The University of Hong Kong
Department of Electrical and Electronic Engineering

ELEC3249/6008 Pattern recognition and machine learning (2018-2019)

Written Assignment: 100 marks
(Counts towards 15% of the overall assessment)

1. Submission Format and Deadline

Solution addressing the problems should be submitted in **PDF format** and **uploaded to the Moodle System by Apr 21, 2019 23:55 (Sun) (GMT +8:00).**

2. Reminder on plagiarism:

Plagiarism is a serious misconduct. Students taking part in plagiarism, whether copying from others or allowing others to copy one's work, will receive heavy penalties. The misconduct will be reported to the University's Disciplinary Committee for disciplinary action.

Answer ALL questions.

Q1.

(a) Let the likelihood of the two classes ω_1 and ω_2 with respect to x be given by

$$p(x|\omega_1)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(x+2)^2}{2}} \quad \text{and} \quad p(x|\omega_2)=\frac{1}{2\sqrt{2\pi}}e^{-\frac{(x-5)^2}{8}}.$$

(Sub-total: 18)

The a priori probabilities for the two classes are given by

$P(\omega_1)=0.8$ and $P(\omega_2)=0.2$.

i) Find the Maximum Likelihood Classifier .

(5 marks)

ii) Using the Bayes rule $P(\omega_i|x)=\frac{p(x|\omega_i)P(\omega_i)}{p(x)}$, find the classifier(s) for the two classes. (If there are more than one decision boundaries, you should find them as well)

(5 marks)

Client X wants to apply the above classifier for bio-medical applications and has suggested the following loss functions for Bayes classification:

is actually ω_1	is actually ω_2	
0	5	choosing ω_1
1	0	choosing ω_2

iii) Write down the 4 different values of the loss function $\lambda(\alpha_1 \mid \omega_1)$, $\lambda(\alpha_1 \mid \omega_2)$, $\lambda(\alpha_2 \mid \omega_1)$ and $\lambda(\alpha_2 \mid \omega_2)$.

(2 mark)

iv) Find the Bayes Minimum Risk Classifier using the new loss function in (iii).

(4 marks)

v) Suggest and explain whether the new classifier in (iv) is still a minimum error rate classifier.

(2 marks)

(b) Consider the following criterion function for finding a hyperplane to separate the two classes of samples, which contain $\boldsymbol{x}_1 = [4,1]^T$, $\boldsymbol{x}_2 = [3,2]^T$ (Class 1) and $\boldsymbol{x}_3 = [6,8]^T$, $\boldsymbol{x}_4 = [9,9]^T$ (Class 2),

$$J_q(\boldsymbol{a}) = \sum_{\boldsymbol{y} \in Y_C} -\boldsymbol{a}^T \boldsymbol{y} . \qquad \text{(Sub-total: 15)}$$

i) The Gradient Descent can be used to solve $J_q(\boldsymbol{a})$. Write down the expression in terms of $\rho^{(k)}$, $\nabla_a J_q(\boldsymbol{a})$, $\boldsymbol{a}^{(k+1)}$ and $\boldsymbol{a}^{(k)}$ that solves $\boldsymbol{a}$ iteratively.

(2 marks)

ii) Suppose the augmented feature vector is defined as $\boldsymbol{y} = [1, x_1, x_2]^T$. Using (i) and (ii), find $\boldsymbol{a}^{(2)}$ and $\boldsymbol{a}^{(3)}$ with an initialization $\boldsymbol{a}^{(1)} = [0,0,0]^T$ and a step size $\rho^{(k)} = 1$.

(6 marks)

iii) Student Y suggests the soft-margin SVM should be employed rather than the perceptron, which is given as $\min_{\boldsymbol{w}, w_0, \xi_i} \|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{N} \max(0, 1 - z_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0))$, $z_i = 1, -1$. Using an initialization $\tilde{\boldsymbol{w}}^{(1)} = [w_0^{(1)}, \boldsymbol{w}^{(1)T}]^T = [0,0,0]^T$, step size $\rho^{(k)} = 0.1$ and regularization parameter $C = 10$, find $\tilde{\boldsymbol{w}}^{(2)}$ and $\tilde{\boldsymbol{w}}^{(3)}$.

(7 marks)

Q2.

(a) Let the likelihood of a parameter θ of the density function given as

$$p(x|\theta)=\begin{cases}\dfrac{4}{3\sqrt{\pi}}\theta^{5/2}x^4\exp(-\theta x^2) & x\geq 0\\ 0 & otherwise\end{cases}. \qquad \text{(Sub-total: 16)}$$

i) Given a set of independent feature samples $\{x_1,x_2,x_3,x_4\}=\{2,5,7,11\}$, determine the maximum likelihood of θ.

(6 marks)

Assume that the parameter θ has an a priori probability

$$p(\theta)=0.5[\delta(\theta-2)+\delta(\theta-3)],$$

where $\delta(.)$ is the ideal unit impulse function informally defined as:

$$\delta(y)=\begin{cases}\infty & y=0\\ 0 & otherwise\end{cases} \text{ and } \int_{-\infty}^{\infty}\delta(y)=1.$$

ii) Determine the posterior probability $p(\theta|x_1,x_2,x_3,x_4)$.

(6 marks)

iii) Find the Maximum A Posteriori (**MAP**) Estimate of θ.

(4 marks)

(b) Consider the following independently drawn samples

$$\boldsymbol{X}=\{1,2,2,4,5,7,8,9,9\},\quad N=9 \qquad \text{(Sub-total: 17)}$$

i) Find $p(x)$ for $x=4.5$ using the Parzen window with a bandwidth $h_d=2$ using the rectangular window.

(5 marks)

ii) The Silverman's Rule is a method to choose the bandwidth. Suggest under what situation the determined bandwidth is optimal.

(1 mark)

iii) Find $p(x)$ for $x=4.5$ using the kNN method with $k_n=3$.

(5 marks)

iv) Suppose $\boldsymbol{X}_1=[1,1,3,4,5,5]$ and $\boldsymbol{X}_2=[8,9,9,11,12,12]$ belongs to class 1 and class 2 respectively, suggest which class does an arbitrary value $x=7$ belongs to if the kNN method with $k_n=3$ is used.

(5 marks)

vi) Explain why an even k_n should not be used in a two-class classification problem.

(1 mark)

Q3.

(a) Consider the following 2-class data samples.

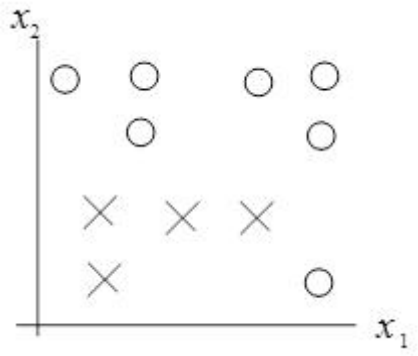

Figure 1. A plot of the two-class samples. (Sub-total: 16)

i) Draw a neural network with fewest units that can separate the samples.

(2 marks)

ii) Given the 10 independent samples above $\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_{10}$ and target values $t_1, t_2, ...t_{10}$,write the risk function for back propagation assuming a quadratic loss in terms of $\boldsymbol{x}_n$ and t_n, $n = 1,2,...,10$.

(2 marks)

iii) What is the dimension of the gradient of the risk function?

(2 marks)

iv) Express the partial derivative $\partial R / \partial w_{jk}$ of the synapse w_{jk} connecting the $j-th$ input to the $k-th$ net activation of the final layer in terms of the targets t_n , output from the previous layers $z_{p,j,n}$, the activation function $f_k(u)$ and its derivative $f_k'(u)$.

(2 marks)

v) Write down the primal formulation of the linear hard-margin SVM in the form a constrained optimization problem.

(2 marks)

Consider the following inequality constrained solver for vi), vii) and viii):

$$\min_{\boldsymbol{x}} . \ f(\tilde{\boldsymbol{w}}) \text{ subject to } \boldsymbol{b}_l \le \boldsymbol{c}(\tilde{\boldsymbol{w}}) \le \boldsymbol{b}_u , \ \boldsymbol{w} = [w_1, w_2, ..., w_I]^T .$$

vii) What should be $f(\tilde{\boldsymbol{w}})$ and $\boldsymbol{c}(\tilde{\boldsymbol{w}})$ if it is quadratic programming solver? Define the necessary input matrices, vectors or scalars (if required)

(3 marks)

viii) Using the results of vii), write down the input matrices, vectors or scalars (if required) if the above quadratic programming solver is used to solve a hard-margin linear SVM.

(3 marks)

(b) Consider the following data samples:

$$\boldsymbol{X}_1 = [[2,4]^T, [3,6]^T, [5,8]^T, [6,6]^T, [8,10]^T]^T$$

$$\boldsymbol{X}_2 = [[7,4]^T, [8,5]^T, [9,7]^T, [10,6]^T, [11,10]^T]^T$$

(Sub-total: 18)

i) Find the 1st and 2nd Principal Components (PCs) of $\boldsymbol{X} = [\boldsymbol{X}_1^T, \boldsymbol{X}_2^T]^T$

(4 marks)

ii) Determine the Fisher's Linear Discriminant (FLD) $\boldsymbol{w}$. Normalize $\boldsymbol{w}$.

(4 marks)

iii) State ONE merit and ONE short-coming for the PCA and the FLD.

(2 marks)

iv) Student A suspects that there is a hidden class among the samples and he wants to perform clustering assuming there are 3 classes. Describe the k-means procedure in words.

(3 marks)

v) Student Z raised concern about Student A's assumption on the number of classes. He suggests one should begin by assigning each sample to the same cluster (so there is only one cluster). Afterwards, the cluster is split into parts until no smaller clusters could be formed. Name this approach.

(1 mark)

vi) Explain the meaning of an impurity measure in decision tree.

(2 marks)

vii) Explain why transformation and filtering techniques are used in feature extraction. Suggest an advantage of HARR transform over a 2-D Fourier Transform.

(2 marks)

END OF ASSIGNMENT