Honor Code of the School of Engineering

"All students taking courses in the School of Engineering agree, individually and collectively, that they will not give or receive unpermitted aid in examinations or other course work that is to be used by the instructor as the basis of grading."

-From the Graduate/Undergraduate Bulletin

I have read, understood, and agree to abide by the Honor Code of the School of Engineering.

Name: _____________________________

ID: _____________________________

Signature: _____________________________

Date: _____________________________

Final Examination

COEN 240 Machine Learning
Department of Computer Engineering
Santa Clara University

Dr. Ming-Hwa Wang
Phone: (408) 526-4844
Email address: mwang2@cse.scu.edu
Course website: http://www.cse.scu.edu/~mwang2/learning/
Office Hours: Monday 9:00-10:00pm

Spring Quarter 2019

Phone: (408) 526-4844
Email address: mwang2@cse.scu.edu
Course website: http://www.cse.scu.edu/~mwang2/learning/
Office Hours: Monday 9:00-10:00pm

1. [40 points] True or false (yes or no, 1 or 0) problems with wrong-answer penalties:
   a) ___ When you analyzing Wikipedia, you prefer using item-based recommender to user-based recommender.
   b) ___ Top-down or bottom-up clustering approaches don’t require the user to input the cluster size.
   c) ___ In hierarchical clustering, the more abstract items are near the root of the hierarchy and the more concrete items are near the leaf of the hierarchy.
   d) ___ Unsupervised learning methods are often based on generative model.
   e) ___ A Gaussian distribution’s entropy increases as the distribution becomes narrower, i.e., as \( \sigma^2 \) decreases.
   f) ___ The higher the dimension we have, the more distances between data points.
   g) ___ K-means algorithm is a point-assignment method.
   h) ___ For a normal distribution, \( \text{E}[x] = \mu \) with variance \( \sigma^2 \), then \( \text{E}[x^2] = \mu^2 - \sigma^2 \).

2. [20 points] Which of the following are clustering, or classification? a) decision tree, b) k-means, c) nearest neighbor, d) support vector machine, and e) canopy.

3. [20 points] How would the clustering of data change if we used for the distance between two clusters:
a) The minimum of the distances between any two points, one from each cluster.
b) The average of the distances between pairs of points, one from each of the two clusters.

4. [40 points] Find the SVD of $A$, $U \Sigma V^T$, where $A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$.

5. [20 points] Consider a histogram-like density model in which the space $x$ is divided into fixed regions for which the density $p(x)$ takes the constant value $h_i$ over the $i^{th}$ region, and that the volume of region $i$ is denoted $\Delta_i$, suppose we have a set of $N$ observations of $x$ such that $n_i$ of these observations fall in region $i$. Using a Lagrange multiplier to enforce the normalization constraint on the density, derive an expression for the maximum likelihood estimator for the $\{h_i\}$.

6. [20 points] Consider the following neural network:

Suppose both of the weights shown in red ($\Theta_{11}^{(2)}$ and $\Theta_{22}^{(2)}$) are equal to 0. After running backpropagation, what can we say about the value of $\delta_1^{(3)}$?
   a) $\delta_1^{(3)} > 0$
   b) $\delta_1^{(3)} = 0$ only if $\delta_1^{(2)} = \delta_2^{(2)} = 0$, but not necessarily otherwise
   c) $\delta_1^{(3)} \leq 0$ regardless of the values of $\delta_1^{(2)}$ and $\delta_2^{(2)}$
   d) There is insufficient information to tell

7. [20 points] Suppose you have two training examples $(x^{(1)}, y^{(1)})$ and $(x^{(2)}, y^{(2)})$. Which of the following is a correct sequence of operations for computing the gradient? (Below, FP = forward propagation, BP = back propagation).
   a) FP using $x^{(1)}$ followed by FP using $x^{(2)}$. Then BP using $y^{(1)}$ followed by BP using $y^{(2)}$.
   b) FP using $x^{(1)}$ followed by BP using $y^{(2)}$. Then FP using $x^{(2)}$ followed by BP using $y^{(2)}$.
   c) BP using $y^{(1)}$ followed by FP using $x^{(1)}$. Then BP using $y^{(2)}$ followed by FP using $x^{(2)}$.
   d) FP using $x^{(1)}$ followed by BP using $y^{(1)}$. Then FP using $x^{(2)}$ followed by BP using $y^{(2)}$.

8. [20 points] Consider applying anomaly detection using a training set $\{x^{(1)}, \ldots, x^{(m)}\}$ where $x^{(i)} \in \mathbb{R}^n$. Which of the following statements are true? Check all that apply.
   a) The original model $p(x; \mu, \Sigma^2) \times \cdots \times p(x; \mu, \Sigma^2)$ corresponds to a multivariate Gaussian where the contours of $p(x; \mu, \Sigma)$ are axis-aligned.
   b) Using the multivariate Gaussian model is advantageous when $m$ (the training set size) is very small ($m < n$).
   c) The multivariate Gaussian model can automatically capture correlations between different features in $x$.
   d) The original model can be more computationally efficient than the multivariate Gaussian model, and thus might scale better to very large values of $n$ (number of features).