform better in the presence of a limited number of training observations, as it does make some assumptions about the form of the decision boundary. • When the true decision boundary is quadratic or moderate non-linear, QDA would outperform LR and LDA.
KNN	• It is a non-parametric method that doesn't need any distributional assumptions. • The algorithm is simple and easy to implement. • The algorithm is versatile. It can be used for classification and regression.	• Choosing the right K value is critical. It can be found using CV. • The algorithm gets significantly slower as n and/or p increase. • No inference results for KNN.	• The KNN-CV is the most robust method. It performs relatively well in most cases. It outperforms the rest when the true decision boundary is very non-linear and the normality assumption is not met.

Decision trees	• Trees are very easy to explain. • Some people believe that decision trees more closely mirror human decision-making than do the other regression and classification approaches. • Trees can be displayed graphically, and are easily interpreted even by a non-expert. • Trees can easily handle qualitative predictors without the need to create dummy variables.	• Trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches. • Trees can be very non-robust. A small change in the data can cause a large change in the final estimated tree.	• Tree pruning is useful in improving the prediction accuracy.
Maximal margin classifier	• The maximal margin hyperplane is a natural choice when separating hyperplanes exist in the given data. • It is effective in high-dimensional spaces. • It is usually memory efficient as only based on support vectors.	• The classes in the given data must be separable, i.e. separating hyperplanes must exist. • The maximal margin hyperplane is extremely sensitive to a change in a single observation. • It often overfits the training data when *p* is large. • It doesn’t directly provide probability estimates.	• The maximal margin hyperplane depends directly on only a small subset of the observations, i.e. the support vectors. • The training error rate is always 0.
Support vector classifier	• It adopts soft margins that leads to a greater robustness to individual observations and better classification of most training observations. • The level of flexibility of the classifier can be adjusted through a tuning parameter *C*. • It is effective in high-dimensional spaces.	• It often has poor performance for data with non-linear decision boundaries. • The tuning parameter C needs to be optimised to get the best prediction performance. • It doesn’t work well when the data set contains more noise. • It doesn’t directly provide probability estimates.	• The tuning parameter C controls the bias-variance trade-off of the support vector classifier. Small C values lead to low bias but high variance, and vice versa.

	• It is usually memory efficient as only based on support vectors.		
Support vector machine (SVM)	• The SVM enlarges the feature space used by the support vector classifier in an efficient way. • It can classify data with non-linear decision boundaries. • It is effective in high-dimensional spaces. • It is usually memory efficient as only based on support vectors.	• Selecting the right type of kernel functions with appropriate parameters needs extra work. • It is computationally expensive when we have a large data set. • It doesn't work well when the data set contains more noise. • It doesn't directly provide probability estimates.	
PCA	• It can decorrelate the original features in the data. • Can help with data visualization.	• The number of PC's to keep is hard to determine. • Finding intuitive interpretations for the PC's can be difficult.	• Standardization is something to consider.
K-means clustering	• The algorithm is simple and easy to implement. • It generates the number of clusters we need.	• The results are not stable for single run. Multiple attempts are required. • It can only generate the number of clusters we specify.	•
Hierarchical clustering	• It has an upside-down tree presentation, i.e. a dendrogram. • A single dendrogram can produce any number of clusters between 1 and *n*. • The clusters generated by this method have a nesting relationship. • We don't need to guess the number of clusters before building the dendrogram.	• If the true clusters in the data don't have a nesting relationship, then Hierarchical clustering won't work properly. • The dissimilarity measure and linkage need to be selected carefully.	• Whether to scale the data or not is something to mind. • Small decisions could have big consequences.