Summary of Marking Guidelines:

- If you want you can use this as a convenient place to keep your overall instructions to your markers. This way, you can find these instructions years later when you are looking back at the exam.
  If you are seeing this and not a cover page, you have 'solutions' and 'markingspace' both set to true in your main `.tex` file.

Question 1. Short Questions [6 marks]

Part (a) KNN [2 marks]

True or False: The 1NN model will typically have higher training accuracy than the 3NN model.

Circle the best answer: TRUE or FALSE

Justify your answer in no more than 3 sentences.

Solution

Answer: True

Justification: 1NN will classify every training data point correctly. 3NN's predictions may be different from the ground truth labels, for example, when the data point is closest to two points with a different label. Therefore, 1NN typically has higher training accuracy than 3NN.

Marking Scheme:

- 0 mark if the answer is incorrect.
- 0 mark for a correct answer with an incorrect explanation.
- 1 mark for a correct answer with insufficient explanation.

Part (b) Decision Trees [2 marks]

In class, we discussed a greedy algorithm to build a decision tree by choosing a feature and a split point to maximize the information gain at each step. Generally, this greedy algorithm does not produce the optimal decision tree (assuming that we have some well-defined criteria to evaluate whether a decision tree is optimal). Explain why this is the case in no more than 3 sentences.

Solution

The greedy algorithm for building decision trees makes the locally optimal choice at each step to maximize information gain. However, a locally optimal choice does not necessarily lead to a globally optimal solution. The algorithm does not consider future splits or the overall structure of the tree when making a decision at a particular node. Consequently, the algorithm might miss a less obvious split early on that could lead to a more optimal overall tree structure.

Marking Scheme:

- look for an explanation of why maximizing information gain at each step may not be globally optimal.
- 0 mark if the answer is incorrect.
- 0 mark for a correct answer with an incorrect explanation.
- 1 mark for a correct answer with insufficient explanation.

Part (c) KNN versus Decision Trees [2 marks]

Suppose that we have very limited time and resources to train our model. Which of KNN and Decision Tree would you choose? Why?

Circle the best answer: KNN or Decision Trees

Justify your answer in no more than 3 sentences.

Solution

Answer: KNN

Justification: KNN does not require any time to train the model, whereas decision trees require some training time to build the tree.

Marking Scheme:

- 0 mark if the answer is incorrect.
- 0 mark for a correct answer with an incorrect explanation.
- 1 mark for a correct answer with insufficient explanation.

Data Point	Feature 1	Feature 2	Label
1	2	2	square
2	1	3	triangle
3	3	3	triangle
4	2	5	square

Table 1: Data Set for KNN

Question 2. KNN [8 marks]

Consider the data set in Table 1. Suppose that we will classify new data points by using the k nearest neighbour algorithm with the Euclidean distance measure. The Euclidean distance between two points (x_1, y_1) and (x_2, y_2) is given by the formula below.

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Part (a) [2 marks]

Draw the decision boundary for 1NN in the Figure below.

Solution The decision boundary for 1NN is given in Figure 1.

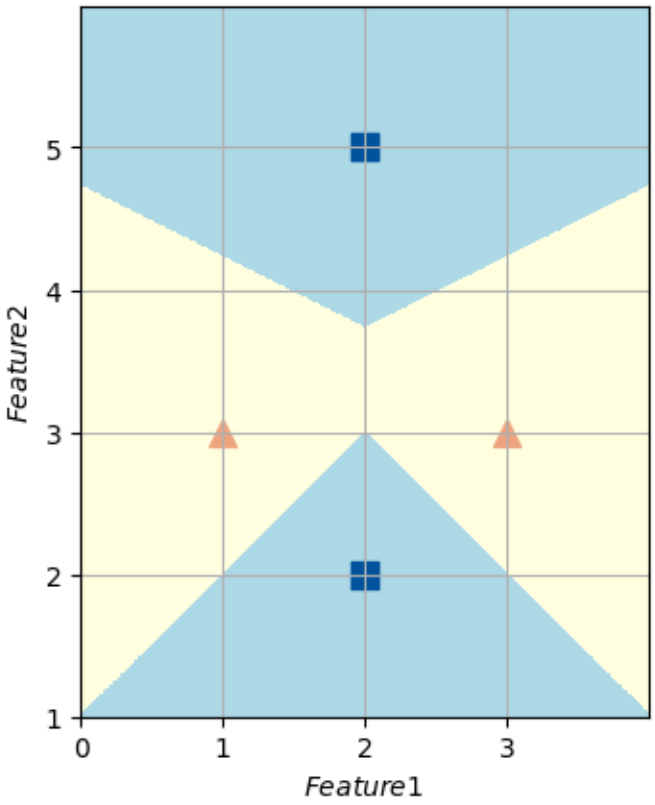

Figure 1: 1NN decision boundry

Marking Scheme:

- 1 mark if the boundary is partially correct. Some examples below.
  - there is a vertical line segment in the middle between the two triangles.
  - missing the bottom part.
  - missing the top part.

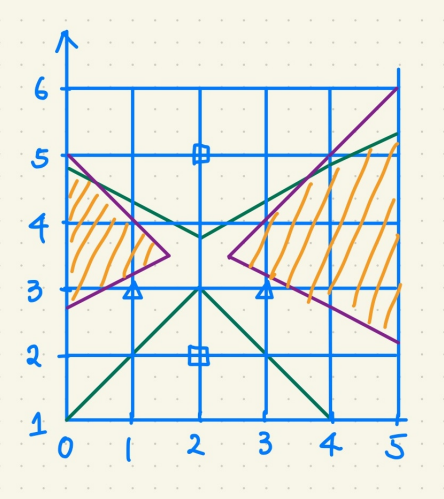

Figure 2: 1NN and 3NN Overlap

Question 2. (CONTINUED)

Part (b) [3 MARKS]

Find one point (x, y) satisfying all of the constraints below. Whenever there is a tie, KNN will always predict 'triangle.'

1. 1NN predicts 'triangle' for (x, y).
2. 3NN predicts 'square' for (x, y).
3. x and y are integers, $1 \leq x \leq 5$, and $1 \leq y \leq 5$.

Show all your work.

SOLUTION

There are five possible solutions: $(4, 3)$, $(4, 4)$, $(5, 3)$, $(5, 4)$, $(5, 5)$. See the decision boundaries in Figure 2. Any point in the shaded region works. Because of the tie-breaking rule favouring triangle, any point on the boundary is an incorrect answer.

For example, consider the point (4,3).

- Squared distance between (4,3) and (2,2) is $2^2 + 1^2 = 5$.
- Squared distance between (4,3) and (1,3) is $3^2 + 0^2 = 9$.
- Squared distance between (4,3) and (3,3) is $1^2 + 0^2 = 1$.
- Squared distance between (4,3) and (2,5) is $2^2 + 2^2 = 8$.

Since (4,3) is closest to (3,3), 1NN predicts 'triangle.'

The three closest points to (4,3) are (3,3), (2,2) and (2,5) with labels triangle, square and square. 3NN predicts 'square.'

Part (c) [3 MARKS]

Compute the training accuracy for the 3NN model on this data set in Figure 1. **Justify your answer**.

SOLUTION

3NN classifies all four points as 'triangle'. Therefore, the training accuracy is 50%.

(2,2) is closest to (1,3) and (3,3) because

- between (2,2) and (1,3), $d^2 = 1^2 + 1^2 = 1$
- between (2,2) and (2,5), $d^2 = 0^2 + 3^2 = 9$

(2,5) is closest to (1,3) and (3,3) because

- between (2,5) and (1,3), $d^2 = 1^2 + 2^2 = 5$

- between (2,5) and (2,2), $d^2 = 0^2 + 3^2 = 9$

(1,3) is closest to (2,2) and (3,3) because

- between (1,3) and (3,3), $d^2 = 2^2 + 0^2 = 4$
- between (1,3) and (2,5), $d^2 = 1^2 + 2^2 = 5$

By similar reasoning, (3,3) is closest to (2,2) and (1,3).

The 3NN decision boundary is given in Figure 3. Students do not need to show the decision boundary to earn marks.

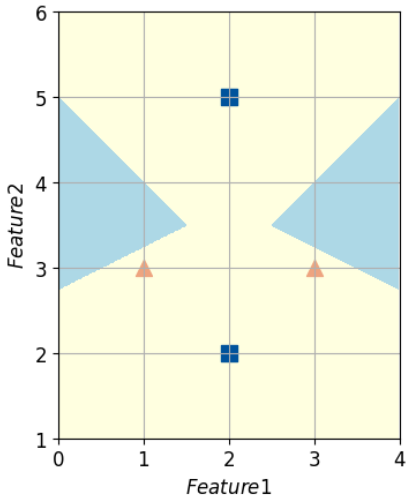

Figure 3: 3NN decision boundry

Marking Scheme:

- 1 mark for the correct answer.
- 2 marks for the correct justification.
- 0 marks for a correct answer with no justification.

Question 3. Decision Trees [9 marks]

Data Point	A	B	Label
1	small	20	yes
2	small	20	no
3	small	10	yes
4	medium	20	no
5	medium	10	yes
6	medium	10	yes
7	medium	10	no
8	large	20	no
9	large	20	no
10	large	20	no
11	large	10	yes

Table 2: Data Set for Decision Trees

Consider the data set in Table 2. There are two features A and B. Feature A is discrete and has three possible values: large, medium, and small. Feature B is real-valued. Assume that we will test feature A by splitting the examples into three groups corresponding to its three possible values. Assume that we will test feature B by making a binary split using $B \leq 15$.

Part (a) [3 marks]

Calculate the information gain of testing feature A at the root node. **Show all your work. Round your final answer to 3 decimal places.**

Whenever you need to calculate the entropy of a probability distribution, it's sufficient to write it in the format $H(p_1, p_2, p_3, ..., p_n)$ where the values $p_1, p_2, ..., p_n$ denote the probability distribution. There is no need to expand the entropy formula.

Feel free to use the quantities below.

$$H\left(\frac{1}{3}, \frac{2}{3}\right) = 0.918 \qquad H\left(\frac{1}{4}, \frac{3}{4}\right) = 0.811$$

Solution:

	yes	no
large	1	3
medium	2	2
small	2	1

$$H(Y) = H\left(\frac{5}{11}, \frac{6}{11}\right) = 0.994$$
$$H(Y|X) = \frac{4}{11}H\left(\frac{1}{4}, \frac{3}{4}\right) + \frac{4}{11}H\left(\frac{2}{4}, \frac{2}{4}\right) + \frac{3}{11}H\left(\frac{2}{3}, \frac{1}{3}\right) = \frac{4}{11}0.811 + \frac{4}{11}1 + \frac{3}{11}0.918$$
$$IG(Y|X) = H(Y) - H(Y|X) = 0.994 - 0.909 = 0.085$$

Marking Scheme:

- 1 mark for the entropy before testing the feature.
- 1 mark for the conditional entropy after testing the feature.
- 1 mark for calculating the information gain.
- overall, 2 marks for writing the formulas and 1 mark for the calculations. If all the formulas are correct except incomplete or incorrect calculation, deduct 1 mark only.

Question 3. (continued)

Part (b) [2 marks]

Write the **complete formula** for calculating the information gain of testing $B \leq 15$ at the root node. There is no need to perform the calculations.

Solution:

	yes	no
10	4	1
20	1	5

$$H(Y) = H\left(\frac{5}{11}, \frac{6}{11}\right)$$
$$H(Y|X) = \frac{5}{11}H\left(\frac{4}{5}, \frac{1}{5}\right) + \frac{6}{11}H\left(\frac{1}{6}, \frac{5}{6}\right)$$
$$IG(Y|X) = H(Y) - H(Y|X)$$

Marking Scheme:

- 1 mark if the formulas are partially correct.

In class, Alice discussed three possible base cases for the decision tree algorithm: (1) no examples left, (2) all the examples have the same label, and (3) no features or split points left.

Part (c) [2 marks]

For the data set in Table 2, do we ever encounter the "no examples left" case when learning the decision tree? **Justify your answer** in no more than 3 sentences.

Solution:

No, we will never encounter the "no examples left" case. For every combination of the feature values, there is at least one example in the data set. Therefore, any branch will have at least one data point.

Marking Scheme:

- 0 marks for an incorrect answer.
- 0 marks for a correct answer with no explanation.
- 1 mark if the answer is correct but the explanation is insufficient.

Part (d) [2 marks]

For the data set in Table 2, do we ever encounter the "no feature or split points left" case when learning the decision tree? **Justify your answer** in no more than 3 sentences.

Solution:

Yes, we will encounter the "no feature or split points left" case. We will encounter this case in two places: for A = medium and B = 10, and for A = small and B = 20. For both cases, there is at least one example labeled yes and one example labeled no. Since there are no more feature or split points, we will need to make the majority decision.

Marking Scheme:

- 0 marks for an incorrect answer.
- 0 marks for a correct answer with no explanation.
- 1 mark if the answer is correct but the explanation is insufficient.

Question 4. Linear Regression with One Feature [2 marks]

Consider a linear regression problem with one feature. For the i^{th} example, the prediction is given below.

$$y^{(i)} = wx^{(i)} + b$$

We will use the squared loss function and the cost function below.

$$L(y^{(i)}, t^{(i)}) = \frac{1}{2}(y^{(i)} - t^{(i)})^2$$
$$\mathcal{E}(w, b) = \frac{1}{2N}\sum_{i=1}^{N}((wx^{(i)} + b) - t^{(i)})^2$$

Suppose that we have three data points below:

$$x^{(1)} = 1, t^{(1)} = 3$$
$$x^{(2)} = 2, t^{(2)} = 2$$
$$x^{(3)} = 3, t^{(3)} = 2$$

Write down the system of linear equations that we need to solve to derive the optimal values of w and b. There is no need to simplify these equations.

Solution

The two linear equations that we derived in class are:

$$\frac{\partial \mathcal{E}(w, b)}{\partial w} = \frac{1}{N}\sum_{i=1}^{N}((wx^{(i)} + b) - t^{(i)})x^{(i)} = 0$$
$$\frac{\partial \mathcal{E}(w, b)}{\partial b} = \frac{1}{N}\sum_{i=1}^{N}((wx^{(i)} + b) - t^{(i)}) = 0$$

Plugging in the three data points, we have the following equations.

$$\frac{1}{3}((w + b - 3) + (2w + b - 2)2 + (3w + b - 2)3) = 0$$
$$\frac{1}{3}((w + b - 3) + (2w + b - 2) + (3w + b - 2)) = 0$$

Marking Scheme:

- Only need to show the final two equations with numbers plugged in. No need to simplify the equations.
- 1 mark if the two equations are partially correct.

Question 5. Linear Regression [11 marks]

Consider a linear regression problem with a L^2 regularizer. The cost function is given below.

$$\mathcal{E}(\mathbf{w}) = \frac{1}{2N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x}^{(i)} - t^{(i)})^2 + \frac{\lambda}{2}\sum_{j=0}^{D} w_j^2 \tag{1}$$

Part (a) Partial Derivative [3 marks]

Derive the expression of $\frac{\partial \mathcal{E}(\mathbf{w})}{\partial w_j}$ given below. **Show all your work.**

$$\frac{\partial \mathcal{E}(\mathbf{w})}{\partial w_j} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x}^{(i)} - t^{(i)})x_j^{(i)} + \lambda w_j \tag{2}$$

Solution:

$$\begin{aligned}
&\frac{\partial \mathcal{E}(\mathbf{w})}{\partial w_j} \\
&= \frac{1}{2N}\sum_{i=1}^{N}\frac{\partial}{\partial w_j}(\mathbf{w}^T\mathbf{x}^{(i)} - t^{(i)})^2 + \frac{\partial}{\partial w_j}\left(\frac{\lambda}{2}\sum_{j=0}^{D} w_j^2\right) \\
&= \left(\frac{1}{2N}\sum_{i=1}^{N}\frac{\partial}{\partial \mathbf{w}^T\mathbf{x}^{(i)}}(\mathbf{w}^T\mathbf{x}^{(i)} - t^{(i)})^2\frac{\partial}{\partial w_j}\left(\mathbf{w}^T\mathbf{x}^{(i)}\right)\right) + \frac{\partial}{\partial w_j}\left(\frac{\lambda}{2}\sum_{j=0}^{D} w_j^2\right) \\
&= \left(\frac{1}{2N}\sum_{i=1}^{N} 2(\mathbf{w}^T\mathbf{x}^{(i)} - t^{(i)})x_j^{(i)}\right) + \lambda w_j \\
&= \frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x}^{(i)} - t^{(i)})x_j^{(i)} + \lambda w_j
\end{aligned}$$

Marking Scheme:

- 2 marks for the derivative of the first term in the sum. This includes 1 mark on using the chain rule.
- 1 mark for computing the derivative of the second term.

Part (b) Vectorized Gradient [2 marks]

Derive the vectorized gradient $\nabla_{\mathbf{w}}\mathcal{E}(\mathbf{w})$. **Show all your work.**

You can use the following quantities in your expression. $\mathbf{X}$ is the $N \times (D+1)$ data matrix. $\mathbf{w}$ is the $(D+1) \times 1$ column vector of weights. $\mathbf{t}$ is the $N \times 1$ column vector of target values.

Solution

Based on the lectures on linear regression, we know that:

$$\frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} - t^{(i)}) = \frac{1}{N}\mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{t})$$

The second term becomes the vector $\mathbf{w}$.

The complete expression for the gradient is below. The result is a column vector.

$$\nabla_{\mathbf{w}}\mathcal{E}(\mathbf{w}) = \frac{1}{N}\mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{t}) + \lambda\mathbf{w}$$

An alternative solution is given below. The result is a row vector.

$$\nabla_{\mathbf{w}}\mathcal{E}(\mathbf{w}) = \frac{1}{N}(\mathbf{X}\mathbf{w} - \mathbf{t})^T\mathbf{X} + \lambda\mathbf{w}^T$$

Marking Scheme:

- 1 mark for the first term.
- 1 mark for the second term.

Question 5. (continued)

Part (c) Direct Solution [2 marks]

Derive the direct solution to this linear regression problem.

Hint: For any $D \times 1$ column vector $\mathbf{v}$, $\mathbf{v} = \mathbf{I}\mathbf{v}$ where $\mathbf{I}$ is the square $D \times D$ identity matrix with entries of ones on the diagonal and zeroes everywhere else.

Solution

$$\begin{aligned}
&\frac{1}{N}\mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{t}) + \lambda\mathbf{w} = \mathbf{0} \\
&\mathbf{X}^T\mathbf{X}\mathbf{w} + N\lambda\mathbf{I}\mathbf{w} - \mathbf{X}^T\mathbf{t} = \mathbf{0} \\
&(\mathbf{X}^T\mathbf{X} + N\lambda\mathbf{I})\mathbf{w} = \mathbf{X}^T\mathbf{t} \\
&\mathbf{w} = (\mathbf{X}^T\mathbf{X} + N\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{t}
\end{aligned}$$

Marking Scheme:

- 1 mark for getting the identity matrix part right.
- 1 mark for solving for $\mathbf{w}$ by inverting the matrix.

Question 5. (continued)

Part (d) Vectorized Gradient for Modified Cost Function [4 marks]

Consider the modified cost function below. There is a potentially different λ_j for every w_j.

$$\mathcal{E}'(\mathbf{w}) = \frac{1}{2N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x}^{(i)} - t^{(i)})^2 + \frac{1}{2}\sum_{j=0}^{D}\lambda_j w_j^2$$

Derive the vectorized gradient $\nabla_{\mathbf{w}}\mathcal{E}'(\mathbf{w})$. **Show all your work.**

This notation may be useful: Suppose that $\mathbf{v}$ denotes a $M \times 1$ column vector. Then, $\mathbf{V} = diag(\mathbf{v})$ denotes the $M \times M$ diagonal matrix where the diagonal entries are the entities of $\mathbf{v}$.

Solution

The first term in the partial derivative $\frac{\partial\mathcal{E}(\mathbf{w})}{\partial w_j}$ is identical to the first term in part (a).

$$\frac{\partial\mathcal{E}(\mathbf{w})}{\partial w_j}\frac{1}{2}\sum_{j=0}^{D}\lambda_j w_j^2 = \lambda_j w_j$$

$$\frac{\partial\mathcal{E}(\mathbf{w})}{\partial w_j} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x}^{(i)} - t^{(i)})x_j^{(i)} + \lambda_j w_j$$

Let λ denote the $(D+1)\times 1$ column vector. Then let $\Lambda = diag(\lambda)$ denote the $(D+1)\times(D+1)$ diagonal matrix where the diagonal entries are the entities of λ.

$$\nabla_{\mathbf{w}}\mathcal{E}(\mathbf{w}) = \frac{1}{N}\mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{t}) + diag(\lambda)\mathbf{w}$$

Marking Scheme:

- 2 marks for deriving the non-vectorized partial derivatives.
- 2 marks for converting to the vectorized expression.
- Give partial marks generously.