Introduction to Machine Learning

New York University

Instructor
Prof. A.G. Wilson

Final Exam
May 17, 2021

There is an honor pledge and four questions, each with multiple parts. The test has 7 pages, including this title page. Read each question part carefully and completely. Some questions may take longer than others. The exam lasts three hours, including any time to upload your answers to GradeScope. **Manage your time carefully and do not spend too much time on any one question.** If you are unclear about how to interpret a question, state your assumptions and do your best in the allotted time. Please write your answers on separate sheets, clearly indicating which question and question part you are answering, and not the exam sheet itself, and upload your answers to GradeScope.

There is a strict honor code for this test. The answers should be entirely your own work. Consulting with anyone about this exam in any way in or out of the class is strictly prohibited. Violations of this policy will be treated very seriously. In order for your exam to be graded you need to upload your signature for the pledge on the following page.

1. HONOR PLEDGE

- This is an open book exam.
- You may NOT use communication tools to collaborate with other humans. This includes but is not limited to, G-Chat, Messenger, E-mail, etc.
- Anyone found sharing questions, resources, or answers for this exam, or communicating about the exam at **any time** in the period from Monday 8 am EDT to Tuesday 8 am EDT, will immediately earn a grade of 0.
- You have three hours to complete the exam and upload your answers to GradeScope, once you have started. No exam will be accepted after Tuesday 8 am EDT. There will be no exceptions.

"I understand these rules and agree to abide by them. I will not communicate about the exam, share questions, resources, or answers, nor will I seek assistance from another person or attempt to view their answers. My answers to this exam are entirely my own work."

Please upload your signature to certify that you have read and will comply with the rules. **Not uploading your signature here will result in a zero grade.**

2. Suppose we observe radioactive particles decay at distances $\mathcal{D} = \{x_j\}_{j=1}^{n}$ from a source. Each distance x_j has an exponential distribution given a rate parameter λ, $p(x_j|\lambda) = \lambda e^{-\lambda x_j}$, up to a distance $x_j = a$ from the source. Outside of this window, the probability of decay is uniform up to $x_j = b$. Suppose that m events occur inside the window $x_j \in [0, a)$, and k events occur in $x_j \in [a, b]$, b $\geq$ a. $n = m + k$. You can assume all decay events happen in $[0, b]$. You may use that $\int_a^\infty \lambda e^{-\lambda x} dx = e^{-a\lambda}$.

(a) Write down the likelihood function $p(\mathcal{D}|\lambda, a, b)$. [6 marks]

(b) Fixing all other variables, find the maximum likelihood estimate for λ, in terms of the variables defined in the question. [6 marks]

(c) Fixing all other variables, what maximum likelihood value do you find for b? [4 marks]

(d) Is the maximum likelihood estimate for b reasonable? Why or why not? [4 marks]

3. A gene represented by a $D = 1000$ dimensional vector $\mathbf{x}$ codes for a latent binary variable $z \in \{0, 1\}$, and the observed binary phenotype $y \in \{0, 1\}$ is a noisy version of z. We start with the model

$$p(z = 1|\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^\top \mathbf{x}) \tag{1}$$

$$p(y = 1|z = 1) = \alpha \tag{2}$$

$$p(y = 0|z = 0) = \lambda \tag{3}$$

where σ is the logistic sigmoid function. Suppose we observe data $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, corresponding to genotype-phenotype pairs. Let $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and $X = \{\mathbf{x}_i\}_{i=1}^n$. Let $\mathbf{z} = (z_1, \ldots, z_n)^\top$.

(a) What is the likelihood of the data $p(\mathbf{y}|\mathbf{w}, \alpha, \lambda, X)$? [10 marks]
(Note we are not conditioning on z).

(b) Suppose we place a prior on $\mathbf{w}$, $p(\mathbf{w}) = \mathcal{N}(0, \gamma^2 I)$. Write down the directed graphical model corresponding to the joint distribution over $\mathbf{y}, \mathbf{z}, \mathbf{w}$ conditioned on $X, \alpha, \lambda, \gamma$. Include all random *and* deterministic variables in your illustration. [5 marks]

(c) Is $y_1 \perp\!\!\!\perp y_2 | \mathbf{w}, X, \alpha, \lambda$? Briefly justify your answer. [2 marks]

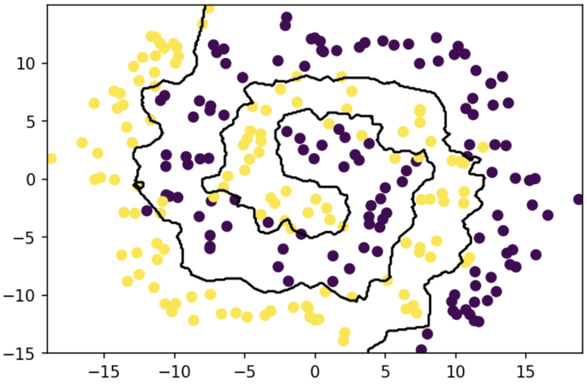
(d) Is $y_1 \perp\!\!\!\perp y_2 | X, \alpha, \lambda, \gamma$, where γ is from part (b). Briefly justify your answer. [2 marks]

(e) Would mini-batch stochastic gradient descent be a reasonable approach for finding $\hat{\mathbf{w}} = \text{argmax}_{\mathbf{w}} p(\mathbf{y}|\mathbf{w}, \alpha, \lambda, X)$? How about $\hat{\gamma} = \text{argmax}_{\gamma} p(\mathbf{y}|\gamma, \alpha, \lambda, X)$? Why or why not? [2 marks]

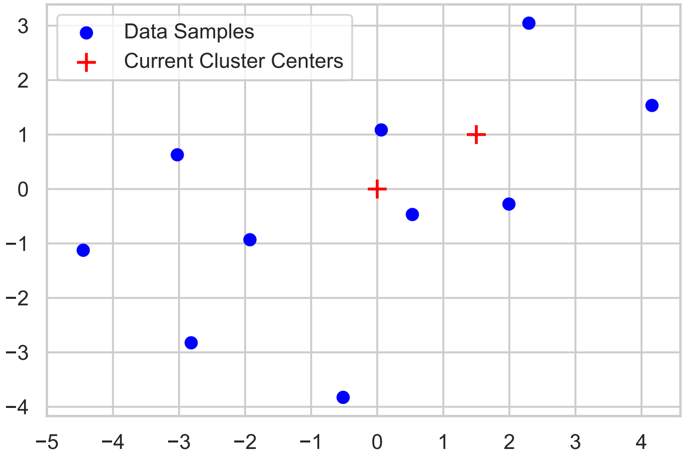
(f) Suppose that only 5 of the dimensions of $\mathbf{x}$ are actually predictive of the label y, such that 995 of the components of $\mathbf{x}$ have no influence on the label y, no matter what their values are. Would PCA be a reasonable approach to help us find that 5 dimensional subspace useful for predicting y? Why or why not? [5 marks].

4. Answer the following questions.

   (a) For the k-nearest neighbor classifier below, where yellow and blue are two different class labels, and the solid curve is a decision boundary: (1) What is the value of k, $k = 1, 2, 4$, or 200? [5 marks]; (2) Which of these k would lead to the highest-bias solution? [2 marks]; (3) The highest variance solution? [2 marks]. Justify your answers.



   (b) Consider the task of clustering using k-means clustering of the data marked using ●, and the initial cluster centers marked by +.



   Compute a single EM update, clearly labeling the E-Step and the M-Step. Graphically mark the final results by appropriately labeling each data point, and also mark the new cluster centers with ×. [5 marks]

   (c) Convolutional neural networks are more flexible, in that they can express a wider array of functions corresponding to different settings of their parameters, than fully connected networks. True or false? Explain your answer. [4 marks]

   (d) Deep neural networks are more flexible than single hidden layer networks are capable of being. True or false? Explain your answer. [2 marks]

Question 3 Continues on the Next Page

(e) Does the depth of a neural network have any significance if we use all linear activation functions? Why or why not? [4 marks]

(f) *Deep ensembles* are a popular approach where a neural network $f(\mathbf{x}, \mathbf{w})$ is retrained J times using maximum likelihood, starting from different initializations, to find parameter settings $\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_J$. Then a model is formed $g(\mathbf{x}) = \frac{f(\mathbf{x},\hat{\mathbf{w}}_1)+\cdots+f(\mathbf{x},\hat{\mathbf{w}}_J)}{J}$ to make predictions at the desired inputs $\mathbf{x}$. (1) Why would the functions $f(\mathbf{x}, \hat{\mathbf{w}}_1)$, ..., $f(\mathbf{x}, \hat{\mathbf{w}}_J)$ be different from one another, if each $\hat{\mathbf{w}}_j$, $j = 1, \ldots, J$, is a valid solution to $\hat{\mathbf{w}}_j = \mathrm{argmax}_{\mathbf{w}} p(\mathcal{D}|\mathbf{w})$? [2 marks]. (2) Explain why it is likely preferable to use $g(\mathbf{x})$ instead of $f(\mathbf{x}, \hat{\mathbf{w}}_1)$. [2 marks]

(g) Fred the Frequentist and Bob the Bayesian are having a conversation. Bob wants to make predictions with a model parametrized by $\mathbf{w}$, alongside a uniform prior $p(\mathbf{w})$. Fred considers the log posterior over parameters $\mathbf{w}$ given data $\mathcal{D}$:

$$\log p(\mathbf{w}|\mathcal{D}) = \log p(\mathcal{D}|\mathbf{w}) + \log p(\mathbf{w}) + \text{constant} \tag{4}$$

"You see, Bob, if I optimize the posterior in Eq. (4) with respect to $\mathbf{w}$ then the prior $p(\mathbf{w})$ acts as a regularizer. Moreover, if $p(\mathbf{w})$ has no preference for any values of $\mathbf{w}$ then this prior becomes irrelevant and only the likelihood $p(\mathcal{D}|\mathbf{w})$ matters". Help Bob explain to Fred why Bayesian methods can make very different predictions than maximum a-posteriori (MAP) approaches even if the prior $p(\mathbf{w})$ is uniform and so has no preference for any $\mathbf{w}$. Use illustrations as part of your answer. In what case would the predictions of a Bayesian approach using a uniform prior $p(\mathbf{w})$ be the same as a maximum likelihood approach? [10 marks]

5. Suppose we have a normalizing flow $\mathbf{x} = f(\mathbf{z}, \mathbf{w})$ and $\mathbf{z} = P\mathbf{u} + \boldsymbol{\epsilon}$, where $p(\mathbf{u}) = \mathcal{N}(\mathbf{m}, A)$ and $p(\boldsymbol{\epsilon}) = \mathcal{N}(0, \sigma^2 I)$, and a dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$. $\mathbf{x}$ is $d \times 1$, $\mathbf{z}$ is $d \times 1$, P is $d \times q$, $\mathbf{u}$ is $q \times 1$, and $\mathbf{w}$ represents ten million parameters of the normalizing flow.

(a) What is $p(\mathbf{z}|P, A, \mathbf{m}, \sigma^2)$? [3 marks]

(b) What is $p(\mathbf{x}|\mathbf{w}, P, A, \mathbf{m}, \sigma^2)$? [2 marks]

(c) What is the likelihood of the data $p(\mathcal{D}|P, A, \mathbf{m}, \mathbf{w}, \sigma^2)$? [2 marks]

(d) Suppose you are worried about over-fitting. Would cross-validation be a reasonable approach for learning $\mathbf{w}$? Why or why not? [2 marks]

For all the following questions, suppose f is the identity mapping, such that $\mathbf{x} = \mathbf{z}$. Further, suppose $\mathbf{m} = \mathbf{0}$, $A = I$. Therefore, $\mathbf{x} = P\mathbf{u} + \boldsymbol{\epsilon}$, $p(\mathbf{u}) = \mathcal{N}(0, I)$, $p(\boldsymbol{\epsilon}) = \mathcal{N}(0, \sigma^2 I)$.

(e) Suppose $\mathbf{x}_i \in \mathbb{R}^{10\times 1}$ and the eigenvalues λ_j of the sample covariance matrix C for the data $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{100}$ satisfy $\lambda_j > 0$ for $j \leq 7$, and $\lambda_j = 0$ for $j > 7$. Now suppose we transform our data as $\tilde{\mathbf{x}}_i = B\mathbf{x}$ where $B \in \mathbb{R}^{4\times 10}$. How many principal components are needed to perfectly reconstruct the data $\mathcal{D} = \{\tilde{\mathbf{x}}_i\}_{i=1}^{100}$? Justify your answer. [5 marks]

(f) Suppose we choose values for $q = 1, \ldots, d$ and form the $d \times q$ matrix P_q for each value of q. Let $\ell(q) = p(\mathcal{D}|\hat{P}_q, \hat{\sigma}^2)$, where $\hat{P}_q, \hat{\sigma}^2 = \text{argmax}_{P_q, \sigma^2} p(\mathcal{D}|P_q, \sigma^2)$. Draw a plot with $\ell(q)$ on the vertical axis and q on the horizontal axis, with q running from $1, \ldots, (d + 10)$. [4 marks]

(g) Explain, step by step, how you can use Bayesian principles to estimate the true value of q from training data alone. As part of your answer, include a diagram that has all possible datasets on the horizontal axis. [Bonus 4 marks].