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): ________ **Reviewer(Signature):** ________

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

This exam paper contains 9 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)

1. **(2 points)** The most suitable loss function for linear regression is ________

A. the sum of squared errors

B. the entropy function

C. the cross entropy function

D. the number of mistakes

2. **(2 points)** The logistic regression is a ________ regression technique that is used to model data having a ________ outcome.

A. linear, numeric

B. linear, binary

C. nonlinear, numeric

D. nonlinear, binary

3. **(2 points)** The most suitable loss functions for neural networks are ________

A. the entropy function and KL divergence

B. the squared error function and cross-entropy function

C. the cross-entropy function and KL divergence

D. the number of mistakes and entropy function

4. **(2 points)** The most suitable loss functions for SVM are ________?

A. the entropy function and KL divergence

B. the squared error function and cross-entropy function

C. the hinge error function and ε–insensitive error function

D. the number of mistakes and entropy function

5. **(2 points)** The most suitable loss functions for GMM are _______.

A. the maximum likelihood function and maximum a posterior function

B. the squared error function and cross-entropy function

C. the cross-entropy function and KL divergence

D. the number of mistakes and entropy function

6. **(2 points)** The three most important problems for HMM are _______.

A. message propagation, expectation and maximization

B. learning, evaluation and decoding

C. belief propagation, parameter learning and state estimation

D. ML learning, MAP learning, and fully Bayesian learning

7. **(2 points)** The reinforcement learning problem can be solved through _______.

A. dynamic programming if the rewards and transition probabilities are known

B. the Monte Carlo method if only reward functions are known

C. the temporal difference method if the online learning is preferred

D. all of the above

8. **(2 points)** Which activation function has the least computational complexity?

A. tanh

B. sigmod

C. ReLu

D. Leaky ReLu

9. **(2 points)** Which of the following is NOT a way to reduce the model under-fitting?

A. increase the amount of training data

B. increase the model complexity

C. decrease the number the model parameters with prior distributions

D. decrease the amount of data augmentation

10. **(2 points)** Which of the following is NOT true for a machine learning system?

A. It has three main components: model, error function and optimization algorithm.

B. It reduces the KL divergence between the distributions of data and the model.

C. It involves the procedure of EM for incomplete data problems.

D. It will achieve the global optimum if the training data is sufficient enough.

Problem II Numerical Calculation (40 Points)

(1) **Linear Regression (5 points)**. For three points {(1, 0), (3, 3), (5, 4)}, what is the linear regression function for the least squared errors (*assuming y = ax + b, using psudo-inverse*)?

(2) **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*) ?

(3) **Clustering (10 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?

(4) **Factor Graph (10 points)**. How to design a factor graph to solve the following linear equation $[2\ 4\ 3]^T = [1\ 0\ 1;\ 1\ 2\ 1;\ 1\ 1\ 1][x_1\ x_2\ x_3]^T$? Assuming the initial value of X is [0 1 1]), show the computation procedure of one iteration.

(5) **Hidden Markov Model (10 points)**. For a HMM, the hidden states are {bull, bear}, the 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 is {fall fall rise}, please show the computation procedure for estimating the most likely state sequence?

Problem III Theoretical Analysis (40 Points)

(1) **Density Mixture Model (20 points).** For a random variable X distributed in a mixture of probability densities, the joint distribution of X and its latent variable Z with the model θ is given by

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

a) Summarize the general EM scheme for DMM (***E***-step and ***M***-step).

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

(2) **Hidden Markov Model (20 Points).** For a finite-state random sequence $\{Z_t\}$with the model of {π, A} and its observation sequence is $\{X_t\}$, the joint distribution of X and Z with the model θ is given by $p(\mathrm{X}, \mathrm{Z}|\theta)$.

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

b) Assuming each observation probability density is Bernoulli, i.e.

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

please derive the corresponding model learning procedure under the EM scheme.

Problem IV Expectation and Maximization (Bonus 10 Points)

(1) What is the EM procedure? When do we need the EM procedure for machine learning?

(2) What is the EM procedure in terms of the Q function?

(3) What is the EM procedure in terms of likelihood and KL divergence?

(4) What is the EM procedure in terms of optimization of non-convex function?

(5) What is the EM procedure for the factor graph network model?