Pennsylvania State University
Data Sciences Program
DS 310: Machine Learning
Vasant Honavar
Fall 2019

Exam II

Instructions

Each problem is worth 25 points.

You are required to solve only 4 out of the 5 problems.

You may solve all 5 problems and receive extra credit for solving the 5th problem. In some cases, the extra credit points *may* marginally improve your grade.

Good Luck!

NAME:

Problem	Score
1	
2	
3	
4	
5	
Total	

1 (a) Recall that each document is an arbitrary length sequence of words taken from a vocabulary W. Consider the bag of words representation of documents where each document d_i is represented by a tuple of word counts $(w_{i,1} \cdots w_{i,|W|})$. Define a kernel function $K_{i,j}$ that computes the similarity between pairs of documents d_i and d_j where the documents are represented by $(w_{i,1} \cdots w_{i,|W|}$ and $w_{j,1} \cdots w_{j,|W|})$ respectively.

(b) Now consider documents that consist of both text and pictures. To keep things simple, assume that each document has no more than one picture in it. Assume that the text in document d_i is represented as before by a tuple of word counts $(w_{i,1} \cdots w_{i,|W|})$. Suppose the picture in document d_i is represented by a bag of image features, i.e., a tuple of word counts $(f_{i,1} \cdots f_{i,|F|})$, where F is a vocabulary of visual features. Define a kernel function $J_{i,j}$ for computing the similarity between pairs of documents d_i and d_j which contain both text and pictures.

2 Consider the problem of learning binary classifiers in a setting where the costs of mis-classification of the two classes are unequal. Recall that the standard 2-class Support Vector Machine (SVM) classifier finds a maximum margin separating hyperplane assuming equal mis-classification costs for the two classes. How would you modify the stochastic gradient descent soft margin update rules (for the parameters $\mathbf{w}$ and b) for the 2-class SVM in a setting where the cost of mis-classifying a sample belonging to class 1 (as class 0) is c_1 and that of mis-classifying a sample belonging to class 0 (as class 1) is c_0? Be sure to provide a detailed justification for your answer.

3 Consider a 3-layer feed-forward network with n input nodes a set of hidden nodes (indexed by j), and a single output node. Assume that the output of the network for pattern $\mathbf{X}_p$ is defined by $o_p = \sum_j u_j z_{jp}$. Here u_j represents the weight between the jth hidden neuron and the output neuron and z_{jp} is the output of the jth hidden neuron on input $\mathbf{X}_p$.

The hidden nodes implement a particular form of radial basis functions defined as follows: $z_{jp} = \frac{1}{\sigma^2 + n_{jp}^2}$ where σ is a constant and $n_{jp} = \sum_i w_{ji} x_{ip}$ is the dot product of input pattern $\mathbf{X}_p$ with the weight vector $\mathbf{W}_j$ for the jth hidden neuron. (As usual, $z_{0p} = 1$; and $x_{0p} = 1$). Define $E_a = \frac{1}{2} \sum_{p=1}^{P} (d_p - o_p)^2$ where P is the number of examples in the training set and d_p is the desired network output for the input sample $\mathbf{X}_p$.

(a) Does the choice of the activation function z_{jp} satisfy the requirements for universal function approximation? Explain.

(b) Derive the update equations for the parameters w_{ji} and u_j that minimize E_a

4 Consider the network described in problem 3 above. Define $E = \lambda E_a + (1 - \lambda) E_b$ where $0 \leq \lambda \leq 1$ is a user-defined non-negative constant and $E_b = \sum_{p=1}^{P} \sum_{i=0}^{N} \left(\frac{\partial E_a}{\partial x_{ip}} \right)^2$.

(a) Clearly explain the impact of minimizing E_b on the sensitivity of the network output to relatively small amounts of noise in the input sample and the tendency of the network to over-fit the training data.

(b) Derive the update equations for the parameters w_{ji} and u_j that minimize E

5 For each of the following problems, describe which of the machine learning algorithms you have encountered in the course you would choose (if any) or how you would modify an existing algorithm (if needed). Explain the rationale behind your choice.

(a) A 2-class classification task where you expect the samples from each class to be approximately linearly separable.

(b) A multi-class classification task where the data set has a limited number of training examples and the features are likely to be independent given the class.

(c) A classification task wherein the features of the training examples are numeric and are corrupted by small amounts of random noise.

(d) A function approximation task in which the number of features (variables) far exceeds the number of training examples and there is little prior knowledge to guide the selection of features to be used to train a neural network function approximator.

(e) A classification task where the data samples are composed of three types of data, namely, images, text, and a set of numeric measurements.