STA 325: Homework 2

DUE: 11:59pm, September 27 (on Sakai)
COVERAGE: ISL Chapters 5.1, 6.1

1. **[15 points]** We discussed in class two common ways of performing model selection. The first is using (i) a "test-error criterion" (e.g., AIC or BIC), and the second is using (ii) cross-validation.

   (a) Compare and contrast (i) and (ii) for model selection. What are the advantages and disadvantages of each type of method? When should a data analyst prefer one over the other?

   Advantages for (i): closed-form criterion which can efficiently computed, interpretable as an adjustment on training error. Advantages for (ii): direct estimate of test error (instead of an approximation), no need to estimate noise variance or flexibility, easily generalizable to more complex models.

   A data analyst may prefer to use K-fold CV to fit more complex models, but may choose to use AIC / BIC for simpler models where model flexibility and noise variance can be more easily estimated. AIC / BIC may also be useful in scenarios with massive data (but this may become a less important factor given the advent of parallel computing systems).

   (b) For (i), compare and contrast AIC and BIC for model selection. When should a data analyst prefer one over the other?

   BIC imposes a higher penalty term than AIC in nearly all practical scenarios ($n \geq 7$). AIC should be used when predictive power is more important than inference; BIC should be used when inference is more important than predictive power, since it returns a more parsimonious model.

   (c) For linear regression, the AIC is equivalent to the classical Mallow's C_p criterion:

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2).$$

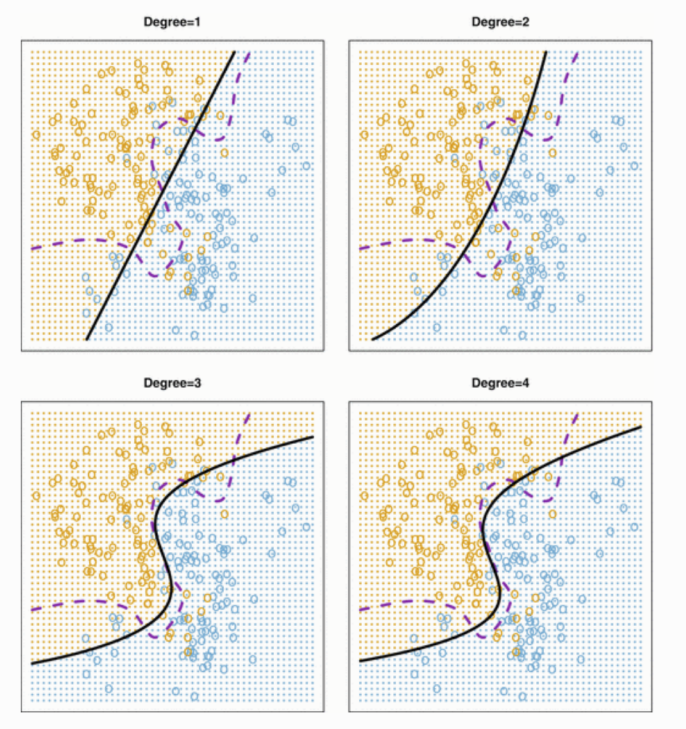
   Intuitively, explain why the adjustment term $2d\hat{\sigma}^2$ should penalize models with more parameters d or larger (estimated) irreducible noise variance $\hat{\sigma}^2$.

   The number of parameters d should be penalized, since a larger d results in higher model complexity. The irreducible noise variance estimate $\hat{\sigma}^2$ should also be penalized, since larger irreducible noise increases the risk of overfitting the model.

   (d) Suppose you have a large dataset, and are choosing between 10 potential models. Let $\hat{f}_1$ be the fitted model selected by AIC, and $\hat{f}_2$ be the fitted model selected by BIC. Which model do you expect to have lower variance, $\text{Var}\{\hat{f}(x)\}$? Which model do you expect to have lower bias, $\text{Bias}\{\hat{f}(x)\}$? Explain your answer.

   Since BIC imposes a greater penalty than AIC, the fitted model from BIC often has less parameters than the fitted model from AIC. Hence, the BIC model will often have higher bias but lower variance.

2. **[20 points]** Consider the following four classifiers, obtained by fitting logistic regression models with different polynomial degrees. The purple dotted lines show the Bayes-optimal classifier, and the black solid lines show the fitted classifier for the four logistic regression models.



More specifically, logistic regression is fit on the following four models:

$$\text{(degree 1):} \quad \text{logit}\{p(x)\} = \beta_0 + \sum_{j=1}^{2} \beta_{j,1} x_j$$

$$\text{(degree 2):} \quad \text{logit}\{p(x)\} = \beta_0 + \sum_{j=1}^{2} \beta_{j,1} x_j + \sum_{j=1}^{2} \beta_{j,2} x_j^2$$

$$\text{(degree 3):} \quad \text{logit}\{p(x)\} = \beta_0 + \sum_{j=1}^{2} \beta_{j,1} x_j + \sum_{j=1}^{2} \beta_{j,2} x_j^2 + \sum_{j=1}^{2} \beta_{j,3} x_j^3$$

$$\text{(degree 4):} \quad \text{logit}\{p(x)\} = \beta_0 + \sum_{j=1}^{2} \beta_{j,1} x_j + \sum_{j=1}^{2} \beta_{j,2} x_j^2 + \sum_{j=1}^{2} \beta_{j,3} x_j^3 + \sum_{j=1}^{2} \beta_{j,4} x_j^4$$

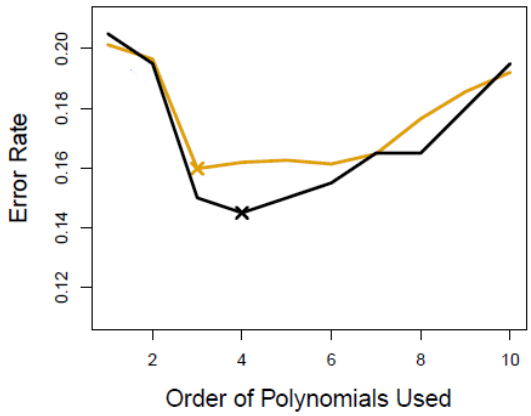
(a) How many parameters are needed for each of the four models? How many parameters are needed for a polynomial model of degree M?

Degree 1: 3; degree 2: 5; degree 3: 7; degree 4: 9. For a polynomial model of degree M, $2M + 1$ parameters are needed.

(b) What can you say about the bias and variance of the fitted model with degree = 1? With degree 2? With degrees 3 and 4?

For degree 1, very high bias and very low variance. For degree 2, high bias and low variance. For degrees 3 and 4, moderate bias and moderate variance. Full marks as long as the answer alludes to the decreasing bias and increasing variance as degree increases.

(c) Plotted below are the true test errors (in black) and the estimated test errors using 10-fold CV (in orange) as a function of model polynomial degree. Comment on these error curves: What is the order of the true optimal model and the estimated optimal model? Why is the black curve (generally) lower than the orange curve?



The optimal model from the true test error curve is the degree-4 polynomial model, whereas the optimal model from the estimated error curve is the degree-3 polynomial model. The true error curve (black) is generally lower than the estimated error curve (orange), since 10-fold CV has a positive bias when estimating the true test error. This is because the testing error for a model fitted with 90% of the data is higher than the testing error for a model fitted with the full data.

(d) In this 0-1 classification problem, model complexity is varied by the order of polynomials used in logistic regression. Give two other ways to vary model complexity for this problem. For each, identify a variable which affects complexity (e.g., polynomial order), and explain how changing this variable affects the complexity of the fitted model.

Several possible answers here:

- *Adding interactions (logistic regression)*: Let M be the maximum order of interactions fit in the model. For example, in the two-predictor problem here,

$M = 2$ would fit the following model:

$$\text{logit}\{p(x)\} = \beta_0 + \sum_{j=1}^{2} \beta_{j,1} x_j + \beta_{j,1\times 2} x_1 x_2.$$

Here, increasing M would increase the complexity of the model (since more parameters are needed to fit higher-order interactions), and vice versa.

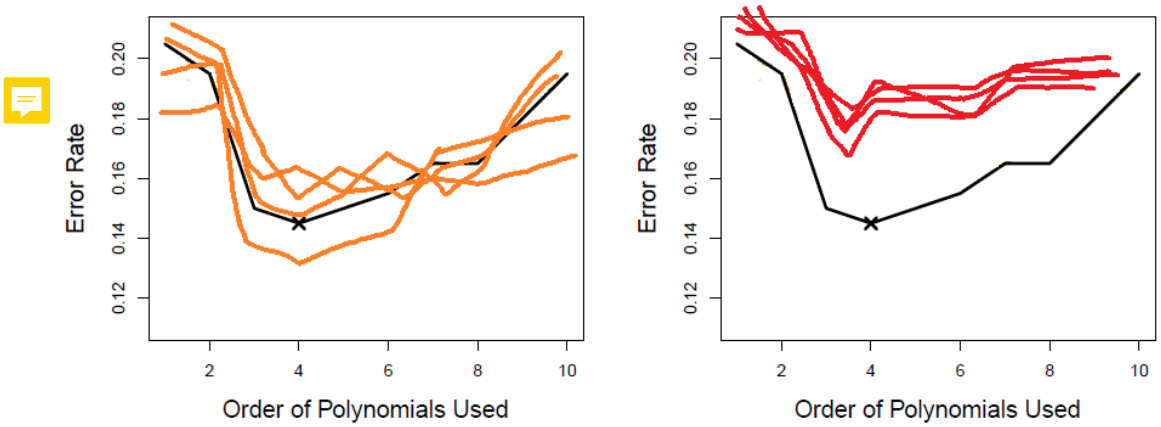
- *Adding predictors (logistic regression)*: Let M be the number of predictors used to fit the model. For example, if we had $M \geq 2$ predictors here, then the fitted model would be:

$$\text{logit}\{p(x)\} = \beta_0 + \sum_{j=1}^{M} \beta_{j,M} x_j.$$

A larger M would increase the complexity of the model (since more parameters are needed to model the effects from additional predictors), and vice versa.

- *Number of neighbours (KNN)*: Let K be the number of neighbours used in KNN. Then a smaller choice of K increases model complexity (since a smaller number of neighbours are used for averaging, which results in a more complex classification model), and vice versa.

(e) [**BONUS: 5 points**] Consider again the plot in part (c). Suppose we repeat the following two steps: (i) collect a new batch of training data from the same population (with same sample size n), and (ii) compute the K-fold CV test curve. Draw several plausible curves for n-fold CV (i.e., LOOCV) on the left plot below, and draw several plausible curves for 2-fold CV on the right plot below. Comment on how changing K (the number of folds) affects the bias-variance trade-off for estimating the true error curve.



Figures: The left plot (for LOOCV) should have low bias with respect to the black curve, but should have very high estimation variance. As a result, the optimum model varies greatly over different folds of the data. The right plot (for 2-fold

CV) should have high bias with respect to the black curve, but should have low estimation variance. As a result, the optimum model will be quite consistent between different folds of the data, but this estimated optimum may not be the true optimum due to high bias.

In general, increasing K (the number of folds) decreases bias but increase variance for test error estimation, and vice versa for decreasing K.

Chapter 5, Exercise 8

(a)

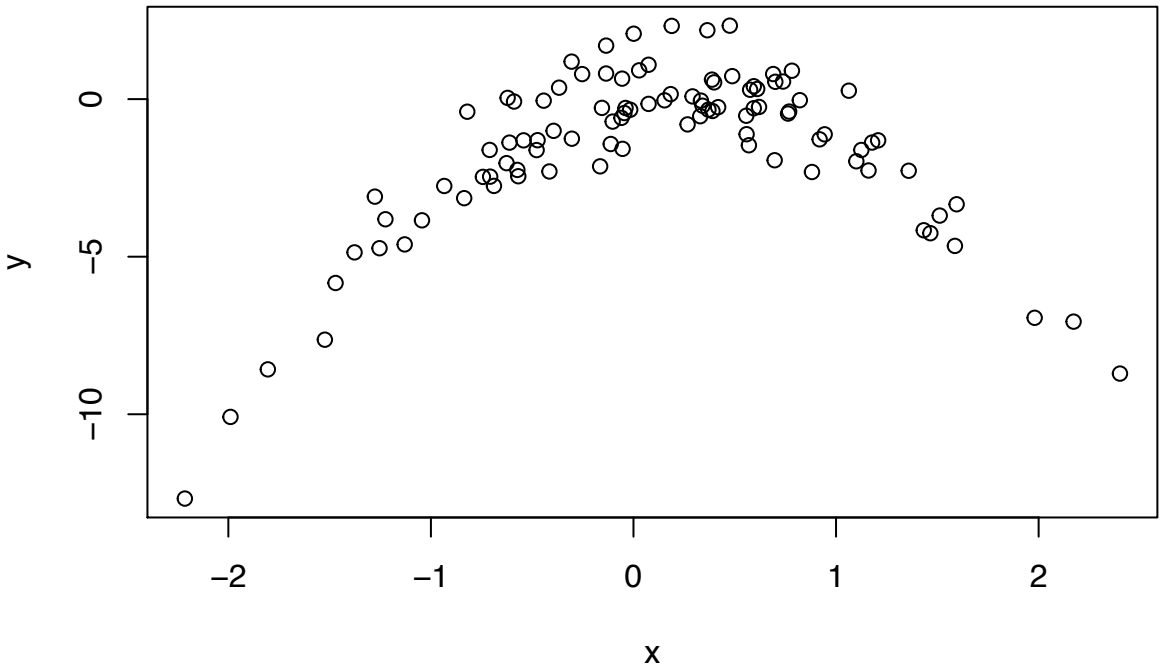
```
set.seed(1)
x = rnorm(100)
y = x - 2*x^2 + rnorm(100)
```

$n = 100$, $p = 2$. Note that p refers to the number of predictors.

$Y = X - 2X^2 + \epsilon$

(b)

```
plot(x, y)
```



In the plot, we can see what appears to be a quadratic relationship and a moderate amount of noise.

(c)

```
library(boot)
set.seed(1)
Data = data.frame(x, y)
```

(i)

```
lm.fit = glm(y ~ poly(x, 1), data=Data)
cv.glm(Data, lm.fit)$delta
```

```
## [1] 7.288162 7.284744
```

(ii)

```
lm.fit = glm(y ~ poly(x, 2), data=Data)
cv.glm(Data, lm.fit)$delta
```

```
## [1] 0.9374236 0.9371789
```

(iii)

```
lm.fit = glm(y ~ poly(x, 3), data=Data)
cv.glm(Data, lm.fit)$delta
```

```
## [1] 0.9566218 0.9562538
```

(iv)

```
lm.fit = glm(y ~ poly(x, 4), data=Data)
cv.glm(Data, lm.fit)$delta
```

```
## [1] 0.9539049 0.9534453
```

(d)

```
set.seed(10)
```

(i)

```
lm.fit = glm(y ~ poly(x, 1), data=Data)
cv.glm(Data, lm.fit)$delta
```

```
## [1] 7.288162 7.284744
```

(ii)

```
lm.fit = glm(y ~ poly(x, 2), data=Data)
cv.glm(Data, lm.fit)$delta
```

```
## [1] 0.9374236 0.9371789
```

(iii)

```
lm.fit = glm(y ~ poly(x, 3), data=Data)
cv.glm(Data, lm.fit)$delta
```

```
## [1] 0.9566218 0.9562538
```

(iv)

```
lm.fit = glm(y ~ poly(x, 4), data=Data)
cv.glm(Data, lm.fit)$delta
```

```
## [1] 0.9539049 0.9534453
```

The results are identical. LOOCV errors are deterministic, because each fold contains just a single point, so there is no randomness in the training/validation splits.

(e)

The model with the quadratic polynomial had the lowest LOOCV test error. This makes sense, since we know that the true relationship is quadratic (since we generated the data).

Chapter 6, Exercise 1

(a)

Best subset selection has the smallest (or is tied for the smallest) training RSS. This is because best subset selection explores every single possible combination of predictors, so it is guaranteed to find the one with the smallest training RSS. In contrast, forward and backward stepwise selection only consider a subset of all possible combinations, so they will not necessarily find the model with the lowest training RSS.

(b)

It's impossible to say, given this information. The performance on the test set will generally depend on the extent to which the learning method overfits on the training data. If we assume that overfitting is not a big problem, we will likely expect the best subset selection to give the lowest test RSS, but this assumption is not necessarily valid.

(c)

(i)

True. The $(k+1)$-variable model is generated by adding one more predictor to the k-variable model.

(ii)

True. The k-variable model is generated by removing one variable from the $(k+1)$ variable model.

(iii)

False. There is no simple relationship between the models generated by backward/forward selection.

(iv)

False. There is no simple relationship between the models generated by backward/forward selection.

(v)

False. These two best subset models are formed by trying all combinations of k or $k+1$ predictors, so there is no guarantee that the k variables will overlap with the $k+1$ variables.