SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

2023-2024 Academic Year
Fall Semester
Midterm Exam Paper

Course Name: Machine Learning **Dept.:** Computer Science and Engineering
Exam Duration: 48 hours

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

This exam paper contains 8 questions and the score is 100 in total (Please hand in your answer sheet in the digital form).

Problem I. Least Square (10 points)

a) Consider $Y = AX + V$ and $V \sim \mathcal{N}(\mathbf{v}|\mathbf{0}, Q)$, what is the least square solution of X ?

b) If there is a constraint of $b^T X = c$, what is the optimal solution of X?

c) If there is an *additional* constraint of $X^T X = d$, in addition to the constraint in b), what is the optimal solution of X?

Problem II. Linear Gaussian System (10 points)

Consider $Y = AX + V$, where X and V are Gaussian, $X \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{m}_0, \boldsymbol{\Sigma}_0)$, $V \sim \mathcal{N}(\boldsymbol{v}|\mathbf{0}, \beta^{-1}\boldsymbol{I})$. What are the conditional distribution, $p(Y \mid X)$, the joint distribution $p(Y, X)$, the marginal distribution, $p(Y)$, the posterior distribution, $p(X|Y = \boldsymbol{y}, \beta, \boldsymbol{m}_0, \boldsymbol{\Sigma}_0)$, the posterior predictive distribution, $p(\hat{Y}|Y = \boldsymbol{y}, \beta, \boldsymbol{m}_0, \boldsymbol{\Sigma}_0)$, and the prior predictive distribution, $p(Y|\beta, \boldsymbol{m}_0, \boldsymbol{\Sigma}_0)$, respectively?

Problem III. Linear Regression (10 points)

Consider $y = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}) + v$, where v is Gaussian, *i.e.*, $v \sim \mathcal{N}(v|0, \beta^{-1})$, and $\boldsymbol{w}$ has a Gaussian *priori*, *i.e.*, $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \alpha^{-1}\boldsymbol{I})$. Assume that $\boldsymbol{\phi}(\boldsymbol{x})$ is known, please derive the posterior distribution, $p(\mathbf{w}|D, \beta, \boldsymbol{m}_0, \alpha)$, the posterior predictive distribution, $p(\hat{y}|\hat{x}, D, \beta, \boldsymbol{m}_0, \alpha)$, and the prior predictive distribution, $p(D|\beta, \boldsymbol{m}_0, \alpha)$, respectively, where $D = \{\phi_n, y_n\}$, $n = 1, \ldots, N$, is the training data set and $\phi_n = \phi(\mathbf{x}_n)$.

Problem IV. Logistics Regression (10 points)

Consider a two-class classification problem with the logistic sigmoid function, $y = \sigma\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}\ (\boldsymbol{x})\right)$, for a given data set $D = \{\phi_n, t_n\}$, where $t_n \in \{0, 1\}$, $\phi_n = \phi(\mathbf{x}_n)$, $n = 1, \ldots, N$, and the likelihood function is given by

$$p(\boldsymbol{t}|\boldsymbol{w}) = \prod_{n=1}^{N} y_n^{t_n}(1 - y_n)^{1-t_n}$$

where $\boldsymbol{w}$ has a Gaussian *priori*, *i.e.*, $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \alpha^{-1}\boldsymbol{I})$. Please derive the posterior distribution, $p(\mathbf{w}|D, \boldsymbol{m}_0, \alpha)$, the posterior predictive distribution, $p(t|x, D, \boldsymbol{m}_0, \alpha)$, and the prior predictive distribution, and $p(D|\boldsymbol{m}_0, \alpha)$, respectively. (*Hint*: using Delta approximation and Laplace approximation properly).

Problem V. Neural Network (10 points)

Consider a two-layer neural network described by following equations:

$$a_1 = \boldsymbol{w}^{(1)}\boldsymbol{x},\ a_2 = \boldsymbol{w}^{(2)}\boldsymbol{z},\ z = h(a_1),\ y = \sigma(a_2)$$

where $\boldsymbol{x}$ and y are the input and output, respectively, of the neural network, $h(\bullet)$ is a nonlinear function, and $\sigma(\bullet)$ is the sigmod function.

(1) Please derive the following gradients: $\frac{\partial y}{\partial \mathbf{w}^{(1)}}, \frac{\partial y}{\partial \boldsymbol{w}^{(2)}}, \frac{\partial y}{\partial a_1}, \frac{\partial y}{\partial a_2}$, and $\frac{\partial y}{\partial \boldsymbol{x}}$.

(2) Please derive the updating rules for $\boldsymbol{w}^{(1)}$ and $\boldsymbol{w}^{(2)}$ given the classification errors between y and t, where t is the ground truth of the output y.

Problem VI. Bayesian Neural Network (20 points)

a) Consider a neural network for regression, $t = y(\boldsymbol{w}, \boldsymbol{x}) + v$, where v is Gaussian, *i.e.*, $v \sim \mathcal{N}(v|0, \beta^{-1})$, and $\boldsymbol{w}$ has a Gaussian *priori*, *i.e.*, $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \alpha^{-1}\boldsymbol{I})$. Assume that $y(\boldsymbol{w}, \boldsymbol{x})$ is the neural network output please derive the posterior distribution, $p(\boldsymbol{w}|D, \beta, \boldsymbol{m}_0, \alpha)$, the posterior predictive distribution, $p(t|x, D, \beta, \boldsymbol{m}_0, \alpha)$, and the prior predictive distribution, $p(D|\beta, \boldsymbol{m}_0, \alpha)$, where $D = \{x_n, t_n\}$, $n = 1,\ldots, N$, is the training data set.

b) Consider a neural network for two-class classification, $y = \sigma(a(\boldsymbol{w}, \boldsymbol{x}))$ and a data set $D = \{x_n, t_n\}$, where $t_n \in \{0,1\}$, $\boldsymbol{w}$ has a Gaussian *priori*, *i.e.*, $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\mathrm{I})$, and $\overset{f}{a}(\boldsymbol{w}, \boldsymbol{x})$ is the neural network model. Please derive the posterior distribution, $p(\boldsymbol{w}|D, \alpha)$, posterior predictive distribution, $p(t|\boldsymbol{x}, D, \alpha)$, and the prior predictive distribution, $p(D|\alpha)$, respectively.

Problem VII. Critical Analyses (20 Points)

a) Please explain, in terms of cost functions, constraints and predictions, **i)** what are the differences between SVM classification and logistic regression; **ii)** what are the differences between ν-SVM regression and least square regression.

b) Please explain why neural network (NN) based machine learning algorithms use *logistic* activation functions ?

c) Please explain **i)** what are the differences between the *logistic* activation function and other activation functions (e.g., *relu*, *tanh*); and **ii)** when these activation functions should be used.

d) Please explain why Jacobian and Hessian matrices are useful for machine learning algorithms.

e) Please explain why exponential family distributions are so common in engineering practice. Please give some examples which are **NOT** exponential family distributions.

f) Please explain why KL divergence is useful for machine learning? Please provide two examples of using KL divergence in machine learning.

g) Please explain why data augmentation techniques are a kind of regularization skills for NNs.

h) Please explain why Gaussian distributions are preferred over other distributions for many machine learning models?

i) Please explain why the MAP model is usually more preferred than the ML model?

Problem VIII. Discussion (10 Points)

What are the generative and discriminative approaches to machine learning, respectively? Can you explain the advantages and disadvantages of these two approaches and provide a detailed example to illustrate your points？