UIN: Name:

CSCE-421 ML Midterm exam

The exam contains 5 questions (2 pages).
You may not use anything but a pen and blank paper. Specifically, no cheatsheet, calculator, computer, or phone. Write your final answers in ink (no pencils) directly on the exam. Submit all pages of this exam. For multiple choice/answer problems, mark an 'x' in front of chosen statements, i.e., '[x]'.
Write your UIN and name at the top of each page.

1. (36%) Prove the closed form solution for Ordinary Least Square Linear Regression? That is, what is $\arg\min_{w} 0.5(xw - y)^2$, for a given training set $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$, $x = [x_1, \dots, x_n] \in \mathbb{R}^{n \times d}$, $y = [y_1, \dots, y_n] \in \mathbb{R}^{n \times 1}$, $w \in \mathbb{R}^d$?

See "8linear_regression.pptx" Slide #7

2. (16%) Mamazon corporation is automating their hiring process using ML. Specifically, they trained a binary classification model to identify promising candidates given their resume. That is, $X \in$ {features from resume (including a gender feature)}, $Y \in$ {should hire, shouldn't hire}. Mamazon used prior hiring decisions as the labeled training set. Post training, the classifier was found to bias against hiring women.
   Assume that (1) there is no real correlation between gender and job proficiency at Mamazon; (2) Mamazon's previous hiring process (which formed the training and test sets) suffered from gender bias against women; and (3) both the training and test error for the classification model were low.
   What is the probable cause for the classifier's bias against women?

   [x] Distribution mismatch between the training set and the real-world.

   [ ] Classification model that suffers from high bias (underfitting/overgeneralization).

   [ ] Classification model that suffers from high variance (overfitting).

   [ ] Attempting to fit noisy data (no function can fit the data).

3. (16%) Which of the following techniques should be considered for bias reduction (select all that apply)?

[ ] L2 regularization

[x] Boosting

[ ] Bagging

[x] Kernelization

4. (16%) Factors that contribute to high test error combined with high training error include (select all that apply).

   [ ] The ML model suffers from high variance.

   [ ] There is a distribution mismatch between the training and test sets and the real-world.

   [x] The ML model suffers from high bias.

   [x] Training and test data is coming from a noisy distribution.

5. (16%) Assume a training set $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with n samples, $x_i \in \mathbb{R}^d$. Assume a Kernel function $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ where $\phi: \mathbb{R}^d \mapsto \mathbb{R}^p$. That is ϕ is a feature map from d dimensions to p dimensions. Select all correct statements from the following.

   [ ] A Kernelized SVM using $K(x_i, x_j)$ trained on D will fit $p + 1$ parameters (p alphas + one bias term).

   [x] A linear (not-kernelized) SVM trained on D (without applying the feature map ϕ) will fit $d + 1$ parameters (d weights + one bias term).

   [x] A Kernelized SVM using $K(x_i, x_j)$ trained on D will fit $n + 1$ parameters (n alphas + one bias term).

   [x] A linear (not-kernelized) SVM trained on $\phi(D) = \{(\phi(x_1), y_1), \ldots, (\phi(x_n), y_n)\}$ will fit $p + 1$ parameters (p weights + one bias term).