CM146: Introduction to Machine Learning **Fall 2021**

Final Exam

Dec 6, 2021

- **Read the instructions below prior to starting the exam!**
- This is an open book exam – you can use course materials posted at CCLE as your reference. Please do not access any other material during the exam. Discussion is strictly prohibited.
- This exam booklet contains **six** problems.
- You have 180 minutes to earn a total of 100 points. We also provide 60 minutes grace period for uploading the answers. In total, you must need to finish the exam in 240 minutes.
- You can ask *private* question in Piazza, but only clarification questions will be answered. If there is any issue of the exam, we will make announcement before before Friday 12/6 11:00aM. If you have a doubt about the question, feel free to write down your explanations. The grading will only based on the solutions written in the exam booklet.
- If you type the answers, it is okay to use latex convention, e.g., x_1, \gamma
- **The deadline to upload your exam to Gradescope is *Dec 7 11:00am.* No late submissions will be accepted, so make sure to submit well before the deadline!**

Good Luck!

Name and ID:

1 Short Answer and Multiple Choices [29 pts]

(a) **(4 pts)** Consider training a Perceptron model with data in the 2-dimensional feature space. Assume after training on t data points, the model is $y = 3x_1 - 2x_2 - 1$. The data $(t+1)$-th and $(t+2)$-th training examples are $((x_1, x_2), y) = ((1, 2), 1)$ and $((-1, -3), -1)$, respectively. Write down the updated model after training with these two data points: $y =$ _____.

Solution: $y = 4x_1$.

(b) **(4 pts)** Given a set of training data, we train a KNN model with $K = 7$. Which of the following statement(s) is/are true?

☐ (A) If we set $K = 3$, the KNN model is more likely to overfit the data than when $K = 7$.

☐ (B) If we set $K = 13$, the KNN model is more likely to overfit the data than when $K = 7$.

☐ (C) The decision boundary of the KNN model is always a linear function.

☐ (D) The hyper-parameter K can be selected by cross validation.

Solution: A, D
Explanation:
KNN with smaller K is more likely to overfit the data.
The decision boundary of KNN can be nonlinear.

(c) **(4 pts)** Which of the following statement(s) about hyper-parameter tuning is/are true?

☐ (A) N-fold cross-validation can be used to tune the hyper-parameters of a model.

☐ (B) We should tune the hyper-parameters on the test dataset.

☐ (C) Using a smaller N for N-fold cross validation always leads to a better estimation of model's performance.

☐ (D) If $|D|$ is the size of the data, when using N-fold cross-validation, N must need to be smaller than $|D|/2$.

Solution: A
Explanation:
C is wrong because larger N leads to a better estimation of model performance, while it takes more time to compute.
D is wrong because N can be as large as $|D|$.

(d) **(4 pts)** Which of the following statement(s) is/are true?

$$m \geq \frac{1}{\epsilon}\left(ln\left(|\mathcal{H}|\right) + \frac{1}{\delta}\right)$$

☐ (A) To achieve an error rate ϵ of 0.1 with probability 0.95 on a hypothesis class with 2^5 hypotheses, we must need $\frac{1}{0.1}\left(5\ln(2)+\frac{1}{0.05}\right)$ examples. Otherwise, the error rate will be higher.

☐ (B) ϵ refers to the training error.

☐ (C) If $\mathcal{H}$ is the class of linear function in n dimensional space, the size of $|\mathcal{H}|$ is 2^{2^n}.

☐ (D) According to the inequality, given $\mathcal{H}$ and δ, with more training data, the error rate ϵ we can guarantee is smaller.

Solution: D

Explanation:
A is wrong because the theorem only suggests that when the training data size is larger than $\frac{1}{0.1}\left(5\ln(2)+\frac{1}{0.05}\right)$, the model is guaranteed to achieve error rate less than 0.1 with probability 0.95. It doesn't mean that we "must" need to have these many examples to achieve the error rate of 0.1.
ϵ represents the error in the underlying distribution (i.e., generalization error or test error)
The size of the linear function hypothesis class is infinity.

(e) **(4 pts)** Suppose we have a dataset, which consists of four points with their class labels on $\mathbb{R}^2$: $((x_1,x_2),y)=((1,1),-1),((2,2),-1),((4,4),1)$, and $((5,5),1)$. If we train a linear hard-SVM in the form of $w_1x_1+w_2x_2+b=0$ on this dataset, what would be the optimal weight vector **w** and bias **b**?

$w_1 =$ blank #1
$w_2 =$ blank #1
$b =$ blank #2

Solution: $w_1 = 0.5$
$w_2 = 0.5$
b = -3

Explanation:
The points ((2,2),-1) and ((4,4), 1) are support vectors. Therefore, we can find the solution by

$$\begin{aligned} \min_{w_1,w_2} \quad & \frac{1}{2}(w_1^2+w_2^2) \\ s.t. \quad & 2w_1+2w_2+b=-1 \\ & 4w_1+4w_2+b=1 \end{aligned} \tag{1}$$

From the constraints, we get $w_1+w_2=1, b=-3$. Plug in $w_2=1-w_1$ to $\frac{1}{2}(w_1^2+w_2^2)$, we can find $w_1=0.5$ is the minimum and $w_2=0.5$.

(f) **(5 pts)** Suppose $\boldsymbol{x}=[x_1,x_2]$ and $\boldsymbol{z}=[z_1,z_2]$. Show that $K(\boldsymbol{x},\boldsymbol{z})=\left(\boldsymbol{x}^T\boldsymbol{z}\right)^3$ is a valid kernel by constructing the corresponding $\Phi(\cdot)$.

$\Phi(x) =$ _____

Solution: $\Phi(\boldsymbol{x}) = \left(x_1^3, \sqrt{3}x_1^2x_2, \sqrt{3}x_1x_2^2, x_2^3\right)$

(g) **(5 pts)** Which of the following statement(s) is/are true?

☐ (A) Using a different initialization may lead to a different clustering in K-means.

☐ (B) Using a different initialization may lead to a different clustering in K-medoids.

☐ (C) Compared with K-medoids, the K-means algorithm is more sensitive to outliers.

☐ (D) K-means algorithm is designed for training a multi-class classification.

☐ (E) It is impossible for the K-means algorithm to find the global optimal solution.

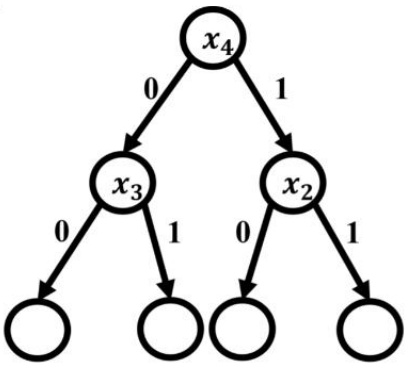
Solution: ABC.

D. K-means is an algorithm for clustering not for classification.

E. It's possible but not guaranteed.

#	x_1	x_2	x_3	x_4	y
1	0	0	0	0	0
2	0	0	0	1	1
3	0	0	1	1	1
4	0	1	0	0	0
5	1	0	0	0	0
6	1	0	1	1	1
7	1	1	0	0	1
8	1	1	1	0	0
9	1	1	1	1	0

(a) Data for Decision Tree



(b) A Decision Tree Example

Table 1: Decision Tree

2 Decision Tree [12 pts]

Consider the following data points listed in Table 1a

(a) **(3 pts)** Does the decision tree in Fig. 1b achieve 0 training error?

Solution: No.

Explanation:
When $x_3 = x_4 = 0$, some data points have label 0 and some have 1. Therefore, the training error cannot be 0.

(b) **(3 pts)** Without any split, the entropy $H(y) =$ _____ blank #1 _____ (please answer with a real number rounded to 2 decimal places in the format of X.XX).

(c) **(6 pts)** If we can only split the data based on one attribute, _____ blank #2 _____ gives the greatest information gain. The information gain is _____ blank #3 _____ (please answer with a real number rounded to 2 decimal places in the format of X.XX).

Solution: blank #1: $H(y) = \frac{4}{9}\log_2\frac{9}{4} + \frac{5}{9}\log_2\frac{9}{5} = 0.99108 \approx 0.99$.
blank #2: x_4.
blank #3:

$$P(y=0|x_4=0) = \frac{4}{5},\ P(y=0|x_4=1) = \frac{1}{4}.$$

$$H(y|x_4) = \frac{5}{9}(\frac{1}{5}\log_2(5) + \frac{4}{5}\log_2(\frac{5}{4})) + \frac{4}{9}(\frac{1}{4}\log_2(4) + \frac{3}{4}\log_2(\frac{4}{3})) = 0.76164.$$

$$0.99108 - 0.76164 = 0.22944 \approx 0.23.$$

3 Bayesian Learning and EM Algorithm [20 pts]

We are testing a set of light bulbs from the same manufacturer, and each light bulb has the same probability p to pass the test.

(a) **(6 pts)** We test 5 bulbs and find that only the first 3 bulbs pass the test. What is the most likely value of p based on MLE? Complete the following derivation.

The likelihood function that describes the observations as a function of p is $L(p) =$ _____. Therefore, the log-likelihood is $\log L(p) =$ _____. Maximizing the log-likelihood, we obtain $p_{MLE} =$ _____ (write down a real number rounded to 2 decimal places in the format of X.XX).

Solution:

blank #1: $p^3(1-p)^2$.
blank #2: $3\log p + 2\log(1-p)$.
blank #3: 0.60.

(b) **(6 pts)** Now, we assume the probability density function of the prior distribution of p is $P(p) = 2p,\ p \in [0,1]$. If we test 5 bulbs and find that only the first 3 bulbs pass the test (represented as the observation D), what is the most likely value of p based on maximum-a-posteriori (MAP) estimation? Complete the following derivation.

The posterior $P(p|D)$ is proportional to _____ (write down as a function of p). Therefore, the MAP estimation of p is $p_{MAP} =$ _____ (write down a real number rounded to 2 decimal places in the format of X.XX).

Solution:

blank #1: $P(p|D) \propto P(D|p)P(p) = p^3(1-p)^2 \times 2p \propto p^4(1-p)^2$.
blank #2: $\frac{4}{4+2} \approx 0.67$.

(c) **(2 pts)** Let us consider another scenario. We assume that the light bulbs come from two different factories, A and B. With probability p_A, a light bulb from factory A passes the test and with probability p_B, a light bulb from factory B passes the test. Now, we receive two boxes of light bulbs : box 1 contains 3 light bulbs and box 2 contains 2 light bulbs. The light bulbs in the same box are from the same factory but we don't know which factory each box comes from. Also, we assume the prior distribution of where the box comes from is uniform.

If we observe that 2 out of 3 light bulbs from box 1 pass the test and 1 out of 2 light bulb from box 2 passes the test (represented as the observation D), we can estimate p_A and p_B based on EM. Complete the following derivation.

We use $Y \in \{AA, AB, BA, BB\}$ to represents the source of boxes. EM is an iterative process, we use $p_A^{(t)}$ and $p_B^{(t)}$ to represent the estimation of p_A and p_B at iteration t. Similarly, Y^t represents the estimation of the source of boxes at iteration t.

For example, $Y = AB$ means the first box is from factory A and the second is from factory B. We first initialize p_A and p_B with $p_A^{(0)} = 0.4$ and $p_B^{(0)} = 0.6$.

Given the observation D and $p_A^{(0)}$ and $p_A^{(0)}$, what is the posterior probability that both boxes are from factory A: $P(Y^{(1)} = AA|D, p_A^{(0)}, p_B^{(0)}) =$ _____ (write down a real number rounded to 2 decimal places in the format of X.XX).

Solution: 0.20

$$P(Y|D,p_A^{(0)},p_B^{(0)}) = \frac{P(D|Y,p_A^{(0)},p_B^{(0)})P(Y)}{P(D|p_A^{(0)},p_B^{(0)})}$$

Since Y has uniform prior, and denominator is independent from Y, we only consider $P(D|Y,p_A^{(0)},p_B^{(0)})$.

$$P(D|Y^{(1)}=AA,p_A^{(0)}=0.4,p_B^{(0)}=0.6) = 0.4^3\times(1-0.4)^2,$$

$$P(D|Y^{(1)}=AB,p_A^{(0)}=0.4,p_B^{(0)}=0.6) = 0.4^2\times(1-0.4)\times 0.6\times(1-0.6).$$

$$P(D|Y^{(1)}=BA,p_A^{(0)}=0.4,p_B^{(0)}=0.6) = 0.6^2\times(1-0.6)\times 0.4\times(1-0.4).$$

$$P(D|Y^{(1)}=BB,p_A^{(0)}=0.4,p_B^{(0)}=0.6) = 0.6^3\times(1-0.6)^2.$$

$$P(Y^{(1)}=AA|D,p_A^{(0)},p_B^{(0)}) = \frac{P(D|Y^{(1)}=AA,p_A^{(0)},p_B^{(0)})}{\sum_{Y'\in\{AA,AB,BA,BB\}}P(D|Y^{(1)}=Y',p_A^{(0)},p_B^{(0)})} = 0.20.$$

Similarly,

$$P(Y^{(1)}=AB|D,p_A^{(0)},p_B^{(0)}) = 0.20$$
$$P(Y^{(1)}=BA|D,p_A^{(0)},p_B^{(0)}) = 0.30$$
$$P(Y^{(1)}=BB|D,p_A^{(0)},p_B^{(0)}) = 0.30$$

(d) **(2 pts)** Similarly, what is the posterior probability that box 1 is from factory B and box 2 is form the factory A? $P(Y^{(1)}=BA|D,p_A^{(0)},p_B^{(0)}) =$ _____ (write down a real number rounded to 2 decimal places in the format of X.XX).

Solution: 0.30
please refer to 3.c.

(e) **(6 pts)** Follow the previous sub-question, based on the $Y^{(1)}$, what is the estimation of $p_A^{(1)}$ and $p_B^{(1)}$ at the first M-step:

$p_A^{(1)} =$ _____ (write down a real number rounded to 2 decimal places in the format of X.XX). $p_B^{(1)} =$ _____ (write down a real number rounded to 2 decimal places in the format of X.XX).

Solution:

$$\begin{aligned}p_A^{(1)},p_B^{(1)} =& \arg\max_{p_A,p_B}\sum_{Y'\in\{AA,AB,BA,BB\}}P(Y^{(1)}=Y'|D,p_A^{(0)},p_B^{(0)})\log P(D|Y^{(1)}=Y',p_A,p_B)\\ =& 1.3\log p_A + 0.9\log(1-p_A) + 1.7\log p_B + 1.1\log(1-p_B).\end{aligned}$$

$p_A^{(1)} = \frac{13}{22}$, $p_B^{(1)} = \frac{17}{28}$.
blank #1: $\frac{13}{22} = 0.59$
blank #2: $\frac{17}{28} = 0.61$

4 Clustering Algorithms [18 pts]

Recall that in the class, we show that K-means algorithm minimizes the average squared Euclidean distance of data points from their cluster prototypes:

$$\min_{r,\mu} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|_2^2$$

In the following, we will derive a clustering algorithm similar to K-Means but using Manhattan distance (L1-distance) instead of squared Euclidean distance as the distance metric. That is, the clustering algorithm minimizes:

$$\min_{r,\mu} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|_1$$

Consider the following five data points in 2-D space: $x_1 = (0, 0)$, $x_2 = (0, 2)$, $x_3 = (1, 0)$, $x_4 = (2, 2)$, $x_5 = (2, 3)$. Answer the following questions.

(a) **(2 pts)** What is the Manhattan distance between x_2 and x_5: _____?

Solution: 3.

(b) **(2 pts)** Assume that two cluster prototypes are $\mu_1 = (0, 1), \mu_2 = (3, 1)$. Which data points will be assigned to cluster 1 based on the Manhattan distance?

Solution: x_1, x_2, x_3.

(c) **(2 pts)** Assume that x_2, x_3, x_4 are assigned to cluster 1, find a point μ_1 that minimize $\sum_{n=2}^{4} \|x_n - \mu_1\|_1$. μ_1 = _____.

Solution: $(1, 2)$.

Explanation: $\sum_{n=2}^{4} \|x_n - \mu_1\|_1 = \sum_{n=2}^{4} |x_{n1} - \mu_{11}| + \sum_{n=2}^{4} |x_{n2} - \mu_{12}|$. Therefore, we can optimize each coordinate separately. To minimize the function $d(\mu_{11}) = \sum_{n=2}^{4} |x_{n1} - \mu_{11}|$, we can draw the function and observe that it is a piece-wise linear function. The minimum is the median of the three points.

(d) **(2 pts)** We will alternatively assign points to clusters (Step 1, 3, ...) and update cluster prototypes (Step 2, 4, ...). Given the initial cluster prototypes $\mu_1 = (0, 1), \mu_2 = (3, 1)$, complete the following table:

Initialization:	$\mu_1 = (0, 1)$	$\mu_2 = (3, 1)$
Step 1:	Cluster 1: _____	Cluster 2: _____
Step 2:	$\mu_1 =$ _____	$\mu_2 =$ _____
Step 3:	Cluster 1: _____	Cluster 2: _____
Step 4:	$\mu_1 =$ _____	$\mu_2 =$ _____

What we get in Step 2? **Solution:** $\mu_1 = (0, 0); \mu_2 = (2, y), y$ can be any number between 2 and 3.

(e) **(3 pts)** What is cluster 1 at Step 3 in Question 4.4?

Solution: x_1, x_2, x_3.

(f) **(3 pts)** What are the prototypes μ_1 and μ_2 in Question 4.4?

Solution: $(0, 0)$; $(2, y), y$ can be any number between 2 and 3.

5 Linear Classification [9 pts]

In the following, we consider training a linear model $\boldsymbol{w}^T\boldsymbol{x} + b$ on the data points in Figure 2. The x-coordinates and y-coordinates of all these points are in the range of $[-3, 3]$.

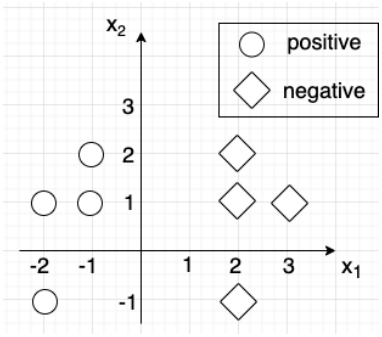


Figure 1: Training data distribution.

Answer the following questions based on the four separating hyper-planes showing in the figure below.

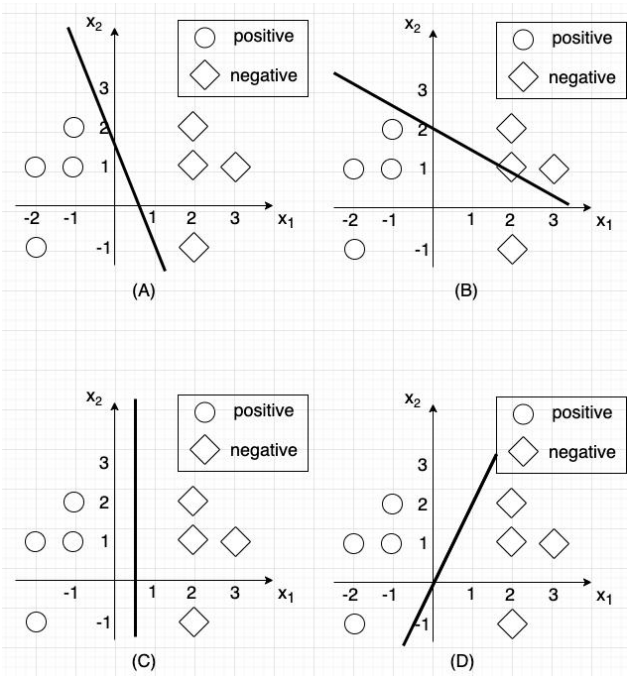


Figure 2: Four separating hyper-planes.

(a) **(3 pts)** Which of the separating hyperplane(s) are likely to be the solution(s) of a hard-SVM? Choose all possible candidates. **Solution:** C
Explanation: Hard SVM finds a linear separating hyper-plane with the maximal margin.

(b) **(3 pts)** Which of the separating hyperplane(s) are likely to be the solution(s) of a Perceptron? Choose all possible candidates. **Solution:** A,C,D
Explanation: Perceptorn finds a model that can separate the data.

(c) **(3 pts)** Consider a variant of SVM that learns a model by solving the following optimization problem on a set of training data $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$.

$$\begin{aligned} \min_{\boldsymbol{w},b} \quad & \frac{1}{2}b^2 \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^T\boldsymbol{x} + b) \geq 1 \end{aligned} \tag{2}$$

Which of the separating hyperplane(s) are likely to be the solution(s) of it? Choose all possible candidates. **Solution:** D
The optimization problem finds a line that passes through the origin and seperates the data.

6 Multi-Class Feed-Forward Neural Networks [11 pts]

In the lectures, we learned multi-class logistic regression. In the following, we will extend it to multi-class feed-forward neural networks.

Given data points $\{\boldsymbol{x}_i, y_i\}, i = 1, \ldots N$, where $\boldsymbol{x}_i \in \mathcal{R}^{D\times 1}$ is an D-dimensional feature vector, $\boldsymbol{y}_i \in \{1, 2, \ldots, K\}$ is the class label.

Consider the following neural network which generates the distribution of the label prediction $\hat{\boldsymbol{y}}$ based on input $\boldsymbol{x}$

$$\begin{aligned}\boldsymbol{h} &= W_1\boldsymbol{x} + \boldsymbol{b}_1\\ \boldsymbol{g} &= W_2\boldsymbol{h} + \boldsymbol{b}_2\\ \hat{\boldsymbol{y}} &= \text{Softmax}(\boldsymbol{g}).\end{aligned}$$

$W_1 \in \mathcal{R}^{H\times D}$, $W_2 \in \mathcal{R}^{K\times H}$ are the weight matrices. $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are the bias terms.

The softmax function takes $\boldsymbol{g} \in \mathcal{R}^K$ and returns a vector $\hat{\boldsymbol{y}}$, where the k-th element of $\hat{\boldsymbol{y}}$ is

$$\hat{y}_k = \frac{\exp(g_k)}{\sum_{j=1}^K \exp(g_j)}, \tag{3}$$

where g_j is the j-th element of $\boldsymbol{g}$.

(a) **(4 pts)** What is the dimensionality of $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ in terms of D, K, H

The dimensionality of $\boldsymbol{b}_1$ is _____. **Solution:** $H \times 1$ (or H)

The dimensionality of $\boldsymbol{b}_2$ is _____. **Solution:** $K \times 1$ (or K)

(b) **(4 pts)** A potential issue of using Softmax is when some item g_i of $\boldsymbol{g}$ are large then computing $\exp(g_i)$ and $\sum_{j=1}^K \exp(g_j)$ might overflow. One trick for resolving it is to let $m = \max_j g_j$ and compute the k-th element of softmax($\boldsymbol{g}$) by

$$\hat{y}'_k = \frac{\exp(g_k - m)}{\sum_{j=1}^K \exp(g_j - m)}. \tag{4}$$

Show that $\hat{y}'_k$ and $\hat{y}_k$ are equivalent. **Solution:**

$$\hat{y}'_k = \frac{\exp(g_k - m)}{\sum_{j=1}^K \exp(g_j - m)} = \frac{\exp(g_k - m)\exp(m)}{\exp(m)\sum_{j=1}^K \exp(g_j - m)} = \frac{\exp(g_k)}{\sum_{j=1}^K \exp(g_j)} = \hat{y}_k.$$

(c) **(3 pts)** Follow the previous question, what is the upper bound value of $\exp(g_k - m)$?

$\exp(g_k - m) \leq$ _____.

Solution: 1

Explanation: $g_k - m \leq 0$ as $g_k \leq m$. Therefore, $exp(g_k - m) \leq 1$