CSE Ph.D. Qualifying Exam, Fall 2020

Data Analysis

Please answer three of the following four questions. All questions are graded on a scale of 10. If you answer all four, all answers will be graded and the three lowest scores will be used in computing your total. Show all your work and write in a readable way.

1. **Graphical Models**

   The Restricted Boltzmann Machine (RBM) is an undirected graphical model over binary vectors. It has “visible” variables v and “hidden” variables h. The jointly distribution is

$$p(v, h) \propto e^{-E(v,h)} \tag{1}$$

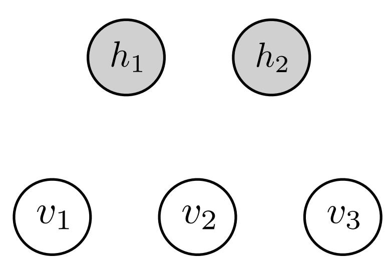
   Note that you need to normalize $e^{-E(v,h)}$ to get the probability.

   The energy function $E(v, h)$ is defined as

$$E(v, h) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_{i,j} v_i h_j w_{ij}$$

   where v_i and h_j are the binary states of the visible variable i and hidden variable j, respectively. a_i and b_i are their biases, and $w_{i,j}$ is the weight between them.

   (a) Consider the RBM with three visible variables and two hidden variables. Complete the graphical model by drawing the edges.



   (b) Mark TRUE or FALSE to the following statements about conditional independence properties in the model

   - $(v_1 \perp v_2 | h_1)$
   - $(h_1 \perp h_2 | v_1)$
   - $(v_1 \perp v_2 | h_1, h_2)$

- $(v_2 \perp v_3 | h_1, h_2)$
- $(h_1 \perp h_2 | v_1, v_2)$
- $(h_1 \perp h_2 | v_1, v_2, v_3)$

(c) What is the marginal distribution $p(v)$ and the corresponding normalization factor.

(d) Assume $a = 0$, $b = 0$, and $w_{2,1} = 1$, and all other entries of w are zero. Compute the probability $P(v_1 = 1, v_2 = 1 | h_2 = 0, v_3 = 0)$.

2. **SVMs and the Kernel trick**

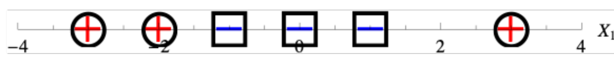


Figure 1: Dataset for SVMs and Kernels.

You are given a data set D (see Fig. 1) with data from a single feature X_1 in $\mathbb{R}^1$ and corresponding label $Y \in \{+, -\}$. The data set contains three positive examples at $X_1 = \{-3, -2, 3\}$ and three negative examples at $X_1 = \{-1, 0, 1\}$.

(a) (1 point) Can this data set (in its current feature space) be perfectly separated using a linear separator? Why or why not? (Explain in 1 line)

(b) (2 points) Lets define the simple feature map $\phi(u) = (u, u^2)$ which transforms points in $\mathbb{R}^1$ to points in $\mathbb{R}^2$. Apply ϕ to the data and plot the points in the new $\mathbb{R}^2$ feature space (i.e. just show the plot). Can a linear separator perfectly separate the points in the new $\mathbb{R}^2$ features space induced by ϕ? Why or why not? (Again, explain in 1 line)

(c) (1 point) Give the analytic form of the kernel that corresponds to the feature map ϕ in terms of only X_1 and X_2. Specifically define $k(X_1, X_2) =< \phi(X_1), \phi(X_2) >$ ($< ., . >$ is the dot-product of two vectors), and give the analytical form of $k(., .)$.

(d) (4 points) Construct a maximum-margin separating hyperplane. This hyperplane will be a line in $\mathbb{R}^2$, which can be parameterized by its normal equation, i.e. $w_1 Y_1 + w_2 Y_2 + c = 0$ for appropriate choices of w_1, w_2 and c. Here, $(Y_1, Y_2) = \phi(X_1)$ is the result of applying the feature map ϕ to the original feature X_1. Give the values for w_1, w_2 and c. Also, explicitly compute the margin for your hyperplane. You do not need to solve a quadratic program to find the maximum margin hyperplane. Instead, let your geometric intuition guide you.

(e) (2 points) Draw the decision boundary separating of the separating hyperplane, in the original $\mathbb{R}^1$ feature space. Also circle the support vectors.

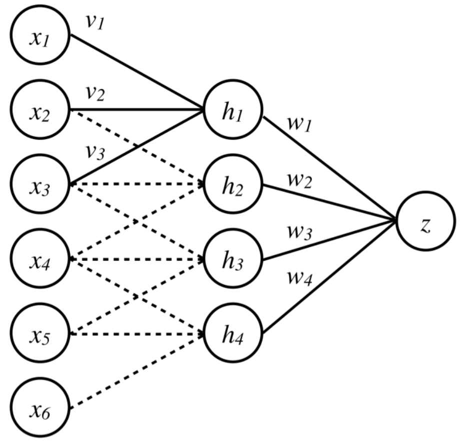


Figure 2: A one-dimensional convolutional network.

3. **Convolutional Neural Networks**

   Consider the following convolutional neural network architecture (Figure 2):

   In the first layer, we have a one-dimensional convolution with a single filter of size 3 such that $h_i = s(\sum_{j=1}^{3} v_j x_{i+j-1})$. The second layer is fully connected, such that $z = \sum_{i=1}^{4} w_i h_i$. The hidden units activation function $s(x)$ is the logistic (sigmoid) function. The output unit is linear (no activation function). We perform gradient descent on the loss function $R = (y - z)^2$, where y is the training label for x.

   (a) [1 pt] What is the total number of parameters in this neural network? Recall that convolutional layers share weights. There are no bias terms.

   (b) [4 pts] Compute $\partial R / \partial w_i$.

   (c) [1 pt] Vectorize the previous expressionthat is, write $\partial R / \partial w$.

   (d) [5 pts] Compute $\partial R / \partial v_j$.

4. **Applied Data Analysis**

   Carol and Bob are data scientists working at a Fortune 100 company that routinely works with petabyte-scale high-dimensional datasets with huge number of data points. They are tasked to solve two machine learning problems: a binary classification problem, and a clustering problem. They are debating what methods to try first. Carol believes that it is a good idea to first try some "simpler," well-known methods (e.g., k-nearest neighbors, random forests, k-means), then try more sophisticated and possibly better-performing methods. Bob, on the other hand, thinks it is good to first try the latest techniques published at top academic conferences (e.g., KDD, NeurIPS) since many of them report state-of-the-art results.

(a) (2 points) Briefly justify why both of their approaches may be reasonable.

(b) (2 points) They decide to first try k-nearest neighbors (k-NN) for their classification problem and k-means for their clustering problem. Briefly describe the scalability challenges they may encounter.

(c) (2 points) Briefly describe how they may determine the value of k in k-NN, and the value of k in k-means. (Both k values can be different.)

(d) (4 points) They believe visualization will play an important role in evaluating and comparing the performance of the machine learning methods that they are going to try. Briefly describe two visualization approaches that can help with such comparison — for each approach:

   i. describe a challenge that could arise when the visualization approach is applied on large datasets that Carol and Bob are working with; and

   ii. propose a method to tackle that challenge.

   For easier discussion, your discussion and examples may center around evaluation metrics of you choosing. You are welcome to include illustrations to support your answers.