EE 5184 Machine Learning Final Exam
Problem Sheet

Date: 2018/12/28

The paper is double-sided, 6 pages, consisting of 9 questions. Total 105 points + 10 bonus points.

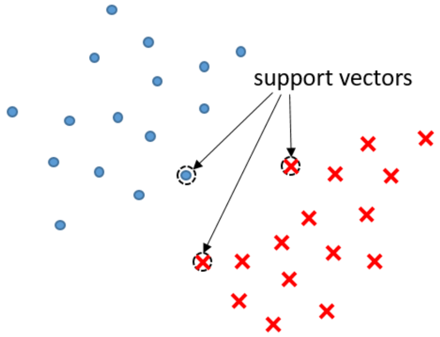
In this exam we denote

- Sigmoid function: $\sigma(z) = \frac{1}{1+e^{-z}}$. You may apply approximate $\sigma(z) \approx \begin{cases} 0, \text{if } z \le -10 \\ 1, \text{if } z \ge 10 \end{cases}$.
- Sign function: $\text{sgn}(z) = \begin{cases} 1, \text{if } z > 0 \\ 0, \text{if } z = 0 \\ -1, \text{if } z < 0 \end{cases}$.
- Unless otherwise specified, all log refer to natural log, i.e., log_e.

Problem 1: (20 pts) Multiple Selection (多選題有倒扣，最多倒扣至本大題零分)

Please answer the following multiple selection questions. Wrong selections will result in inverted scores. ***No derivation required.***

(1) Suppose you are using a hard margin linear SVM classifier on 2 class classification problem. Now you have been given the following data in which some points are dash-circled that are representing support vectors.



(A) Removing any dash-circled points from the data will change the decision boundary.
(B) Removing any dash-circled points from the data will not change the decision boundary.
(C) Removing any non-dash-circled points from the data will change the decision boundary.
(D) Removing any non-dash-circled points from the data will not change the decision boundary.
(E) Removing all non-dash-circled points from the data will not change the decision boundary.

(2) If we increase parameter C in soft margin linear SVM classifier, what may happen?
(A) The training error decreases.
(B) The training error increases.
(C) The margin decreases.
(D) The margin increases.
(E) The testing error decreases.

(3) Suppose you are using a kernel SVM to 2 class classification problem, where the data points are distributed on the x-y plane (i.e., data points are 2 dimensional). Suppose we choose kernel function as $\text{k}\big((\text{x},\text{y}),(\text{x}',\text{y}')\big) = (xx' + yy')^2$, which of the following decision boundaries, as described by equation $\text{f}(\text{x},\text{y}) = 0$, are possible?
(A) $\text{f}(\text{x},\text{y}) = \text{x} + \text{y}$.
(B) $\text{f}(\text{x},\text{y}) = x^2 + y^2$.
(C) $\text{f}(\text{x},\text{y}) = (x + y)^2$.
(D) $\text{f}(\text{x},\text{y}) = (x-1)^2 + 3(y+2)^2$.
(E) $\text{f}(\text{x},\text{y}) = x^2 - 4y$.

(4) Suppose you are using a kernel SVM to 2 class classification problem, where the data points are distributed on the x-y plane (i.e., data points are 2 dimensional). Suppose we choose kernel function as $\mathrm{k}\big((\mathrm{x},\mathrm{y}),(\mathrm{x}',\mathrm{y}')\big) = (1 + xx' + yy')^2$, which of the following decision boundaries, as described by equation $\mathrm{f}(\mathrm{x},\mathrm{y}) = 0$, are possible?

(A) $\mathrm{f}(\mathrm{x},\mathrm{y}) = \mathrm{x} + \mathrm{y}$.

(B) $\mathrm{f}(\mathrm{x},\mathrm{y}) = x^2 + y^2$.

(C) $\mathrm{f}(\mathrm{x},\mathrm{y}) = (x + y)^2$.

(D) $\mathrm{f}(\mathrm{x},\mathrm{y}) = (x - 1)^2 + 3(y + 2)^2$.

(E) $\mathrm{f}(\mathrm{x},\mathrm{y}) = x^2 - 4y$.

(5) Suppose a SVM classifier is trained from data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, where $y_i \in \{+1, -1\}$ denotes the labels, and the classifier classifies $\boldsymbol{x}$ as positive label if $\mathrm{f}(\mathbf{x}) = \boldsymbol{w}^T\boldsymbol{x} + b \geq 0$.

The primal problem for solving $\boldsymbol{w}$ is given by

Minimize	$\frac{1}{2}\lVert\boldsymbol{w}\rVert^2 + C\sum_{i=1}^{N}\xi_i$
Subject to	$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i, \forall i = 1, \dots, \mathrm{N}$
Variables	$\boldsymbol{w} \in \mathbb{R}^{\boldsymbol{d}},\ b \in \mathbb{R}, \xi_1, \dots, \xi_N \geq 0$

The dual problem for solving α_i's in $\boldsymbol{w} = \sum_{i=1}^{N}\alpha_i y_i \boldsymbol{x}_i$ is given by

Maximize	$\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (\boldsymbol{x}_i{}^T\boldsymbol{x}_j)$
Subject to	$\sum_{i=1}^{N}\alpha_i y_i = 0$
Variables	$0 \leq \alpha_i \leq C$

Upon achieving optimal in both primal and dual problems,

(A) If $\alpha_i > 0$ then $\xi_i > 0$.

(B) If $\xi_i > 0$ then $\alpha_i > 0$.

(C) If $\alpha_i = C$ then $\xi_i > 0$.

(D) If $\xi_i > 0$ then $\alpha_i = C$.

(E) If $\alpha_i = 0$ then $\xi_i = 0$.

(6) Suppose the neural network was trained with dropout rate p=0.2, in the sense that each neuron had probability p of only passing zero to the consecutive neurons. After the neural network is trained, how should we modify the weights in the neural network, so it can be applied without dropout?

(A) Multiply each weight by 0.2.

(B) Multiply each weight by 0.8.

(C) Multiply each weight by 1.2.

(D) Multiply each weight by 1.25.

(E) No modification is needed.

(7) Suppose you have an input volume of dimension 48x48x3. That is, the inputs are images of size 48x48 with 3 channels (RGB). How many parameters would a single 5x5 convolutional filter have (not including bias)?

Note: Since there is just a single 5x5 convolutional filter, the output has only 1 channel.

(A) 3

(B) 25

(C) 75

(D) 2304

(E) 6912

(8) In the context of ensemble methods, which of the following statements are true?

(A) In bagging, each weak classifier is independent of each other.

(B) In boosting, each weak classifier is independent of each other.

(C) In case of under-fitting, we expect bagging as a better remedy over boosting.

(D) In case of over-fitting, we expect bagging as a better remedy over boosting.

(E) AdaBoost (Adaptive boosting) considers hinge loss.

(9) Which of the following are convex functions on $\mathbb{R}^2$?

(A) $\mathrm{f}(\mathrm{u},\mathrm{v}) = u^2 - 4uv + v^2$

(B) $\mathrm{f}(\mathrm{u},\mathrm{v}) = \mathrm{u} - 3v$

(C) $\mathrm{f}(\mathrm{u},\mathrm{v}) = \log(u^2 + 1)$

(D) $\mathrm{f}(\mathrm{u},\mathrm{v}) = u^2 + v^2 + \max(-u, 0)$

(E) $\mathrm{f}(\mathrm{u},\mathrm{v}) = \mathrm{sgn}(\mathrm{u})$

(10) Select all that belong to unsupervised learning algorithms.

(A) Deep auto-encoder

(B) Hierarchical Agglomerative Clustering

(C) K-means

(D) Linear regression

(E) Logistic regression

(F) Locally Linear Embedding (LLE)

(G) Principle Component Analysis (PCA)

(H) Random forest

(I) Support Vector Machine (SVM)

(J) t-Distributed Stochastic Neighbor Embedding (t-SNE)

Problem 2: (10 pts) Linear Regression

Consider regression function $f_{\boldsymbol{w}}(\mathrm{x}) = w_0 + w_1 x + w_2 x^2$. Consider 10 data points as follows

i	1	2	3	4	5	6	7	8	9	10
x_i	0.89	0.03	0.49	0.17	0.98	0.71	0.50	0.47	0.06	0.68
y_i	3.03	-1.14	0.96	-0.53	3.90	2.21	1.09	0.78	-0.77	1.97

Find the values of w_0, w_1, w_2 that minimizes the following loss function

$$L(\boldsymbol{w}) = \sum_{i=1}^{10} |y_i - f_{\boldsymbol{w}}(x_i)|^2$$

Problem 3: (12 pts) Fitting Single Gaussian Distribution

Denote $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

(1) (2 pts) Write down the probability density function for $\mathrm{X} \sim \mathcal{N}\left(\begin{bmatrix} 3 \\ -2 \end{bmatrix}, \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}\right)$.

(2) (10 pts) Suppose the following 10 data points

$$(4.48, 1.27),\ (2.36, 1.78),\ (4.21, -1.10),\ (5.42, 9.42),\ (3.48, -1.91),$$
$$(1.56, -2.39),\ (3.71, -2.97),\ (3.37, -1.13),\ (3.35, 1.04),\ (4.26, -1.65)$$

are independently generated from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Find the maximum likelihood estimator of the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Problem 4: (3 pts) Cross-entropy

Let $\mathrm{X} = \{$*滷肉飯, 牛肉麵, 大波蘿, 壽司, 素食*$\}$. Consider two probability distributions P_X, Q_X as follows:

x	*滷肉飯*	*牛肉麵*	*大波蘿*	*壽司*	*素食*
$P_X(x)$	0.3	0.4	0.15	0.1	0.05
$Q_X(x)$	0.2	0.1	0.05	0.25	0.4

Find the cross entropy

$$H(P_X, Q_X) = \sum_{x \in X} P_X(x) ln\left(\frac{1}{Q_X(x)}\right)$$

Problem 5: (10 pts) Logistic regression

A group of 10 students spend various hours studying for the machine learning (ML) exam. The following table shows the number of hours each student spent studying, and whether they passed (1) or failed (0).

Hours (X)	0.5	1	1.5	1.75	2.5	2.75	3.25	4	4.5	5
Pass (Y)	0	0	0	1	0	1	1	1	1	1

Consider the logistic model that predicts the probability of passing the exam by the hours spent studying

$$P(Y=1|X=x)=\sigma(wx+b)$$

Find the cross entropy loss should we fit the data with logistic model with parameter $\mathrm{w}=1.5, \mathrm{b}=-4$.

Problem 6: (5 pts + 10 pts Bonus) Gaussian Mixture Model and Expectation Maximization

Suppose we wish to fit the following 10 data points (distributed in 1-D space)

-12.72, -2.05, -6.56, 2.55, -1.77, 9.19, 8.85, -3.34, -3.74, 3.63

by Gaussian mixture model $p_\theta(x)$ parameterized by $\theta=(\pi_1,\mu_1,\sigma_1,\pi_2,\mu_2,\sigma_2)$, as given as follows

$$p_\theta(x)=\pi_1\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}+\pi_2\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$$

Suppose the initial guess of parameter is $\theta^{(0)}=\left(\pi_1^{(0)},\mu_1^{(0)},\sigma_1^{(0)},\pi_2^{(0)},\mu_2^{(0)},\sigma_2^{(0)}\right)=(0.3,-1,2,0.7,2,3)$

(a) **(5 pts)** Compute the log likelihood of parameter $\theta^{(0)}$.

(b) **(10 pts Bonus)** Apply expectation maximization algorithm, find the next update of parameters

$$\theta^{(1)}=\left(\pi_1^{(1)},\mu_1^{(1)},\sigma_1^{(1)},\pi_2^{(1)},\mu_2^{(1)},\sigma_2^{(1)}\right)$$

Problem 7: (14 pts) Principle Component Analysis

Consider the following 10 data points distributed in 2-D space as follows

(1.91, -0.11), (-2.24, -1.09), (1.36, -0.20), (0.33, 0.13), (-0.33, 0.37),
(0.00, -0.63), (-3.10, -0.47), (-0.34, 2.38), (2.43, -3.00), (-0.02, 2.62)

(a) **(10 pts)** Find the first and second principle axes.

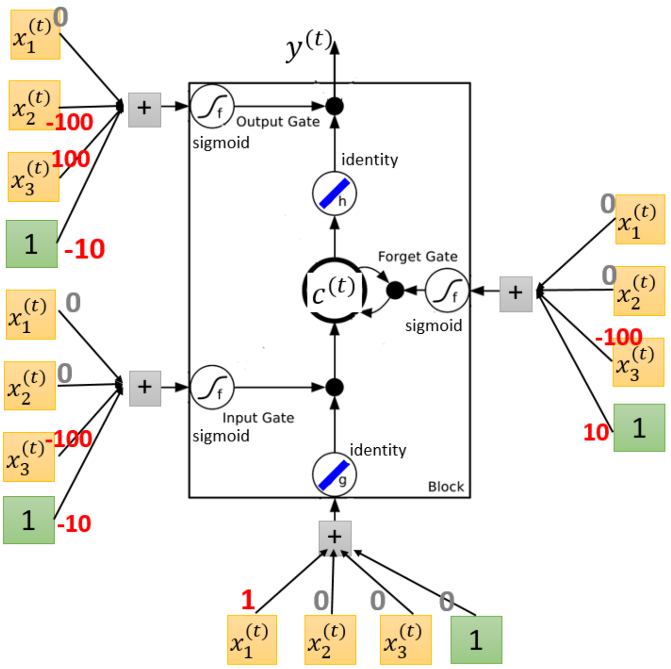
(b) **(4 pts)** Find the first and second principle components for data point (0.96, 0.28).

Note: The data points have zero mean.

Problem 8: (9 pts) LSTM

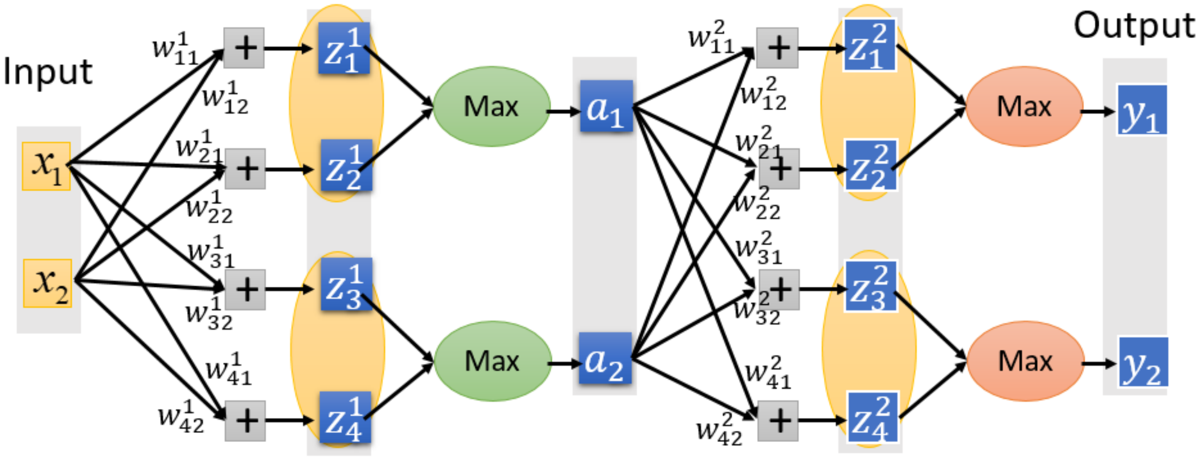
Consider a LSTM node as follows. ***Please fill out the following table in the answer sheet. No derivation required.***

Time	t	0	1	2	3	4	5	6	7	8	9	10
Input	$x_1^{(t)}$	0	1	3	2	4	-3	7	2	3	-5	8
	$x_2^{(t)}$	0	0	1	-2	-3	-1	0	0	0	0	-2
	$x_3^{(t)}$	0	0	-1	-1	-1	0	0	1	1	-1	-1
Memory cell	$c^{(t)}$	0										
Output	$y^{(t)}$	0										



Problem 9: (22 pts) Feedforward and Back Propagation

Consider the following neural network

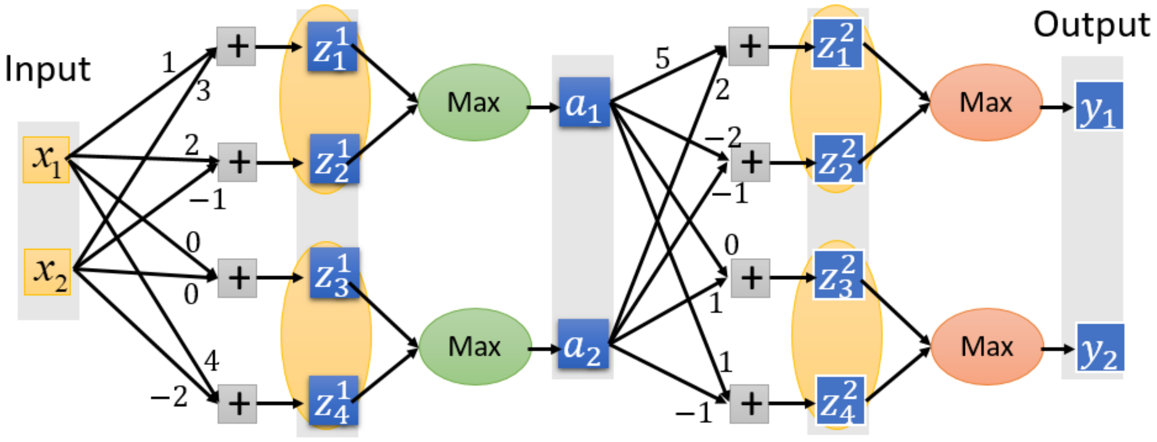


The above neural network can be represented as a function f_θ, namely

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = f_\theta\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right),$$

where parameter θ records all the weights w_{ij}^k.

(a) **(6 pts)** Suppose the weights are initialized as follows



If $(x_1, x_2) = (1, -1)$, ***please fill out the following table in the answer sheet. No derivation required.***

Variable	z_1^1	z_2^1	z_3^1	z_4^1	a_1	a_2	z_1^2	z_2^2	z_3^2	z_4^2	y_1	y_2
Value												

(b) **(16 pts)** Continuing (a), if the ground truth is $(\hat{y}_1, \hat{y}_2) = (-10, 7)$, and the loss function is defined as

$$\mathrm{L}(\theta) = \left\| \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} - f_\theta\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) \right\|^2$$

Perform back propagation and ***fill out the following table in the answer sheet. No derivation required.***

Variable	$\frac{\partial \mathrm{L}}{\partial w_{11}^1}$	$\frac{\partial \mathrm{L}}{\partial w_{12}^1}$	$\frac{\partial \mathrm{L}}{\partial w_{21}^1}$	$\frac{\partial \mathrm{L}}{\partial w_{22}^1}$	$\frac{\partial \mathrm{L}}{\partial w_{31}^1}$	$\frac{\partial \mathrm{L}}{\partial w_{32}^1}$	$\frac{\partial \mathrm{L}}{\partial w_{41}^1}$	$\frac{\partial \mathrm{L}}{\partial w_{42}^1}$
Value								
Variable	$\frac{\partial \mathrm{L}}{\partial w_{11}^2}$	$\frac{\partial \mathrm{L}}{\partial w_{12}^2}$	$\frac{\partial \mathrm{L}}{\partial w_{21}^2}$	$\frac{\partial \mathrm{L}}{\partial w_{22}^2}$	$\frac{\partial \mathrm{L}}{\partial w_{31}^2}$	$\frac{\partial \mathrm{L}}{\partial w_{32}^2}$	$\frac{\partial \mathrm{L}}{\partial w_{41}^2}$	$\frac{\partial \mathrm{L}}{\partial w_{42}^2}$
Value								

END OF PROBLEM SHEET

EE 5184 Machine Learning Final Exam

Answer Sheet

Date: 2018/12/28

系級:__________ 學號: __________ 姓名: __________
Department/Year Student ID no. Name

Problem 8:

Time	t	0	1	2	3	4	5	6	7	8	9	10
Input	$x_1^{(t)}$	0	1	3	2	4	-3	7	2	3	-5	8
	$x_2^{(t)}$	0	0	1	-2	-3	-1	0	0	0	0	-2
	$x_3^{(t)}$	0	0	-1	-1	-1	0	0	1	1	-1	-1
Memory cell	$c^{(t)}$	0										
Output	$y^{(t)}$	0										

Problem 9:

(a)

Variable	z_1^1	z_2^1	z_3^1	z_4^1	a_1	a_2	z_1^2	z_2^2	z_3^2	z_4^2	y_1	y_2
Value												

(b)

Variable	$\frac{\partial L}{\partial w_{11}^1}$	$\frac{\partial L}{\partial w_{12}^1}$	$\frac{\partial L}{\partial w_{21}^1}$	$\frac{\partial L}{\partial w_{22}^1}$	$\frac{\partial L}{\partial w_{31}^1}$	$\frac{\partial L}{\partial w_{32}^1}$	$\frac{\partial L}{\partial w_{41}^1}$	$\frac{\partial L}{\partial w_{42}^1}$
Value								
Variable	$\frac{\partial L}{\partial w_{11}^2}$	$\frac{\partial L}{\partial w_{12}^2}$	$\frac{\partial L}{\partial w_{21}^2}$	$\frac{\partial L}{\partial w_{22}^2}$	$\frac{\partial L}{\partial w_{31}^2}$	$\frac{\partial L}{\partial w_{32}^2}$	$\frac{\partial L}{\partial w_{41}^2}$	$\frac{\partial L}{\partial w_{42}^2}$
Value								

END OF ANSWER SHEET