Pennsylvania State University
Data Sciences Program
Machine Learning (DS 310)
Vasant Honavar
Fall 2018

Take Home Final

Instructions

1. Please write your name in the space provided.
2. The test contains 5 problems each of which is worth 20points.
3. Please consult with the instructor or the TA if you have difficulty *understanding* any of the problems.
4. Please be brief and precise in your answers. Write your solutions in the space provided.
5. Please show all the major steps in calculations or proofs to get partial credit where appropriate.
6. Good luck.

Name:

Problem	Score
1	
2	
3	
4	
5	
Total	

1. (**20 pts.**)

   (a) (**5 pts.**) State the *universal approximation theorem* for real-valued functions defined on the N-dimensional unit hypercube. Comment on the implications of the theorem for the design of artificial neural networks (and corresponding learning algorithms) for function approximation and pattern classification.

(b) (**15 pts.**) Consider a 3-layer feed-forward network with N input nodes, H hidden nodes (where H is assumed to be as large as necessary, but finite), and a single output node. Suppose we index the input nodes by i and hidden nodes by j. Let x_{ip} be the ith component of the N-dimensional input pattern $\mathbf{X}_p$ and z_{jp} the corresponding output of the hidden node j. Suppose the corresponding output of the network is given by: $o_p = \sum_{j=0}^{H} u_j z_{jp}$. Define $n_{jp} = \sum_{j=0}^{N} w_{ji} x_{ip}$. Assume w_{ji} and u_j are real-valued parameters and $\forall p \;\; z_{0p} = x_{0p} = 1$. For each of the following choices for the functions computed by the hidden neurons, state whether or not the conditions required by the *universal approximation theorem* are satisfied. Briefly justify your conclusions.

i. $z_{jp} = 1$ iff $n_{jp} \geq 0$ and $z_{jp} = 0$ otherwise.

ii. $z_{jp} = n_{jp}$.

iii. $z_{jp} = \frac{1}{1+e^{-n_{jp}}}$.

iv. $z_{jp} = tanh(n_{jp}) = \frac{1-e^{-n_{jp}}}{1+e^{-n_{jp}}}$

v. $z_{jp} = \frac{2}{\pi} arctan(n_{jp})$

2. (**20 pts.**) Some learning problems lend themselves to solution using a *divide-and-conquer* strategy where the network structure is designed to exploit inherent modularity in the problem. As an illustration of this approach, consider a 3-layer feed-forward network with n input nodes (indexed by i), two sets (H_1 and H_2) of hidden nodes (indexed by j and k respectively), and a single output node. Assume that the output of the network for pattern $\mathbf{X}_p$ is defined by $o_p = \sum_{j=0}^{|H_1|} u_j z_{jp} + \sum_{k=0}^{|H_2|} u_k z_{kp}$. Here u_j (u_k) represents the weights between the jth (kth) hidden neuron and the output neuron and z_{jp} (z_{kp}) is the output of the jth (kth) hidden neuron on input $\mathbf{X}_p$. $\mid H_1 \mid$ and $\mid H_2 \mid$ denote the number of hidden neurons in the two sets H_1 and H_2 respectively.

   The hidden nodes in H_1 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 for the jth hidden neuron. w_{ji} denotes the weight from the ith input neuron and the jth hidden neuron. As usual, $z_{0p} = 1$; and $x_{0p} = 1$.

   The hidden nodes in H_2 implement a sigmoid function defined as follows: $z_{kp} = \frac{1}{1+e^{n_{kp}}}$ where $n_{kp} = \sum_i w_{ki} x_{ip}$; w_{ki} denotes the weight from the ith input neuron and the kth hidden neuron. As usual, $z_{0p} = 1$ and $x_{0p} = 1$.

   Define $E = \frac{1}{2} \sum_{p=1}^{P} (d_p - o_p)^2$ where P is the number of patterns in the training set and d_p is the desired network output for the input pattern $\mathbf{X}_p$. *Derive from first principles*, the update equations for u_j and w_{ji} so as to minimize E.

.

3. (**20 pts.**) Consider the problem of learning binary classifiers in a setting where the costs of misclassification of the two classes are unequal. Recall that the standard soft marhin Support Vector Machine classifier finds a maximum margin separating hyperplane assuming equal misclassification cost for both classes (class 1 and class 0). Precisely formulate the problem of learning soft margin Support Vector Machine Classifier to minimize the total cost of misclassification over the training set when λ_{01} as the cost of incorrectly labeling a sample belonging to class 0; and λ_{10} is the cost of misclassifying a sample belonging to class 1. Assume that the cost of correctly labeled a sample of either class is 0. Show how you can adapt the soft margin SVM algorithm derived in class can be adapted so as to minimize the misclassification cost over the training set.

.

4. (**20 pts.**) In this problem, we consider the design of kernel functions for kernel machines e.g., support vector machines, regularized kernel logistic regression, etc.

   (a) Specify a kernel function for text documents, keeping in mind that documents can be of variable length. Justify your choice.

   (b) Specify a kernel function for color images, keeping in mind that images can be of variable resolution. Justify your choice.

   (c) Specify a kernel function for graphs. Justify your choice.

(d) Specify a kernel function for documents consisting of images and text. Justify your choice.

(e) Specify a kernel function for graphs of inter-linked documents (hypertext). Justify your choice.

5. (**20 pts.**) Suppose you have been hired by an Data Science consulting firm. You have been asked to recommend a promising machine learning approach in each of the following application scenarios. In each case, briefly justify your recommendation.

   (a) Your client, has a database of automobile repair records containing symptoms and expert diagnosis. She would like to build a diagnosis system. The attributes can be numeric (e.g., sensor readings), as well as categorical (e.g., the results of specific diagnostic tests). Your client would like to use the database to come up with a recommended protocol for diagnosing the problem. Because different diagnostic tests can have different costs, she would like your diagnosis protocol to minimize the overall cost of diagnosis.

   (b) Your client, a major online book seller, is interested in a personalized book recommendation system. The book seller has access to previous purchase histories of customers, a list of keywords that describe the customer's interests, etc.

(c) Your client is a commodities trader who is interested in accurate prediction of the market price of a commodity (e.g., wheat) based on a large historical database of commodity (e.g., wheat) prices and the relevant variables (weather, satellite imagery, production and demand related variables, geopolitical events, etc.)

(d) Your client is a major online digital media vendor that is interested in developing a personalized movie recommendation system. The vendor has access to a large database of client profiles, their movie viewing histories, and whenever available, their ratings for the movies they have watched. Note that the total number of movies in the database far exceeds the number of movies any individual can possibly expected to have watched.