Algorithms for Learning and Inference 2019: Final Exam

Instructor: Morteza H. Chehreghani

Due: See Canvas

- NOTE 1. You must submit your solution to Canvas, in the same way as the assignments.
- NOTE 2. The exam must be done individually. You may not receive help from anyone else.
- NOTE 3. Your submission must be in pdf format. You may either type your solutions in latex/word and submit a pdf file, or take the photo/scanning of the handwritten solutions and upload the pdf file. If you take photos, make sure that it is easy to read and that you combine photos into a single pdf file such that each page appears in the right order. There are both command line and online tools to do this.
- NOTE 4. Read the questions carefully such that you do not miss any question and ensure you clearly give the answer required for each (sub)question.
- NOTE 5. You do not need to write (Python) code for any question.

1. (14 points) Consider a dataset consisting of i.i.d. observations, x_1, x_2, $\cdots$, x_N, where each observation is a discrete value drawn form the set: $x_i \in \{0, 1, 2, \cdots\}$. The observations are generated from the following distribution with parameter $\lambda > 0$, $\lambda \in \mathbb{R}$

$$Pr(X = x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{1}$$

   (a) Explain in few words the concepts of maximum likelihood estimation and Bayesian estimation. (4 points)

   (b) Derive the mathematical expression for the likelihood function. (3 points)

   (c) Calculate the Maximum Likelihood Estimator (MLE) for λ, λ_{MLE}, in terms of the observations and the number of observations N. (3 points)

   (d) Assume that the prior distribution of λ follows a gamma distribution with parameters $\alpha > 0$, $\beta > 0$:

$$g(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \tag{2}$$

   Show that the posterior distribution for λ follows also a gamma distribution. (4 points)

2. (16 points) This question is concerned with regression models.

   (a) Consider the following table:

x	4	6	10	15
t	15	17	32	25

   Consider the linear model in the form of $t = w_1 x + w_0$ applied to this dataset. Write down the cost (loss) function and compute the model on this dataset, where the parameters are obtained by minimizing the mean squared error. (4 points)

(b) Given are a dataset of N rows (data points, observations) and D features $\boldsymbol{X}$ and the respective target variables $\mathbf{t}$. This dataset can be represented as $(\boldsymbol{X}, \boldsymbol{t})$ with $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ and $\boldsymbol{t} \in \mathbb{R}^{N \times 1}$. One can solve the linear regression exactly by setting the weight vector $\boldsymbol{w}$ as $\boldsymbol{w} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{t}$

(i) Explain (in at most two lines) the main issue that can arise when solving linear regression using the above-mentioned formula. (2 points)

(ii) Bayesian approach is another way to solve the linear regression problem. Mention two main advantages of the Bayesian linear regression. (4 points)

(c) Answer by True or False and justify your answer in at most two lines.

(i) Given the following two linear models $t_1 = w_0 + w_1 x$ and $t_2 = w_0 + w_1 x + w_2 x^2$, the mean squared error (loss) of t_2 on training data will always be at least as low as the error with t_1, regardless of the data. (3 points)

(d) Explain in at most three lines why one performs cross validation. (3 points)

3. (19 points) This question is about classification methods.

(a) Draw the picture of the decision boundaries for the following (train) data points for a 1-nearest neighbor classifier (i.e., $K = 1$). In this dataset, each data point is shown by two features, where $x^{(1)}$ shows the first feature and $x^{(2)}$ shows the second feature. Use ℓ_1-distance (i.e., $\text{dist}(x_i, x_j) = |x_i^{(1)} - x_j^{(1)}| + |x_i^{(2)} - x_j^{(2)}|$) and draw the picture with an $x^{(1)}$-axis of $0 \leq x^{(1)} \leq 7$ and $x^{(2)}$-axis of $0 \leq x^{(2)} \leq 7$ and use a clear labeling of the axes. Include the data points in the picture and also indicate which region belongs to which class. (4 points)

$x^{(1)}$	$x^{(2)}$	Class label
0	0	1
0	1	1
0	5	1
1	1	1
2	2	2
2	5	2
4	4	2
1	7	3
5	2	3
5	5	3
7	1	3

80

(b) When labeling a new data point using K-nearest neighbor classification (i.e., during test time) there might be areas where labeling the new data point is ambiguous (not clear which label it should be given). What would be **one** good way of handling these situations (there are multiple reasonable strategies)? Give a short explanation (maximum three sentences) why the method would be reasonable. (3 points)

(c) Use the kernel

$$k(x_n, x_m) = \exp\{-\frac{1}{\gamma}(x_n - x_m)^T (x_n - x_m)\} \tag{3}$$

to calculate the value of the kernel for the following data points with $\gamma = 1$. Comment on why this kernel seems suitable (or not) based on the values you have obtained. (6 points)

Table 1: Data for kernel SVM

	$x^{(1)}$	$x^{(2)}$	Class label
x_1	0.37431335	1.16780523	1
x_2	0.73829051	-0.56552444	1
x_3	-0.37910848	0.79427185	1
x_4	0.27019947	0.51053891	-1
x_5	0.39835632	0.1232738	-1
x_6	0.39835632	0.1232738	-1

$$k(x_1, x_2) =$$
$$k(x_2, x_3) =$$
$$k(x_4, x_5) =$$
$$k(x_5, x_6) =$$
$$k(x_1, x_6) =$$
$$k(x_2, x_5) =$$

(d) When using the kernel in Eq (3) to train an SVM on the data in Table 1 we obtain that the classifier is supported on $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_6 = 1$, $\alpha_5 = 0$ (the different alphas correspond to the support vectors). Calculate the value for b (the bias in SVM) and write down the formula for prediction. Use this formula to classify a new data point at $x^{(1)} = x^{(2)} = 1$, show your calculations. (6 points)

4. (16 points) This question is about deep learning.

(a) Consider a 2D convolutional layer in a neural network that is applied to an RGB image of size $h \times w$. The convolutional layer has 8 filters of size $k = 2h_f + 1$. The convolutional layer is applied with stride 1 and without padding to the input image. What are the dimensions of the output of the convolutional layer? How many learnable parameters does the convolutional layer have? (4 points)

(b) If instead of the convolutional layer you wanted to use a fully connected layer to produce equally as many outputs as the convolutional layer from the flattened input image, what would be the size of the weight matrix of the layer? Compare to the number of learnable parameters of the convolutional layer! What does this imply for the training process? (4 points)

(c) Let $\mathbf{x} \in \mathrm{R}^m$ denote the input vector of a fully connected layer without any activation function. Furthermore, let $\mathbf{y} \in \mathrm{R}^n$ denote the output vector of the layer. The components y_i of $\mathbf{y}$ are thus computed according to

$$y_i = \sum_{j=1}^{m} W_{i,j} x_j \tag{4}$$

where $W_{i,j}$ are the elements of the weight matrix $\mathbf{W} \in \mathrm{R}^{n \times m}$ of the layer. Given the derivatives $\frac{dE}{dy_i}$ of an error term E with respects to the elements y_i of the output vector $\mathbf{y}$, derive an expression for the derivatives $\frac{dE}{dx_i}$ of E with respect to the elements of the input vector $\mathbf{x}$. (4 points)

(d) Assume that you have derived an algorithm to compute the gradients of the loss term with respect to all trainable parameters in your neural network model. Outline a full procedure (algorithm), including initialization, on how the model can be trained using these gradients. (4 points)

5. (15 points) We consider K-means and Gaussian Mixture Model (GMM) clustering methods, with the number of clusters specified by K. We investigate these methods on a dataset of N data points (samples) and the dimensionality of data is D.

   (a) Mention one main difference between K-means and GMM. (2 points)

   (b) Describe the generative approach/viewpoint for a Gaussian Mixture Model (GMM). (3 points)

   (c) What is the number of the free (unknown) parameters of a GMM with a selected K? Assume all the co-variance matrices are known in advance. (2 points)

   (d) How can you use Bayesian Information Criterion (BIC) to estimate the optimal K for a GMM? (2 points)

   (e) Determine whether the following statement is True or False. Explain your answer. (3 points)
   If we apply K-means with $K = 5$ and $K = 7$, then the optimal cost for $K = 7$ is never larger than the optimal cost for $K = 5$.

   (f) In the Sum of Norms (SON) formulation of K-means, what happens if $\lambda \to \infty$? Explain the answer. (3 points)