HW 1 – STAT 627
Homayoon Fotros

Question 1

Briefly describe what distinguishes a "flexible statistical learning method" versus an "inflexible" method. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

Flexible methods are less restrictive, have a lower level of bias, and provide a wider range of possibilities for the shape of f -the relationship between the model's inputs and outcomes. Despite having a higher level of bias, however, less flexible methods are often easier to use and interpret because they simplify the relationship between the inputs and outcomes. For example, a linear model is inflexible because it fits the input data on a linear plate. In contrast, a non-linear method is more flexible as it can better capture the variation in the data, thus potentially generating more accurate estimates. However, a high level of flexibility may lead to overfitting issue and following the noise in the data. Flexible models such as neural networks and GAMs also require estimating more parameters and perform better on a larger data set.

a. The sample size n is extremely large, and the number of predictors p is small.

A flexible method can perform better because we can capture complex relationships with a limited number of dimensions in a large sample while maintaining a low risk of overfitting issue.

b. The number of predictors p is extremely large, and the number of observations n is small.

An inflexible method will likely perform better in this case. First, the number of observations is small, so the risk of overfitting and finding spurious relationships is high. Second, flexible models usually perform worse than inflexible ones when there is a large set of predictors. In contrast, an inflexible method like linear regression can simplify the relationship and show which predictors are actually significant to obtain a good fit.

c. The relationship between the predictors and response is highly non-linear.

The non-linear relationship between the predictors and response signals using a flexible model. If we use an inflexible model such as linear regression, we should expect the fitted values to be far from the true observed values. In contrast, a non-linear flexible method (e.g., a non-parametric approach) can provide a higher level of accuracy as they avoid having a linear assumption about the shape of f.

d. The variance of the error terms, i.e., $\sigma^2 = \text{Var}(\varepsilon)$, is extremely high.

Using flexible model often generate a high level of variance. If we already know the variance of error terms is high, using a flexile method will increase the overall variance (MSE). Therefore, we may choose an inflexible method to avoid fitting to the noise.

Question 2

Explain whether each scenario is a classification or regression problem and indicate whether we are most interested in inference or prediction. Identify the output/response variable (y), the sample size (n), the inputs/explanatory variables (x), and the number of inputs (p) (provided these are described in a given scenario).

a. We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry, and the CEO salary. We are interested in understanding which factors affect CEO salary.

The outcome variable (y) is the CEO salary. There are 3 inputs (explanatory variables): *firm's profit*, *number of employees*, and *industry*. The sample size consists of 500 firms in the US (n=500). Since the CEO's salary is a continuous quantity, this is a regression problem. We are interested in inference—finding the existence and magnitude of the relationship between the predictors and the outcome.

b. We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

Our data set consists of 20 units (n=20). The response (y) is success or failure (1 or 0). The features (inputs) are *price charged for the product*, *marketing budget*, *competition price*, and ten other variables. which implies that the total number of predictors (p) is 13. This is a classification problem, and we are interested in predicting the class of product (success or failure) based on its features.

Question 3

Come up with three separate real-life applications of statistical learning, (1) one in which classification might be useful, (2) one in which regression might be useful, and (3) one in which cluster analysis might be useful. For each of your examples, describe the response and the predictors and state the goal - inference or prediction.

Classification: A political campaign in UK is preparing an email to reach out to potential voters in the upcoming election. The campaign is interested in knowing the political orientation of the email recipient, so the message can be curtailed accordingly. Five types of voters are: Conservative, Labor, Liberal, Greens, and UKIP. The campaign has collected the age, race, zip code, employment sector, and 7 other attributes of the people in the mailing list. This is a classification problem, since we want to categorize the email recipients into 5 groups. We are interested in prediction, and we have 11 predictors that include a person's demographic information among other variables.

Regression: A pizza restaurant wants to know to what extent the thickness of the dough is associated with pizza leftover in a given day. They have collected a data set including the price, dough thickness in

millimeter, the leftover amount in grams, the topping weight in grams, whether the pizza is served indoors or in the restaurant outdoor patio, the day of the week, outside temperature, whether there was a sport match at the time of serving the pizza, and five other variables. This is a regression problem as the outcome (y)—the amount of leftover is a continuous variable and is estimated based on the mentioned input variables. The restaurant is interested in inference, that is, the relationship between dough thickness and leftover weight when other factors are controlled for.

Clustering: YouTube wants to send customized advertisements to its users based on what they watch and profiles. Based on this information, YouTube aims to create groups of customers and assigns a unique label to them. Each group then will receive a particular set of "ad. treatment" that is deemed more effective on the receiving side. The "cluster of users" are unknown in the first place, but YouTube is interested in bundling users together in a way to simplify its marketing strategies. This is a clustering problem, and the inputs are YouTube users' data and watching history. The goal is to identify user clusters in a way that could predict the most relevant cluster for a given user.