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:

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.

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

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

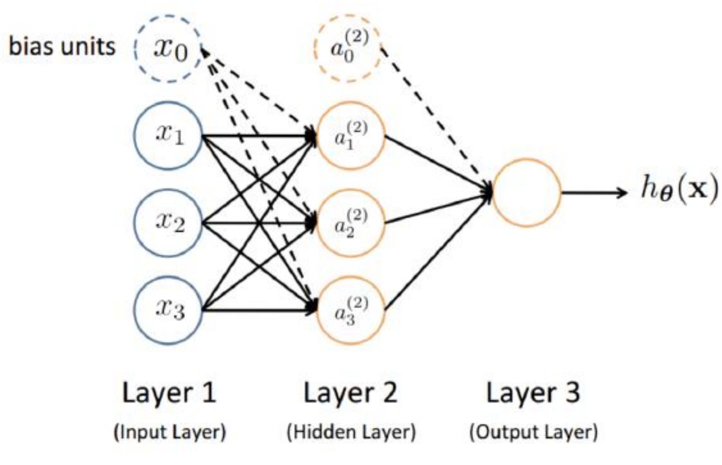
$$\xi_i \geq 0$$

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 $\mathbf{x} = (0, 1)$?

$\xi =$ _________

Solution: The corresponding ξ needs to satisfy the following constraint
$-1(0 + 2 + 1) \geq 1 - \xi$
$\Rightarrow \xi \geq 4$
Therefore, in the optimal solution $\xi = 4$.

(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.



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

Solution: For 1 vs 1 approach, we need 6 binary neural networks for 4 classes. Each binary neural network has 16 parameters. Therefore, the answer is 6*16=96.

number of parameters = _________

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

number of parameters = ________

Solution: For one against all approach, we need 4 binary neural networks for 4 classes. Each binary neural network has 16 parameters. Therefore, the answer is 4*16=64

(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 $D = (\mathbf{x}_i, y_i), i = 1 \dots 10$.

Algorithm 1 Perceptron with learning rate α

```
Initialize w = 0
for i = 1 ... 10 do
    if y_i ≠ sgn(w^⊤ x_i) then
        w ← w + η y_i 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.

$\mathbf{w}$ = ________________

Solution: $\eta(y_3\mathbf{x}_3 + y_5\mathbf{x}_5 + y_7\mathbf{x}_7)$

(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:

Solution: Yes. Based on the observation in 1 (d), at any step, $w = \eta \sum_i \alpha_i y_i x_i$, where $\alpha_i = 1$ if and only if the model made mistake on instance i in previous steps. Changing learning rate η to 2η, 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, $sgn(w^T x)$ remains the same. As the prediction at every step is the same, the model updates on the same set of instances.

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

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

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 = ____________________

Explanation:

Solution: Ans: sgn(b)

$$\begin{aligned} & ||\boldsymbol{x}_{far} - \boldsymbol{x}_i|| \gg 0 \;\forall\; i \in SV \\ \implies\; & k(\boldsymbol{x}_{far}, \boldsymbol{x}_i) \approx 0 \;\forall\; i \in SV \\ \implies\; & \sum_{i \in SV} y_i \alpha_i K(\boldsymbol{x}_{far}, \boldsymbol{x}_j) \approx 0 \\ \implies\; & h(\boldsymbol{x}_{far}; \alpha, b) \approx b \end{aligned}$$

(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.

Solution: A,B

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

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 $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$.

(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 $K(\mathbf{x}, \mathbf{y})$.

$d(\phi(\mathbf{x}), \phi(\mathbf{y})) =$ ____________________________________

Solution: $d(\phi(\mathbf{x}), \phi(\mathbf{y})) = \sqrt{\phi(x)^T \phi(x) - 2\phi(x)^T \phi(y) + \phi(y)^T \phi(y)} = \sqrt{K(\mathbf{x}, \mathbf{x}) - 2K(\mathbf{x}, \mathbf{y}) + K(\mathbf{y}, \mathbf{y})}$

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

$n =$ _________

Solution: 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 $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\top \mathbf{y} + 16)^2$?

Solution: $\phi(x) = [16, 4\sqrt{2}x_1, 4\sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2]^T$

3 EM with Naive Bayes [26 pts]

Consider the following binary classification dataset with 2 binary features.

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

(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?

$a =$ ________________

Solution: 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?

$a =$ ________________

Solution: Maximize $3a(1-a)^2$, we get $a = 1/3$.

(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:

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

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

Solution: Based on the MLE of Naive Bayes, we get

$P(Y = 1) = 0.25, P(X_1 = 1|Y = 1) = 1, P(X_1 = 1|Y = 0) = 1/3, P(X_2 = 1|Y = 1) = 0, P(X_2 = 1|Y = 0) = 2/3.$

(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.

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

Solution: Based on the posterior probability, we get 0, 0, 3/4, 0

For example, for the case $X_1 = 1, X_2 = 0$,
$P(X_1 = 1, X_2 = 0, Y = 1) = 0.25 \times 1 \times 1$
$P(X_1 = 1, X_2 = 0, Y = 0) = 0.75 \times 1/3 \times 1/3$
Therefore, $P(Y = 1|X_1 = 1, X_2 = 0 = 0.25/(0.25{+}0.75/9) = 0.75$

(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

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

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

Solution: $P(Y = 1) = 1.5/4 = 0.38$ (or $3/8$ or 0.37), $P(X_1 = 1|Y = 1) = 0, P(X_1 = 1|Y = 0) = 2/2.5 = 0.8, P(X_2 = 1|Y = 1) = 0.5/1.5 = 0.33, P(X_2 = 1|Y = 0) = 1.5/2.5 = 0.6.$

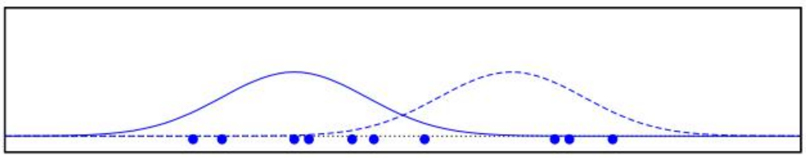
4 Gaussian Mixture Model [25 pts]

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 $\mathcal{N}(.)$.

$P(x_n; \Theta) =$ ____

Solution: $\omega\mathcal{N}(x_n; \mu_1, 1) + (1-\omega)\mathcal{N}(x_n; \mu_2, 1)$

(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 $\mathcal{N}(.)$.

$P(z_n = 1|x_n; \Theta^{(t)}) =$ ____

Solution: $\frac{\omega\mathcal{N}(x_n;\mu_1,1)}{\omega\mathcal{N}(x_n;\mu_1,1)+(1-\omega)\mathcal{N}(x_n;\mu_2,1)}$

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

$\log P(z_n = 1, x_n; \Theta) =$ ____

Solution: $\log P(z_n = 1, x_n; \Theta) = \log \omega - \log \sqrt{2\pi} - \frac{(x_n - \mu_1)^2}{2}$

(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 ☒.

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

Solution: A,C

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

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

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

$\omega =$ ________________; $\mu_1 =$ ________________; $\mu_2 =$ ________________

Solution: $\omega = 0.5, \mu_1 = 0, \mu_2 = 1$