南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Course Name: Machine Learning **Exam Duration:** 2 hours

Dept.: Department of Computer Science and Engineering

Exam Paper Setter(Signature): ______________

Question No.	1	2	3	4	5	6	7	8	9	10
Score	20	50	30	10						

This exam paper contains 4 questions and the score is 110 in total. (Please hand in your exam paper, answer sheet, and your scrap paper to the proctor when the exam ends.)

Problem I Multiple Choice (20 Points)

(*only one correct answer for each question*)

B

1. **(2 points)** Three essential components of a learning system are _____.
   A. model, gradient descent, learning algorithm
   B. error function, model, learning algorithm
   C. accuracy, sensitivity, specificity
   D. model, error function, cost function

A

2. **(2 points)** The objective of machine learning is to minimize _____.
   A. the KL divergence between real-world data and the trained probabilistic model
   B. the KL divergence between training data and the trained probabilistic model
   C. the KL divergence between real-world data and training data
   D. the KL divergence between training data and prediction data

B

3. **(2 points)** What is the loss function most suited for linear regression?
   A. the entropy function
   B. the squared error function
   C. the cross-entropy function
   D. the number of mistakes

D

4. **(2 points)** What is the loss function most suited for probabilistic density mixture model based clustering?
   A. the entropy function
   B. the squared error function

C. the likelihood function of complete data
D. the likelihood function of incomplete data

C

5. **(2 points)** The differences between the generative and discriminative approaches include ________.
A. that the former has less parameters
B. that the former cannot add a new class
C. that the latter emphasizes the boundary among classes
D. that the latter can be trained faster

D

6. **(2 points)** Neural networks can NOT be regularized by using ________.
A. using a prior on model parameters
B. data augmentation
C. node dropping out
D. ReLU activation

A

7. **(2 points)** The advantages of the hidden Markov model DO NOT include ________.
A. global convergence
B. fast estimation algorithm
C. unsupervised learning
D. capability of modeling both continuous and discrete data

B

8. **(2 points)** The advantages of using ReLU as activation functions DO NOT include ________.
A. reducing gradient vanishing
B. reducing gradient explosion
C. encouraging model sparsity
D. increasing computational efficiency

D

9. **(2 points)** Which of the following is NOT a way to reduce the model overfitting?
A. increase the amount of training data
B. improve the optimization algorithm being used for error minimization
C. decrease the model complexity
D. reduce the noise in the training data

C

10. **(2 points)** Which of the following statements is NOT true for Bellman equations?
A. it can be used to estimate state value functions
B. it is can be solved by using dynamic programing, Monti Carlo, and temporal difference approaches
C. solving Bellman equation requires environment models
D. its fixed point is the optimal policy

Problem II Numerical Calculation (50 Points)

(1) **Linear Regression (5 points)**. For three points {(1, 4), (2, 8), (3, 14)}, what is the linear regression function for the least squared errors (*assuming* $y = a_2x^2 + a_1x + a_0$)?

(2) **Supervised Classification (5 points)**. For class A of two points {(1, 2) (2, 1)} and class B of two points {(4, 1) (3, 4)}, what are the labels for points {(2,2) (3,3)} using the K-NN algorithm (*where* K=3)?

(3) **Maximum margin classifier (5 points)**. For one class of two points {(1, 2) (2, 2)} and another class of two points {(4, 4) (5, 6)}, what are the support vectors and what is the decision boundary's function (*plot your answer*) ?

(4) **Clustering (5 points)**. For four points with two classes, {(1, 2) (2, 2) (4, 4) (5, 6)}, how to achieve two cluster centers using the K-means algorithm (*outline the algorithm and show the details of one iteration*) ?

(5) **Factor Graph (15 points)**. How to design a factor graph to solve the following linear Gaussian system: $[3\ \ 3]^T = [1\ \ 1\ \ 1;\ 0\ \ 2\ \ 1][x_1\ \ x_2\ \ x_3]^T$? Assuming the initial Gaussian distributions of X is $\{[m_1, \sigma_1], [m_2, \sigma_2], [m_3, \sigma_3]\}$, outline the whole computation procedure and show the details of one iteration.

(6) **Hidden Markov Model (15 points)**. For a HMM, the states of latent variables are {bull, bear}, the states of observation variables are {rise, fall}, the initial state probability distribution $\boldsymbol{\pi}$ is $[0.5\ 0.5]^T$, the transition probability distribution $\boldsymbol{A}$ is [0.4 0.7; 0.6 0.3], and the observation probability distribution $\boldsymbol{B}$ is [0.8 0.1;0.2 0.9]. If the observation sequence $\boldsymbol{X}$ is {fall fall rise}, please show the computation procedure for $p(z_1|\boldsymbol{X}, \boldsymbol{\theta})$ and $p(z_1, z_2|\boldsymbol{X}, \boldsymbol{\theta})$ using the forward-backward algorithm, where z_n is the latent variable at time n and $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}\}$?

Problem III Theoretical Analysis (30 Points)

For a finite-state random sequence $\{Z_t\}$with the model of $\{\pi, A\}$ and its observation sequence is $\{X_t\}$, the joint distribution of X and Z with the model θ is given by

$$p(X, Z|\theta) = \prod_{i=1}^{K} [p(z_i)p(X|\theta_i)]^{z_i}$$

(1) Summarize the general forward-backward EM scheme for HMM (***E***-step and ***M***-step).

(2) Assuming each observation probability density is Bernoulli, *i.e.* $p(X|\theta_i) = \theta_i^X(1-\theta_i)^{1-X}$, please derive the corresponding model learning procedure under the EM scheme.

(3) Use message passing to derive the forward-backward algorithms.

Problem IV Expectation-Maximization Learning (Bonus 10 Points)

(1) What is the EM procedure? When do we need the EM procedure for machine learning? Please give a specific example.

(2) What is the EM procedure in terms of the Q function? Please give the detailed equations assuming that X is the observed variable, Z is the latent variable and θ is the model parameter.

(3) What is the EM procedure in terms of likelihood and KL divergence? Please give the detailed equations and plots to illustrate the procedure.

(4) What is the EM procedure in terms of optimization of non-convex function? Please give a plot to illustrate the procedure.

(5) What is the EM procedure for the factor graph network model? Please give an example.