CSCI 5521: Intro to Machine Learning
Sample Exam

Name:

Student ID:

Question 1: There might be multiple correct choices in each question (20 points)

1. Which of the following methods are un-supervised learning methods_____.
   A. Perceptron B. PCA C. Linear regression D. k-means

2. Which of the following methods can be used to control model complexity in model selection _____.
   A. Cross-validation B. Regularization by 2-norm C. Dimension reduction D. Kernel selection for kernel methods

3. Which of the following dimension reduction methods allows non-linear mapping of data _____.
   A. LLE (Local Linear Embedding) B. Isomap C. PCA D. Feature (subset) selection

4. Give a kernel smoother $g(x) = \frac{\sum_t K(\frac{x-x^t}{h})r^t}{\sum_t K(\frac{x-x^t}{h})}$, where $K(y)$ is RBF function, which of the following might lead to overfitting in regression with the kernel smoother _____.

A. Use a small h B. Use a large h C. Apply the method to small dataset D. Apply the method to large dataset

5. Which error function does soft-margin SVM use _____.

   A. 0/1 loss B. root mean squared error C. Hinge loss D. cross entropy

6. Which of the following regularization on w will lead to sparse solution _____.

   A. 2-norm $||w||_2$ B. 0-norm $||w||_0$ C. 1-norm $||w||_1$ D. infinte-norm $||w||_{+\infty} = max(w)$

Question 2 (20 points)

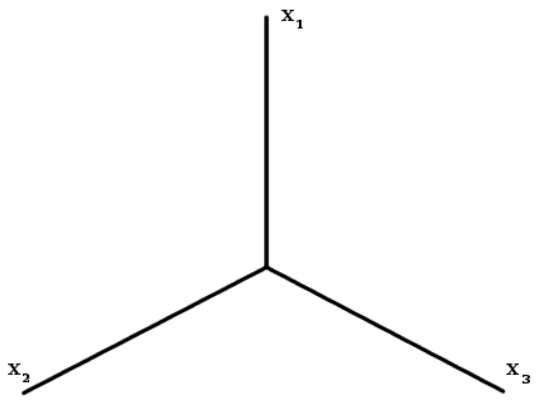
Using a logistic regression model defined as

$$P(C_1|\mathbf{x}, \mathbf{w}) = sigmoid(w_0 + w_1x_1 + w_2x_2 + w_3x_3) = \frac{1}{1 + \exp(-(w_0 + w_1x_1 + w_2x_2 + w_3x_3))}$$

train a model using the following 40 binary samples $\mathbf{x}^t \in \{0, 1\}^3$ with binary output values $y^t \in \{0, 1\}$ for $t = 1, \ldots, 40$

Number of samples	x_1	x_2	x_3	y
10	1	1	0	1
10	1	0	1	0
10	0	1	0	0
10	0	0	1	1

1. What is the best training error rate (classification error of the training examples) that we could achieve by using all features (x_1, x_2, and x_3) in the logistic regression model? Explain your answer. (Hint: Use the following axis to plot the samples and best $\mathbf{w}$).



Training error rate = _____
Solution: Error rate = 0.25 (10/40 misclassified points).

Question 2 continued

2. Design a feature mapping function to map (x_1, x_2, and x_3) to a new feature space in which we can perfectly classify all the training samples using logistic regression. Write down the mapping function and justify your answer by showing the mapping for each of the 40 training samples given above.

   Solution: x_1 xor $x_2 \rightarrow z_1$, $x_3 \rightarrow z_2 = 0$.
   This will give $(z_1, z_2, y) = (0, 0, 1), (0, 0, 1), (1, 0, 0), (1, 0, 0)$.

Question 3 (20 points)

In a particular class, let the probability that a student gets an A be $P(A) = 1/2$, a B be $P(B) = \mu$, a C be $P(C) = 2\mu$, and a D be $P(D) = 1/2 - 3\mu$. We are told that c students got a C, d students got a D and h students got either an A or B. Let variable a be the number of students get A. Use the expectation maximization (EM) algorithm to obtain a maximum likelihood estimates (MLE) of μ.

1. First, define the complete log-likelihood function $L(\mu|h, a, c, d)$ (Hint: it is a multinomial with $a, h - a, c$ and d occurrences of each grade).
   $L(\mu|h, a, c, d) = log((\frac{1}{2})^a * \mu^{h-a} * (2\mu)^c * (\frac{1}{2} - 3\mu)^d) = ...$

2. **E-step:** Compute the expected value of a given μ. (Hint: $P(A|A \text{ or } B) = \frac{P(A \text{ and } (A \text{ or } B))}{P(A \text{ or } B)} = \frac{P(A)}{P(A)+P(B)}$. The expectation is over the h students getting A or B.)
   Solution: The expectation for a student among the h students getting "A" or "B" to get an "A' is $\frac{\frac{1}{2}}{\frac{1}{2}+\mu}$. Thus,
   $\hat{a} = \frac{\frac{1}{2}}{\frac{1}{2}+\mu}h.$
   $h - \hat{a} = \frac{\mu}{\frac{1}{2}+\mu}h.$

3. **M-step:** Compute the MLE of μ, assuming the unobserved variable is replaced by the expectation a.
   Solution:
   $\hat{\mu} = \frac{h-a+c}{6(h-a+c+d)}.$

Question 4 (20 points)

The back-propagation algorithm for a RBF neural network for regression with error function $E(\{m_h, s_h, w_i\}_{i,h}|X) = \frac{1}{2}\sum_{t=1}^{N}\sum_{i=1}^{K}(r_i^t - y_i^t)^2$ is given below.

Forward :

$$p_h^t = \exp\left[-\frac{\|x^t - m_h\|_2^2}{2s_h^2}\right]$$

$$y_i^t = \sum_{h=1}^{H} w_{ih} p_h^t + w_{i0}$$

Backward :

$$\Delta w_{ih} = \eta \sum_{t=1}^{N}(r_i^t - y_i^t)p_h^t$$

$$\Delta m_{hj} = \eta \sum_{t=1}^{N}\left[\sum_{i=1}^{K}(r_i^t - y_i^t)w_{ih}\right] p_h^t \frac{(x_j^t - m_{hj})}{s_h^2}$$

$$\Delta s_h = \eta \sum_{t=1}^{N}\left[\sum_{i=1}^{K}(r_i^t - y_i^t)w_{ih}\right] p_h^t \frac{\|x^t - m_h\|_2^2}{s_h^3}$$

A regularization term on the RBF centers m_h and weights w_i can be introduced into the error function as

$$E(\{m_h, s_h, w_i\}_{i,h}|X) = \frac{1}{2}\sum_{t=1}^{N}\sum_{i=1}^{K}(r_i^t - y_i^t)^2 + \alpha\sum_{h=1}^{H}||m_h||^2 + \beta\sum_{i=1}^{K}||w_i||^2.$$

1. Derive the forward steps of the regularized RBF neural network, i.e., derive equations for p_h^t and y_i^t.

$$p_h^t = \exp\left[-\frac{\|x^t - m_h\|_2^2}{2s_h^2}\right]$$

$$y_i^t = \sum_{h=1}^{H} w_{ih} p_h^t + w_{i0}$$

2. Derive the backward steps of the regularized RBF neural network, i.e., derive equations for Δw_i, Δm_h and Δs_h.

$$
\begin{aligned}
\Delta w_{ih} =& \eta(\sum_{t=1}^{N}(r_i^t - y_i^t)p_h^t - 2\beta w_{ih}) \\
\Delta m_{hj} =& \eta(\sum_{t=1}^{N}\left[\sum_{i=1}^{K}(r_i^t - y_i^t)w_{ih}\right] p_h^t \frac{(x_j^t - m_{hj})}{s_h^2} - 2\alpha m_{hj}) \\
\Delta s_h =& \eta \sum_{t=1}^{N}\left[\sum_{i=1}^{K}(r_i^t - y_i^t)w_{ih}\right] p_h^t \frac{\|x^t - m_h\|_2^2}{s_h^3}
\end{aligned}
$$

Question 5 (20 points)

For a matrix of n samples in d dimensions $\mathbf{X} \in \mathbb{R}^{n \times d}$ with target values $\mathbf{y} \in \mathbb{R}^n$ and $\lambda \geq 0$, the ridge regression problem is defined as

$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \ .$$

1. Prove the solution to the above ridge regression problem to find the optimal $\hat{\mathbf{w}}$ estimate is $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$.

   Solution: See Homework 0.

2. Using part 1, derive the kernel ridge regression solution of $w = \sum_i \alpha_i x_i$ in terms of α using the identity $(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = PB^T(BPB^T + R)^{-1}$ where P, B and R are matrices. Use K to denote the kernel matrix (Hint: Let $P = \frac{I}{\lambda}; B = X; R^{-1} = I$).

   Solution:

$$\begin{aligned}
\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} &= (P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} \\
\Rightarrow PB^T(BPB^T + R)^{-1} &= \frac{I}{\lambda} X^T (X * \frac{I}{\lambda} * X^T + I)^{-1} * y \\
&= X^T (X * X^T + \lambda I)^{-1} * y \\
\text{Let,} & \\
w &= \sum_i \alpha_i x_i \text{ then} \\
\alpha &= (X * X^T + \lambda I)^{-1} * y \\
&= (K(X, X) + \lambda I)^{-1} * y
\end{aligned}$$