

# Homework 04

STAT 430, Fall 2017

Due: Friday, October 6, 11:59 PM

Please see the [homework instructions document](#) for detailed instructions and some grading notes. Failure to follow instructions will result in point reductions.

---

## Exercise 1 (Comparing Classifiers)

[8 points] This exercise will use data in [hw04-trn-data.csv](#) and [hw04-tst-data.csv](#) which are train and test datasets respectively. Both datasets contain multiple predictors and a categorical response  $y$ .

The possible values of  $y$  are "dodgerblue" and "darkorange" which we will denote mathematically as  $B$  (for blue) and  $O$  (for orange).

Consider four classifiers.

$$\hat{C}_1(x) = \begin{cases} B & x_1 > 0 \\ O & x_1 \leq 0 \end{cases}$$

$$\hat{C}_2(x) = \begin{cases} B & x_2 > x_1 + 1 \\ O & x_2 \leq x_1 + 1 \end{cases}$$

$$\hat{C}_3(x) = \begin{cases} B & x_2 > x_1 + 1 \\ B & x_2 < x_1 - 1 \\ O & \text{otherwise} \end{cases}$$

$$\hat{C}_4(x) = \begin{cases} B & x_2 > (x_1 + 1)^2 \\ B & x_2 < -(x_1 - 1)^2 \\ O & \text{otherwise} \end{cases}$$

Obtain train and test error rates for these classifiers. Summarize these results using a single well-formatted table.

- Hint: Write a function for each classifier.
  - Hint: The `ifelse()` function may be extremely useful.
- 

## Exercise 2 (Creating Classifiers with Logistic Regression)

[8 points] We'll again use data in [hw04-trn-data.csv](#) and [hw04-tst-data.csv](#) which are train and test datasets respectively. Both datasets contain multiple predictors and a categorical response  $y$ .

The possible values of  $y$  are "dodgerblue" and "darkorange" which we will denote mathematically as  $B$  (for blue) and  $O$  (for orange).

Consider classifiers of the form

$$\hat{C}(x) = \begin{cases} B & \hat{p}(x) > 0.5 \\ O & \hat{p}(x) \leq 0.5 \end{cases}$$

Create (four) classifiers based on estimated probabilities from four logistic regressions. Here we'll define  $p(x) = P(Y = B | X = x)$ .

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0$$

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2$$

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$

Note that, internally in `glm()`, R considers a binary factor variable as 0 and 1 since logistic regression seeks to model  $p(x) = P(Y = 1 | X = x)$ . But here we have "dodgerblue" and "darkorange". Which is 0 and which is 1? Hint: Alphabetically.

Obtain train and test error rates for these classifiers. Summarize these results using a single well-formatted table.

---

### Exercise 3 (Bias-Variance Tradeoff, Logistic Regression)

[8 points] Run a simulation study to estimate the bias, variance, and mean squared error of estimating  $p(x)$  using logistic regression. Recall that  $p(x) = P(Y = 1 | X = x)$ .

Consider the (true) logistic regression model

$$\log\left(\frac{p(x)}{1-p(x)}\right) = 1 + 2x_1 - x_2$$

To specify the full data generating process, consider the following R function.

```
make_sim_data = function(n_obs = 25) {  
  x1 = runif(n = n_obs, min = 0, max = 2)  
  x2 = runif(n = n_obs, min = 0, max = 4)  
  prob = exp(1 + 2 * x1 - 1 * x2) / (1 + exp(1 + 2 * x1 - 1 * x2))  
  y = rbinom(n = n_obs, size = 1, prob = prob)  
  data.frame(y, x1, x2)  
}
```

So, the following generates one simulated dataset according to the data generating process defined above.

```
sim_data = make_sim_data()
```

Evaluate estimates of  $p(x_1 = 1, x_2 = 1)$  from fitting three models:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0$$

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2$$

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1^2 + \beta_4x_2^2 + \beta_5x_1x_2$$

Use 1000 simulations of datasets with a sample size of 25 to estimate squared bias, variance, and the mean squared error of estimating  $p(x_1 = 1, x_2 = 1)$  using  $\hat{p}(x_1 = 1, x_2 = 1)$  for each model. Report your results using a well formatted table.

At the beginning of your simulation study, run the following code, but with your nine-digit Illinois UIN.

```
set.seed(123456789)
```

---

## Exercise 4 (Concept Checks)

[1 point each] Answer the following questions based on your results from the three exercises.

- (a) Based on your results in Exercise 1, do you believe that the true decision boundaries are linear or non-linear?
- (b) Based on your results in Exercise 2, which of these models performs best?
- (c) Based on your results in Exercise 2, which of these models are underfitting?
- (d) Based on your results in Exercise 2, which of these models are overfitting??
- (e) Based on your results in Exercise 3, which models are performing unbiased estimation?
- (f) Based on your results in Exercise 3, which of these models performs best?