Statistical Data Mining
QUIZ

1. **BRIEFLY** describe the following methods for regression. In your description, **include any complexity parameters, 1 advantages, and 1 disadvantage** of the method.:

a) kNN

kNN is the non-parametric regression approach where the output is predicted by measuring the mean of the distance of k nearest neighbours. The parameter k depends on the data and higher value of k reduces variance and increases bias. An advantage of using kNN is that the cost of learning process is zero and is robust to noisy training data. One disadvantage is that the model cannot be interpreted and it has poor performance for high dimensional data.

b) Ordinary Least Squares

OLS is one of the simplest linear regression model which chooses the weights of the variable by minimizing the sum of the squares of the differences between the observed dependent variables and that predicted using linear function. One advantage of OLS is that it is computationally very fast and easy to interpret. But it is very susceptible to outliers(disadvantage).

c) Forwards Subset Selection

Forward subset selection is the method to select the best variables that would best describe(fit the model) the response variable. It starts with 0 variables and starts including variables into the model one-by-one by including one variable at a time into the model and calculating their respective error. It then includes the variable whose model have minimum error. This continues until it traverses all variables. One advantage is that it does not have to iterate over all the possible (2^p)subsets and give almost similar result with respect to best subset selection. One disadvantage is that it may not select the best model all the time.

d) Backwards Subset Selection

Backward and forward subset selection are similar in algorithm structure. But it starts with a complete set of variables as against forward subset selction which starts with null set of variables. It shares the same advantages and disadvantages as forward subset selection. But addition disadvantage of backward subset selection is that to fit the starting model it requires that the number of observations be more than the number of variables.

e) Ridge Regression

Ridge regression is a linear model that minimizes the least square error subject to a maximum value that the summations of the squares of weights can take. One advantage is that it is more flexible than OLS and restricts variance when we increase its parameter lambda. One disadvantage is that it does not ever discard unimportant variables from the model.

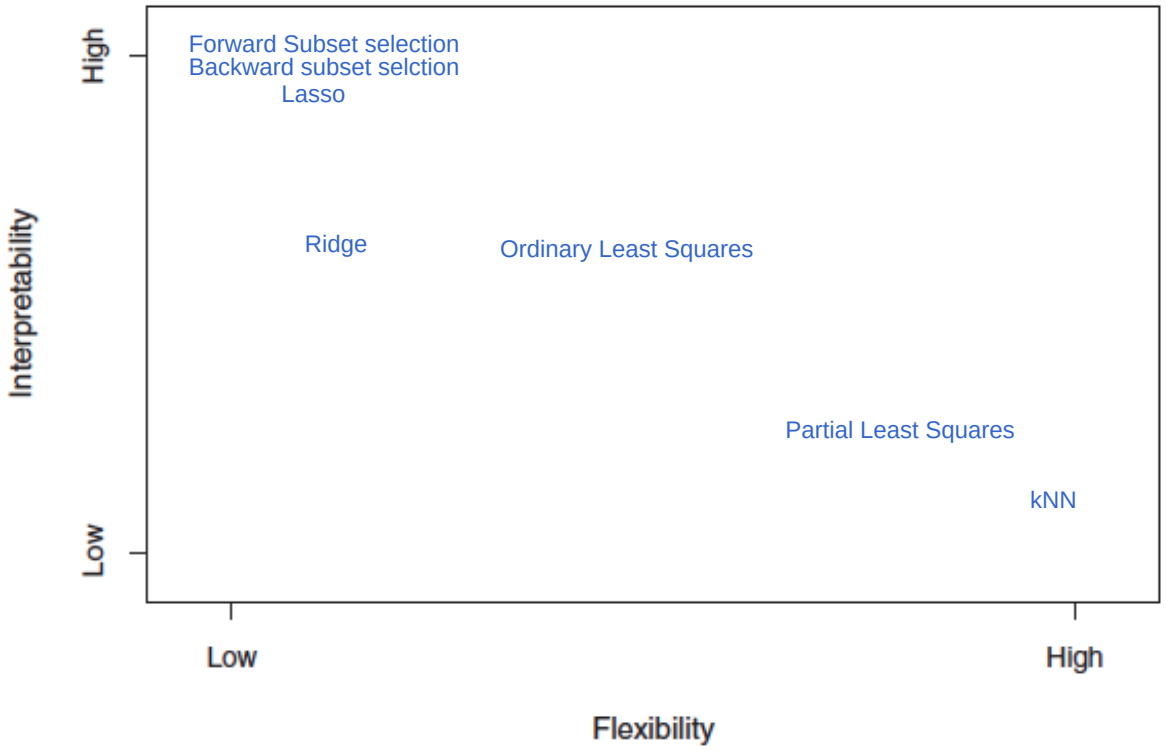
f) LASSO regression

Lasso regression is a linear model that minimizes the least square error subject to a maximum value that the summation of absolute value of weights can take. One advantage is that it is more interpretable as it eliminates out unimportant variables and reduces the weights of the less important variables if the parameter lambda is sufficiently large. One disadvantage is that while eliminating variables it may lose out some relevant independent variables

g) Partial Least Squares

PLS is a supervised alternative to Principal Component Regression which attempts to find directions that explains both the response and the predictors as against PCR that tries to find directions that only explains the predictors. The number M (reduced dimension) of partial least squares directions used in PLS is a tuning parameter and is typically chosen by cross-validation. One advantage is that it is very flexible but has poor interpretability(disadvantage).

2) Arrange the methods from Question 1 on the following diagram.



BONUS: sketch the decision boundary for a 1NN classifier.