22F-COM SCI-M146-LEC-1 Final Exam

CHARLES ZHANG

TOTAL POINTS

68 / 100

QUESTION 1

short question 28 pts

1.1 (a) xi 3 / 3

✓ **- 0 pts** Correct (4)

- 1 pts wrong sign (-4)

- 3 pts Incorrect. (other answer.)

1.2 (b) 1 vs 1 1 / 3

- 0 pts Correct

✓ **- 2 pts Answer is a multiple of C(4,2)=6 or 16**

- 3 pts Incorrect.

1.3 (c) 1 against all 1 / 3

- 0 pts Correct

✓ **- 2 pts the wrong answer is a multiplier of 4 or 16**

- 3 pts Incorrect. (No partial credit for this question)

1.4 (d) perceptron 4 / 5

- 0 pts Correct

✓ **- 1 pts Minor mistakes (e.g., omit w is initialized as 0)**

- 2 pts small mistake

- 5 pts Incorrect

1.5 (e) kernel 1 / 5

✓ **- 0 pts The answer is Yes and provides the correct explanation: need to argue that at any step, w = \eta \sum_i \alpha_i y_i, x_i, where \alpha_i=1 iff the model made mistake on instance i in previous steps. Changing learning rate \eta to 2 \eta (or \eta+2), we get w =(2 \eta) \sum_i \alpha_i y_i, x_i, as \eta>0, the direction of w is the same but only the size is different. Therefore, sgn(w^Tx) remains the same.**

- 2 pts minor mistake or incomplete proof. Note that if \eta is not a fixed value, the perception may make mistakes on different points than x_3, x_5, x_7. Therefore, only saying the updated points do not depend on eta value is not correct.

✓ **- 4 pts Answer is No, but provide some reasonable explianation**

- 5 pts incorrect

1.6 (f) support vector 6 / 6

✓ **- 0 pts** Correct

- 2 pts a minor mistake in explanation or answer (e.g., missing sgn() function)

- 4 pts Incorrect answer but mention the b term dominates w^T\phi(x)+b

- 6 pts Incorrect

- 0 pts Click here to replace this description.

1.7 (g) learning theory 2 / 3

- 0 pts Correct

✓ **- 1 pts One wrong answer**

- 2 pts two wrong answers

- 3 pts Incorrect or unrecognizable

QUESTION 2

KNN 21 pts

2.1 (a) distance 0 / 4

- 0 pts Correct

- 1 pts minor mistake (e.g., forgot square root), represent kernel as \phi function, sqrt(K(x,x)+ K(y,y)+2K(x,y)) instead

✓ **- 4 pts Incorrect**

2.2 (b) a 3 / 3

✓ **- 0 pts** Correct

- 3 pts incorrect (no partial credit)

2.3 (c) linear model 6 / 6

✓ **- 0 pts** Correct

- 1 pts one mistake

- 2 pts two mistakes

- 3 pts 3 mistakes

- 6 pts Too many mistakes or incorrect one unselected

2.4 (d) kernel 2 / 8

- 0 pts Correct

- 2 pts minor mistake (e.g., small computation error, wrong coefficients, forgot the constant term)

- 4 pts major mistake (e.g,. some incorrect terms)

✓ **- 8 pts incorrect**

+ 2 Point adjustment

QUESTION 3

Naive Bayes 26 pts

3.1 (a) MLE 3 / 3

✓ **- 0 pts** Correct

- 3 pts Incorrect (no partial credit)

3.2 (b) MAP 5 / 5

✓ **- 0 pts** Correct

- 3 pts Partial correct (mention maximizing 3a(1-a)^2)

- 5 pts Incorrect

3.3 (c) M-Step 5 / 5

✓ **- 0 pts All Correct**

- 1 pts One answer incorrect

- 2 pts Two answers incorrect

- 3 pts three answers incorrect

- 4 pts four answers incorrect

- 5 pts Unattempted / All answers incorrect

3.4 (d) E-step 6 / 8

- 0 pts Correct

✓ **- 2 pts One answer incorrect or minor mistake**

- 4 pts two answers incoorect or major mistake

- 6 pts three answers incorrect

- 8 pts Unattempted / Incorrect answers

3.5 (e) M-step 2 5 / 5

✓ **- 0 pts** Correct

- 1 pts one wrong answers

- 2 pts two wrong answers (or a minor error leads to multiple wrong answers)

- 3 pts three wrong answers

- 4 pts four wrong answers

- 5 pts Unattempted/incorrect

QUESTION 4

GMM 25 pts

4.1 (a) definition 3 / 5

- 0 pts Correct

✓ **- 2 pts Partially derive the formula with **only one** of the following mistakes: missed one term; mistake on specifications for normal dist; did not use one of the given variables (e.g. not inferring the value of $$\omega_1$$ and $$\omega_2$$); one incorrect coefficient; incorrect operator...**

- 5 pts Incorrect.

May have multiple mistakes including: not using any of the provided variables but only listing a common (and partly wrong) formula; using unclear/incorrect variables like $$N(x_n|\Theta)$$, $$N(x_n|\mu_n,\sigma^2)$$; unclear/incorrect summations $$\sum_{k=1}^{k}$$, $$\sum_{n=1}^N$$; not inferring the number of terms and the value of $$\omega_k$$; did not derive the probability, ...

e.g. $$\sum \omega N(\cdot)$$, the formula is not clear on what it sums on and does not consider any of the provided variables.

4.2 (b) conditional probabiility 5 / 5

✓ **- 0 pts Correct.**

Full credit is also given for having the correct nominator and stating/having the denominator comes from part (a), regardless of whether (a) is correct.

- 2 pts partially derive the formula: for example, incorrect nominator, incorrect denominator. missed one term in denominator, wrong specifications for normal dist, part of the terms not derived, incorrect sum $$\sum_n$$...

- 5 pts incorrect

4.3 (c) EM definition **2 / 4**

- 0 pts Correct

✓ **- 2 pts partially derive the formula: for example, missed one term; not using normal dist to simplify, ...**

- 4 pts incorrect

4.4 (d) M-step definition. **5 / 5**

✓ **- 0 pts A and C selected**

- 2 pts Either A or C;
A and C and either B or D

- 5 pts All selected;
none of A or C is selected

4.5 (e) M-Step **0 / 6**

- 0 pts Correct

- 2 pts one value is incorrect

- 4 pts two values are incorrect

✓ **- 6 pts All incorrect**

CM146: Introduction to Machine Learning **Fall 2022**

Final Exam

Dec 9th, 2022

- **Read the instructions below prior to starting the exam!**
- This is a close book exam, but a letter/A4 size cheat sheet is allowed. Please do not access any other material during the exam.
- Please write your answers clear.
- Please double check your answers. We might not be able to give partial credits for some questions.
- This exam booklet contains **four** problems.

Good Luck!

Name and ID: Charles Zhang 305-413-659

1 Short Question[28 pts]

(a) **(3 pts)** Given a training data set $\{\mathbf{x}_i, y_i\}_{i=1}^N, \mathbf{x}_i \in \mathbb{R}^2, y_i = \{1, -1\}$, soft SVM identifies a hyper-plane $\mathbf{w}^T\mathbf{x} + b = 0$ by solving the following optimization problem.

$$\min_{w,b,\xi_i} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C \cdot \sum_{i=1}^{N} \xi_i$$

$$s.t \ \forall i, \ y(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0 \qquad 4 \leq \xi_i$$

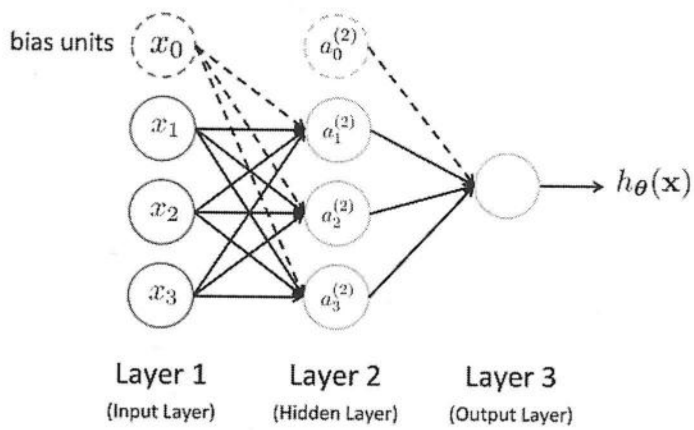
Let $\mathbf{w} = [0, 2]$ and $b = 1$ be the solution of the above optimization problem. What is the value of the slack variable ξ for a *negative* training data point $\mathbf{x} = (0, 1)$?

$\xi =$ 4

$-1(0 \cdot 0 + 2 + 1) \geq 1 - \xi_i$

$-3 \geq 1 - \xi_i \qquad -4 \geq -\xi_i$

(b) **(3 pts)** Consider a multi-class classification problem with 4 classes and 3 features. We use the following neural network to build binary classifiers, where x_0 is the bias term.



$\binom{10}{2}$ $\quad \frac{10!}{8!2!} = \frac{90}{2} = 45$

What is the total number of parameters when using *one-vs-one* strategies for classification?

$\frac{4!}{2!2!} = \frac{24}{4} = 6$

number of parameters = 24

$\binom{K}{2}$ models $\qquad \binom{4}{2}$ models $\qquad \binom{4}{2} \times 4$

~~4 features/model~~ $\qquad 6 \times 4 = 24$

4 params/model =

(c) **(3 pts)** Follow the previous question, what is the total number of parameters when using *one-against-all* strategies for classification?

K models

number of parameters = 16

$4K = 16$

(d) **(5pt pts)** Consider training a Perceptron model ($y = \mathbf{w}^\top \mathbf{x}$, $\mathbf{w} \in \mathbb{R}^d$) with a **learning rate** η on a dataset $D = (\mathbf{x}_i, y_i), i = 1 \ldots 10$.

Algorithm 1 Perceptron with learning rate α

Initialize $\mathbf{w} = \mathbf{0}$
for i = 1 ... 10 **do**
  if $y_i \neq sgn(\mathbf{w}^\top \mathbf{x}_i)$ **then**
    $\mathbf{w} \leftarrow \mathbf{w} + \eta y_i \mathbf{x}_i$
  end if
end for
return w

If the model makes mistakes exactly on data points $(\mathbf{x}_3, y_3), (\mathbf{x}_5, y_5), (\mathbf{x}_7, y_7)$ during training, write down $\mathbf{w}$ in terms of $\mathbf{x}_i, y_i$ and η.

$\mathbf{w} =$ $w + \sum_{i=1}^{10} \eta \alpha_i y_i x_i = w + \eta y_3 x_3 + \eta y_5 x_5 + \eta y_7 x_7$

(e) **(5pt pts)** Follow the previous question, if we increase the learning rate η by 2, will the model still only update on $(\mathbf{x}_3, y_3), (\mathbf{x}_5, y_5), (\mathbf{x}_7, y_7)$ during training? Select your answer by marking a cross in the box ☒, and then explain your answer.

☐ Yes ☒ No

Explanation: This is not guaranteed. It is possible that, since the amount w is updated by is now greater, the update on (x_3, y_3) caused the model to overcorrect, making it predict incorrectly and update on (x_4, y_4). This is one of many examples of things that may change the points the model updates on.

(f) **(6pt pts)** Consider training a SVM model with RBF kernel function $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2}\underbrace{||\mathbf{x}_i - \mathbf{x}_j||^2}_{>>0}\right)$. As we learned in class, the trained model is given by $\text{sgn}(h(\boldsymbol{x}; \boldsymbol{\alpha}, b))$ where

$$\mathbf{w}^T\phi(\mathbf{x}) + b = \sum_{i \in SV} y_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b = h(\mathbf{x}; \boldsymbol{\alpha}, b). \qquad e^{-L} \to 0$$

SV is the set of support vectors and α_i is the corresponding coefficient. sgn is a sign function that returns the sign of a real number. Assume that there is a test point $\boldsymbol{x}_{far}$ that is far away from any training point $\boldsymbol{x}_i$ in the original space $\mathbb{R}^d$ (i.e., $||\mathbf{x}_{far} - \mathbf{x}_i|| \gg 0$), what is the prediction of the SVM model at this data point? Briefly prove your answer.

Prediction = $\underline{\text{sgn}(b)}$

Explanation: $||x_{far} - x_i|| >> 0 \Rightarrow \exp(-\frac{1}{2}(\infty)) = \exp(-\infty) = 0$

$\therefore k(x_{far}, x_i) = 0 = k(x_i, x_{far})$

$\text{sgn}(h(x;\alpha, b)) = \text{sgn}(\sum_{i \in SV}' 0 + b) = \text{sgn}(\sum_{i \in SV}' b) = \boxed{\text{sgn}(b)}$

(g) **(3 pts)** In the lecture, we derive the sample complexity of the monotone conjunction concept with n-dimensional Boolean variables is:

$$m > \frac{n}{\epsilon}\left(\log(n) + \log(1/\delta)\right)$$

.

Which of the following statement(s) is/are true?

☒ Given $\delta = 0.05$ and $n = 10$, to reduce the error rate from 10% to 5%, we will need more training examples.

☐ Given $\delta = 0.05$, if we increase the number of variables from 10 to 100, we will need more training examples to achieve the same error rate.

☐ ~~ϵ refers to the training error~~.

~~$x^T y \to (x_1^2 y_1^2) * (x_2 y_2)$~~ (x, y) $\sqrt{(a-x)^2+(b-y)^2}$

x_1^2 $x_1(-y_1) + x_2(-y_2)$ (a, b)

2 K-NN with a polynomial kernel [21 pts]

$\sqrt{\phi(x)}$

Consider a polynomial kernel $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\top \mathbf{y} + c)^n$, where $c \in \mathbb{R}$ is a real number and $n \in \mathbb{N}$ is a positive integer. As we learned in class, $K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y})$, and $\phi(\mathbf{x})$ is a function that maps the input vector $\mathbf{x}$ into a higher dimensional space.

In the following, we consider a K-NN model with Euclidean distance $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$.

x^2y^2

(a) **(4 pts)** Write the Euclidean distance $d(\phi(\mathbf{x}), \phi(\mathbf{y})) = \|\phi(\mathbf{x}) - \phi(\mathbf{y})\|_2$ with mapping function ϕ using the corresponding kernel function $K(\mathbf{x}, \mathbf{y})$.

$\|\phi(x) - \phi(y)\|_2 = \sqrt{\quad}$

$d(\phi(\mathbf{x}), \phi(\mathbf{y})) =$ $\|(x + \ldots + \sqrt[n]{c})^n - (y + \ldots + \sqrt[n]{c})^n\|_2$

$K(x, y) = \phi(x)\phi(y)$

(b) **(3 pts)** If $c = 0$, at what value of n for the polynomial kernel will we have $d(\phi(\mathbf{x}), \phi(\mathbf{y})) = \|\mathbf{x} - \mathbf{y}\|_2$?

$n =$ 1

.

(c) **(6 pts)** Which of the following models are linear models (i.e., even with parameter tuning, they cannot achieve 0 training error if the data are not linearly separable). Select all that apply by marking a cross in the box ☒.

- ☐ 1-NN with linear kernel.
- ☒ SVM with linear kernel.
- ☐ 1-NN with polynomial kernel.
- ☐ 3-NN with linear kernel.
- ☐ SVM with polynomial kernel.
- ☐ 3-NN with polynomial kernel.

(d) **(8 pts)** Let n=2, c = 16, and $\mathbf{x} \in \mathbb{R}^2$, and $\mathbf{y} = [y_1, y_2] \in \mathbb{R}^2$, what is the corresponding feature map $\phi(\mathbf{x})$ for the kernel $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\top \mathbf{y} + 16)^2$?

x^n

$K(x,y) = x^2y^2 + 32xy + 256$

$$\boxed{\phi(x) = \begin{bmatrix} x^2 \\ \sqrt{32}\,x \\ 16 \end{bmatrix}}$$

c

$\sqrt{\phi(x)},$

$x^2 \; (x_1, y_1), (x_2, y_2)$ x^2y^2

$d(x,y) = \|x-y\|_2$ $K(x,y) = \phi(x)\phi(y) = (x^Ty + c)^n$ $(x^Ty+c)^n$

$\sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$ $\phi(x) = \frac{K(x,y)}{\phi(y)}$ $\phi(x) = \frac{(x^Ty+c)^n}{\phi(y)}$

$\sqrt{x_2^2 - 2x_1x_2 + x_1^2 + y_2^2 - 2y_1y_2 + y_1^2}$

 $\phi(x) =$

x_1^2

3 EM with Naive Bayes [26 pts]

Consider the following binary classification dataset with 2 binary features.

$h_{MAP} =$ likelihood × prior

$3(1-a)^2$

$3a(1-a)^2$

X_1	X_2	Y
0	0	0
0	1	0
1	0	1
1	1	0

$P(X_1 | Y) = 1$

(a) **(3 pts)** Let the parameter $a = P(X_1 = 1|Y = 1)$ be one of the parameters of Naive Bayes. Based on the above data, what is the value of a estimated by MLE?

$a =$ 1

(b) **(5 pts)** Follow the previous question (a). If the parameter a follows a prior distribution $P(a) = 3(1-a)^2$, based on the data in Table 1, what is the value of a using MAP?

$a =$ 1/3

$P(X_1 = 1 | Y = 1) P(a)$

$a P(a)$

$3a(1-a)^2$

$-6a(1-a) + 3(1-a)^2 = 0$

$-6a + 3 - 3a = 0$

$a = \frac{1}{3}$

(c) **(5 pts)** Next, we consider apply Naive Bayes in an unsupervised learning setting (i.e., label Y is not given). In this case, we will use Expectation-Maximization (EM) to learn the model parameters for Naive Bayes. Assume that in the initial step we randomly assign the label Y to the training instances as shown in Table 1. Write down the value of all the parameters of Naive Bayes based on MLE:

$P(Y = 1) =$ $\frac{1}{4}$; $P(X_1 = 1|Y = 1) =$ 1; $P(X_1 = 1|Y = 0) =$ $\frac{1}{3}$.

$P(X_2 = 1|Y = 1) =$ 0; $P(X_2 = 1|Y = 0) =$ $\frac{2}{3}$.

$$P(Y=1|\cancel{x_1 x_2},\theta) = \frac{P(\cancel{x_1 x_2},\theta|Y=1)\,P(Y=1)}{P(\cancel{x_1 x_2},\theta)}$$

(d) **(8 pts)** Let Θ be the set of parameter you evaluated in (c). Based on them reassign the label distribution to the four points. Write down your answer in the following table.

$$\frac{P(x_1=0|Y=1)P(x_2=0|Y=1)P(Y=1)}{P(}$$

X_1	X_2	$P(Y=1\|X_1,X_2;\Theta)$
0	0	0
0	1	0
1	0	1
1	1	0

$$P(Y|x_1,x_2) = \frac{P(x_1,x_2|Y)P(Y)}{P(x_1,x_2)} = \frac{P(x_1|Y)P(x_2|Y)P(Y)}{P(x_1|Y)P(x_2|Y)}$$

(e) **(5 pts)** Assume after several EM-steps, we obtain the label distribution for these 4 examples as shown in the following table. What are the model parameters after performing M-Step on these 4 examples (round up your answer to 2 decimal places)?

X_1	X_2	$P(Y=1\|X_1,X_2;\Theta)$
0	0	1
0	1	0.5
1	0	0
1	1	0

$0 + \frac{0}{2}$
$0.5 + \frac{1}{2}$
$1 + \frac{2}{2}$
$1 + \frac{2}{2} = \frac{5}{8}$

$P(Y=1) =$ 0.375; $P(X_1=1|Y=1) =$ 0; $P(X_1=1|Y=0) =$ 0.8.

$P(X_2=1|Y=1) =$ 0.33; $P(X_2=1|Y=0) =$ 0.6.

$$P(Y=1) = \frac{1.5}{4} = \frac{3}{8}$$

$$P(x_1=1|Y=1) = \frac{P(Y=1|x_1=1)P(x_1=1)}{P(Y=1)} = 0$$

$$P(x_1=1|Y=0) = \frac{P(Y=0|x_1=1)P(x_1=1)}{P(Y=0)} = \frac{1(0.5)}{0.625} = \frac{\frac{1}{2}}{\frac{5}{8}} = \frac{\frac{4}{8}}{\frac{5}{8}} = 0.8$$

$$P(x_2=1|Y=1) = \frac{P(Y=1|x_2=1)P(x_2=1)}{P(Y=1)} = \frac{0.25(0.5)}{0.375} = 0.33$$

$$P(x_2=1|Y=0) = \frac{P(Y=0|x_2=1)P(x_2=1)}{P(Y=0)} = \frac{0.75(0.5)}{0.625} = 0.6$$

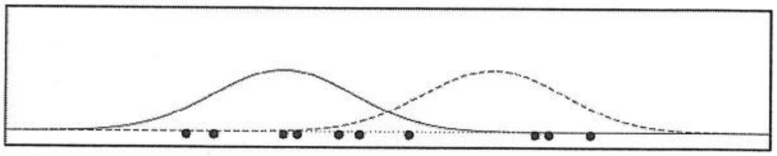
4 Gaussian Mixture Model [25 pts]

In this question, we will derive a simplified version of GMM. We assume that the data set consists of N one dimensional data points $\{x_n\}_{n=1}^{N}, x_n \in \mathbb{R}$. Our goal is to cluster the data points into 2 groups (denoted as $z_n = 1$, and $z_n = 2$). We model the likelihood $P(x_n|z_n)$ using 2 unit-variance Gaussian distribution: $\mathcal{N}(x_n; \mu_1, 1)$ and $\mathcal{N}(x_n; \mu_2, 1)$, where μ_1 and μ_2 are the cluster centers of the cluster 1 and 2, respectively. The probability density function for the Gaussian distribution:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

We assume the prior distribution $P(z_n = 1) = \omega$. We use $\Theta = \{\mu_1, \mu_2, \omega\}$ to represent the collection of all the model parameters, and $\Theta^{(t)}$ represents the parameters at step t.

The following figure illustrates the 1-dimensional GMM.



(a) **(5 pts)** Based on the GMM assumptions described above, what is $P(x_n; \Theta)$ (i.e., $P(x_n)$ based on the GMM assumptions and parameters) Hint: You can write your answer in terms of ω and the normal distribution $\mathcal{N}(.)$.

$\cancel{P(x_n;\Theta)} = P(x_n|z_n)P(z_n)$

$P(x_n; \Theta) = \sum_{n=1}^{N} N(x_n|\mu_1, 1)\omega + \sum_{n=1}^{N} N(x_n|\mu_2, 1)(1-\omega)$

(b) **(5 pts)** Assume at step t, we obtain $\Theta^{(t)} = \{\mu_1^{(t)}, \mu_2^{(t)}, \omega^{(t)}\}$, what is $P(z_n = 1|x_n; \Theta^{(t)})$? Hint: You can write your answer in terms of ω and the normal distribution $\mathcal{N}(.)$.

$$P(z_n = 1|x_n; \Theta^{(t)}) = \frac{N(x_n|\mu_1^{(t)}, 1)\,\omega^{(t)}}{\sum_{n=1}^{N} N(x_n|\mu_1^{(t)}, 1)\,\omega^{(t)} + \sum_{n=1}^{N} N(x_n|\mu_2^{(t)}, 1)(1-\omega^{(t)})}$$

$$P(z_n=1|x_n;\theta^{(t)}) = \frac{P(x_n|z_n;\theta^{(t)})\,P(z_n)}{P(x_n)}$$

~~$\log w + \log \frac{1}{\sqrt{2\pi}} - \frac{(x_n - \mu_1)^2}{2}$~~ $\sum_{n=1}^{N}$

$\log P(z_n = k | x_n) + \log P(x_n)$

(c) **(4 pts)** Recall in the EM algorithm, the M-step maximizing the following function:

$$\max_{\Theta} \sum_{n} \sum_{k=1,2} P(z_n = k | x_n; \Theta^{(t)}) \log P(z_n = k, x_n; \Theta) \tag{1}$$

What is $\log P(z_n = 1, x_n; \Theta)$? Simplify your answer using $\mathcal{N}(x|\mu, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right)$

$\log P(z_n = 1, x_n; \Theta) =$ $\log w + \log \frac{1}{\sqrt{2\pi}} - \exp\left(-\frac{(x-\mu)^2}{2}\right)$

$= \log P(z_n = 1 | x_n) + \log P(x_n)$?

(d) **(5 pts)** Let $\gamma_{nk} = P(z_n = k | x_n; \Theta^{(t)})$. Which of the following optimization problems are equivalent to Eq. (1)? Select all that apply by marking a cross in the box ☒.

- ☒ $\max_\Theta \sum_n \sum_{k=1,2} \gamma_{nk} \log P(z_n = k, x_n; \Theta)$
- ☐ $\max_\Theta \sum_n \left[\gamma_{n1} \log w + \frac{\gamma_{n1}(x-\mu_1)^2}{2} + \gamma_{n2} \log(1-w) + \frac{\gamma_{n2}(x-\mu_2)^2}{2}\right]$
- ☒ $\max_\Theta \sum_n \left[\gamma_{n1} \log w - \frac{\gamma_{n1}(x-\mu_1)^2}{2} + \gamma_{n2} \log(1-w) - \frac{\gamma_{n2}(x-\mu_2)^2}{2}\right]$
- ☐ $\max_\Theta \sum_n \left[\gamma_{n1} w - \frac{\gamma_{n1}(x-\mu_1)^2}{2} + \gamma_{n2}(1-w) - \frac{\gamma_{n2}(x-\mu_2)^2}{2}\right]$

(e) **(6 pts)** Assume we have the following 4 data points, and after step t, the corresponding γ_{nk} are listed in the following.

γ

x_n	$\gamma_{n1} = P(z_n = 1 \| x_n; \Theta^{(t)})$
-1	0.8
0	0.6
1	0.4
2	0.2

$\frac{\partial}{\partial \mu_1} = \sum_{n}^{N} \gamma_{n1}(x_n - \mu_1)$

$0.8(-1-\mu_1) + 0.6(-\mu_1) + 0.4(1-\mu_1) + 0.2(2-\mu_1)$

$-0.8 - 2\mu_1 + 0.4 + 0.2 = 0$

$\mu_1 = -0.1$

$\frac{\partial}{\partial \mu_2} = \sum^{N} \gamma_{n2}(x_n - \mu_2)$

$0.2(-1-\mu_2) + 0.4(-\mu_2) + 0.6(1-\mu_2) + 0.8(2-\mu_2)$

$-2\mu_2 + 1.2 = 0$

$\mu_2 = \frac{1.2}{2} = \frac{3}{5}$

What are the ω, μ_1, μ_2 based on solving Eq. (1)?

$\omega =$ 1 ; $\mu_1 =$ -0.1 ; $\mu_2 =$ 0.6

$\frac{\partial}{\partial w} = \sum^{N} \frac{\gamma_{n1}}{w} - \frac{\gamma_{n2}}{1-w} = 0$

$0.8 \log w - \frac{0.8(-1-\mu_1)^2}{2} + 0.2 \log(1-w) - \frac{0.2(-1-\mu_2)^2}{2}$

$\frac{2}{w} - \frac{2}{1-w} = 0$

$w = 1$

$0.8 \log w + 0.2 \log(1-w)$

$\frac{0.8}{w} - \frac{0.2}{1-w} = 0$

0.8

$\frac{0.8 + 0.6 + 0.4 + 0.2}{2} = 1$

$\frac{\gamma_{n1}}{w} - \frac{\gamma_{n2}}{1-w} = 0$

$\frac{2}{w} - \frac{2}{1-w} = 0$