CS551: Introduction to Deep Learning

Mid Semester, Spring 2019

IIT Patna

Attempt all questions. Do not write anything on the question paper.

Time: 2 Hrs

Full marks: 30

1. Consider real-valued variables X and Y. The Y variable is generated, conditional on X, from the following process: $\epsilon \sim N(0, \sigma^2)$, $Y = aX + \epsilon$ where every ϵ is an independent variable. This is a one-feature linear regression model, where a is the only weight parameter. The conditional probability of Y has distribution $p(Y|X,a) \sim N(aX, \sigma^2)$, so it can be written as $p(Y|X,a) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(Y - aX)^2)$. The following questions are all about this model. Assume we have a training dataset of n pairs (X_i, Y_i) for $i = 1, \ldots, n$, and σ is known.

   (a) Which one of the following equations correctly represent the maximum likelihood problem for estimating a? Say *yes* or *no* to each one.

      i. $\arg\max_a \sum_i \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2)$,

      ii. $\arg\max_a \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2)$,

      iii. $\arg\max_a \sum_i \exp(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2)$,

      iv. $\arg\max_a \prod_i \exp(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2)$,

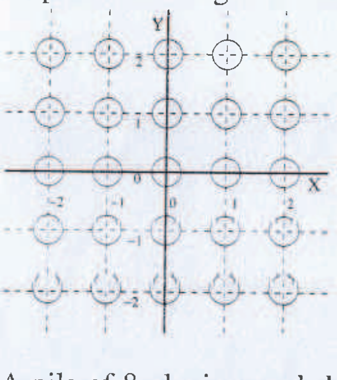
      v. $\arg\max_a \sum_i \frac{1}{2}(Y_i - aX_i)^2$,

      vi. $\arg\min_a \sum_i \frac{1}{2}(Y_i - aX_i)^2$.

   (b) Derive the maximum likelihood estimate of the parameter a in terms of the training example X_i's and Y_i's. (Hints: Start with the simplest form of the problem you found above.) (2+2)

2. Consider a binary classification problem which is solved using a deep feed-forward network. The final output layer has a sigmoid function. What is the maximum derivative that can be contributed by the sigmoid function irrespective of the inputs and other parameters? (3)

3. Consider the set of points in $\mathbb{R}^2$ as shown in figure below. Both X and Y lie between -2 and 2. These set of points ($\boldsymbol{v}_i = [x_i \ y_i]^T$ say) are transformed using a 2×2 matrix $\boldsymbol{A}$ that is $\boldsymbol{A}\boldsymbol{v}_i$. (a) If $\boldsymbol{A} = [[1 \ \frac{1}{2}]^T \ [\frac{1}{2} \ 1]^T]$ holds and (b) the eigenvectors and the corresponding eigenvalues for a given matrix are as follows - Eigenvectors for the eigenvalues 0 and 2 are $[-\frac{1}{\sqrt{2}} \ -\frac{1}{\sqrt{2}}]^T$ and $[-\frac{1}{\sqrt{2}} \ \frac{1}{\sqrt{2}}]^T$ respectively. Draw the shape of the transformed points for both the cases with respect to the original axis X and Y. (3)



4. A pile of 8 playing cards has 4 aces, 2 kings and 2 queens. A second pile of 8 playing cards has 1 ace, 4 kings and 3 queens. You conduct an experiment in which you randomly choose a card from the first pile and place it on the second pile. The second pile is then shuffled and you randomly choose a card from the second pile. If the card drawn from the second deck was an ace, what is the probability that the first card was also an ace? (4)

5. You are training maximal margin classifier on a tiny dataset with 4 points shown below. This dataset consists of two examples with class label 0, and two examples with class label 1. (a) Find the weight vector $\boldsymbol{w}$ and bias b. What is the equation corresponding to the decision boundary? (b) Circle the support vectors and draw the decision boundary. (4)

Points	1	2	3	4
x_1	1	2	4	5
x_2	4	3	5	6
Label	0	0	1	1

6. Consider the following design matrix. You need to use PCA to reduce the dimension from 2 to 1. For this determine the following. (a) Compute the covariance matrix for the sample points. (Note: x_i need to be centered.) (b) Then compute the unit eigenvectors, and the corresponding eigenvalues, of the covariance matrix. (c) Suppose we use PCA to project the sample points onto a one-dimensional space. What one-dimensional subspace are we projecting onto? For each of the four sample points in X (original, not the centered version of X!), write the coordinate (in principal coordinate space, not in $\mathbb{R}^2$) that the point is projected to. (1+2+4)

Points	1	2	3	4
x_1	6	-3	-2	7
x_2	-4	-5	-6	-3

7. Consider a neural network for a binary classification which has one hidden layer as shown in the figure below. We use a linear activation function $h(z) = cz$ at hidden units and a sigmoid activation function $g(z) = \frac{1}{1+e^{-z}}$ at the output unit to learn the function for $P(y = 1|\boldsymbol{x}, \boldsymbol{w})$ where $x = (x_1, x_2)$ and $w = (w_1, w_2, \ldots, w_9)$. (a) What is the output $P(y = 1|\boldsymbol{x}, \boldsymbol{w})$ from the above neural net? Express it in terms of x_i, c and weights w_i. What is the final classification boundary? (b) Draw a neural network with no hidden layer which is equivalent to the given neural network, and write weights $\tilde{w}$ of this new neural network in terms of c and w_i. (2+3)