1. (a) better - When n is large and p is small, the data is likely representative of the true pattern, so we want a method that fits better to the data; a more flexible approach will fit the data closer and with the large sample size a better fit than an inflexible approach would be obtained

(b) worse - a flexible method would overfit the small number of observations

(c) better - with more degrees of freedom, a flexible model would obtain a better fit for the non-linearity that an inflexible model

(d) worse - flexible methods fit to the noise in the error terms and increase variance

2. (a) regression. inference. quantitative output of CEO salary based on CEO firm's features.

n - 500 firms in the US

p - profit, number of employees, industry

(b) classification. prediction. predicting new product's success or failure.

n - 20 similar products previously launched

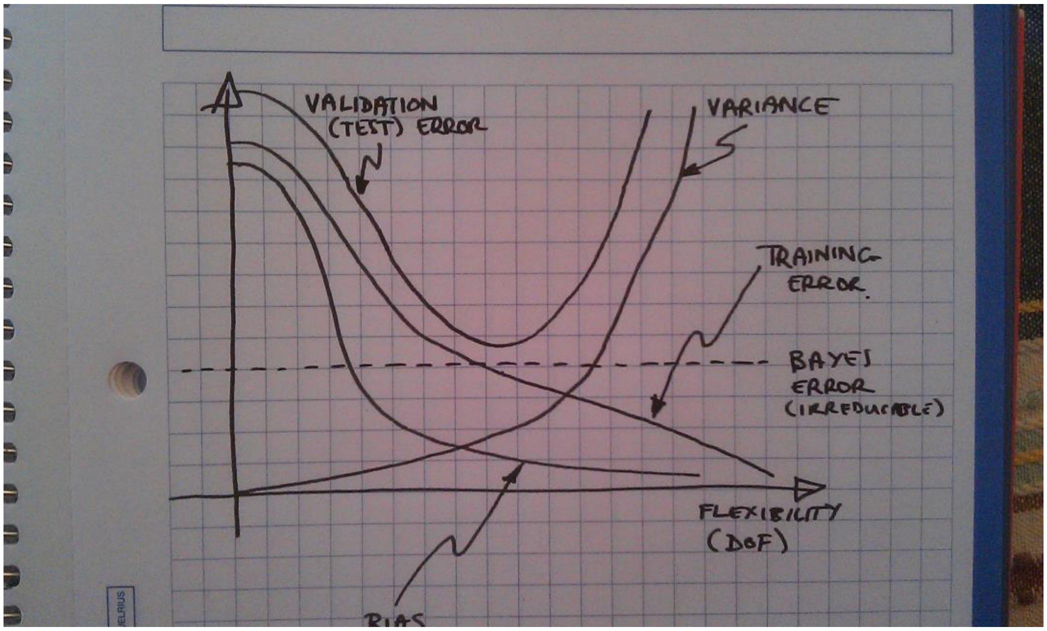
p - price charged, marketing budget, comp. price, ten other variables

(c) regression. prediction. quantitative output of % change

n - 52 weeks of 2012 weekly data

p - % change in US market, % change in British market, % change in German market

3. (a) See 3a.jpg.



(b)

all 5 lines >= 0

i. (squared) bias - decreases monotonically because increases in flexibility yield a closer fit

ii. variance - increases monotonically because increases in flexibility yield overfit

iii. training error - decreases monotonically because increases in flexibility yield a closer fit

iv. test error - concave up curve because increase in flexibility yields a closer fit before it overfits

v. Bayes (irreducible) error - defines the lower limit, the test error is bounded below by the irreducible error due to variance in the error (epsilon) in the output values (0 <= value). When the training error is lower than the irreducible error, overfitting has taken place.

The Bayes error rate is defined for classification problems and is determined by the ratio of data points which lie at the 'wrong' side of the decision boundary, (0 <= value < 1).

4. (a) i. stock market price direction, prediction, response: up, down, input: yesterday's price movement % change, two previous day price movement % change, etc.

ii. illness classification, inference, response: ill, healthy, input: resting heart rate, resting breath rate, mile run time

iii. car part replacement, prediction, response: needs to be replace, good, input: age of part, mileage used for, current amperage

(b) i. CEO salary. inference. predictors: age, industry experience, industry, years of education. response: salary.

ii. car part replacement. inference. response: life of car part. predictors: age of part, mileage used for, current amperage.

iii. illness classification, prediction, response: age of death, input: current age, gender, resting heart rate, resting breath rate, mile run time.

(c) i. cancer type clustering. diagnose cancer types more accurately.

ii. Netflix movie recommendations. recommend movies based on users who have watched and rated similar movies.

iii. marketing survey. clustering of demographics for a product(s) to see which clusters of consumers buy which products.

5. The advantages for a very flexible approach for regression or classification are obtaining a better fit for non-linear models, decreasing bias. The disadvantages for a very flexible approach for regression or classification are requires estimating a greater number of parameters, follow the noise too closely (overfit), increasing variance.

A more flexible approach would be preferred to a less flexible approach when we are interested in prediction and not the interpretability of the results.

A less flexible approach would be preferred to a more flexible approach when we are interested in inference and the interpretability of the results.

6. A parametric approach reduces the problem of estimating f down to one of estimating a set of parameters because it assumes a form for f.

A non-parametric approach does not assume a functional form for f and so requires a very large number of observations to accurately estimate f.

The advantages of a parametric approach to regression or classification are the simplifying of modeling f to a few parameters and not as many observations are required compared to a non-parametric approach.

The disadvantages of a parametric approach to regression or classification are a potential to inaccurately estimate f if the form of f assumed is wrong or to overfit the observations if more flexible models are used.

7. (a)

Obs.	X1	X2	X3	Distance(0, 0, 0)	Y
1	0	3	0	3	Red
2	2	0	0	2	Red
3	0	1	3	sqrt(10) ~ 3.2	Red
4	0	1	2	sqrt(5) ~ 2.2	Green
5	-1	0	1	sqrt(2) ~ 1.4	Green
6	1	1	1	sqrt(3) ~ 1.7	Red

(b) Green. Observation #5 is the closest neighbor for K = 1.

(c) Red. Observations #2, 5, 6 are the closest neighbors for K = 3. 2 is Red, 5 is Green, and 6 is Red.

(d) Small. A small K would be flexible for a non-linear decision boundary, whereas a large K would try to fit a more linear boundary because it takes more points into consideration.