COMP-SCI5565: Exam 1 Mar.4, 2018 **NAME:** _____

This exam has six questions (for 70 points). You are to write legibly, show all work. I cannot give partial credit if no work is shown. References are made to definitions and methods as used in class and/or book. If you use additional assumptions, you have to state your definitions and assumptions clearly, otherwise they may be misinterpreted. This exam is closed book, closed notes, no calculators, and to be taken without help or assistance. Suspected academic dishonesty will be reported.

You may leave the answers in terms of factorials, permutations, combinations, binomial coefficients, logarithms, exponentials, summations, products, ..., whichever is appropriate. GOOD LUCK

---

1. (10 points) Suppose we collect data for a group of students in a statistics class with variables $X_1 =$ hours studied, $X_2 =$ undergrad GPA, and $Y =$ receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6, \hat{\beta}_1 = 0.05, \hat{\beta}_2 = 1$.

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...}}{1+e^{\beta_0+\beta_1 X_1+...}}$$

   (a) (5 points) Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.

   (b) (5 points) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?.

2. (20 points) Assume two data sets sampled from the same distribution where the number of observations for each data is 8,000 and 100,000 respectively.

   (a) (5 points) Draw two curves for training error test error and true error for each data set with y-axis denoting the error and x-axis denoting the model complexity. You should have a total of 4 curves: one training error and one test error curve for each data set. Draw all 4 of them in the same diagram. Also, Draw the True Error. Clearly *mark* all your curves.

   (b) (5 points) Give an explanation for each curve drawn.

3. (20 points) Suppose you are solving a regression problem and have the choice of the following two functions:

   (a) $Y = \widehat{\beta}_0 + \widehat{\beta}_1 X + \epsilon$

   (b) $Y = \widehat{\beta}_0 + \widehat{\beta}_1 X + \widehat{\beta}_2 X^2 + \widehat{\beta}_3 X^3 + \epsilon$

   For each question below, determine which function would be the best choice, (or neither if it doesn't matter). To get credit, you must explain your choice.

   (a) (5 points) The sample size n is extremely large, and the number of predictors p is small.Explain:

   (b) (5 points) The number of predictors p is extremely large, and the number of observations n is small. Explain:

   (c) (5 points) The relationship between the predictors and response is highly non-linear. Explain:

   (d) (5 points) The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high. Explain:

4. (16 points) Circle the correct answer and give justification in the space provided: (justification must be given and must advocate circled answer to earn any credit)

   (a) (2 points) Using a model with less bias is always better than using a model with more bias **T F** Explain:

   (b) (2 points) The Variance of a model typically decreases as the number of features increases. **T F** Explain:

   (c) (2 points) To predict the probability of an event, one would prefer a linear regression model trained with squared error to a classifier trained with logistic regression. **T F** Explain:

   (d) (2 points) When we have less data and the model complexity remains the same, overfitting is more likely **T F** Explain:

   (e) (2 points) When our data points have fewer predictors(features), overfitting is more likely **T F** Explain:

   (f) (2 points) As linear regression is given more and more data, its training data will eventually approach 0 (assuming there is no noise in the data) **T F** Explain:

   (g) (2 points)(circle the correct answers) Underfitting is generally a symptom of high **Bias** | **Variance**, while Overfitting is generally a symptom of high **Bias** | **Variance** Explain:

   (h) (2 points)The following model can be learned by linear regression: $y_i = e^{\beta_0+\beta_1 X_1+\ldots+\epsilon}$ where $\epsilon \sim N(0, \sigma^2)$ is iid Gaussian noise **T F** Explain:

5. (20 points) Suppose we have a data set with five predictors, X_1 =GPA, X_2 = IQ, X_3 = Gender (1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = 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 $\widehat{\beta}_0 = 50$, $\widehat{\beta}_1 = 20$, $\widehat{\beta}_2 = 0.07$, $\widehat{\beta}_3 = 35$, $\widehat{\beta}_4 = 0.01$, $\widehat{\beta}_5 = -10.0$

   (a) (8 points) Which answer is correct, and why?

      i. (2 point) For a fixed value of IQ and GPA, males earn more on average than females. **True** or **False** Explain:

      ii. (2 point) For a fixed value of IQ and GPA, females earn more on average than males. **True** or **False** Explain:

      iii. (2 point) For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.**True** or **False** Explain:

      iv. (2 point) For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.**True** or **False** Explain:

   (b) (6 points) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

   (c) (6 points) **True** or **False**: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. (To get credit you must justify your answer.)

6. (20 points total) Training and Validation
You are a reviewer for the International Mega-Conference on Algorithms for Radical Learning of Outrageous Stuff, and you read papers with the following experimental setups. Would you accept or reject each paper? Provide a one sentence justification. (This conference has short reviews.)

(a) (5 points) **accept**|**reject** "My algorithm is better than yours. Look at the training error rates!"
Explain:

(b) (5 points) **accept**|**reject** "My algorithm is better than yours. Look at the test error rates!
(Footnote: reported results for $\lambda = 1.789489345672120002$.)"
Explain:

(c) (5 points) **accept**|**reject** "My algorithm is better than yours. Look at the test error rates!
(Footnote: reported results for best value of λ.)" Explain:

(d) (5 points) **accept**|**reject** "My algorithm is better than yours. Look at the test error rates!
(Footnote: reported results for best value of λ, chosen with 10-fold cross validation.)" Explain:

- Boosting algorithm version 1.
  (a) F_0 Start with the mean of the data set
  (b) Root of the first tree is the residual

$$Ps_0 = x_i - F_0$$

  (c) fit the tree to the training data.
  (d) h_0 = mean of all residuals in each leaf .
  (e) $F_1 = F_0 + h_0$
  (f) $Ps_1 = Ps_0 - h_0$
  (g) h_1 = mean of all values in each leaf.
  (h) $F_2 = F_1 + h_1$
- Dot product squared

$$<\bar{x}, \bar{y}>^2 = <\bar{x}, \bar{y}><\bar{x}, \bar{y}> = (x_1y_1 + x_2y_2 + \ldots + x_ny_n) \cdot (x_1y_1 + x_2y_2 + \ldots + x_ny_n)$$

  For two dimensions this is equivalent to

$$<\bar{x}, \bar{y}>^2 = (x_1y_1)^2 + 2x_1y_1x_2y_2 + (x_2y_2)^2$$

- Confusion Matrix

n	Predicted: NO	Predicted YES
Actual: No	TN	FP
Actual: Yes	FN	TP

  – Accuracy rate (correct classification)

$$\frac{TP + TN}{TP + TN + FP + FN}$$

  – True Positive (Recall,sensitivity)

$$\frac{TP}{TP + FN}$$

  – (Precision) correct positive divided all labeled positive

$$\frac{TP}{TP + FP}$$

  – (Misclassification)

$$\frac{FP + FN}{TP + TN + FP + FN}$$

  – False positive. False positive over actual no

$$\frac{FP}{TN + FP}$$

  – False negative. False negative over actual yes

$$\frac{FN}{FN + TP}$$