南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

2020-2021 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 | |
|---|---|---|---|---|---|---|---|---|
| **Score** | 15 | 20 | 10 | 20 | 20 | 20 | 10 | |

This exam paper contains 7 questions and the score is 110 in total (Please hand in your answer sheet in the digital form).

**Problem I. Least Square and Gaussian (15 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) Consider $Y = AX + V$, where $X$ and $V$ are Gaussian, $X \sim \mathcal{N}(\mathbf{x}|\mathbf{m}_0, \mathbf{\Sigma}_0)$, $V \sim \mathcal{N}(\mathbf{v}|\mathbf{0}, \beta^{-1}\mathrm{I})$.

What are the conditional distribution, $p(Y\,|\,X)$, the marginal distribution, $p(Y)$, the posterior distribution, $p(X|Y)$, and the posterior predictive distribution, $p(\hat{Y})$, respectively?

**Problem II. Regression and Classification (20 points)**

a) Consider $y = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) + v$, where $v$ is Gaussian, *i.e.*, $v \sim \mathcal{N}(v|0, \beta^{-1})$, and $\mathbf{w}$ has a Gaussian *priori*, *i.e.*, $\mathbf{w} \sim \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$. Assume that $\boldsymbol{\phi}(\mathbf{x})$ is known, please derive the posterior distribution, $p(\hat{\mathbf{w}}|y)$.

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

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

where $\mathbf{w}$ has a Gaussian *priori*, *i.e.*, $\mathbf{w} \sim \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$. Please derive the Gaussian posterior distribution, $p(\hat{\mathbf{w}}|\mathbf{t})$ (*Hint*: using Laplace approximation).

**Problem III. Neural Network (10 points)**

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

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

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

(2) Please derive the updating rules for $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ for classification errors between $y$ and $t$.

**Problem IV. Bayesian Neural Network (20 points)**

a) Consider a neural network for regression, $t = y(\mathbf{w}, \mathbf{x}) + v$, where $v$ is Gaussian, *i.e.*, $v \sim \mathcal{N}(v|0, \beta^{-1})$, and $\mathbf{w}$ has a Gaussian *priori*, *i.e.*, $\mathbf{w} \sim \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathrm{I})$. Assume that $y(\mathbf{w}, \mathbf{x})$ is the neural network model, please derive the posterior distribution, $p(\hat{\mathbf{w}}|\mathbf{t})$, and the predictive distribution, $p(t|D, \beta, \alpha)$, where $D = \{\mathbf{x}, \mathbf{t}\}$.

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

**Problem V. Sparse Vector Machine (15 points)**

a）Please explain why the dual problem formulation is used to solve the SVM machine learning problem.

b）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.

c） Consider a two-class classification problem with the logistic sigmoid function $y = \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}))$, with a *priori* on $\mathbf{w}$, $p(\mathbf{w}|\alpha) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{w}_i|0, \alpha_i)$, where the data set is $\{\phi_n, t_n\}$, $t_n \in \{0, 1\}$, $\phi_n = \phi(\mathbf{x}_n)$, and $n = 1,\ldots, N$, please derive the analysis of sparsity and the fast learning algorithm which fully optimizes the single hyper-parameter $\alpha_i$.

**Problem VI. Critical Analyses (20 Points)**

a) Explain why neural network (NN) based machine learning algorithms use *logistic* activation functions ?

b) 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.

c) Explain what are the differences between the hinge cost function and other cost functions (*e.g.*, *softplux*, *binary*)?

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

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

f) Explain why the data learning efficiency of RVM is better than that of SVM?

g) Explain why KL divergence is useful for machine learning? Please provide two examples.

h) Explain why data augmentation techniques are a kind of regularization skills for NNs.

**Problem VII. Bonus (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？