Midterm 1 Practice for COMP6321 Fall 2019

The questions in this practice midterm are suggestive only of the *style* and *difficulty* of questions that will be asked on on the real midterm. The length and the particular course content evaluated will be different.

Q1. [10 marks total] This question is about *logistic regression models*.

a) [2 marks] What kind of learning task is logistic regression used for?

b) [1 mark] Can the optimal parameter vector $\boldsymbol{w}$ for a logistic regression problem be solved for 'directly'?

c) [2 marks] Is the decision boundary of logistic regression linear or non-linear within the feature space $\boldsymbol{\phi}$? Explain.

d) [3 marks] Assume you are given data set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$. Write the *logistic regression* loss function with respect to this data set. For full marks include the feature transformation $\boldsymbol{\phi}(\cdot)$.

e) [2 marks] Assume you are given training set in matrix format $\begin{bmatrix} 1 & x_1 & x_2 \end{bmatrix}$ where X and $\boldsymbol{y}$ are

$$X = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -2 \\ 1 & 0 & 1 \\ 1 & 2 & 0 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Plot the data in two dimensions and draw the decision boundary that would result from applying logistic regression. Be sure to indicate which side corresponds to predicting $y = 1$.

Q2. [10 marks total] Assume we have samples $\{x_1, \ldots, x_N\}$ from a univariate normal distribution $\mathcal{N}(\mu, \sigma^2)$. The likelihood $p(x \mid \mu, \sigma)$ having observed a single point x is therefore

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

a) [2 marks] The likelihood is a function of which variable(s)?

b) [2 marks] Write the likelihood $p(x_1, \ldots, x_N \mid \mu, \sigma)$ having observed all x_i jointly.

c) [2 marks] Write the negative log likelihood of $p(x_1, \ldots, x_N \mid \mu, \sigma)$.

d) [2 marks] Write the gradient of the negative log likelihood of $p(x_1, \ldots, x_N \mid \mu, \sigma)$.

e) [2 marks] Use your answer from part (d) to derive a maximum likelihood estimate of the normal distribution parameters.

Q3. [8 marks total] This question is about programming machine learning concepts with Numpy. You can assume that `import numpy as np` has already been run.

a) [2 marks] You are given the following incomplete function:

```
def linear_model_predict(X, w):
    """
    Returns predictions from linear model y(X, w) at each point X[i,:] using parameters w.
    Given X with shape (N,D) and w with shape (D,), returns predictions with shape (N,).
    """
```

Complete the function in the space below. (No need to copy the above function signature.) For full marks, your answer should be fully vectorized.

b) [2 marks] You are given the following incomplete function:

```
def sigmoid(z):
    """Returns the element-wise logistic sigmoid of array z."""
```

Complete the function in the space below. (No need to copy the above function signature.) For full marks, your answer should be fully vectorized.

c) [4 marks] You are given the following incomplete function:

```
def linear_regression_by_gradient_descent(X, y, w_init, learn_rate=0.05, num_steps=500):
    """
    Fits a linear model by gradient descent.

    Given X, y, and w_init with shapes (N,D), (N,), and (D,) respectively,
    returns a new parameter vector w that minimizes the squared error to the targets
    by running num_steps of gradient descent.
    """
```

The gradient of a linear model can be expressed mathematically as

$$\nabla \ell_{\mathrm{LS}}(\boldsymbol{w}) = (X^T X)\boldsymbol{w} - X^T \boldsymbol{y}$$

Complete the function in the space below. (No need to copy the above function signature.) For full marks, your answer should be fully vectorized.

Q4. [6 marks total] This question is about the assigned reading: the 2001 paper by Leo Breiman.

a) [2 marks] How are what Breiman calls "data models" typically validated? How are what he calls "algorithmic models" typically validated?

b) [3 marks] Give three examples of what Breiman calls "data models" and three examples of what he calls "algorithmic models"

c) [1 mark] What is Breiman's argument against doing dimensionality reduction?

Q5. [8 marks total] This question is about the *K-means* clustering algorithm.

a) [4 marks] What is the optimization problem that the K-means clustering algorithm tries to solve? Express your answer mathematically, then explain what each symbol means.

b) [1 mark] In the K-means optimization problem, the objective function has a name. What is it?

c) [1 mark] Why is K-means considered a "coordinate descent" algorithm?

d) [1 mark] Write the formula for computing the "assignment step" of the K-means algorithm

e) [1 mark] Prove that your answer to (d) is an optimal assignment with respect to the current means.

Q6. [3 marks] The probability of a point $x \in \mathbb{R}$ under the univariate normal distribution $\mathcal{N}(\mu, \sigma^2)$ is

$$p(x \mid \mu, \sigma) = \frac{1}{Z} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \text{ where } Z = \sqrt{2\pi}\sigma$$

Write the probability of a point $\boldsymbol{x} \in \mathbb{R}^D$ under the multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. (It is OK to just write Z for the new normalization constant without specifying its value.) Explain how your answer reduces to the univariate case when $D = 1$.

Q7. [6 marks total] This question is about the *expectation maximization* (EM) algorithm for fitting a K-component Gaussian mixture model to a data set $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^N$.

a) [4 marks] Write the formula for the density $p(\boldsymbol{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ of a Gaussian mixture model with parameters $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$, $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$, and $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$.

b) [2 marks] The EM algorithm is derived from a probablistic model of how each $\boldsymbol{x}$ is generated. This model is based on a component selection vector $\boldsymbol{z}$. Prove that $p(z_k = 1 \mid \boldsymbol{x}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, *i.e.* the probability that component k was used to generate data point $\boldsymbol{x}$, is equal to

$$\frac{\pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Q8. [12 marks total] This question is about *support vector machines*. Suppose you are given the following training set for 2-class classification:

$$\boldsymbol{X} = \begin{bmatrix} -2 & -1 \\ 2 & -2 \\ 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad \boldsymbol{t} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix}$$

a) [5 marks] For the above training set, write the quadratic program for the primal hard-margin SVM learning problem. You should not have symbols $\boldsymbol{x}$ or t in your answer: instead substitute the coefficients ($2w_1$, etc).

b) [3 marks] Plot the data on a set of axes. Plot the optimal decision boundary (as closely as you can). Indicate which points are the support vectors.

c) [2 marks] The dual formulation of an SVM is defined in terms of kernels $k(\boldsymbol{x}, \boldsymbol{x}')$. Given kernel $k(\boldsymbol{x}, \boldsymbol{x}') = 1 + e^{-x_1 - x_1'} + e^{-x_2 - x_2'} + \cdots + e^{-x_D - x_D'}$, what feature space does this kernel correspond to? Explain.

d) [2 marks] The dual formulation of hard-margin SVM is defined in terms of dual variables $\boldsymbol{\alpha} \in \mathbb{R}^N_{\geq 0}$. Write the formula for the SVM decision function $y(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) + b$ in terms of the α_i, training labels t_i, training samples $\boldsymbol{x}_i$, kernel function $k(\cdot, \cdot)$, and the bias term b which you can assume is known. Show your steps.

Q9. [6 marks total] This question is about *boosting*.

a) [2 marks] What makes boosting an "ensemble" method?

b) [2 marks] When the AdaBoost algorithm is training $y_r(\boldsymbol{x})$, the r^{th} weak learner, does it compute the current ensemble's predictions $y(\boldsymbol{x}_i)$ over the training set, where $y(\boldsymbol{x}) = \sum_{j=1}^{r-1} \alpha_j y_j(\boldsymbol{x})$? Explain.

c) [2 marks] In the AdaBoost from lecture, if the weak learners are always "better than random guessing" then each sample weight w_i will either stay the same or will increase. *True or False?* Explain.