CSCI-567 Spring 2019 Midterm Exam [Rubric]

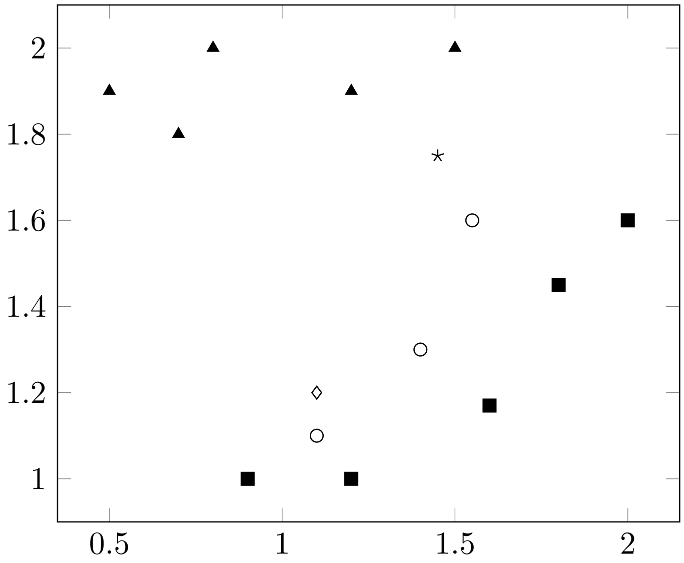
Problem	1	2	3	4	5	Total
Points	28	20	20	15	17	100

Please read the following instructions carefully:

- The exam has a total of **15 pages** (including this cover and two blank pages in the end). Each problem have several questions. Once you are permitted to open your exam (and not before), you should check and make sure that you are not missing any pages.
- Duration of the exam is **3 hours**. Questions are not ordered by their difficulty. Budget your time on each question carefully.
- Select **one and only one answer** for all multiple choice questions.
- Answers should be **concise** and written down **legibly**. All questions can be done within 5-12 lines.
- You must answer each question on the page provided. You can use the last two blank pages as scratch paper. Raise your hand to ask a proctor for more if needed.
- This is a **closed-book/notes** exam. Consulting any resources is NOT permitted.
- Any kind of cheating will lead to **score 0** for the entire exam and be reported to SJACS.
- You **may not** leave your seat **for any reason** unless you submit your exam at that point.

1 Multiple Choice (28 points)

Questions 1 through 3 deal with the following data, where squares, triangles, and open circles are three different classes of data in the training set and the diamond (◇) and the star (*) are test points. The distance metric is L_2 distance.

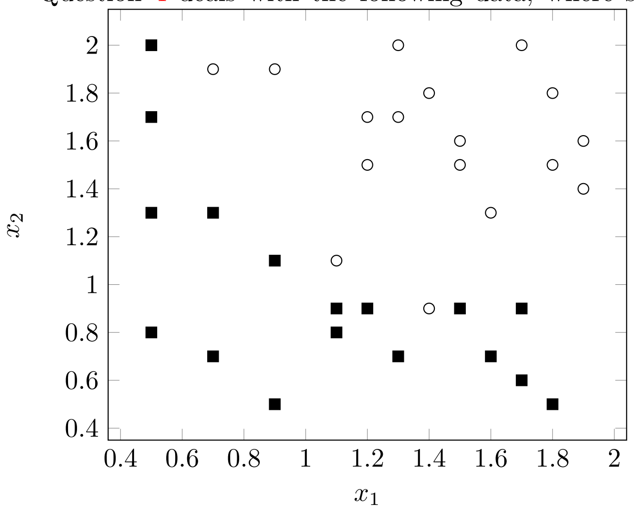


1. What label will we predict for diamond with $k = 1$? Answer: B **(2 points)**
   (a) Triangle
   (b) Circle
   (c) Square
   (d) Cannot be determined from data available

2. What label will we predict for diamond with $k = 5$? Answer: C **(2 points)**
   (a) Triangle
   (b) Circle
   (c) Square
   (d) Cannot be determined from data available

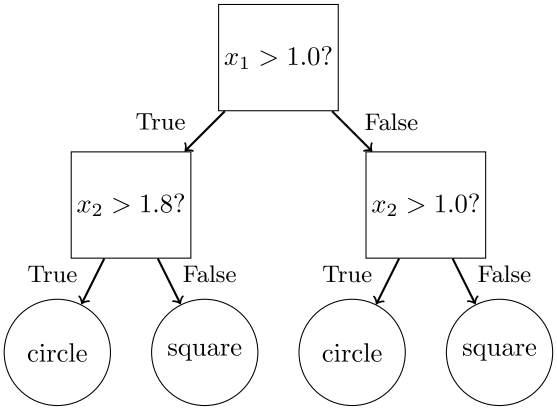
3. Which of the following values of k can be used to classify the star as triangle? Answer: B **(2 points)**
   (a) 1
   (b) 3
   (c) 5
   (d) None of above

Question 4 deals with the following data, where squares and circles are two different classes of data.

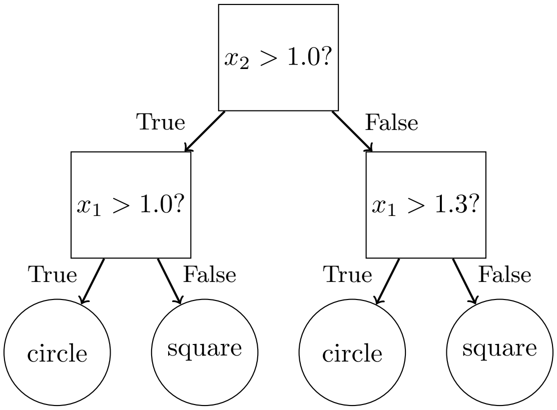


4. Suppose we use the data as training data, which of the following decision trees has the smallest training error? Answer: C **(2 points)**

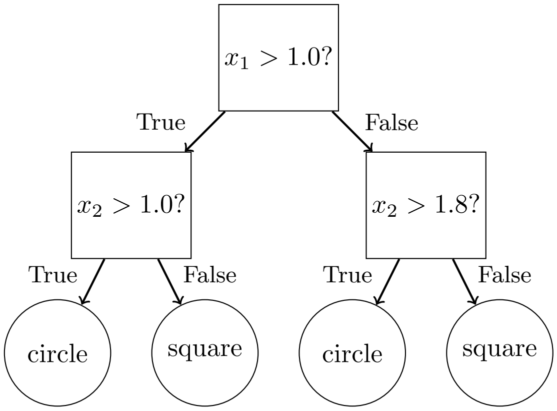
(a)



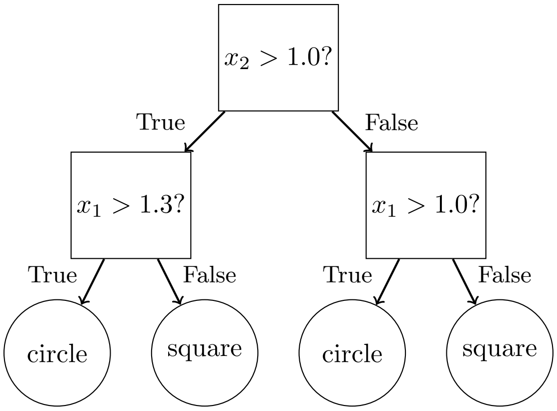
(b)



(c)



(d)



5. K-fold cross-validation runtime complexity is Answer: A **(2 points)**

   (a) linear in K

   (b) quadratic in K

   (c) cubic in K

   (d) exponential in K

6. The multiclass perceptron algorithm is essentially minimizing the multiclass perceptron loss via SGD with learning rate 1. Based on this information, which of the following is the multiclass perceptron loss? Answer: C **(2 points)**

   (a) $\sum_{n=1}^{N} \max_{k \neq y_n} \boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{x}_n - \boldsymbol{w}_{y_n}^{\mathrm{T}} \boldsymbol{x}_n$

   (b) $\sum_{n=1}^{N} \max \left\{0, \max_{k \neq y_n} \boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{x}_n \right\}$

   (c) $\sum_{n=1}^{N} \max \left\{0, \max_{k \neq y_n} \boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{x}_n - \boldsymbol{w}_{y_n}^{\mathrm{T}} \boldsymbol{x}_n \right\}$

   (d) $\sum_{n=1}^{N} \max \left\{0, \max_{k \in [C]} \boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{x}_n \right\}$

7. Which is *not* an advantage of using kernels? Answer: A **(2 points)**

   (a) They can scale well to large number of data instances.

   (b) They can convert linear models into nonlinear models.

   (c) They can efficiently represent high dimensional inputs.

   (d) They can be analyzed with statistical learning theory.

8. Which is *not* a valid kernel function, for samples x and y and kernels $k_1(\cdot, \cdot)$ and $k_2(\cdot, \cdot)$? Answer: B **(2 points)**

   (a) $k(x, y) = 5$

   (b) $k(x, y) = x + y$

   (c) $k(x, y) = e^{x+y}$

   (d) $k(x, y) = \langle x, y \rangle^3 + (\langle x, y \rangle + 1)^2$

9. Suppose we apply the kernel trick with a kernel function k to the nearest neighbor algorithm (with L2 distance in the new feature space). What is the nearest neighbor of a new data point $\boldsymbol{x}$ from a training set $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$? Answer: D. **(2 points)**

   (a) $\arg\min_{\boldsymbol{x}_n \in S} \; k(\boldsymbol{x}_n, \boldsymbol{x}_n) + k(\boldsymbol{x}, \boldsymbol{x}) + 2k(\boldsymbol{x}_n, \boldsymbol{x})$

   (b) $\arg\min_{\boldsymbol{x}_n \in S} \; k(\boldsymbol{x}_n, \boldsymbol{x})$

   (c) $\arg\min_{\boldsymbol{x}_n \in S} \; (k(\boldsymbol{x}_n, \boldsymbol{x}) - k(\boldsymbol{x}, \boldsymbol{x}_n))^2$

   (d) $\arg\min_{\boldsymbol{x}_n \in S} \; k(\boldsymbol{x}_n, \boldsymbol{x}_n) - 2k(\boldsymbol{x}_n, \boldsymbol{x})$

10. Which of the following is wrong about neural nets? Answer: C. **(2 points)**

    (a) A fully connected feedforward neural net without nonlinear activation functions is the same as a linear model.

    (b) Dropout technique prevents overfitting.

    (c) A neural net with one hidden layer and a fixed number of neurons can represent any continuous function.

    (d) A max-pooling layer has no parameters to be learned.

11. Suppose a convolution layer takes a $4 \times 6 \times 3$ image as input and outputs a $3 \times 4 \times 6$ tensor. Which of the following is a possible configuration of this layer? Answer: D **(2 points)**

    (a) Two $2 \times 4 \times 3$ filters, stride 1, no zero-padding.

    (b) Two $2 \times 2 \times 3$ filters, stride 2, 1 zero-padding.

    (c) Six $2 \times 4 \times 3$ filters, stride 1, no zero-padding.

    (d) Six $2 \times 2 \times 3$ filters, stride 2, 1 zero-padding.

12. Recall that the Gini impurity of a distribution p over a set of K items is defined as $\sum_{k=1}^{K} p(k)(1-p(k))$. Which of the following distributions has the largest Gini impurity? Answer: A. **(2 points)**

    (a) (0.25, 0.25, 0.25, 0.25)

    (b) (0.2, 0.3, 0.5, 0)

    (c) (1, 0, 0, 0)

    (d) (0.5, 0.5, 0, 0)

13. Which of the following cannot be used as regularization to control model complexity? Answer: C **(2 points)**

    (a) $R(\boldsymbol{w}) = \sum_{d=1}^{D} |w_d|$

    (b) $R(\boldsymbol{w}) = \sum_{d=1}^{D} |w_d|^3$

    (c) $R(\boldsymbol{w}) = \sum_{d=1}^{D} w_d^3$

    (d) $R(\boldsymbol{w}) = \sum_{d=1}^{D} |w_d|^4$

14. In K-Fold cross-validation, a large value of K could result in which of the following? Answer: A **(2 points)**

    (a) low bias, high variance

    (b) high bias, low variance

    (c) it could result in (a) or (b)

    (d) none of the above

2 Naïve Bayes (20 points)

Assume we have a data set with three binary input attributes, A, B, C, and one binary outcome attribute Y. The three input attributes, A, B, C take values in the set $\{0, 1\}$ while the Y attribute takes values in the set $\{True, False\}$.

A	B	C	Y
0	1	1	True
1	1	0	True
1	0	1	False
1	1	1	False
0	1	1	True
0	0	0	True
0	1	1	False
1	0	1	False
0	1	0	True
1	1	1	True

In this problem, a set S of input values which consists of A, B, C with $P(S) > 0$ will have an unambiguous predicted classification of $Y = True \leftrightarrow P(Y = True|S) > P(Y = False|S)$

15. How would you classify the record $S = (A = 1, B = 1, C = 0)$ based on the data given in the table above? Write down both $P(Y = True|A = 1, B = 1, C = 0)$ and $P(Y = False|A = 1, B = 1, C = 0)$, and explain if Y should be True or False. **(10 points)**

$$P(Y = True|A = 1, B = 1, C = 0) = \frac{P(A = 1, B = 1, C = 0|Y = True)P(Y = True)}{P(A = 1, B = 1, C = 0)}$$

$$P(Y = False|A = 1, B = 1, C = 0) = \frac{P(A = 1, B = 1, C = 0|Y = False)P(Y = False)}{P(A = 1, B = 1, C = 0)}$$

(2 points)

$$\begin{aligned}&P(A = 1|Y = True)P(B = 1|Y = True)P(C = 0|Y = True)P(Y = True)\\&= \frac{2}{6} * \frac{5}{6} * \frac{3}{6} * \frac{6}{10}\\&= \frac{1}{12}\end{aligned}$$

(3 points)

$$\begin{aligned}&P(A = 1|Y = False)P(B = 1|Y = False)P(C = 0|Y = False)P(Y = False)\\&= \frac{3}{4} * \frac{2}{4} * \frac{0}{4} * \frac{4}{10}\\&= 0\end{aligned}$$

(3 points)

$\because P(A=1|Y=True)P(B=1|Y=True)P(C=0|Y=True)P(Y=True) > P(A=1|Y=False)P(B=1|Y=False)P(C=0|Y=False)P(Y=False)$ **(1 points)**

$\therefore Y = True.$ **(1 points)**

16. How would you classify the record $S = (A = 0, B = 0, C = 1)$ based on the data given in the table above? Write down both $P(Y = True|A = 0, B = 0, C = 1)$ and $P(Y = False|A = 0, B = 0, C = 1)$, and explain if Y should be True or False. **(10 points)**

$$P(Y = True|A = 0, B = 0, C = 1) = \frac{P(A = 0, B = 0, C = 1|Y = True)P(Y = True)}{P(A = 0, B = 0, C = 1)}$$
$$P(Y = False|A = 0, B = 0, C = 1) = \frac{P(A = 0, B = 0, C = 1|Y = False)P(Y = False)}{P(A = 0, B = 0, C = 1)}$$

(2 points)

$$\begin{aligned}&P(A = 0|Y = True)P(B = 0|Y = True)P(C = 1|Y = True)P(Y = True)\\&= \frac{4}{6} * \frac{1}{6} * \frac{3}{6} * \frac{6}{10}\\&= \frac{1}{30}\end{aligned}$$

(3 points)

$$\begin{aligned}&P(A = 0|Y = False)P(B = 0|Y = False)P(C = 1|Y = False)P(Y = False)\\&= \frac{1}{4} * \frac{2}{4} * \frac{4}{4} * \frac{4}{10}\\&= \frac{1}{20}\end{aligned}$$

(3 points)

$\because P(A = 0|Y = True)P(B = 0|Y = True)P(C = 1|Y = True)P(Y = True) < P(A = 0|Y = False)P(B = 0|Y = False)P(C = 1|Y = False)P(Y = False)$ **(1 points)**

$\therefore Y = False.$ **(1 points)**

3 Logistic Regression and Kernels (20 points)

Review Recall that the logistic regression model is defined as:

$$p(y=1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}+b) \tag{1}$$

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{2}$$

Given a training set $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$, where $\mathbf{x}_n \in \mathbb{R}^{K\times 1}$ and $y_n \in \{0,1\}$, we will minimize the cross-entropy error function to solve $\mathbf{w}$.

$$\min_{\mathbf{w},b} L(\mathbf{w},b) = \min_{\mathbf{w},b} -\sum_n \{y_n \log[p(y_n=1|\mathbf{x}_n)] + (1-y_n)\log[p(y_n=0|\mathbf{x}_n)]\} \tag{3}$$

$$= \min_{\mathbf{w},b} -\sum_n \{y_n \log[\sigma(\mathbf{w}^T\mathbf{x}_n+b)] + (1-y_n)\log[1-\sigma(\mathbf{w}^T\mathbf{x}_n+b)]\} \tag{4}$$

17. Consider the dataset consisting of points (x,y), where $x \in \mathbb{R}$ and $y \in \{0,1\}$. Suppose we have three training points $(x_1,y_1)=(0,0)$, $(x_2,y_2)=(1,1)$ and $(x_3,y_3)=(-1,1)$.

    Suppose our logistic regression model is $p(y=1|x) = \sigma(wx+b)$ with $b=1$. What would be the optimal logistic regression classifier using the training data provided and what is the training accuracy?

$$\min_w L(w) = \min_w -\sum_n \{y_n \log[\sigma(wx_n+b)] + (1-y_n)\log[1-\sigma(wx_n+b)]\} \ \textbf{(2 points)}$$

$$\frac{\partial L(w)}{\partial w} = \sum_n \{[y_n - \sigma(wx_n+b)]x_n\}$$

Set gradient to 0,

$$\Rightarrow 0 = \sum_n \{[y_n - \sigma(wx_n+b)]x_n\}$$

Substitute the data points,

$$0 = 0 + [1-\sigma(w+1)] + [1-\sigma(-w+1)]*(-1) \ \textbf{(2 points)}$$

Therefore $w^* = 0$ and the optimal logistic regression classifier is $\hat{y} = 1$ **(2 points)**

Predictions on training data:

$$\hat{y}_1 = \mathbb{I}[\sigma(w^*x_1+b) > 0.5] = 1 \neq y_1$$
$$\hat{y}_2 = \mathbb{I}[\sigma(w^*x_2+b) > 0.5] = 1 = y_2$$
$$\hat{y}_3 = \mathbb{I}[\sigma(w^*x_3+b) > 0.5] = 1 = y_3$$

The training accuracy is $\frac{2}{3}$. **(4 points)**

18. Consider the following function:

$$k(x, x') = \begin{cases} 1 & \text{if } x = x' \\ 0 & \text{otherwise} \end{cases}$$

Prove this is a valid kernel. **(10 points)**

Kernel is valid because because gram matrix positive semi-definite, since it is the identity matrix with eigenvalue 1.

i) Explanation of why Gram matrix is positive semi-definite (PSD), either from the definition of PSD or the fact that eigenvalues are 1 → +4 points

ii) Gram matrix is PSD → +2 points

iii) PSD → +2 points

iv) Any logic → +2 points

v) Any errors with Explanation in i) → see note in exam for deduction

4 K-means clustering **(15 points)**

Recall the K-means clustering algorithm. Given a training set $\{\boldsymbol{x}_n \in \mathbb{R}^D\}_{n=1}^N$ of N samples, the K-means algorithm aims to minimize the distortion measure J,

$$J(\{\boldsymbol{\mu}_k\}, \{\gamma_{nk}\}) = \sum_n \sum_k \gamma_{nk} \left\| \boldsymbol{x}_n - \boldsymbol{\mu}_k \right\|_2^2 = \sum_n \sum_k \gamma_{nk} \sum_d (x_{nd} - \mu_{kd})^2 \tag{5}$$

w.r.t. $\{\boldsymbol{\mu}_k \in \mathbb{R}^D\}_{k=1}^K$ and $\{\gamma_{nk} \in \{0,1\}\}_{n=1,k=1}^{N,K}$ iteratively, where $\sum_{k=1}^K \gamma_{nk} = 1, \forall n$. Alg. 1 outlines the algorithm, where $\{\boldsymbol{\mu}_k^{(t)}\}_{k=1}^K$ and $\{\gamma_{nk}^{(t)}\}_{n=1,k=1}^{N,K}$ are the learned parameters after iteration t.

Algorithm 1: The K-means algorithm (with K clusters)

```
1 Initialization $t = 0, \{\boldsymbol{\mu}_k^{(0)} \in \mathbb{R}^D\}_{k=1}^K$
  while $J$ is strictly decreasing do
2     $t \leftarrow t + 1$
3     (a) $\gamma_{nk}^{(t)} = \begin{cases} 1, & \text{if } k = \arg\min_j \left\| \boldsymbol{x}_n - \boldsymbol{\mu}_j^{(t-1)} \right\|_2^2, \\ 0, & \text{otherwise.} \end{cases}$
4     (b) $\boldsymbol{\mu}_k^{(t)} = \frac{\sum_{n=1}^N \gamma_{nk}^{(t)} \boldsymbol{x}_n}{\sum_{n=1}^N \gamma_{nk}^{(t)}} = \frac{\sum_{n:\gamma_{nk}^{(t)}=1} \boldsymbol{x}_n}{\sum_{n:\gamma_{nk}^{(t)}=1} 1}, \quad \forall k \in \{1, \cdots, K\}.$
5 Output $\{\boldsymbol{\mu}_k^{(t)}\}_{k=1}^K$ and $\{\gamma_{nk}^{(t)}\}_{n=1,k=1}^{N,K}$
```

19. **K-means with L_1 norm:** Now we use the L_1 norm instead of L_2 norm in the distortion measure J, that is,

$$J_{\ell_1}(\{\boldsymbol{\mu}_k\}, \{\gamma_{nk}\}) = \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \| x_n - \mu_k \|_1 = \sum_n \sum_k \gamma_{nk} \sum_d |x_{nd} - \mu_{kd}|,$$

where we sum up the absolute difference across all coordinates. Please derive the corresponding update rules of γ_{nk}, and $\boldsymbol{\mu}_k$ (Line 3 and 4 in Alg. 1) for the distortion J_{ℓ_1}. Answer for step (b) might not be unique, please write out one feasible answer. **(10 points)**

(a) $\gamma_{nk}^{(t)} = \begin{cases} 1, & \text{if } k = \arg\min_j \| \boldsymbol{x}_n - \boldsymbol{\mu}_j^{(t-1)} \|_1, \\ 0, & \text{otherwise.} \end{cases}$ **(5 points)**

i) $\gamma_{nk} = 1$ correct → 3

ii) $\gamma_{nk} = 0$ correct → 2

iii) unclear distance L_1 or L_2 → -1

(b) $\boldsymbol{\mu}_k^{(t)} = \mathsf{median}_{n:\gamma_{nk}^{(t)}=1} \{\boldsymbol{x}_n\}, \ \forall k \in \{1, \cdots, K\}$ (coordinate-wise median) **(5 points)**

i) mention the idea of median and provide final form → 5

ii) mention the idea of median → 2

iii) not mention median of data “with $\gamma_{nk} = 1$” → -1

iv) not provide final form → -2

v) wrong update form → -3

20. ***K*-means with outliers:** If your data contains outliers, which version of K-means would you use: the L_1 norm one or the original one with L_2 norm? Explain your answer. **(5 points)**

L_1 norm is better for data with outliers, because the median is robust to outliers. The median only cares about picking the middle point in the ordering. And the furthest points, possibly outliers, would not effect the cluster centroid.

i) L_1 is better → 3.

ii) The *median* is robust to outliers → 2. (To get full credit, the answer must explicitly mention median.)

5 Neural Network (17 points)

21. Consider the following convolutional neural network. A $32 \times 32 \times 3$ image input, followed by a convolution layer with 2 filters of size 5×5 (stride 1, no zero-padding), then another convolution layer with 3 filters of size 5×5 (stride 1, no zero-padding), and finally an average-pooling layer with a 2×2 filter (stride 2, no zero-padding).

    i. How many parameters do we need to learn for this network? **(3 points)**

    $(5 \times 5 \times 3 + 1) \times 2 + (5 \times 5 \times 2 + 1) \times 3 = 152 + 153 = 305$

    i) 305 or 300 (without bias) → 3 points

    ii) any errors → -1 point for each

    ii. What is the picture final dimension? **(3 points)**

    $12 \times 12 \times 3$

    i) $12 \times 12 \times 3 \rightarrow 3$ points

    ii) any errors → -1 point for each

22. Suppose we have a neural network defined below: **(11 points)**

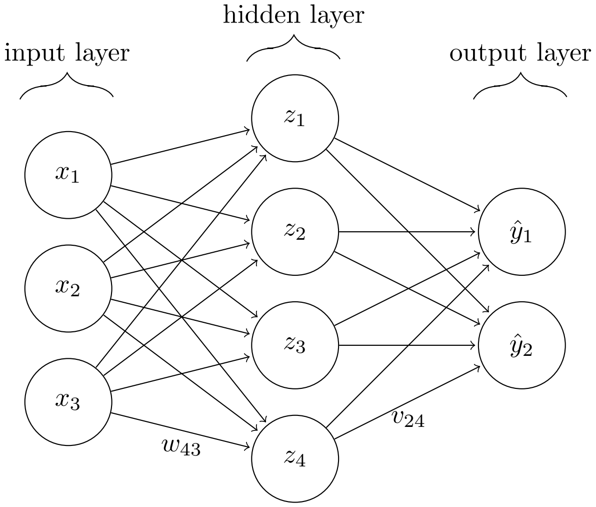


Figure 1: A neural network with one hidden layer.

$$\text{input layer} \qquad x_i, \tag{6}$$

$$\text{hidden layer} \qquad z_k = \arctan\left(\sum_{i=1}^{3} w_{ki} x_i\right), \tag{7}$$

$$\text{output layer} \qquad \hat{y}_j = \sum_{k=1}^{4} v_{jk} z_k, \tag{8}$$

$$\text{loss function} \qquad L(y, \hat{y}) = \sqrt{\frac{1}{2}\left((\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2\right)}. \quad \hat{y}_j \text{ is prediction, } y_j \text{ is ground truth} \tag{9}$$

Write down $\frac{\partial L}{\partial v_{jk}}$ and $\frac{\partial L}{\partial w_{ki}}$ in terms of only x_i, z_k, y_j, $\hat{y}_j$, w_{ki}, and/or v_{jk}.

Hint: $\frac{\partial \arctan(\alpha)}{\partial \alpha} = \frac{1}{\alpha^2 + 1}$.

$$
\begin{aligned}
\frac{\partial L}{\partial v_{jk}} &= \frac{\partial L}{\partial \hat{y}_j}\frac{\partial \hat{y}_j}{\partial v_{jk}} \\
\beta &= \sqrt{\frac{1}{2}\left((\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2\right)} \\
\frac{\partial L}{\partial \hat{y}_1} &= \frac{\partial}{\partial \hat{y}_1}\beta(\hat{y}_1) \\
&= \frac{1}{2\beta}(\hat{y}_1 - y_1) \\
\frac{\partial \hat{y}_j}{\partial v_{jk}} &= \sum_{k=1}^{4} v_{jk} z_k = z_k \\
\rightarrow \frac{\partial L}{\partial v_{jk}} &= \frac{1}{2\beta}(\hat{y}_i - y_i) z_k
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial w_{ki}} &= \frac{\partial L}{\partial z_k}\frac{\partial z_k}{\partial w_{ki}} \\
\frac{\partial L}{\partial z_k} &= \frac{\partial L}{\partial \hat{y}_1}\frac{\partial \hat{y}_1}{\partial z_k} + \frac{\partial L}{\partial \hat{y}_2}\frac{\partial \hat{y}_2}{\partial z_k} \\
&= \sum_{j=1}^{2} \frac{1}{2\beta}(\hat{y}_j - y_j) v_{jk} \\
\frac{\partial z_k}{\partial w_{ki}} &= \frac{\partial}{\partial w_{ki}} \arctan\left(\sum_{i=1}^{3} w_{ki} x_i\right) \\
&= \frac{x_i}{\tan^2(z_k) + 1} \\
\rightarrow \frac{\partial L}{\partial w_{ki}} &= \left(\sum_{j=1}^{2} \frac{1}{2\beta}(\hat{y}_j - y_j) v_{jk}\right) \frac{x_i}{\tan^2(z_k) + 1}
\end{aligned}
$$

The rationale behind your grade for this question is reported on your exam.

You may use this page as scratch paper

You may use this page as scratch paper