ECS171 Winter 2025
Midterm Study Guide

Cheatsheet:

Sigmoid and derivative of sigmoid:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\sigma'(x) = \frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x))$$

ReLU and Derivative for ReLU:

$$f(x) = max(0, x) =$$

ReLU'(x)= $$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Gradient descent weight update: $W^{t+1} = W^t - \lambda \cdot \nabla J(W^t)$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

$$SSE = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

Variance = $$\sigma^2 = \frac{\sum(xi - \bar{x})^2}{N}$$

$$h(x) = f(g(x))$$

Chain rule: $$h'(x) = f'(g(x)) \cdot g'(x)$$

Q1:
In machine learning, what is the dropout technique, and how does it help prevent overfitting in neural networks?

Q2:
Given the hours of exercising per week measured in hours (x_1) and time from last Covid-19 infection measured in weeks(x_2), the model predicts the probability that the person will be re-infected with Covid-19 in the next 5 months. The model follows the Linear Regression $\hat{y} = 0.4 - 0.05x_1 + 0.07x_2$, for each of the data pairs in the training and testing sets, do the following:

(a) Compute the predicted output for the given regression model for each set of input
(b) Compute the bias (SSE) for each set of inputs
(c) Compute the variance.
(d) Is the model overfit, underfit, or a good fit? Justify your answer using the following metrics for the base case:
SSE_train = 0.050
SSE_test = 0.049
Variance = 0.01

Where needed, round your answer to 4 d.p.

Training set:

x_1	x_2	y	$\hat{y}$	
0.0	4.0	0.70		
3.0	10.0	0.95		
2.0	3.0	0.60		
5.0	1.0	0.15		
8.0	5.5	0.25		
12.0	7.5	0.23		
10.0	4.0	0.20		
3.0	2.0	0.30		

Testing set:

x_1	x_2	y	$\hat{y}$	
1.0	9.0	0.95		

9.0	6.0	0.20		
7.0	3.0	0.25		
5.0	5.0	0.45		

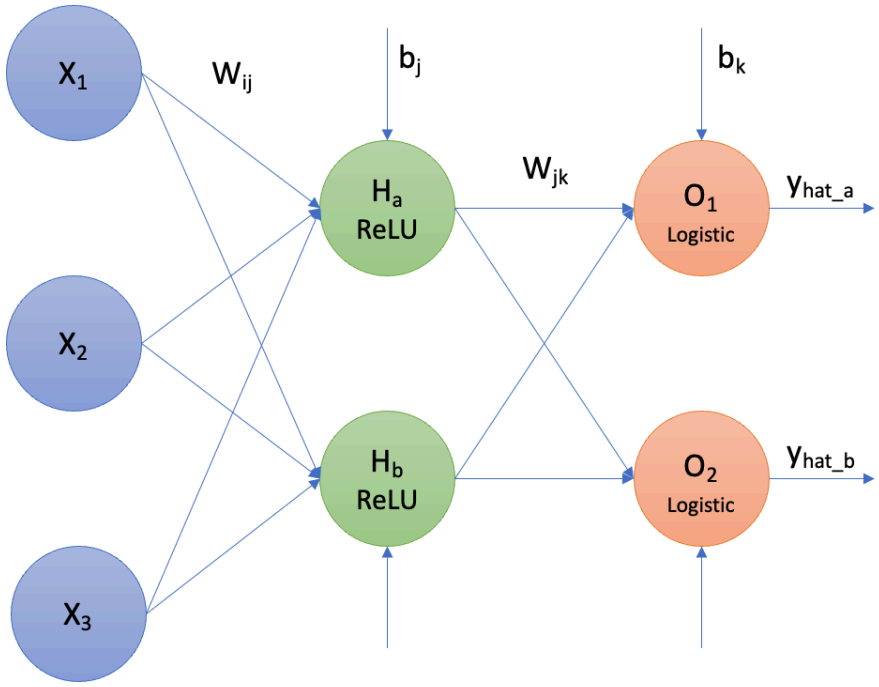
Testing set:

x_1	x_2	y	$\hat{y}$	
1.0	9.0	0.95		
9.0	6.0	0.20		
7.0	3.0	0.25		
5.0	5.0	0.45		

Q3:
Find the coefficients of a polynomial with degree 2 which gives the lowest mean square error.
y_predicted = ax^2 + bx + c
x= [1,2,3]
y_actual = [4, 13, 20]
Coefficient Set 1: a= 1, b=2, c=3
Coefficient Set 2: a=3, b=5, c=2
Coefficient Set 3: a=1, b=1, c=5

Answer:

Use the graph below to answer Q4-6:



Q4:
If we designed this neural network with ReLU add after each hidden neural as activation function, and logistic(sigmoid) add in the end before we get the predicted value(y_hat), use the table below to calculate that which class does x1 = 7, x2 = 10, and x3 = 9 belongs. (Assume that as the result of one output has a value over a threshold τ = 0.9, it will be classified into that class.)

W_1a = 0.4	W_a1=0.2	b_a=0.7
W_1b =-0.7	W_a2=0.5	b_b=0.4
W_2a=0.6	W_b1=-0.1	b_1=0.9
W_2b=-0.5	W_b2=0.6	b_2=-0.8
W_3a=0.3		
W_3b=0.7		

Q5:
if the data points mentioned in Q7 have y_a= 1 and y_b = 0, update weight w_a1, w_a2, w_1a, and w_3a, consider the learning rate η =0.2. Assume we use SSE for the loss of this question.

Q6:
Similarly, update all the biases.

Q7:
In logistic regression, what is the hypothesis function and what does the predicted output of the hypothesis function represent, given an input data point x(1)?

Q8:
Joshua claims that in Machine Learning, most of the data are used in testing because accuracy in predicting unseen data is more important than the accuracy of the seen data in the training set. Do you agree with this claim? Justify your answer.

Q9:
Can OLS method be used to train a logistic regression model as a common practice? Explain your answer.

Q10:
Show mathematically how to obtain the weight of a linear regression model with attribute X using the OLS method.

Q11:
What is the number of neurons in the input layer of an ANN if the number of attributes in the dataset is 3?

Q12:
Classify the feature based on the weight and threshold given below:

Feature 1	**Classification**

0.5	
0.7	
1.1	
1.5	
1.3	

The weight is as follows:
Weight combination 1: w1 = 0.5.
Threshold: t = 0.6
Use the sigmoid function to compute the probability.

Q13:
Given a dataset D that has 6 data points shown in the table below, you are going to develop a least mean square regression model in the form of $y\hat{} = w_0 + w_1 \cdot x_1 + w_2 \cdot x_2$ where the weights $W = [w_0, w_1, w_2]$ makes the model have minimum error.

x_1	x_2	y
4	1	2
2	8	-14
1	0	1
3	2	-1
1	4	-7
6	7	-8

Use the gradient descent method to update the weights W of the regression model. The current weights are $W^t = [0.1, 0.2, 0.3]$, learning right $\lambda = 0.02$.
Include all your calculations for one round of weight updates. Round to 2 decimal places.
Hint: use the formula:

$$\frac{\partial J}{\partial w_j} = -\sum_{i=1}^{n} \left(y_i - y_i\hat{}\right) x_{ij}$$

to calculate the gradient of the cost function and to update the weights use the formula:

$$W^{t+1} = W^t - \lambda \cdot \nabla J\left(W^t\right)$$

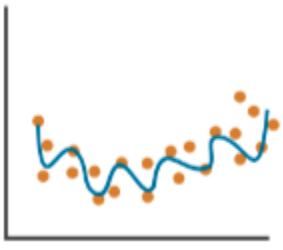
Q14:
What is the difference between Batch GD (Gradient Descent) and Stochastic GD? Why is Stochastic more widely used compared to Batch GD and Newton's method?

Q15:
In batch gradient descent, if the number of batches is equal to the number of observations in the training dataset, the gradient descent approach is the same as “Stochastic gradient descent ”.

Q16:
In the context of training artificial neural networks, which of the following best describes the role of gradient descent?
a) It's a type of activation function applied to the neurons.
b) It's the process of adding layers to the neural network.
c) It's an optimization algorithm used to minimize the error by adjusting the weights.
d) It's a method to regularize the network and prevent overfitting.

Q17:
Which of the following statements about L1 and L2 regularization is correct?
A. L1 regularization adds the squared value of the loss to the weight update
B. L2 regularization adds the absolute value of the weight to the loss function
C. L1 regularization adds the absolute value of the weight to the loss function
D. L2 regularization adds the squared value of the loss to the weight update

Q18:
Ture or False: This graph on the right is an example of overfitting



Q19:
Ture or False: Newton's method is a method that completely outperforms gradient descent in any setting because it can always find the minimum loss function value with fewer weight updates.

Q20:
Which of the following is the correct formula to find the weight of a linear regression model with the OLS method.

A. $X^T Y$

B. $Y^T Y$

C. $X^T X$

D. $X^T X$

Q21:
A company conducted a study to analyze the performance of two machine learning models, Model A and Model B, on a dataset of 500 instances. The confusion matrices for both models are as follows:

Model A:	Predicted Positive	Predicted Negative
Actual Positive	50	10
Actual Negative	20	420

Model B:	Predicted Positive	Predicted Negative
Actual Positive	60	40
Actual Negative	30	370

Using these confusion matrices, calculate and compare the following performance metrics for Model A and Model B: Accuracy, Precision for the positive class, Recall (Sensitivity) for the positive class, F1 score for the positive class

Based on these metrics, determine which model (A or B) performed better on the dataset