NAME: ID:

About CS412 Exam

Some sample questions from many (but not all) of the topics are given below. **No more samples will be given, as these are sufficient to give an idea about depth and style of the exam.**

In general, question points reflect amount of time spent on that topic in the course. I.e. if I have spent a lot of time on a topic or emphasized it (e.g. gradient descent), I will ask more about it. Roughly, 10-15 points on each main topic and some general True/False questions at the end.

I ask from all topics covered, so that the exam reflects the full content and not just focus on a narrow part.

Cheat sheet is not allowed, but some reminders will be given (e.g. for sigmoid function, threshold activation function (only for the issue of thr(0)), or Normal distribution formula) **but no reminders will be given for many other things -** things that you should know from the top of your head. (e.g. definition of entropy of a rand. var, binary cross entropy loss, convolution, net input and activation of a neuron,…)

Note: These were from exams in the last few years and questions in this year should be similar for the most part, but there may be more challenging other questions.

NAME: ID:

1) 12pt – Basic Math for ML

a) 3pts – You have trained a linear regression model $\hat{f}(x) = a + bx$ that minimizes the squared distance between the target value y and predicted value $\hat{f}(x)$, for a sample x.

What is the **mean square error** of the estimates $\hat{f}$(x) over the given test set T ={(x_i,y_i)}, i=1..M.

Give as a formula, no partial for errors/missing details.

MSE = ...

b) 3pts – Apply gradient descent to **minimize** the function f(x,y) = x^3 + 2xy + 4 starting at the point p_0 = (x y)= $[1\ 2]^T$. Use a learning rate of 0.1 and just go one step.

Gradient =

p_1 =

.....

2) Decision trees

You *may* use this log table in some part of this question:

x	$\log_2$ x
1/4	-2
1/3	-1.6
1/2	-1
2/3	-0.6
3/4	-0.4
1	0

a) 2pt – What is the entropy of a random variable that represents the output of an **X-sided fair dice?** (hence; possible outputs are 1,2,…,X with **equal probability**; if X is 6, we have the usual dice). Do not just give the final answer, show your work.

NAME: ID:

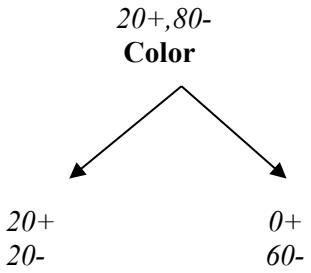
b) 2pt – **How would the entropy in a) change if the dice was biased towards one of the outcomes** (for example, the probability of having a 1 is higher than the rest of the outputs)?

The entropy would…….

a) decrease.
b) increase.
c) remain unchanged.

Circle the appropriate answer. No explanation necessary.

c) 6pt – What is the **remaining entropy** after the tree is split according to the feature "**Color**". The +s represent one class, and –s represent another class. There are a total of 100 samples at the root of the tree.
Give numeric answers. You can directly write the entropy values without showing your work for the left and right entropies, but show your work for remaining entropy.



2pt - Entropy at left leaf = …….. *2pt - Entropy at right leaf = ……..*

2pt – Remaining Entropy = ……

3) 10pt – K-NN

a) 4pts – You have samples (x,y) coming from the underlying function: y= 10x + e where e is zero-mean noise. Assume the samples are (10, 90), (20, 200) and (30, 320).

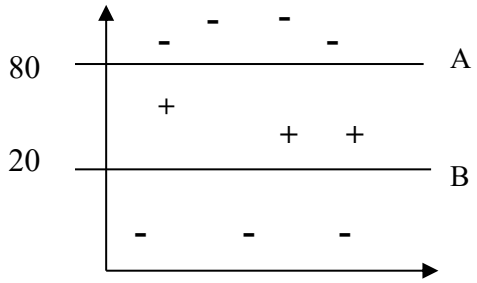
What is the predicted y value for x=22, …

- When k=1 : **Answer**: ……
- When k=2 : **Answer**: ……

NAME: ID:

4) Neural Networks

a) 8pt – We are given the following binary classification problem, where **+ is the positive class and – is the negative class.**



i) 4pt –What is the **weight vectors w_A and w_B,** corresponding to the following two decision boundaries A and B, respectively?
Draw them on the figure as well.

w_A = w_B =

ii) 4pts – Give the *architecture, weights and biases* of the *full network* that uses the above weight vectors and can classify a given **input (x,y)** appropriately, as belonging to the + or – class.

b) 4pt - Assume we are dealing with a <u>regression</u> problem. Complete the following derivative that shows how the squared error $E_p=(t_p - o)^2$ for a pattern p ,with target t_p, changes with changes in weight w_i of the **output unit**. o is the output of the system. Make sure to show your derivation. Hint: Oneline differentiation.

$\delta E_p/\delta w_i$ =

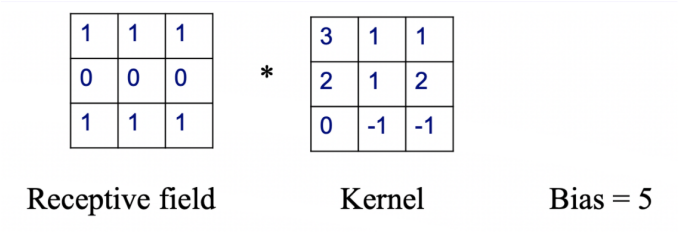
c) 8pts - True/False (2pt each) – -1pt for each wrong answer.

- **T / F** A neuron with a saturated sigmoid activation (close to 1 or 0) will learn very quickly.
- **T / F** Linearly non-separable problems (e.g. XOR) can be solved with two layers of *linear* units (neurons with linear activations).
- **T / F** A shortcoming of gradient descent based methods, such as backpropagation, is that they may get stuck in local minima.

NAME: ID:

5) Deep Learning

a) 3pts - What is the **net input and output of a node with RELU activation**, for the given receptive field, filter weights (kernel) and bias?



2pts - Net input:: ..... (no partial so be careful) 1pt - **Output**: .....

b) 3pts – Convolutional Networks: Consider an input size of 200x200x3 to a convolutional network. What is the size of the first conv layer (right after the input layer), if it uses 5x5 filters with a stride of 1 and that there are 100 feature maps in the layer?

Write as WxHxDepth without spaces.

Answer: .....

NAME: ID:

6) Bayes Classifiers

a) **4pt –** Assume you are given the following conditional probability distributions for left handedness with respect to gender.

P(LeftHanded |Male) = 0.3. P(LeftHanded |Female) = 0.2

Compute the probability of picking/encountering a LeftHanded person from the whole population. You must show your work. Assume:

P(Male) = 0.4 P(Women)=0.6

Answer: ……

NAME: ID:

7) General ML Understanding

a) **8pts –** 1pts each - Each incorrect answer cancels **half a correct answer**.

Note: A sentence that does not use the terms **typically**, **generally** or **expected etc**, **claims to always be true/valid.** If that is not the case, do not choose it as correct.

1. **T / F** A decision tree or random forest does not need **scale-normalized features**.
2. **T / F** A K-NN classifier using Euclidean distance does not need **scale-normalized features**.
3. **T / F** If there is no intrinsic error in the training data, k-NN with k=1 will always have zero **training set error**.
4. **T / F** More complex models with larger number of parameters will be more likely to better model/learn the **training data** compared to smaller models.
5. **….**

For Grading: Num. Correct: ..... Num. Wrong: ..... Score: .....