**Pennsylvania State University**
**College of Information Sciences and Technology**
**DS 310 Midterm Exam**
**Vasant Honavar**

**Instructions**

1. There are 4 problems each of which is worth 25 points.

2. Please consult the instructor if you have difficulty *understanding* any of the problems.

1. (**25 pts.**) Consider a $M$-class pattern classification problem in which each pattern $\mathbf{X}$ to be classified belongs to exactly one of $M$ mutually exclusive classes $\omega_1 \cdots \omega_M$. Suppose that $\mathbf{X}$ is represented using a vector of $N$ binary features $\mathbf{X} = [x_1, x_2 \cdots x_N]$. Let $p_{ji} = P(x_i = 1|\omega_j)$. Assume that the features are independent given the class label. Let $P(\omega_1) \cdots P(\omega_M)$ be the prior probabilities of each class.

   (a) Describe in words, the meaning of $p_{ji}$.

   (b) Show that a classifier that assigns $\mathbf{X}$ to class $\omega_k$ $g_k(\mathbf{X}) \geq g_j(\mathbf{X}) \forall j \neq k$. where $g_i(\mathbf{X})$ can be written in the form $\sum_{i=1}^{N} w_{ji}x_i + w_{j0}$ is a minimum error (Bayes optimal) classifier. Express $w_{ji}$, in terms of $p_{ji}$ and $P(\omega_j)$ where ($1 \leq j \leq M$; $0 \leq i \leq N$).

   (c) Generalize the above solution for (b) above to obtain minimum loss classification where $\lambda_{ij}$ is the loss incurred in assigning a pattern to class $\omega_i$ when it in fact belongs to class $\omega_j$.

2. (**25 pts.**) Suppose you have been hired by an AI consulting firm. Indicate which algorithm you would choose in each of the following data-driven knowledge acquisition scenarios. In each case, briefly justify your recommendation.

   (a) Your client, has a database of patient records containing symptoms and expert diagnosis. She would like to build a diagnosis system. The attributes can be numeric (e.g., patient's temperature), as well as categorical (e.g., whether the patient is pregnant). In addition to performing accurate diagnosis of patients, your client would like to use the system to obtain insight regarding the relationships between attributes for different diseases.

   (b) Your client is designing an email organizer to be integrated into an email program. The SPAM detector is to be trained using information that is obtained by observing the user's actions on each email (read and respond, mark as junk, ignore).

   (c) Your client is a large company which has managed to gather a large database of information about past applicants to jobs in various categories (e.g., software engineer, data scientist, etc.) and the hiring decisions on each applicant. Experts in the organization are convinced that they can automate the process of shortlisting applicants for further consideration. You are told that it is important that the decision-making process be transparent – that is, it should be easy to understand why an applicant was shortlisted (or not).

(d) Your client is a large hospital that is interested in reducing the workload on radiologists working on breast cancer diagnosis from radiological images. The hospital has a large database of radiological images labeled with expert diagnosis. The goal here is to automate diagnosis on the cases where the AI system can produce high confidence results, and identify a subset of cases that the AI system is not so confident about for further examination by expert radiologists.

3. (**25 pts.**) Recall that the perceptron learning algorithm that was described in class is an additive weight update algorithm - that is, we add or subtract a fraction of the misclassified sample to the weight vector at each iteration. Consider instead, a multiplicative weight update algorithm for an $n$-input neuron defined as follows: Consider a neuron defined by two weight vectors $\mathbf{w}^+$ and $\mathbf{w}^-$. Suppose both weight vectors are initialized with a value 1 for each of their components. Consider a training example $(\mathbf{x}_p, d_p)$ where $\mathbf{x}_p \in \{-1, 1\}^n$ is an input pattern and $d_p \in \{-1, 1\}$ is its class label. Let $y_p$, the output of the classifier be 1 if $\mathbf{w}^+ \cdot \mathbf{x}_p > \mathbf{w}^- \cdot \mathbf{x}_p$ and $y_p = -1$ otherwise. Suppose the weights are updated as follows: $w_i^+ \leftarrow w_i^+ \beta^{-(d_p - y_p)x_{ip}}$ and $w_i^- \leftarrow w_i^- \beta^{(d_p - y_p)x_{ip}}$ where $0 < \beta < 1$ is a learning rate. Comment on the potential advantages of such a multiplicative weight update algorithm over its additive counterpart. Prove that this algorithm is guaranteed to converge to a pair of weight vectors $(\mathbf{w}_\star^+, \mathbf{w}_\star^-)$ that correctly classify the training data whenever such weight vectors exist. Hint: Show that the multiplicative weight update algorithm as an instance of the standard perceptron algorithm operating on a suitably transformed weight space.

4. (a) (**12.5 pts.** )

   Suppose you have designed (or trained) a $n$-input perceptron with the weight vector $\mathbf{W}$ and threshold $\theta$ to correctly classify a set $S$ of $n$-dimensional patterns. Further assume that all the weight values of the perceptron so designed have been hard-wired but the threshold $\theta$ can be set under user control. When the perceptron is later used on a factory floor, suppose you find that the source of the patterns (say the camera or the digitizer) adds a fixed amount of noise (given by the $n$-dimensional noise vector $\mathbf{V}$) to each pattern. Assuming that you can somehow measure (or can calculate) $\mathbf{V}$, how would you change $\theta$ so that the perceptron with the same weight vector $\mathbf{W}$ continues to correctly classify all the patterns in $S$?

   (b) (**12.5 pts**) Briefly explain the significance of the following properties of the error functions used to derive gradient-descent based learning algorithms of the type considered in part (a).

   i. Continuity and Differentiablilty (with respect to the weights or other parameters of interest)

   ii. Convexity with respect to weights or other parameters of interest