

COMPSCI 4ML3
Instructor: Hassan Ashtiani
Final Exam
April 2023

Surname: _____

Given Name: _____

Student Number _____

- This examination paper includes 9 pages (including this cover page and the last blank page) and 5 questions. **You are responsible for ensuring that your copy of the papers is complete. Bring any discrepancy to the attention of your invigilator.**
- Examination duration is 120 minutes.

Exam Instructions:

- SINGLE VERSION exam

Materials Permitted in the Exam Venue:

- No electronic aids are permitted, e.g. laptops, phones
- McMaster Standard Calculator (Casio FX-991MS/MS+)

Materials to be supplied to the students:

- Scrap paper

Instructions to the students:

- If you think there is an issue with one of the questions, make the best sensible assumption and write your assumption down along with your solution.
-

Grade Table

Question	Points	Score
1	45	
2	15	
3	15	
4	10	
5	15	
Total:	100	

1. For each multiple choice question, there is only one correct answer. Mark/check your choice clearly.

- (a) (5 points) We have a labeled data set $(x^i, y^i)_{i=1}^m$ of m points where $y^i \in \{-1, +1\}$. The following is the optimization formulation is for which model?

$$\min_W \left(\lambda \|W\|_2^2 + \frac{1}{m} \sum_{i=1}^m \max\{0, 1 - y^i W^T x^i\} \right)$$

- Regularized Least Squares
 Hard Margin SVM
 Soft Margin SVM
 Perceptron
- (b) (5 points) Which one is the most reasonable choice for linear classification when data is not linearly separable?

- Using linear programming
 Hard Margin SVM
 Soft Margin SVM
 Perceptron

- (c) (5 points) Let TP, TN, FP, FN be the number of true positives, true negatives, false positives, and false negatives respectively. The “precision” of a model is defined by

- $\frac{TP}{TP+TN+FP+FN}$
 $\frac{TP}{TP+FP}$
 $\frac{TP}{TP+TN}$
 $\frac{TP+TN}{TP+TN+FP+FN}$
 $\frac{TP+FP}{TP+TN+FP+FN}$

- (d) (5 points) Which one of these choices is the most accurate about multiclass SVM?
- The number of linear classifiers trained for one-versus-all is more than the number of linear classifiers trained for the all pairs strategy
 There is an end-to-end version of all-pairs SVM that can be trained in end-to-end fashion
 There is an end-to-end version of one-versus-all SVM that can be trained in an end-to-end fashion

- (e) (5 points) Which one is more accurate?
- Universal approximation theorem indicates that you need RELU activation function rather than sigmoid to approximate any function with neural networks.
 - Any bounded continuous function can be approximated (with any desired accuracy) with a neural network with one layer (i.e., one input layer and no hidden layers).
 - Any sigmoid neural network with 10 layers and any number of parameters can be approximated by a sigmoid network with only 1 hidden layer.
- (f) (5 points) Which is more accurate regarding gradient descent (GD) and stochastic gradient descent (SGD)?
- Unlike GD, SGD does not require backpropagation since the gradient is computed as the sum of gradients of individual data points.
 - SGD is more likely to get stuck in a local minimum since each update only uses a few data points.
 - SGD makes more updates in an epoch (compared to GD)
- (g) (5 points) Which one is NOT a usual strategy for improving the “vanishing gradient” issue?
- Using backpropagation for computing the gradient
 - Using RELU activations rather than sigmoid
 - Using a residual architecture (ResNET)
- (h) (5 points) Which one is NOT a usual strategy for regularization in neural networks?
- Early stopping
 - Dropout
 - Weight sharing
 - Using a ResNET architecture
- (i) (5 points) We apply a 3x3 convolution kernel to a 10x10 image with no padding, and stride=2. What would be the size of the output image?

Student Name:

Student Number:

2. (15 points) Clearly write TRUE or FALSE besides each choice.

- Boosting is likely to reduce overfitting.

- Logistic regression is often used for classification rather than regression.

- Random Forests combine the idea of boosting and bagging

- Vapnik-Chrvoenkis dimension (VC dimension) of the class of all linear classifiers in \mathbb{R}^d is infinite.

- We have trained a fully connected neural network for classifying images and observe that its accuracy on test data is 50%. We would like to improve this to 70%. We also observe that the accuracy on training data is 85%. A sensible solution is to use dropout.

3. In this questions we want to study the relationship between smoking and having a heart attack.

Assume that the ratio of people who smoke to the whole population is C (in other words, $\Pr[x \text{ is a smoker}] = C$ where x is a random person generated from the population).

Assume that the probability of heart attack for smokers during their lifetime is 0.3, but this probability is 0.1 for non-smokers.

- (a) (5 points) What is the probability of having a heart-attack for a randomly selected individual (uniform over the whole population, including smokers and nonsmokers) assuming $C = 0.3$? Show your work and simplify the final solution.

Student Name:

Student Number:

- (b) (10 points) We observe that from a randomly selected set of 100 people, 20 of them have experienced heart attack during their lifetime. What is the maximum likelihood estimate for the value of C in this case? Show your work and simplify the final solution.

Student Name:

Student Number:

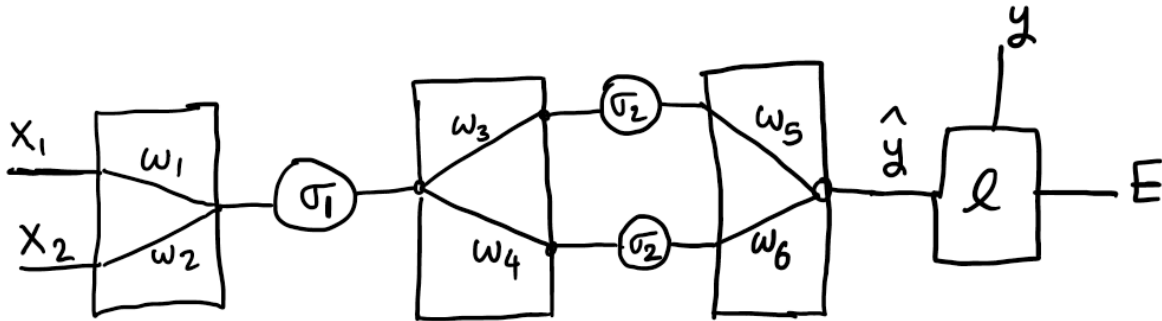
4. (10 points) We are trying to fit the homogeneous line $y = a.x$ to a non-linear curve, $f(x) = x^2 + 1$ where $a, x, y \in \mathbb{R}$.

Assume that the x values are uniformly distributed on $[0, 1]$. What should we pick for the value of a in order to minimize the squared error? Show your work and simplify your final answer.

5. (15 points) The following neural network has three layers and is used for regression. Here are the details:

1. The activation functions of the first and second layers are ReLU (i.e., $\sigma_1(x) = \sigma_2(x) = \max(0, x)$) but there is no activation function in the third layer.
2. The loss function is the squared loss: $l(y, \hat{y}) = (y - \hat{y})^2$.
3. $w_1 = 1, w_2 = 1, w_3 = 0.1, w_4 = 0.2, w_5 = -2, w_6 = 5$.
4. The given input is specified by $x = [x_1, x_2]^T = [5, 5]^T, y = 10$

Compute the partial derivative of error E with respect to w_1 for the given input point ($x = [5, 5]^T, y = 10$). Show your work.



Student Name:

Student Number:

The End