

**SDSC 3006 Fundamentals of Machine Learning I**  
**Assignment #1**

**Deadline: October 2, Sunday@ 10:00 PM**

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  $n$  is extremely large, and the number of predictors  $p$  is small.
- (b) The number of predictors  $p$  is extremely large, and the number of observations  $n$  is small.
- (c) The relationship between the predictors and response is highly non-linear.
- (d) The variance of the error terms, i.e.  $\sigma^2 = \text{Var}(\epsilon)$ , is extremely high.

2. We now revisit the bias-variance decomposition.

- (a) Provide a sketch of typical (squared) bias, variance, training error, and test error, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be four curves. Make sure to label each one.
- (b) Explain why each of the four curves has the shape displayed in part (a).

3. Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Gender}$  (1 for Female and 0 for Male),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Gender}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\hat{\beta}_0 = 50$ ,  $\hat{\beta}_1 = 20$ ,  $\hat{\beta}_2 = 0.07$ ,  $\hat{\beta}_3 = 35$ ,  $\hat{\beta}_4 = 0.01$ ,  $\hat{\beta}_5 = -10$ .

- (a) Which answer is correct, and why?
  - i. For a fixed value of IQ and GPA, males earn more, on average, than females.
  - ii. For a fixed value of IQ and GPA, females earn more, on average, than males.
  - iii. For a fixed value of IQ and GPA, males earn more, on average, than females provided that the GPA is high enough.
  - iv. For a fixed value of IQ and GPA, females earn more, on average, than males provided that the GPA is high enough.
- (b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.
- (c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

4. Using the **Carseats** data set to answer the following questions.

- (a) Fit a multiple regression model to predict **Sales** using **Price**, **Urban**, and **US**.
- (b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!
- (c) Write out the model in equation form, being careful to handle the qualitative variables properly.
- (d) For which of the predictors can you reject the null hypothesis  $H_0: \beta_j = 0$ ?
- (e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.
- (f) How well do the models in (a) and (e) fit the data?
- (g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).
- (h) Is there evidence of outliers or high leverage observations in the model from (e)?