Instructions:

- This exam contains 15 pages (including this cover page) and 5 questions. Total of points is 100.
- Unless otherwise stated, answers should include all intermediate steps leading to the final result.
- All answers should be placed in the spaces given. If you need more space to complete an answer, you may use the blank pages at the end.
- Cheating is an academic offense. Your signature on this exam indicates that you understand and agree to the University's policies regarding cheating on exams.

Name: _____

Signature: _____

Distribution of Marks

Question	Points	Score
1	40	
2	15	
3	15	
4	15	
5	15	
Total:	100	

1 Part I: Knowledge.

1. Respond to each of these question parts or tasks using full sentences, working mathematical expressions and equations into those sentences where required. Use precise terminology appropriately for full marks.

   (a) (2 points) What mathematical property would the solution to the Lasso regression problem exhibit?

   (b) (2 points) In general, the joint probability distribution of two random variables X_1 and X_2, $p(X_1, X_2)$, may be decomposed as $p(X_2|X_1)p(X_1)$. What does one assume when making this simplification $p(X_1, X_2) = p(X_2)p(X_1)$?

   (c) (2 points) What density kernel would you propose if you wanted to most closely emulate the standard procedure of building a histogram?

   (d) (2 points) Your colleague is discussing a numerical optimization techique for a classical (not deep learning) machine learning model that is converging after relatively few iterations, but for which each iteration is taking a long time to complete. Which optimization technique do you suspect they are using, and what is the basis of this speculation?

(e) (2 points) A student in machine learning wants to sketch the decision boundary of a 2D binary classification problem, given the weight vector w and bias b. Provide a step-by-step analytical procedure for manually sketching the decision boundary.

(f) (2 points) How would one modify the linear regression objective to combine it with a prior belief that weights follow an isotropic Gaussian distribution, centered at the origin?

(g) (2 points) Why might one select Entropy over Gini Index as the loss function for a decision tree?

(h) (2 points) What must be true about the base models in an ensemble to achieve an average mean square error

$$\frac{1}{E}\sum_{i=1}^{E}\mathbb{E}[\varepsilon_i(x)^2]$$

where E is the number of base models in the ensemble, and $\mathbb{E}[\varepsilon(x)^2]$ is the average error of base models acting independently?

(i) (2 points) In the expectation maximization algorithm, what is the name of the quantity that is optimized directly?

(j) (2 points) In the expectation maximization algorithm, state the names and mathematical expressions of the distributions for which we seek to minimize the Kullback-Leibler divergence in the E-step.

(k) (2 points) Suppose that a trained fully-connected neural network with a single hidden layer consisting of 8 units is behaving identically on some nonlinear, 1-dimensional classification task to a multilayer perceptron with a single hidden layer with a width of one. What do you suspect the practitioners have failed to ensure in the training process?

(l) (2 points) A colleague of yours has implemented a convolutional neural network from scratch, to be applied to a dataset of black and white images with 1 Megapixel resolution. However, after experiencing issues training it, you inspect the code and learn that its first layer, which has 1000 hidden units, has 1 billion trainable parameters. Name the fundamental principle of the architecture of convolutional neural networks that your colleague violated.

(m) (2 points) What constraints, if any, must be placed on the weights of a reservoir computer?

(n) (2 points) What is the justification underlying summing positional and data embedding matrices prior to input to a transformer?

(o) (2 points) Describe one advantage of the RMSProp optimizer over Adagrad.

(p) (2 points) Why is the reparametrization trick necessary for training a variational autoencoder with backpropagation of error?

(q) (2 points) What is the forward process in diffusion models, and what role does it play in enabling the model to generate meaningful images?

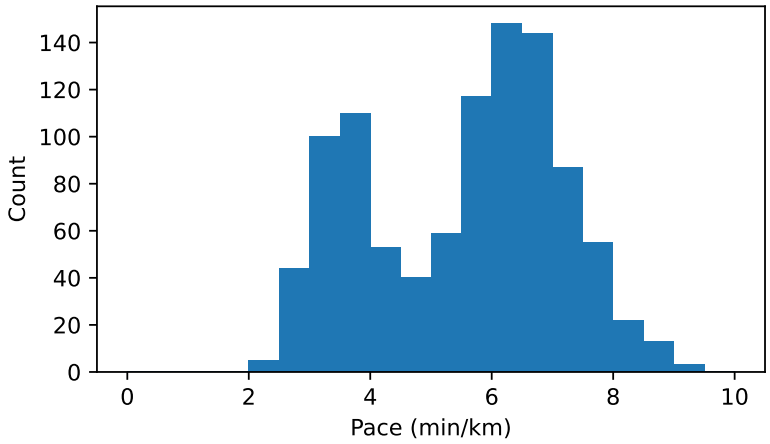
(r) (2 points) In the context of robustness against adversarial attack, why is it safe to use clipping as a projection function in the Fast Sign Gradient Method?

(s) (2 points) In (ϵ, δ)-Differential Privacy, what do the magnitudes of ϵ and δ reveal about the privacy guarantee?

(t) (2 points) Argue in favor of, or against this statement, using the case studies discussed in class: Early AI systems for natural language tasks exhibited hallucinations, especially when deployed on tasks for which they were not intended. However, these issues have now been addressed, and there is no evidence of this behavior persisting in such systems as of Fall 2024.

2 Part II: Application.

2. *Generative Models.* A colleague of yours has just completed a study that involved collecting data on the maximum achievable running pace among the general population, and has generated the following histogram:



(a) (3 points) Write the expression for an appropriate probability distribution to describe this data, such that new samples could be generated from it. Define all terms.

(b) (3 points) Write the log-likelihood function that relates the unknown parameters of the distribution to the observed samples.

(c) (3 points) State the name of the two-stage, iterative algorithm that can be used to solve for the parameters of this distribution. Write the corresponding objective. Define all terms used and state whether the objective concerns maximizing or minimizing.

(d) (6 points) Suppose you have access to a subset of 3 samples from the full dataset, presented in this table:

x_1	x_2	x_3
3.0	5.0	7.0

For the algorithm you proposed in part c, compute the parameter estimates of your model, after one iteration. State any assumptions required in order to do so.

3. *Perceptron.*

   (a) (15 points) For a linearly separable dataset of n pairs of feature vectors and class labels that lie within a d-ball of radius R and margin γ, demonstrate that the maximum number of errors k made by the Perceptron algorithm is independent of the number of samples in the dataset, the dimensionality d and the sequence of arriving samples. That is, starting from first principles, derive an upper bound on the Perceptron error.

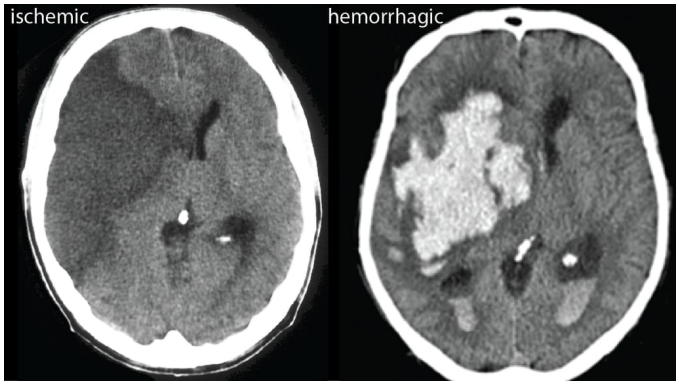
4. *Differential Privacy.* Let XYZ be an embarrassing or illegal behaviour.

   (a) (15 points) Faced with the query "have you ever engaged in XYZ in the past?", a respondent is instructed to perform the following steps to randomize their response.

   1. Flip a coin
   2. If the coin is tails, then respond truthfully.
   3. If the coin is heads, then flip a second coin.
   4. If the second coin is tails, then respond truthfully. Otherwise, respond randomly.

   Derive the privacy guarantee that this random response provides with respect to the true answer of any respondent.

5. *Computer Vision.* The images below depict brain scans of patients that have suffered two different types of stroke. Rapid and accurate diagnosis of the type of stroke that has occurred is essential for developing an appropriate treatment plan, and machine learning systems have been proposed to provide a second opinion for diagnosis in emergency room settings.



(a) (2 points) Defend the use of deep learning to solve this problem, with reference to at least one other family of machine learning methods covered in class.

(b) (2 points) Propose a particular family of neural networks, among those covered in class, to apply to this problem. Provide a brief justification for your decision.

(c) (6 points) Design a neural network architecture that you could train to solve the problem, by making a sketch. Provide enough detail to enable its precise implementation using a Deep Learning library like PyTorch.

(d) (5 points) Suppose an input image has resolution of 1000 pixels $\times$ 1000 pixels. How many trainable parameters are in the first layer of the architecture you proposed in part (c)?

This page is intentionally left blank to accommodate work that wouldn't fit elsewhere and/or scratch work.

This page is intentionally left blank to accommodate work that wouldn't fit elsewhere and/or scratch work.

This page is intentionally left blank to accommodate work that wouldn't fit elsewhere and/or scratch work.

This page is intentionally left blank to accommodate work that wouldn't fit elsewhere and/or scratch work.