1. 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.

   (a) The sample size m is extremely large, and the number of features n is small

   (b) The number of features n is extremely large, and the number of observations (samples) m is small.

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

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

   SOL:

   (a) flexible model is better. If the data is sufficient, flexible model is likely to better capture relation (even high-order relation) between features and response

   (b) flexible model is worse. It is likely to overfit to the given data

   (c) flexible model is better. Inflexible (simple) model is not likely to capture complex relation between features and responses

   (d) flexible model is worse. It is likely to overfit to the noise in the data

2. Explain whether each scenario is a classification or regression problem. Also provide m, the number of samples, and n, the number of features.

   (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.

   (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.

   (c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

   SOL:

   (a) **TO BE GRADED:**response (CEO salary) is numerical variable: regression. $m = 500$, $n = 3$

   (b) response (success/failure) is categorical variable: classification. $m = 20$, $n = 13$

   (c) response (% change) is numerical variable: regression. $m = 52$, $n = 3$

3. The grades of a class of 9 students on a midterm report x and on the final examination y are as follows:

x	77	50	71	72	81	94	96	99	67
y	82	66	78	34	47	85	99	99	68

(a) Estimate the linear regression line.

(b) Estimate the final examination grade of a student who received a grade of 85 on the midterm report.

SOL:

(a) **TO BE GRADED:**Let $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \\ 1 & x_9 \end{bmatrix}$, $y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_9 \end{bmatrix}$ Then

$$w = (X^T X)^{-1} X^T y$$

gives $w_0 = 12.062$, $w_1 = 0.777$. So

$$\hat{y} = 12.062 + 0.777x$$

(b) When $x = 85$ we have

$$\hat{y} = 12.062 + 0.777 * 85 = 78$$

4. Suppose we would like to classify input x given by

$$x = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$$

Consider parameter W is given by

$$W = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

and bias b given by

$$b = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$$

We would like to use linear score $s = Wx + b$, where the 1st, 2nd and 3rd element of s represents the score for "cat" , "dog" and "ship", respectively.

(a) Calculate the score for "cat" category

(b) Calculate the score for "dog" category

(c) Calculate the score for "ship" category

SOL:

(a) 3

(b) 6

(c) 7

5. In certain areas of ocean, two kinds of fish are caught: salmon and sea bass. Let salmon be class 1 (c_1) and sea bass be class 2 (c_2). The size x of salmon has normal distribution with mean 1 and variance 1, and that of sea bass has normal distribution with mean 3 and variance 1. In other words:

$$f(x|c_1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-1)^2}{2}\right), \quad f(x|c_2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-3)^2}{2}\right)$$

where the probability that is 0.4 and 0.6. Suppose we caught a fish of size x. Let the probability that the fish is salmon is $p(c_1|x)$. Show that $p(c_1|x)$ has the sigmoid form

$$p(c_1|x) = \sigma(w_0 + w_1 x),$$

where

$$\sigma(t) := \frac{1}{1+e^{-t}}$$

Also, find w_0 and w_1.
SOL: Using Bayes' theorem, we have that

$$p(c_1|x) = \frac{p(x|c_1)p(c_1)}{p(x|c_1)p(c_1) + p(x|c_2)p(c_2)} = \frac{1}{1 + \frac{p(x|c_2)p(c_2)}{p(x|c_1)p(c_1)}} = \frac{1}{1 + \exp\left(2x - 4 + \log(3/2)\right)} = \sigma(-2x+4-\log(3/2))$$

Thus, $w_0 = 4 - \log(3/2)$ and $w_1 = -2$.

6. **Naive Bayes classification.** This example will guide you through building a Naive Bayes (NB) classifier for discrete-valued input data. Note that this model is different from logistic regression, and does not use parameterized estimates of probabilities, but uses simple estimates directly from data.

   Consider the data in Table 1 of data regarding some illness in a hospital. Note $+$ and $-$ of three symptoms, Fever, Cough, Nose (runny nose) means the presence of symptom is positive and negative respectively. These $+$ or $-$ values of symptoms are input data. The output data is the presence of Illness, again $+$ or $-$ as positive or negative.

   Suppose a doctor received new patient, Tommy, with symptoms: Fever:$-$, Cough:$-$, Nose: $+$. We would like to decide whether this patient is ill using NB classifier.

   In this binary decision problem, we would like to decide which one is greater of

$$P(I_-|F_-, C_-, N_+) \text{ vs } P(I_+|F_-, C_-, N_+)$$

   Note I_+ and I_- denotes the event of illness positive and negative respectively. Similarly F_+ and F_- denote the event of Fever positive or negative: other notations are defined similarly for C_- (cough) and N_+ (nose). In other words, which one is greater: probability of being ill or not, given that the patient's symptom is Fever:$-$, Cough:$-$, Nose: $+$? From Bayes' rule, this is equivalent to asking

$$\frac{P(F_-, C_-, N_+|I_+)P(I_+)}{P(F_-, C_-, N_+)} \text{ vs } \frac{P(F_-, C_-, N_+|I_-)P(I_-)}{P(F_-, C_-, N_+)}$$

   or, because the denominators are the same,

$$P(F_-, C_-, N_+|I_+)P(I_+) \text{ vs } P(F_-, C_-, N_+|I_-)P(I_-)$$

   From conditional independence assumption of NB,

$$P(F_-, C_-, N_+|I_+) = P(F_-|I_+)P(C_-|I_+)P(N_+|I_+)$$
$$P(F_-, C_-, N_+|I_-) = P(F_-|I_-)P(C_-|I_-)P(N_+|I_-)$$

   There are several probabilities that needs to be estimated from the data:

Fever	Cough	Nose	Illness
+	+	+	+
+	-	-	+
-	+	-	+
-	+	+	-
+	-	+	-
-	-	-	-

Table 1:

- prior probabilities $P(I_+), P(I_-)$
- posterior probabilities $P(F_-|I_+), P(C_-|I_+), P(N_+|I_+), P(F_-|I_-), P(C_-|I_-), P(N_+|I_-)$.

(a) We want to estimate the probabilities $P(I_+)$ and $P(I_-)$. Make a maximum likelihood estimate (MLE) of $P(I_+)$. Note $P(I_-)$ is simply $1 - P(I_+)$.

(b) We want to estimate the probability $P(F_-|I_+)$. Make a MLE of $P(F_-|I_+)$.

(c) Use similar approach to make MLE of $P(C_-|I_+), P(N_+|I_+), P(F_-|I_-), P(C_-|I_-)$ and $P(N_+|I_-)$.

(d) Based on your estimates, what is the doctor's decision on whether Tommy is ill(+) or not(-)?

SOL: Let us define the following:

$$p_{i+} = P(I_+),\ p_{F+} = P(F_+|I_+), p_{F-} = P(F_+|I_-),$$
$$p_{N+} = P(N_+|I_+), p_{N-} = P(N_+|I_-),\ p_{C+} = P(C_+|I_+), p_{C-} = P(C_+|I_-),$$

The likelihood function for the data in Table 1 can be represented using these probabilities. For example, for the likelihood of the first sample

$$P(I_+|F_+, C_+, N_+) \propto P(F_+, C_+, N_+|I_+)P(I_+) = P(F_+|I_+)P(C_+|I_+)P(N_+|I_+)P(I_+) = p_{F+}p_{C+}p_{N+}p_{i+}$$

Similarly, the likelihood for the second sample is $p_{F+}(1 - p_{C+})(1 - p_{N+})p_{i+}$, etc. Thus the overall likelihood Λ of data can be represented to be proportional to the product of these probabilities. Now, given the likelihood (or logarithm of it), we know how to maximize it: (partial) differentiate it and set to zero. For example, we know that the likelihood will have the factor

$$p_{i+}^3(1 - p_{i+})^3$$

because there are three + and three - in the illness data. By taking the partial differentiation of λ with respect to p_{i+} and setting it to zero, we see that the optimal $p_{i+} = 0.5$. We can find similar maximum likelihood estimates for other probabilities $p_{i+},\ p_{F+}, p_{F-},\ p_{N+}, p_{N-}\ p_{C+}, p_{C-}$

(a) This is the optimal value of p_{i+} which is the maximum likelihood estimate. $P(I_+) = 3/6 = 0.5$. $P(I_-) = 1 - P(I_+) = 0.5$

(b) The maximum likelihood estimate for $p_{F+} = 2/3$, because there are two cases of $+$ and 1 case of $-$ Fever, given that Illness is $+$. Thus $P(F_-|I_+) = 1 - p_{F+} = 1/3$,

(c) $P(C_-|I_+) = 1/3, P(N_+|I_+) = 1/3$ and $P(F_-|I_-) = 2/3, P(C_-|I_-) = 2/3$ and $P(N_+|I_-) = 2/3$

(d) Using our estimates, we have that

$$P(F_-, C_-, N_+|I_+)P(I_+) = P(F_-|I_+)P(C_-|I_+)P(N_+|I_+)P(I_+) = 1/54$$

where

$$P(F_-, C_-, N_+|I_-)P(I_-) = P(F_-|I_-)P(C_-|I_-)P(N_+|I_-)P(I_-) = 8/54$$

So our decision is, the new patient is not ill or negative $(-)$

7. Suppose we have the score

$$s = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

What is the softmax applied to s?
SOL:

$$\begin{bmatrix} \frac{\exp(2)}{\exp(2)+\exp(1)+\exp(0)} \\ \frac{\exp(1)}{\exp(2)+\exp(1)+\exp(0)} \\ \frac{\exp(0)}{\exp(2)+\exp(1)+\exp(0)} \end{bmatrix}$$