SINGAPORE UNIVERSITY OF
TECHNOLOGY AND DESIGN

Established in collaboration with MIT

01.112 Machine Learning
2017 Midterm

03 Nov 2017

2:30-4:30pm

Student Name:

Student ID:

Student Pillar:

Instructions to Candidates

1. There are 7 questions with 12 printed pages. (This title page counts as the first page.)
2. This is a closed book examination.
3. Cheat sheets are not allowed.
4. Attempt all the questions in this booklet.
5. Write your answers in this question booklet.
6. Do not turn over the title page until you are given further instructions.
7. All the best!

Problem	Score
1	/11
2	/22
3	/08
4	/10
5	/07
6	/05
7	/07
Total	/70

Q1. Major Themes [1+2+2+3x2=11pt]

(1a) The ultimate goal of every machine learning algorithm is

A. Computational speed
B. Generalization
C. Bias reduction
D. Dimensionality reduction

Answer B

(1b) From the list below, pick the top two dangers of treating machine learning as a black box.

A. Difficult to debug if something goes wrong.
B. An algorithm may be applied to data which do not fulfill the assumptions of the algorithm, leading to incorrect conclusions.
C. Difficult to integrate with other components of the software system.
D. Difficult to discern if the machine learning outcomes are due to statistically-significant relationships in the data, or if they are just consequences of the algorithmic design.

Answers B D

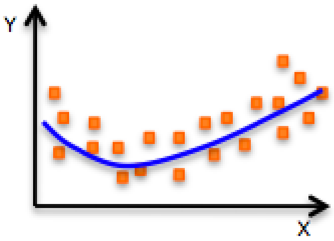
(1c) From the list below, pick the top two roles of unsupervised learning in supervised learning.

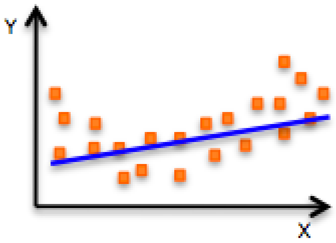
A. To find better features for supervised learning.
B. To reduce the dimensionality of the supervised learning problem.
C. To produce labels and responses, when they are not available, for supervised learning.
D. To validate the outcomes of a supervised learning algorithm.

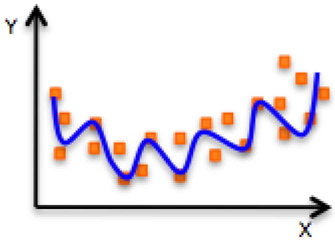
Answers A B

(1d) For each row, match the words 'overfitting', 'underfitting' and 'just right' to the three pictures, and write your answers in the boxes below the pictures.

Regression





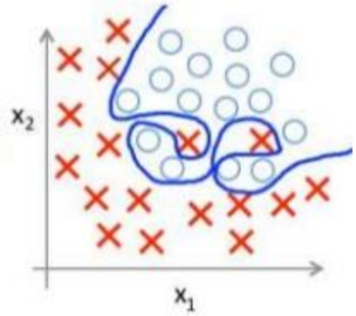


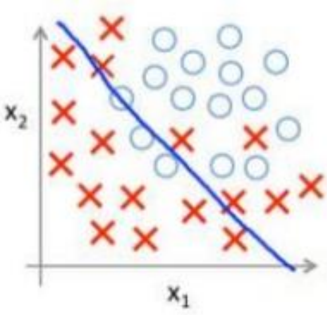
Just Right

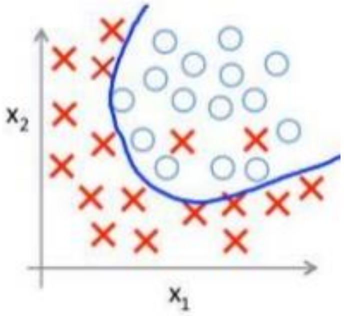
Underfitting

Overfitting

Classification





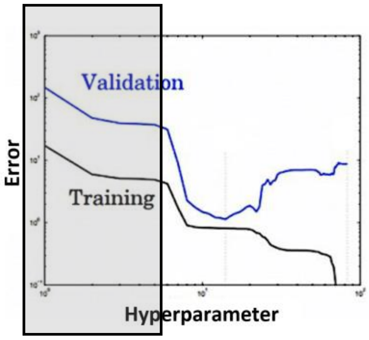


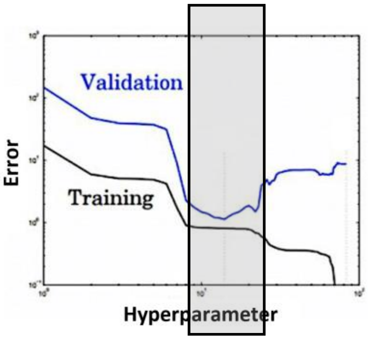
Overfitting

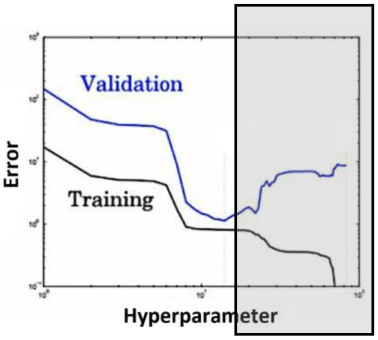
Underfitting

Just Right

Validation







Underfitting

Just Right

Overfitting

Q2. Classification and Regression [6x3+2+2=22pt]

For each point loss in the table below, fill in the corresponding predictor (i.e. classification or regression function), learning technique and learning algorithm using the options (e.g. P1, T1, A1) listed below.

Recall that the training loss is the average of the point losses over the training data, i.e.

$$\mathcal{L}_n(\theta, \theta_0; \mathcal{S}_n) = \frac{1}{n} \sum_{(x,y) \in \mathcal{S}_n} \mathcal{L}_1(\theta, \theta_0; x, y).$$

You may assume that the training data $\mathcal{S}_n$ is a set of n pairs (x, y) where $x \in \mathbb{R}^d$ is a feature vector and y is either a signed label $y \in \{-1,1\}$ or a real-valued response $y \in \mathbb{R}$, depending on whether the problem-of-interest is classification or regression. The model parameters are $\theta \in \mathbb{R}^d$ and $\theta_0 \in \mathbb{R}$.

Here are the loss functions that we will use in the table:

- Hinge loss $\mathcal{L}_H(z) = \max\{1 - z,\ 0\}$
- Logistic loss $\mathcal{L}_L(z) = \log(1 + e^{-z})$
- Squared loss $\mathcal{L}_S(z) = \frac{1}{2} z^2$
- Zero-one loss $\mathcal{L}_Z(z) = [\![z \leq 0]\!]$

where $[\![\cdot]\!]$ is the indicator function.

Predictor

P1. $f(x; \theta,\ \theta_0) = \theta^\top x + \theta_0$
P2. $h(x; \theta, \theta_0) = \mathrm{sign}(\theta^\top x + \theta_0)$
P3. $p(y|x, \theta, \theta_0) = \mathrm{sigmoid}\big(y(\theta^\top x + \theta_0)\big)$

Technique

T1. Ridge Regression
T2. Linear Classification using Hinge Loss
T3. Linear Regression
T4. Logistic Regression
T5. Perceptron (with Offset)
T6. Support Vector Machine with Slack Variables

Algorithm

A1. Exact Solution
A2. Gradient Descent
A3. Perceptron Algorithm

(2a)

Point Loss $\mathcal{L}_1(\boldsymbol{\theta}, \boldsymbol{\theta_0}; \boldsymbol{x}, \boldsymbol{y})$	**Predictor**	**Technique**	**Algorithm***
$\mathcal{L}_S(y - (\theta^\top x + \theta_0))$	P1	T3	A1, A2
$\mathcal{L}_S(y - (\theta^\top x + \theta_0)) + \frac{\lambda}{2}\|\theta\|^2$	P1	T1	A1, A2
$\mathcal{L}_H\big(y(\theta^\top x + \theta_0)\big)$	P2	T2	A2
$\mathcal{L}_H\big(y(\theta^\top x + \theta_0)\big) + \frac{\lambda}{2}\|\theta\|^2$	P2	T6	A2
$\mathcal{L}_Z\big(y(\theta^\top x + \theta_0)\big)$	P2	T5	A3
$\mathcal{L}_L\big(y(\theta^\top x + \theta_0)\big)$	P3	T4	A2

* It is possible to have more than one algorithm for each row.

(2b) Which of the following is the gradient for linear regression without the offset θ_0?

A. $\nabla_\theta \mathcal{L}_n(\theta; \mathcal{S}_n) = \frac{1}{n}\sum_{(x,y)\in\mathcal{S}_n} x(\theta^\top x - y)$

B. $\nabla_\theta \mathcal{L}_n(\theta; \mathcal{S}_n) = \frac{1}{n}\sum_{(x,y)\in\mathcal{S}_n} (\theta^\top x - y)$

C. $\nabla_\theta \mathcal{L}_n(\theta; \mathcal{S}_n) = -\frac{1}{n}\sum_{(x,y)\in\mathcal{S}_n} x(\theta^\top x - y)$

D. $\nabla_\theta \mathcal{L}_n(\theta; \mathcal{S}_n) = -\frac{1}{n}\sum_{(x,y)\in\mathcal{S}_n} (\theta^\top x - y)$

Answer $\boxed{A}$

(2c) Which of the following is the gradient for logistic regression without the offset θ_0?

A. $\nabla_\theta \mathcal{L}_n(\theta; \mathcal{S}_n) = \frac{1}{n}\sum_{(x,y)\in\mathcal{S}_n} x\big(\mathrm{sigmoid}\big(y(\theta^\top x)\big) - \llbracket y = 1 \rrbracket\big)$

B. $\nabla_\theta \mathcal{L}_n(\theta; \mathcal{S}_n) = \frac{1}{n}\sum_{(x,y)\in\mathcal{S}_n} x(\mathrm{sigmoid}(\theta^\top x) - \llbracket y = 1 \rrbracket)$

C. $\nabla_\theta \mathcal{L}_n(\theta; \mathcal{S}_n) = -\frac{1}{n}\sum_{(x,y)\in\mathcal{S}_n} x\big(\mathrm{sigmoid}\big(y(\theta^\top x)\big) - \llbracket y = 1 \rrbracket\big)$

D. $\nabla_\theta \mathcal{L}_n(\theta; \mathcal{S}_n) = -\frac{1}{n}\sum_{(x,y)\in\mathcal{S}_n} x(\mathrm{sigmoid}(\theta^\top x) - \llbracket y = 1 \rrbracket)$

Answer $\boxed{B}$

Q3. Clustering [2+2+4=8pt]

The k-means algorithm iteratively computes the set of centroids given a clustering of the data points, and the clustering of the data points given a set of centroids. In this question, you will analyze the performance of the k-means algorithm on a one-dimensional problem.

Suppose that we have six data points $1, 2, 3, 50, 100, 150 \in \mathbb{R}$.

(3a) If $k = 1$ in your k-means algorithm, where would the centroid of the single cluster be?

A. 50
B. 51
C. 75.5
D. 100

Answer B

(3b) If $k = 2$ in your k-means algorithm, there will be two Voronoi regions corresponding to the two clusters. The boundary between the Voronoi regions will be:

A. Closer to the centroid with the tightly-clustered points.
B. Closer to the centroid with the sparsely-clustered points.
C. Exactly halfway between the two centroids.
D. Exactly halfway between the point 1 and the point 150.

Answer C

(3c) If $k = 2$ in your k-means algorithm, which of the following are possible clusters when the algorithm has converged? Circle 'Yes' if it is a possible clustering, and 'No' otherwise.

- {1, 2} and {3, 50, 100, 150} Yes / [✓ No]
- {1, 2, 3} and {50, 100, 150} Yes / [✓ No]
- {1, 2, 3, 50} and {100, 150} [✓ Yes] / No
- {1, 2, 3, 50, 100} and {150} Yes / [✓ No]

Q4. Collaborative Filtering [2+2+3+3=10pt]

Suppose that you are given a data set from Amazon in the form of a partially-observed matrix Y whose entry Y_{ai} represents the rating of customer a for a product i. You decide to try both the k-nearest-neighbors algorithm and matrix factorization to predict the values of unknown ratings Y_{ai}.

In k-nearest-neighbors, to predict the *unbiased* rating $Y_{ai} - \bar{Y}_a$ where $\bar{Y}_a$ is the average of all observed ratings by customer a, we use the weighted sum of the unbiased ratings of neighbors b which are nearest to a. These neighbors are ranked according to a cosine similarity function $\mathrm{sim}(a, b)$.

(4a) If we predict $Y_{ai} - \bar{Y}_a$ using $\sum_b w_b r_b$, which weights w_b and values r_b should we use?

A. $w_b = \frac{\mathrm{sim}(a,b)}{\sum_{b'} \mathrm{sim}(a,b')}$ and $r_b = Y_{bi} - \bar{Y}_b$

B. $w_b = \frac{|\mathrm{sim}(a,b)|}{\sum_{b'} \mathrm{sim}(a,b')}$ and $r_b = Y_{bi} - \bar{Y}_b$

C. $w_b = \frac{\mathrm{sim}(a,b)}{\sum_{b'} |\mathrm{sim}(a,b')|}$ and $r_b = \mathrm{sign}\big(\mathrm{sim}(a,b)\big)\,(Y_{bi} - \bar{Y}_b)$

D. $w_b = \frac{|\mathrm{sim}(a,b)|}{\sum_{b'} |\mathrm{sim}(a,b')|}$ and $r_b = \mathrm{sign}\big(\mathrm{sim}(a,b)\big)\,(Y_{bi} - \bar{Y}_b)$

Answer

A or D

In matrix factorization, we choose vectors $U_a \in \mathbb{R}^k$ for each customer a, and vectors $V_i \in \mathbb{R}^k$ for each product i, such that $U_a^\top V_i$ is a good prediction for the rating Y_{ai}. To optimize for these vectors, we use the alternating least-squares algorithm which takes the following form:

1. Initialize vectors $V_i \in \mathbb{R}^k$ randomly.
2. Repeat until convergence:
    a. While fixing the V_i, for each customer a, find the optimal U_a minimizing

$$\sum_{(a,i)\text{ observed}} \frac{1}{2}(Y_{ai} - U_a^\top V_i)^2 + \frac{\lambda}{2}\|U_a\|^2$$

    b. While fixing the U_a, for each product i, find the optimal V_i minimizing

$$\sum_{(a,i)\text{ observed}} \frac{1}{2}(Y_{ai} - U_a^\top V_i)^2 + \frac{\lambda}{2}\|V_i\|^2$$

For step (2a) of the algorithm, for a given customer a, let Z be the vector of all product ratings observed from customer a. Let X be the matrix whose j-th row is V_i if the entry Z_j is a rating for product i.

(4b) Which of the following is the exact solution for step (2a) of the algorithm?

A. $(X^\top X - \lambda I)^{-1} X^\top Z$
B. $(X^\top Z - \lambda I)^{-1} X^\top X$
C. $(X^\top X + \lambda I)^{-1} X^\top Z$
D. $(X^\top Z + \lambda I)^{-1} X^\top X$

Answer C

(4c) Which of the following is the best reason for using a nonzero value for λ?

A. We regularize to pick the most relevant features for prediction.
B. We do this to convert the problem into a convex optimization problem.
C. The matrix $X^\top X$ is always non-invertible, so we use $\lambda > 0$ to ensure invertibility.
D. There are infinitely many solutions for the original training loss of the matrix factorization problem, so we use $\lambda > 0$ to ensure that we get a unique solution.

Answer D

To improve the prediction accuracies, we now introduce additional parameters and approximate

$$Y_{ai} \approx U_a^\top V_i + \beta_a + \gamma_i + \mu$$

where β_a, γ_i are parameters that represent the bias in the ratings from customer a and from product i respectively, and μ is the average of all the observed ratings. We will pre-compute μ from the data, but the parameters β_a, γ_i will be learned by optimization. The training loss of this new model is

$$\sum_{(a,i)\text{ observed}} \frac{1}{2}(Y_{ai} - U_a^\top V_i - \beta_a - \gamma_i - \mu)^2 + \frac{\lambda}{2}\left(\sum_a \|U_a\|^2 + \sum_i \|V_i\|^2 + \sum_a \beta_a^2 + \sum_i \gamma_i^2\right).$$

By applying coordinate descent to this problem, we get the following algorithm.

1. Initialize V_i, β_a, γ_i randomly.
2. Repeat until convergence:
   a. While fixing the V_i, β_a, γ_i, find the optimal U_a.
   b. While fixing the U_a, β_a, γ_i, find the optimal V_i.
   c. While fixing the U_a, V_i, find the optimal β_a, γ_i.

(4d) In step (2c) of the algorithm, what is the exact solution for β_a?

A. $\frac{1}{1+\lambda}\sum_{(a,i)\text{ observed}}(Y_{ai} - U_a^\top V_i - \gamma_i - \mu)$

B. $\frac{1}{1-\lambda}\sum_{(a,i)\text{ observed}}(Y_{ai} - U_a^\top V_i - \gamma_i - \mu)$

C. $\frac{1}{1+\lambda}\sum_{(a,i)\text{ observed}}(U_a^\top V_i + \gamma_i + \mu - Y_{ai})$

D. $\frac{1}{1-\lambda}\sum_{(a,i)\text{ observed}}(U_a^\top V_i + \gamma_i + \mu - Y_{ai})$

Answer A

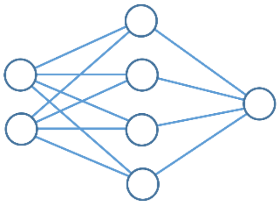
Q5. Support Vector Machines [1+1+1+2+2=7pt]

(5a) For each of the following statements, circle 'True' or 'False'.

The size of the margin of the SVM classifier is proportional to $\|\theta\|^{-1}$.	True
The SVM without slack variables applies to data that is not linearly separable.	False
For the SVM without slack variables, if a data point (x, y) lies on the edge of the margin, then it must be a support vector.	True
Consider the SVM with slack variables, whose training loss is given by $\frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{(x,y)\in\mathcal{S}_n} \max\{1 - y(\theta^\top x + \theta_0),\ 0\}.$ To increase the margin, one must increase the hyperparameter C.	False
For the SVM with slack variables, if the multiplier $\alpha_{x,y}$ for a data point (x, y) is equal to C, then the point (x, y) must lie on the edge of the margin.	False

6. Deep Learning [2+3 = 5pt]

Consider the three-layer neural network shown below, where $x = (x_1, x_2) \in \mathbb{R}^2$ are input neurons, $y = (y_1, y_2, y_3, y_4) \in \mathbb{R}^4$ are ReLU neurons, and $z \in \mathbb{R}$ is a linear neuron. Let $\tilde{z}$ be the desired output, as given by the training data.



The forward propagation of the neural network may be written as

$$u = W^{(1)}x + b^{(1)}, \qquad y = \mathrm{ReLU}(u), \qquad z = W^{(2)}y + b^{(2)}$$

where $\mathrm{ReLU}(u) = \max\{0, u\}$. The point loss $\mathcal{L}_1$ is the squared loss $\frac{1}{2}(\tilde{z} - z)^2$. Using backpropagation, the gradients of the point loss are given by

$$\nabla_{W^{(1)}}\mathcal{L}_1 = \delta^{(2)}x^\top, \qquad \nabla_{W^{(2)}}\mathcal{L}_1 = \delta^{(3)}y^\top$$

$$\nabla_{b^{(1)}}\mathcal{L}_1 = \delta^{(2)}, \qquad \nabla_{b^{(2)}}\mathcal{L}_1 = \delta^{(3)}$$

where $\delta^{(3)}, \delta^{(2)}$ are the backpropagating error signals. Let $*$ represent the element-wise multiplication of two vectors, and let H be the function where $H(u) = 1$ if $u > 0$, and $H(u) = 0$ if $u \leq 0$.

(6a) Given that $\delta^{(3)} = z - \tilde{z}$, which of the following formulas compute the error signal $\delta^{(2)}$?

A. $\delta^{(2)} = \left(W^{(2)\top}\delta^{(3)}\right) * H(y)$
B. $\delta^{(2)} = \left(W^{(2)\top}\delta^{(3)}\right) * H(u)$
C. $\delta^{(2)} = \left(W^{(2)\top}\delta^{(3)}\right) * ReLU(y)$
D. $\delta^{(2)} = \left(W^{(2)\top}\delta^{(3)}\right) * ReLU(u)$

Answer B

(6b) One can prove that ReLU networks give rise to continuous piecewise-linear functions. In other words, the space of inputs can be cut up into regions where the output of the network is a linear function within each region, and where neighboring regions have different linear functions. For our three-layer ReLU network, what is the maximum number of regions that can be obtained?

A. 4
B. 8
C. 11
D. 16

Answer C

7. Generative Methods [2+2+3=7pt]

Suppose that you are a myrmecologist (a person who studies ants), and you are currently exploring two large ant nests which are near each other in a forest.

You label the two nests '+' and '-', and let their locations be $\mu^+, \mu^- \in \mathbb{R}^2$ respectively. You observe that the positions of the ants around their own nest seem to follow spherical Gaussian distributions,

$$\mathcal{N}(\mu^+, \sigma^2 I), \qquad \mathcal{N}(\mu^-, \sigma^2 I),$$

with the same variance σ^2 for both nests because all the ants are of the same species. Let p^+ be the probability that any given ant is from the '+' nest, and let $p^- = 1 - p^+$ be the probability for the '-' nest.

(7a) Using a video (and with help from friends who are computer-vision experts), you painstakingly tracked 1000 ants, and determined which nest they are from. At some snapshot of the video, the positions and nests of these 1000 ants are given by

$$\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(1000)}, y^{(1000)}\right)$$

where each $x^{(i)} \in \mathbb{R}^2$ and each $y^{(i)} \in \{+, -\}$. Let n^+, n^- be the number of ants observed from each nest respectively. Let $[\![\cdot]\!]$ be the indicator function, and let

$$\hat{p}^+ = \frac{n^+}{1000}, \qquad \hat{p}^- = \frac{n^-}{1000},$$
$$\hat{\mu}^+ = \frac{1}{n^+}\Sigma_i \, [\![y^{(i)} = +]\!] x^{(i)}, \;\; \mu^- = \frac{1}{n^-}\Sigma_i \, [\![y^{(i)} = -]\!] x^{(i)}$$

be the maximum likelihood estimates (MLEs) of the nest probabilities and locations. Which of the following is the correct MLE of σ^2 from the data?

A. $\sigma^2 = \frac{1}{999} \; \Sigma_i \left\| x^{(i)} - \hat{\mu}^{y^{(i)}} \right\|^2$

B. $\sigma^2 = \frac{1}{999} \left(\Sigma_i \left\| x^{(i)} - \hat{\mu}^{y^{(i)}} \right\| \right)^2$

C. $\sigma^2 = \frac{1}{1000} \; \Sigma_i \left\| x^{(i)} - \hat{\mu}^{y^{(i)}} \right\|^2$

D. $\sigma^2 = \frac{1}{1000} \left(\Sigma_i \left\| x^{(i)} - \hat{\mu}^{y^{(i)}} \right\| \right)^2$

Answer $\boxed{\text{C}}$

(7b) You find a new ant wandering around at position $x \in \mathbb{R}^2$. By substituting the MLEs above into the log likelihood ratio, you determine that the ant is from the '+' nest if $\alpha^\top x + \alpha_0 > 0$ for some α, α_0. What is the value of α and α_0?

A. $\alpha = \frac{1}{\sigma^2}(\mu^+ - \mu^-),\ \alpha_0 = \frac{1}{2\sigma^2}(\|\mu^+\|^2 - \|\mu^-\|^2)$

B. $\alpha = \frac{1}{\sigma^2}(\mu^+ - \mu^-),\ \alpha_0 = \frac{1}{2\sigma^2}(\|\mu^-\|^2 - \|\mu^+\|^2)$

C. $\alpha = \frac{1}{\sigma^2}(\mu^+ - \mu^-),\ \alpha_0 = \frac{1}{2\sigma^2}(\|\mu^+\|^2 - \|\mu^-\|^2) + \log\frac{p^+}{p^-}$

D. $\alpha = \frac{1}{\sigma^2}(\mu^+ - \mu^-),\ \alpha_0 = \frac{1}{2\sigma^2}(\|\mu^-\|^2 - \|\mu^+\|^2) + \log\frac{p^+}{p^-}$

Answer D

(7c) You find a new ant wandering around at position $x \in \mathbb{R}^2$. The probability that the ant is from the '+' nest may be given by a sigmoid function

$$\mathbb{P}(+|x) = \text{sigmoid}(\theta^\top x + \theta_0) = \frac{1}{1 + e^{-(\theta^\top x + \theta_0)}}.$$

This shows that our model is a special case of logistic regression. What is the value of θ and θ_0?

A. $\theta = \frac{1}{\sigma^2}(\mu^+ - \mu^-),\ \theta_0 = \frac{1}{2\sigma^2}(\|\mu^+\|^2 - \|\mu^-\|^2) + \log\frac{p^+}{p^-}$

B. $\theta = \frac{1}{\sigma^2}(\mu^+ - \mu^-),\ \theta_0 = \frac{1}{2\sigma^2}(\|\mu^-\|^2 - \|\mu^+\|^2) + \log\frac{p^+}{p^-}$

C. $\theta = \frac{1}{\sigma^2}(\mu^- - \mu^+),\ \theta_0 = \frac{1}{2\sigma^2}(\|\mu^-\|^2 - \|\mu^+\|^2) + \log\frac{p^-}{p^+}$

D. $\theta = \frac{1}{\sigma^2}(\mu^- - \mu^+),\ \theta_0 = \frac{1}{2\sigma^2}(\|\mu^+\|^2 - \|\mu^-\|^2) + \log\frac{p^-}{p^+}$

Answer B

END OF PAPER