22417.202210 Midterm

Pankaj Kumar Jatav

TOTAL POINTS

73.5 / 100

QUESTION 1

Decision Trees 8 pts

1.1 Numerical answer 0 / 2

- **0 pts** Correct

✓ **- 2 pts** *Incorrect*

- **1 pts** Incorrect but still positive.

1.2 True or False 0 / 2

- **0 pts** Correct

✓ **- 2 pts** *Incorrect*

1.3 True or False 2 / 2

✓ **- 0 pts** *Correct*

- **2 pts** Incorrect

1.4 True or False 2 / 2

✓ **- 0 pts** *Correct*

- **2 pts** Incorrect

QUESTION 2

K Nearest Neighbors 5 pts

2.1 True or False 2 / 2

✓ **- 0 pts** *Correct*

- **2 pts** Incorrect

2.2 Select all that apply 0 / 2

- **0 pts** Correct (options 1 and 4)
- **0.7 pts** One correct option missing

✓ **- 1.4 pts** *Both correct options missing*

- **0.3 pts** One incorrect option selected

✓ **- 0.6 pts** *Two incorrect options selected*

2.3 Select one 0 / 1

- **0 pts** Correct (first option)

✓ **- 1 pts** *Incorrect*

QUESTION 3

Model Selection and Errors 15 pts

3.1 Select all that apply 2.5 / 4

- **0 pts** Correct (options 1 and 4)

✓ **- 1 pts** *One correct answer missing*

- **2 pts** Two correct answers missing

✓ **- 0.5 pts** *One incorrect answer selected*

- **1 pts** Two incorrect answers selected
- **1.5 pts** Three incorrect answers selected
- **2 pts** Four incorrect answers selected

3.2 Explain your choices 4 / 5

- **0 pts** Correct: The model is overfitting. (1,2) increasing training data size should help generalization. (3-4) decreasing model complexity helps avoid overfitting. (5): we should never do this (negative points).

✓ **- 1 pts** *One incorrect argument.*

3.3 What is this scenario called? 2 / 2

✓ **- 0 pts** *Correct (overfitting)*

- 2 pts Incorrect

3.4 Fill in the circle for (a) or (b) 1 / 1

✓ **- 0 pts** *Correct*

- 1 pts Incorrect (correct answer is (b))

3.5 Training sample size - q1 0 / 1

- 0 pts Correct (curve ii is the training error)

✓ **- 1 pts** *Incorrect*

3.6 Training sample size - q2 0.7 / 1

- 0 pts Correct

✓ **- 0.3 pts** *Not really justifying the choice above, or wrong reasoning.*

- 1 pts Blank

3.7 Training sample size - q3 0 / 1

- 0 pts Correct (overfitting)

✓ **- 1 pts** *Incorrect*

- 0.6 pts Incorrect but still somewhat related to overfitting

QUESTION 4

Perceptron 14 pts

4.1 Select all that apply 2 / 4

- 0 pts Correct (options 3 and 4)

✓ **- 1 pts** *One correct answer missing*

✓ **- 1 pts** *One incorrect answer selected*

- 2 pts Two incorrect answers selected

4.2 True or False 1 / 1

✓ **- 0 pts** *Correct (false)*

- 1 pts Incorrect

4.3 Perceptron Calculation - Numerical Answer (a) 1 / 1

✓ **- 0 pts** *Correct: [1, -2]*

- 1 pts Incorrect

- 0.6 pts Wrong sign

4.4 Perceptron Calculation - Numerical Answer (b) 1 / 1

✓ **- 0 pts** *Correct: [-1]*

- 1 pts Incorrect

- 0.6 pts Wrong sign

4.5 Perceptron Calculation - Numerical Answer (c) 1 / 1

✓ **- 0 pts** *Correct: [1, -2]*

- 1 pts Incorrect

- 0.6 pts Wrong sign

- 0.5 pts Half correct (somehow)

4.6 Perceptron Calculation - Numerical Answer (d) 1 / 1

✓ **- 0 pts** *Correct: [-1]*

- 1 pts Incorrect

- 0.6 pts Wrong sign

4.7 Perceptron Calculation - Numerical Answer (e) 1 / 1

✓ **- 0 pts** *Correct: [1, -2]*

- 1 pts Incorrect

- 0.6 pts Wrong sign

- 0.5 pts Half correct (somehow)

4.8 Perceptron Calculation - Numerical Answer (f) 1 / 1

✓ **- 0 pts** *Correct: [-1]*

- 1 pts Incorrect

- 0.6 pts Wrong sign

4.9 Perceptron Calculation - Numerical Answer (g) 0 / 1

- 0 pts Correct

✓ **- 1 pts** *Incorrect*

4.10 True or False 2 / 2

✓ **- 0 pts** *Correct*

- 2 pts Incorrect

QUESTION 5

Linear Regression 8 pts

5.1 True or False 3 / 3

✓ **- 0 pts** *Correct*

- 3 pts Incorrect

5.2 Fill-in the blanks 4 / 5

- 0 pts Correct (b,c,b,a,a)

✓ **- 1 pts** *One incorrect*

- 2 pts Two Incorrect

- 3 pts Three incorrect

- 4 pts Four incorrect

- 5 pts Five incorrect

QUESTION 6

Optimization 7 pts

6.1 Select all that apply 1 / 2

- 0 pts Correct (only second option is correct)

✓ **- 1 pts** *Correct answer missing*

- 0.5 pts One incorrect answer selected

- 1 pts Two incorrect answers selected

6.2 Select all that apply 1.4 / 2

- 0 pts Correct (only option 4)

- 1 pts Correct option not selected

- 0.3 pts One incorrect option selected

✓ **- 0.6 pts** *Two incorrect options selected*

- 1 pts All incorrect options selected

6.3 Range of all values 0.9 / 1

- 0 pts Correct: (0,1)

✓ **- 0.1 pts** *Incorrect bound (including 1 or 0)*

- 1 pts Blank, or not a range, or not meaningful range (e.g. including negative learning rate)

- 0.6 pts Half the correct range

- 0.5 pts Including all positive learning rates

6.4 Select all that apply 1.4 / 2

- 0 pts Correct: only option 1 is correct

✓ **- 0.3 pts** *Confusing option 1 and 2*

✓ **- 0.3 pts** *Option 3 selected (wrong sign)*

- 0.5 pts Option 4 selected (not true)

- 1 pts Not selecting either option 1 or 2

QUESTION 7

MLE/MAP 10 pts

7.1 True or False 1 / 1

✓ **- 0 pts** *Correct*

- 1 pts Incorrect

7.2 True or False 1 / 1

✓ **- 0 pts** *Correct*

- 1 pts Incorrect or blank

7.3 True or False **1 / 1**

✓ **- 0 pts** *Correct*

- 1 pts Incorrect or blank

7.4 Likelihood of L **1.5 / 2**

✓ **- 0 pts** *Correct. (Could have expanded)*

- 1 pts Where does the $$x^2$$ come from?

- 1 pts Where does the log come from?

- 0.4 pts Where do the $$n$$s in the exponent come from?

- 2 pts Not answering what is asked.

- 0 pts Where does this $$\frac{1}{n}$$ come from?

- 0.5 pts Using sum instead of product.

✓ **- 0.5 pts** *Sum (in second part) should be in the exponents*

7.5 Log likelihood **2 / 2**

✓ **- 0 pts** *Correct. (Could have expanded)*

- 1 pts Where does the $$y_i$$ come from?

- 0.4 pts Where do the $$n$$s in the products come from?

- 2 pts Not answering what is asked.

- 0 pts Where does this $$\frac{1}{n}$$ come from?

- 2 pts Incorrect

- 0.3 pts Technically correct, could have expanded

7.6 MLE **3 / 3**

✓ **- 0 pts** *Correct*

- 1 pts Minor error

- 3 pts Incorrect

QUESTION 8

Neural Networks 32 pts

8.1 Numerical answer **3 / 3**

✓ **- 0 pts** *Correct (all are 0.5)*

- 3 pts Not providing numerical answer

- 3 pts All incorrect

8.2 Derivative (1) **2 / 2**

✓ **- 0 pts** *Correct: expression is $$o_5-t$$ and value is 0.5*

- 1 pts Wrong expression.

- 0.2 pts Wrong sign

- 0.5 pts Value not computed

8.3 Derivative (2) **3.4 / 4**

- 0 pts Correct expressions and values

- 2 pts Wrong expression.

- 0.4 pts Wrong sign

- 1 pts Value not computed

- 1.6 pts Expressions not expanded

✓ **- 0.6 pts** *Incorrect values*

8.4 Derivative (3) **3 / 4**

- 0 pts Correct

- 1.5 pts Incorrect or not expanded expression, correct value

✓ **- 1 pts** *Incorrect value*

- 2 pts Blank values

8.5 Derivative (4) **1.7 / 2**

- 0 pts Correct

- **0.7 pts** Incorrect or not expanded expression, correct value

✓ - **0.3 pts** *Incorrect value*

- **0.5 pts** Blank value

- **2 pts** blank

8.6 Derivative (5) 0 / 3

- **0 pts** Correct: $$-\frac{1}{160} = -0.00625$$ for the first two, $$0$$ for the rest.

- **1 pts** One of three incorrect or blank

- **2 pts** Two of three incorrect or blank

✓ - **3 pts** *All three incorrect or blank*

- **0.2 pts** Magnitude error (e.g. $$10^{-3}$$ instead of correct $$10^{-5}$$.

- **0.2 pts** Sign error

- **1 pts** Correct formulas but not computed value

8.7 Derivative (5) - second part 0 / 3

- **0 pts** Correct: $$0$$ for all three.

- **1 pts** One of three incorrect or blank

- **2 pts** Two of three incorrect or blank

✓ - **3 pts** *All three incorrect or blank*

- **0.2 pts** Magnitude error (e.g. $$10^{-3}$$ instead of correct $$10^{-5}$$.

- **0.2 pts** Sign error

- **1 pts** Correct formulas but not computed value

8.8 Activation function types 6 / 6

✓ - **0 pts** *Correct: L, L, S*

- **2 pts** one incorrect

- **4 pts** two incorrect

- **6 pts** All incorrect or blank

8.9 Beta_1 derivation 5 / 5

✓ - **0 pts** *Correct: $$c(w_5w_1 + w_2w_6)$$*

- **5 pts** Blank

- **1 pts** Incorrect, but close to solution and with some effort

- **4 pts** Incorrect, but far from the solution

QUESTION 9

9 Extra credit 1 / 1

✓ - **0 pts** *Correct*

Extra credit

MIDTERM EXAM[1]

CS 688 MACHINE LEARNING (SPRING 2022)
https://nlp.cs.gmu.edu/course/cs688-spring22/

OUT: March 24, 2022

Your name: Pankaj Kumar Jatav

Your GID: 01338769

Please read the instructions (pages 1 and 2) carefully

Instructions for Specific Problem Types

For "Select One" questions, please fill in the appropriate bubble:

Select One: Who is teaching this course?

- ● Antonios Anastasopoulos
- ○ Marie Curie
- ○ Noam Chomsky

If you need to change your answer, you may cross out the previous answer and bubble in the new answer:

Select One: Who is teaching this course?

- ● Antonios Anastasopoulos
- ○ Marie Curie
- ⊗ Noam Chomsky

For "Select all that apply" questions, please fill in all appropriate squares completely:

Select all that apply: Which are scientists?

- □ Stephen Hawking
- ■ Albert Einstein

[1]Compiled on Wednesday 23rd March, 2022 at 17:34 Most questions created by Matt Gormley.

☐ Isaac Newton

☐ None of the above

Again, if you need to change your answer, you may cross out the previous answer(s) and bubble in the new answer(s):

Select all that apply: Which are scientists?

■ Stephen Hawking

■ Albert Einstein

■ Isaac Newton

■ Elon Musk

For questions where you must fill in a blank, please make sure your final answer is fully included in the given space. You may cross out answers or parts of answers, but the final answer must still be within the given space.

Fill in the blank: What is the course number?

688

~~788~~ 688

For questions that require a **numerical answer**, make sure to only write the result in the box. For example, if the question asks you what is the value of some intermediate output o_k, and e.g. you compute that $o_k = 5 * 1 + 3 * (-1) + 0.5 * (-1) = 5 - 3 - 0.5 = 1.5$ then simply write the number:

Correct:

Your Answer

1.5

Wrong:

Your Answer

$o_k = 5 - 3 - 0.5 = 1.5$

Similarly, for **questions asking for an expression**, simply write the final, most simplified expression. e.g. for a question asking what is the partial derivative of $\sigma(z)$ where $z = wx^2$ wrt x, then write:

Correct:

Your Answer

$\sigma(z)(1 - \sigma(z))2x$

Wrong:

Your Answer

$\frac{\partial \sigma(z)}{\partial x} = \frac{\partial \sigma(z)}{\partial z}\frac{\partial z}{\partial x} = \sigma(z)(1 - \sigma(z))\frac{\partial wx^2}{\partial x} = \sigma(z)(1 - \sigma(z))2x$

Feel free to use any white space throughout the exam (or request additional whitepaper if you need to) for any purpose (notes, derivations, etc).

1 Written Questions [99 pts]

1.1 Decision Trees

Perceptron Trees To exploit the desirable properties of decision tree classifiers and perceptrons, Adam came up with a new algorithm called "perceptron trees", which combines features from both. Perceptron trees are similar to decision trees, however each leaf node is a perceptron, instead of a majority vote.

To create a perceptron tree, the first step is to follow a regular decision tree learning algorithm (such as ID3) and perform splitting on attributes until the specified maximum depth is reached. Once maximum depth has been reached, at each leaf node, a perceptron is trained on the remaining attributes which have not been used up in that branch. Classification of a new example is done via a similar procedure. The example is first passed through the decision tree based on its attribute values. When it reaches a leaf node, the final prediction is made by running the corresponding perceptron at that node.

Assume that you have a dataset with 6 binary attributes **(A, B, C, D, E, F)** and two output labels **(-1 and 1)**. A perceptron tree of depth 2 on this dataset is given below. Weights of the perceptron are given in the leaf nodes. Assume bias=1 for each perceptron:

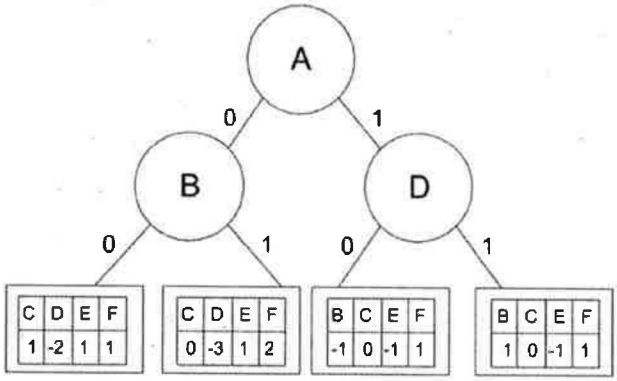


Figure 1.1: Perceptron Tree of max depth=2.

1. (2 points) **Numerical answer**: Given a sample $\mathbf{x}$ = [1, 1, 0, 1, 0, 1], predict the output label for this sample:

Answer

-1

2. (2 points) **True or False**: The decision boundary of a perceptron tree will always be linear.

● True

○ False

3. (2 points) **True or False**: For small values of max depth d (e.g. $d < 6$), decision trees are more likely to underfit the data than perceptron trees.

● True

○ False

4. (2 points) **True or False**: The ID3 algorithm is guaranteed to find the optimal solution for decision trees.

○ True

● False

2 K Nearest Neighbors

5. (2 points) **True or False**: Consider a binary (two classes) classification problem using k-nearest neighbors. We have n 1-dimensional training points $x_1, x_2, \ldots, x_n$ with $x_i \in \mathbb{R}$, and their corresponding labels $y_1, y_2, \ldots, y_n$ with $y_i \in 0, 1$.

Assume the data points $x_1, x_2, \ldots, x_n$ are sorted in ascending order, we use Euclidean distance as the distance metric, and a point can be it's own neighbor. **True or False: We CAN** build a decision tree (with decisions at each node having the form "$x \geq t$" and "$x < t$", for $t \in \mathbb{R}$) that behave exactly the same as the 1-nearest neighbor classifier, on this dataset.

● True

○ False

6. (2 points) **Select all that apply**: Please select all that apply about kNN in the following options: Assume a point can be its own neighbor.

☐ k-NN works great with a small amount of data, but struggles when the amount of data becomes large.

☑ k-NN is sensitive to outliers; therefore, in general we decrease k to avoid over- fitting.

☑ k-NN can only be applied to classification problems, but it cannot be used to solve regression problems.

☐ We can always achieve zero training error (perfect classification) with k-NN, but it may not generalize well in testing.

7. (1 point) **Select one**. Imagine you are using a k-Nearest Neighbor classifier on a data set with lots of noise. You want your classifier to be less sensitive to the noise. Which is more likely to help and with what side-effect?

○ Increase the value of $k \longrightarrow$ Increase in prediction time

● Increase the value of $k \longrightarrow$ Decrease in prediction time

○ Decrease the value of $k \longrightarrow$ Increase in prediction time

○ Decrease the value of $k \longrightarrow$ Decrease in prediction time

2.1 Model Selection and Errors

8. **Train and test errors:** In this problem, we will see how you can debug a classifier by looking at its train and test errors. Consider a classifier trained till convergence on some training data $\mathcal{D}^{\text{train}}$, and tested on a separate dataset $\mathcal{D}^{\text{test}}$. You look at the test error, and you find it to be very high. You then compute the training error, and find it close to 0.

(a) (4 points) **Select all that apply.** Which of the following is expected to help?

- ☐ Increase the training data size.
- ■ Decrease the training data size.
- ☐ Increase model complexity (For example, if your classifier is a neural model, add another hidden layer. Or if it is a decision tree, increase the depth).
- ■ Decrease model complexity.
- ◪ Train on a combination of $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{test}}$ and test on $\mathcal{D}^{\text{test}}$.
- ☐ Conclude that Machine Learning does not work.

(b) (5 points) Explain your choices (only with about one sentence per choice).

Answer

1) The model will be less train so so it will learn less & model become simple (

2) Making the model simple solve the problem to memorize the ~~dat~~ train dataset.

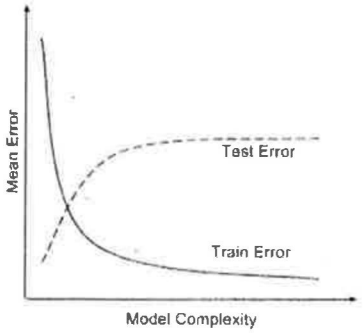
(c) (2 points) What is this scenario called?

Answer

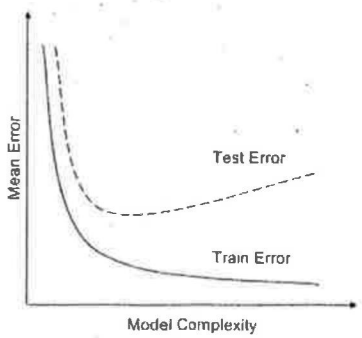
Overfitting

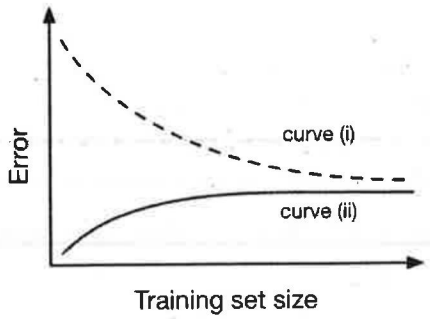
(d) (1 point) Say you plot the train and test errors as a function of the model complexity. Which of the following two plots is your plot expected to look like? **Fill in the circle for (a) or (b).**

○ (a)



● (b)





9. **Training Sample Size:** In this problem, we will consider the effect of training sample size n on a logistic regression classifier with d features. The classifier is trained by optimizing the conditional log-likelihood. The optimization procedure stops if the estimated parameters perfectly classify the training data or they converge.

   The following plot shows the general trend for how the training and testing error change as we increase the sample size $n = |S|$. Your task in this question is to analyze this plot and identify which curve corresponds to the training and test error.

   (a) (1 point) **Select one:** Which curve represents the training error?

   - [x] Curve (i) is the training error, curve (ii) is the test error.
   - [ ] Curve (i) is the test error, curve (ii) is the training error.

   (b) (1 point) Please provide 1–2 sentences of justification for your above choice:

   Answer

   The graph shows that the model is overfitting with loss of data. So the training rate will decrease & test rate will be increase

   (c) (1 point) In one word, what does the gap between the two curves represent?

   Answer

   Model Training

2.2 Perceptron

10. (4 points) **Select all that apply:** Let $S = (\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})$ be n linearly separable points by a separator through the origin in $\mathbb{R}^d$. Let S' be generated from S as: $S' = (c\mathbf{x}^{(1)}, y^{(1)}), \ldots, (c\mathbf{x}^{(n)}, y^{(n)})$, where $c \in \mathbb{Z}^+$ is a constant. Suppose that we would like to run the perceptron algorithm on both data sets separately, and that the perceptron algorithm converges on S. Which of the following statements are true?

    ☐ The mistake bound of perceptron on S' is larger than the mistake bound on S.

    ■ The mistake bound of perceptron on S' will be smaller than the mistake bound on S if $c < 1$, and larger if $c > 1$.

    ☐ The perceptron algorithm when run on S and S' returns the same classifier, modulo constant factors (i.e., if $\mathbf{w}_S$ and $\mathbf{w}_{S'}$ are outputs of the perceptron for S and S', then $\mathbf{w}_S = c_1\mathbf{w}_{S'}$ for some constant c_1).

    ■ The perceptron algorithm converges on S'.

11. (1 point) **True or False:** We know that if the samples are linearly separable, the perceptron algorithm finds a separating hyperplane in a finite number of steps. Given such a dataset with linearly separable samples, select whether the following statement is True or False: The running time of the perceptron algorithm depends on the sample size n.

    ○ True

    ● False

12. **Perceptron Calculation** Suppose you're given the following dataset:

Example Number	X_1	X_2	Y
1	-1	2	-1
2	-2	-2	+1
3	1	-1	+1
4	-3	1	-1

You wish to perform the Perceptron algorithm on this data. Assume you start with initial weights $\theta^T = [0, 0]$, bias $b = 0$, and that we pass all of our examples through in order of their example number. Use a learning rate of 1.

(a) (1 point) **Numerical Answer:** What would be the updated weight vector θ after we pass example 1 through the perceptron algorithm?

Answer

$\theta^T =$ [1, -2]

(b) (1 point) **Numerical Answer:** What would be the updated bias b after we pass example 1 through the perceptron algorithm?

Answer

$b =$ -1

(c) (1 point) **Numerical Answer:** What would be the weight vector θ after we pass example 2 through the perceptron algorithm?

Answer

$\theta^T =$ [1, -2]

(d) (1 point) **Numerical Answer:** What would be the updated bias b after we pass example 2 through the perceptron algorithm?

Answer

$b =$ -1

(e) (1 point) **Numerical Answer:** What would be the updated weight vector θ after we pass example 3 through the perceptron algorithm?

Answer

$\theta^T =$ [1, -2]

(f) (1 point) **Numerical Answer:** What would be the updated bias b after we pass example 3 through the perceptron algorithm?

Answer

$b =$ -1

(g) (1 point) **True or False:** Your friend stops you here and tells you that you do not need to update the Perceptron weights or the bias anymore. Is this true or false?

○ True

● False

13. (2 points) **True or False:** Dataset (X, Y) has a non-linear decision boundary. Fortunately, there is a function $\mathcal{F}$ that maps (X, Y) to $(\mathcal{F}(X), Y)$, which is linearly separable. We have tried to build a modified perceptron algorithm to classify (X, Y). Is the given (modified) perceptron update rule correct?

$$\text{if } \text{sign}(w\mathcal{F}(x^i) + b) \neq y^i :$$
$$w' = w + y^i \mathcal{F}(x^i)$$
$$b' = b + y^i$$

● True

○ False

2.3 Linear Regression

14. (3 points) **True or False:** Given data $D = (x_1, y_1), \ldots, (x_n, y_n)$, we obtain $\hat{w}$, the parameters that minimize the training error cost for the linear regression model $y = w^T \mathbf{x}$ we learn from D.

    Consider a new dataset D_{new} generated by duplicating the points in D and adding 10 points that lie along $y = \hat{w}^T \mathbf{x}$. Then the $\hat{w_{new}}$ that we learn for $y = w^T \mathbf{x}$ from D_{new} is equal to $\hat{w}$.

    ● True

    ○ False

15. (5 points) **Fill-in the blanks** Given that we have an input x and we want to estimate an output y, in linear regression we assume the relationship between them is of the form $y = wx + b + \epsilon$, where w and b are real-valued parameters we estimate and ϵ represents the noise in the data. When the noise is Gaussian, maximizing the likelihood of a dataset $S = (x_1, y_1), \ldots, (x_n, y_n)$ to estimate the parameters w and b is equivalent to minimizing the squared error:

$$\arg\min_w \sum_{i=1}^{n} (y_i - (wx_i + b))^2.$$

Consider the dataset S plotted in Fig. 2.1 along with its associated regression line. For each of the altered data sets S^{new} plotted in Fig. 2.2, indicate which of the three regression line (a,b, or c; relative to the original one) in Fig. 2.3 corresponds to the regression line for the new data set. Write your answers in the table below.

Fill in this table

Dataset S^{new}	(a)	(b)	(c)	(d)	(e)
Regression Line	b	c	c	a	a

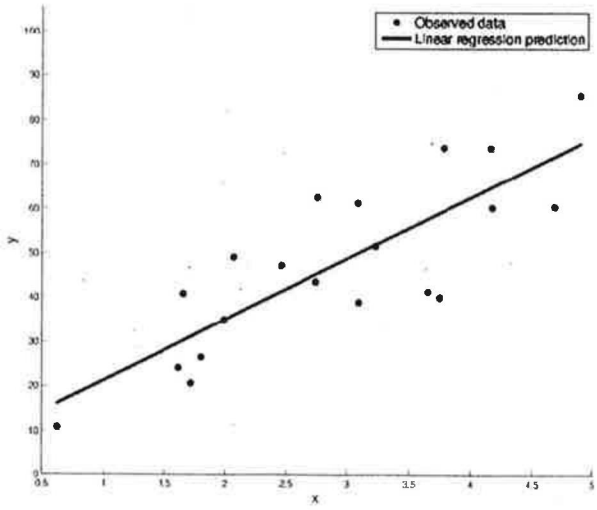
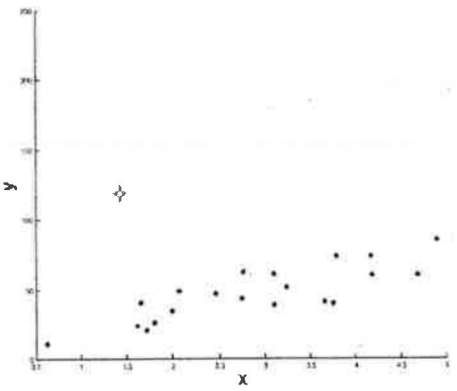
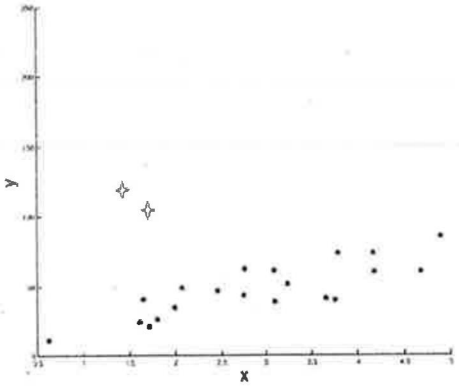


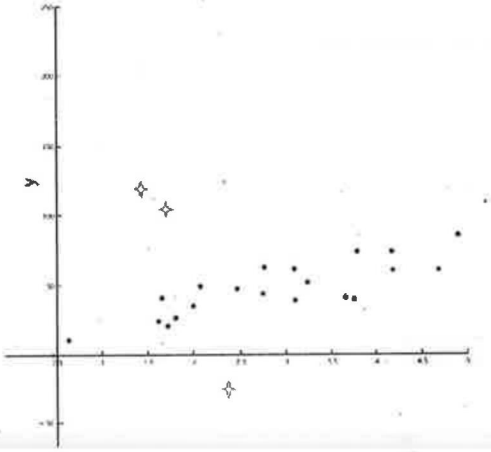
Figure 2.1: An observed data set and its associated regression line.



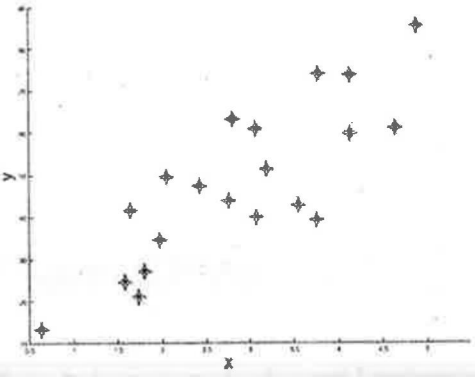
(a) Adding one outlier to the original data set.



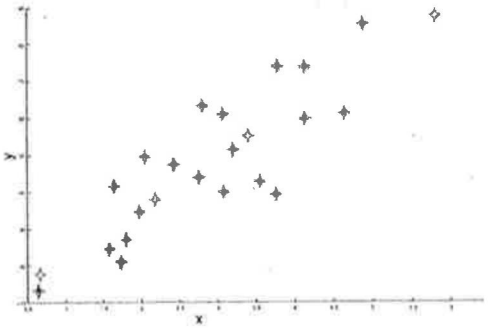
(b) Adding two outliers to the original data set.



(c) Adding three outliers to the original data set. Two on one side and one on the other side.

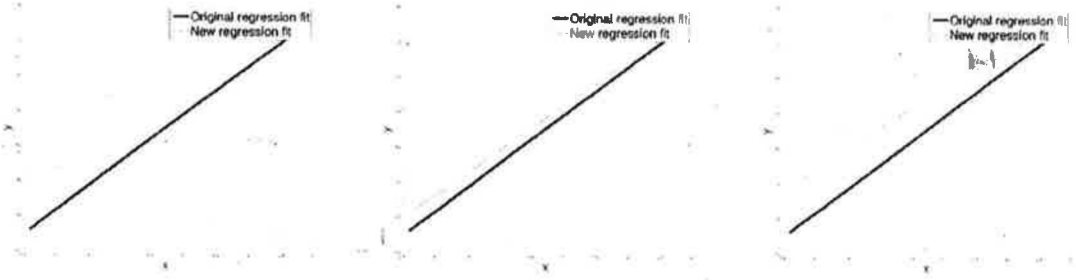


(d) Duplicating the original data set.



(e) Duplicating the original data set and adding four points that lie on the trajectory of the original regression line.

Figure 2.2: New datasets S^{new}.



(a) Old and new regression lines. (b) Old and new regression lines. (c) Old and new regression lines.

Figure 2.3: New regression lines for altered data sets S^{new} compared to the original regression.

2.4 Optimization

16. (2 points) **Select all that apply:** Which of the following are correct regarding Gradient Descent (GD) and stochastic gradient descent (SGD)

    ■ Each update step in SGD pushes the parameter vector closer to the parameter vector that minimizes the objective function.

    ☐ The gradient computed in SGD is, in expectation, equal to the gradient computed in GD.

    ☐ The gradient computed in GD has a higher variance than that computed in SGD, which is why in practice SGD converges faster in time than GD.

17. Consider the convex function $f(z) = z^2$. Let α be our learning rate in gradient descent.

    (a) (2 points) For which values of α will $\lim_{t\to\infty} f(z^{(t)}) = 0$, assuming the initial value of z is $z^{(0)} = 1$ and $z^{(t)}$ is the value of z after the t-th iteration of gradient descent? **Select all that apply:**

    ■ $\alpha = 0$

    ■ $\alpha = 1$

    ☐ $\alpha = 2$

    ■ $\alpha = \frac{1}{2}$

    (b) (1 point) Also, give the range of all values for $\alpha \geq 0$ such that $\lim_{t\to\infty} f(z^{(t)}) = 0$, assuming the initial value of z is $z^{(0)} = 1$. Be specific.

    Answer

    $1 \geq \alpha > 0$

18. (2 points) **Select all that apply:** Which of the following are correct regarding Gradient Descent (GD). Assume data log-likelihood is $L(\theta|X)$, which is a function of the parameter θ, and the objective function is negative log-likelihood.

    ☐ GD requires that $L(\theta|X)$ is concave with respect to parameter θ in order to converge

    ■ GD requires that $L(\theta|X)$ is convex with respect to parameter θ in order to converge

    ■ GD update rule is $\theta \leftarrow \theta - \alpha\nabla_\theta L(\theta|X)$.

    ☐ Given a fixed small learning rate (say $\alpha = 10^{-10}$), GD will always reach the optimum after infinite iterations (assume that the objective function satisfies the convergence condition).

2.5 MLE/MAP

19. (1 point) **True or False:** The MAP estimate is always better than the MLE.

    ○ True

    ● False

20. (1 point) **True or False:** In the limit as n (the number of samples) increases, the MAP and MLE estimates become the same.

- [x] True
- [ ] False

21. (1 point) **True or False:** The bias of the Maximum Likelihood Estimate (MLE) is typically less than or equal to the bias of the Maximum A Posteriori (MAP) Estimate.

- [x] True
- [ ] False

22. Assume we have a random variable that is Bernoulli distributed: $X_1, \ldots, X_n \sim Bernoulli(\theta)$. We are going to derive its MLE. Recall that in a Bernoulli $X = 0, 1$ and the pdf of a Bernoulli is

$$p(X; \theta) = \theta^x (1-\theta)^{1-x}.$$

(a) (2 points) Derive the likelihood $L(\theta; X_1, \ldots, X_n)$

Answer

$$L(\theta, x_1 \ldots x_n) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}$$

$$\sum_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}$$

(b) (2 points) Derive the formula for the log likelihood $l(\theta; X_1, \ldots, X_n)$:

Answer

$$G = Ln(L(\theta, x_1 \ldots x_n)) = Ln\left(\sum_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}\right)$$

$$= \sum_{i=1}^{n} \left[x_i Ln(\theta) + (1-x_i) Ln(1-\theta) \right]$$

(c) (3 points) Derive the MLE, $\hat{\theta}$, and show that it is $\hat{\theta} = \frac{1}{n}(\Sigma_{i=1}^{n} X_i)$.

G = log-like. from last question

Answer

$$\frac{\partial G}{\partial \theta} = \sum_{i=1}^{n} \left[x_i \frac{1}{\theta} + \frac{1}{1-\theta}(1-x_i)(-1) \right]$$

$$0 = \sum_{i=1}^{n} \left[(1-\theta)x_i - \theta(1-x_i) \right]$$

$$= \sum_{i=1}^{n} \left[x_i - \theta x_i - \theta + \theta x_i \right]$$

$$= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \theta$$

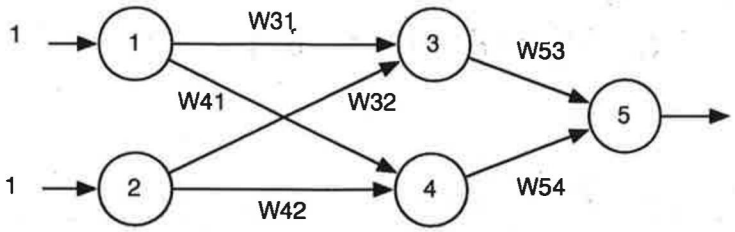
$$n\theta = \sum_{i=1}^{n} x_i \Rightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

[Space intentionally left empty (for calculations etc). More questions in next page.]

2.6 Neural Networks

23. Apply the backpropagation algorithm to update the weight values of a feedforward neural network (given below) after providing in input the vector $\mathbf{x} = (1, 1)$ with target value $t = 0$.

The network has a single hidden layer with two units, one output unit, and all weights are initialized to zero. All hidden and output units use the sigmoid activation function. Assume the learning rate $\eta = 0.1$.



Details:

- *Input units*: Each input unit j simply propagates its input. i.e. the output of the unit 1 is $o_1 = x_1$ and the output of the unit 2 is $o_2 = x_2$.
- **Write your numerical answers in full precision, i.e. do not round any answers. You may use fractions (e.g. you may write either $\frac{1}{32}$ or 0.03125).**
- Note that there are no bias parameters in this simple network.

(a) (3 points) **Numerical answer.** *Forward pass*: For each hidden and output unit j, compute: $o_j = \sigma(\sum_i w_{ji} o_i)$;

o_3	o_4	o_5
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$

(b) (2 points) Assume we're using the quadratic loss function, such that $J(\theta) = \frac{1}{2}(t - o_5)^2$. First, let's derive and calculate the derivative with respect to the output. What is the expression for $\frac{\partial J}{\partial o_5}$? What does it evaluate to?

Expression of $\frac{\partial J}{\partial o_5}$	Value of $\frac{\partial J}{\partial o_5}$
$J(\theta) = \frac{1}{2}(t - o_5)^2$ $\frac{\partial J}{\partial o_5} = (t - o_5)(-1) = (o_5 - t)$ $= o_5$ $(t = 0$	$\frac{1}{2}$

(c) (4 points) Now, we will derive the gradients associated with the last layer, i.e. the ones for w_{53} and w_{54}. Use the chain rule. You don't need to expand anything that you have computed before. Derive and calculate the derivative with respect to w_{53}. What is the expression for $\frac{\partial J}{\partial w_{53}}$? What does it evaluate to?

Expression of $\frac{\partial J}{\partial w_{53}}$	Value of $\frac{\partial J}{\partial w_{53}}$
Let $g = \sigma(o_3 w_{53} + o_4 w_{54})$ $\frac{\partial J}{\partial w_{53}} = \frac{\partial J}{\partial o_5} \cdot \frac{\partial o_5}{\partial g} \frac{\partial g}{\partial w_{53}}$ $= o_5 \cdot o_3 \cdot (g)(1-g)$	$\frac{1}{8}$

Let's do the same process for w_{54}:

Expression of $\frac{\partial J}{\partial w_{54}}$	Value of $\frac{\partial J}{\partial w_{54}}$
Let $g = \sigma(o_3 w_{53} + o_4 w_{54})$ $\frac{\partial J}{\partial w_{54}} = \frac{\partial J}{\partial o_5} \cdot \frac{\partial o_5}{\partial g} \frac{\partial g}{\partial w_{54}}$ $o_5 \cdot o_4 \cdot (g)(1-g)$	$\frac{1}{8}$

(d) (4 points) Backpropagate through the hidden units. Let's derive and calculate the derivative with respect to the intermediate outputs o_3 and o_4. Use the chain rule. You don't need to expand anything that you have computed before.

What is the expression for $\frac{\partial J}{\partial o_3}$? What does it evaluate to?

Expression of $\frac{\partial J}{\partial o_3}$	Value of $\frac{\partial J}{\partial o_3}$
Let $g = \sigma(o_3 w_{53} + o_4 w_{54})$ $\frac{\partial J}{\partial o_3} = \frac{\partial J}{\partial o_5} \cdot \frac{\partial o_5}{\partial g} \cdot \frac{\partial g}{\partial o_3}$ $= o_5 \cdot w_{53} \cdot (g)(1-g)$	$\frac{1}{8}$

What is the expression for $\frac{\partial J}{\partial o_4}$? What does it evaluate to?

Expression of $\frac{\partial J}{\partial o_4}$	Value of $\frac{\partial J}{\partial o_4}$
Let $g = \sigma(o_3 w_{53} + o_4 w_{54})$ $\frac{\partial J}{\partial o_4} = o_5 \cdot w_{54} \cdot (g)(1-g)$	$\frac{1}{8}$

(e) (2 points) Last, let's compute the gradients with respect to the weights of the first layer. You will only have to write this our completely for one of the weights: w_{32}. Use the chain rule. You don't need to expand anything that you have computed before.

Derive and calculate the derivative with respect to w_{32}. What is the expression for $\frac{\partial J}{\partial w_{32}}$? What does it evaluate to?

Expression of $\frac{\partial J}{\partial w_{32}}$	Value of $\frac{\partial J}{\partial w_{32}}$
let $g = \sigma(o_3 w_{53} + o_4 w_{54})$ let $h = \sigma(w_{32} o_2 + w_{31} o_1)$ $\frac{\partial J}{\partial w_{31}} = o_5 \cdot w_{53} \cdot (g)(1-g)(h)(1-h) o_2$	$\frac{1}{16}$

(f) (6 points) Now let's compute the new weights, using the update rule: $w_{ji}^{new} = w_{ji}^{old} - \eta \frac{\partial J}{\partial w_{ji}^{old}}$. Compute the new values for all weights:

new w_{54}

$$\frac{1}{8} - 0.1 \times \frac{1}{8} = \frac{1}{8}(0.9)$$

new w_{53}

$$\frac{0.9}{8}$$

new w_{42}

$$\frac{0.9}{16}$$

new w_{41}

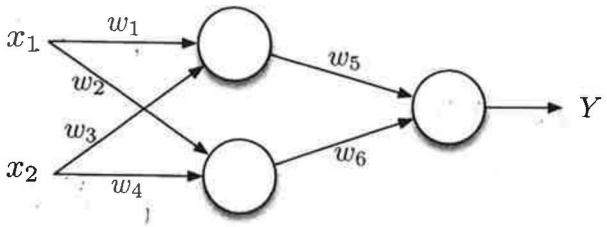
$\frac{0.8}{8}$ $\frac{0.9}{16}$

new w_{32}

$\frac{0.9}{16}$

new w_{31}

$\frac{0.8}{8}$ $\frac{0.9}{16}$



24. Consider the above neural network. There are two possible choices for the activation function, as implemented by each unit in this network:

    S: The signed sigmoid function $S(\alpha) = sign[\sigma(\alpha) - 0.5] = sign\left[\frac{1}{1+e^{-\alpha}} - 0.5\right]$;

    L: A linear function $L(\alpha) = c\alpha$, for some constant c,

    where in both cases α is the intermediate output of the unit (after the dot product with its corresponding weights but before the activation function).

    Now, assume that you want your neural network to simulate exactly a two-class logistic regression classifier i.e., one where the output $Y = \text{argmax}_y\, P(Y = y|\mathbf{x})$, where $P(Y = 1|\mathbf{x}) = \frac{1}{1+e^{-(\beta_1 x_1 + \beta_2 x_2)}}$, and $P(Y = -1|\mathbf{x}) = \frac{e^{-(\beta_1 x_1 + \beta_2 x_2)}}{1+e^{-(\beta_1 x_1 + \beta_2 x_2)}}$.

    (a) (6 points) What activation function would you use for the activation functions (simply write S or L) for each unit in the given graph?

Hidden Unit 1 (top)	Hidden Unit 2 (bottom)	Output Unit
L	L	S

    (b) (5 points) For your above answer, derive β_1 in terms of $w_1, w_2, \ldots, w_6$ and the c constants. Show your calculations:

β_1 derivation

$HU1 = (x_1 w_1 + w_3 x_2)\, c$

$HU2 = (x_1 w_2 + x_2 w_4)\, c$

$out = sign\left(\sigma\left(\frac{1}{1+e^{-(HU1 \cdot w_5 + HU2 \cdot w_6)}} - 0.5 \right) \right)$

$$\frac{1}{1+e^{-\left(w_5 (x_1 w_1 + w_3 x_2) c + w_6 (x_1 w_2 + x_2 w_4) c - 0.5\right)}}$$

$$\frac{1}{1+e^{-\left((c(w_5 w_1 + w_6 w_2) x_1 + c(w_5 w_3 + w_4 w_6) x_2 - 0.5\right)}}$$

$\beta_1 = c(w_5 w_1 + w_6 w_2)$

[This page is intentionally left blank]

25) 70 (Around)

26) 60% - 70%.