Department of Computer Science & Engineering
Indian Institute of Technology Kharagpur
CS60050: Machine Learning
Mid-semester Examination, Spring 2017-18

Full Marks: 40 **Time: 2 hours**

Note:
Attempt all questions. All parts of the same question must be answered together.
No clarifications can be provided during the exam. Make reasonable assumptions if necessary, and state any assumptions made.
All workings must be shown. You can use calculators.

1. State whether the following statements are true or false. Explain briefly with reasons.

   (i) We can get multiple local optimum solutions if we solve a linear regression problem by minimizing the sum of squared errors using gradient descent.

   (ii) In a least-squares linear regression problem, adding regularization can decrease the error of the solution on the training data.

   (iii) When the feature space is larger, overfitting is less likely.

   (iv) A model having lower in-sample/training error and higher out-of-sample/testing error should be preferred to another model having higher in-sample error but lower out-sample error.

   (v) The difference between in-sample error and out-of-sample error of a hypothesis function is usually independent of the sample size, but depends on the number of parameters in the hypothesis function. $[2 \times 5 = 10]$

2. Answer the following Multiple Choice Questions. Each question may have any number of correct answers, including zero. List all choices that you believe to be correct. **There is 1 mark for each correct answer, -0.5 for each incorrect answer.**

   (i) Which of the following problems are best suited to applying Machine Learning?
   (a) Classifying new emails received by a user into two folders 'long' and 'short', as soon as the email comes into his inbox. An email is long if it has more than 50 lines.
   (b) Classifying new emails received by a user into two folders 'important' and 'non-important', as soon as the email comes into his inbox. An email is important if the user is likely to read it within an hour of it reaching his inbox.
   (c) Predicting the time it would take a bus to travel from source to destination in a city on a given day
   (d) Predicting the time it would take a falling object to hit the ground

   (ii) Regarding bias and variance, which of the following statements are true? (Here 'high' and 'low' are relative to an 'ideal' model.)
   (a) Models which overfit have a high variance
   (b) Models which overfit have a low variance
   (c) Models which underfit have a high bias
   (d) Models which underfit have a low bias

(iii) Suppose your model is overfitting. Which of the following is a valid way to try and reduce the overfitting?
(a) Increase the amount of training data
(b) Improve the optimization algorithm being used for error minimization
(c) Decrease the model complexity
(d) Using regularization on the model paremeters

(iv) You are asked to develop a ML algorithm to predict the number of views that a news article is likely to get, using several features like the author, the topic, and so on. Which of these evaluation measures would you use for evaluating the algorithm?
(a) Precision (b) Recall (c) Mean Squared Error (d) Percentage accuracy

(v) While using gradient descent, which of the following statements are correct?
(a) Using a high learning rate can lead to slow convergence
(b) Using a high learning rate can result in the model not converging
(c) If the cost function has multiple local minima, the initialization decides which local minima is reached
(d) Adjusting the learning rate is an effective way to reach the global minima in case the cost function has multiple local minima

3. A bank wants to decide whether a customer can be given a loan, based on two features related to the monthly salary of the customer, and his/her account balance. For simplicity, we model the two features with two binary variables $X_1, X_2 \in \{0, 1\}$ and the class $Y \in \{0, 1\}$ where $Y = 1$ indicates that the customer can be given loan, and $Y = 0$ indicates otherwise. Consider the following dataset having four instances:
$(X_1 = 0, X_2 = 0, Y = 0)$, $(X_1 = 0, X_2 = 1, Y = 0)$
$(X_1 = 1, X_2 = 0, Y = 0)$, $(X_1 = 1, X_2 = 1, Y = 1)$
(a) Which model is better for the said application - logistic regression or linear regression? Explain briefly.
(b) Is there any logistic regression classifier using X_1 and X_2 as features, that can perfectly classify the given data?
(c) If we change the first instance to $(X_1 = 0, X_2 = 0, Y = 1)$, can there be any logistic regression classifier using X_1 and X_2 as features, that perfectly classifies the data? **[6]**

4. Consider the following data with one input (x) and one output (y): $\{(x = 1, y = 2), (x = 2, y = 1), (x = 3, y = 2)\}$. Apply linear regression on this data, using the hypothesis $h_\theta(x) = \theta_0 + \theta_1 x$, where θ_0 and θ_1 represent the parameters to be learned. Considering learning rate α, write the iterative steps showing how values of θ_0 and θ_1 are updated in each iteration. Assuming $\alpha = 0.1$, and the initial values $\theta_0 = 1.0$, $\theta_1 = 0.0$, perform the first **three** iterations and state the resulting model. Show the steps clearly. **[8]**

5. Derive a gradient descent training algorithm that minimizes the sum of squared error cost function, for the following hypothesis:
$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_1 x_1^2 + \theta_2 x_2 + \theta_2 x_2^2 + \ldots + \theta_n x_n + \theta_n x_n^2$
where $(x_1, x_2, \ldots, x_n)$ represents an instance having n features and $\theta_i, 0 \leq i \leq n$ represent the parameters to be learned. Assume that the number of training instances is m. Give your answer in the form $\theta_j := \theta_j + \ldots$ for $1 \leq j \leq n$. **[6]**