

Iowa State University
Department of Computer Science
Machine Learning (Com S 573)
Vasant Honavar
Spring 2007
Due February 26, 2007
Problem Set 3

1. **(25 pts.)** Consider a variation on the Perceptron training algorithm in which the weight update equation is given by equation given by:
$$\mathbf{W} \leftarrow \mathbf{W} + (d_p - o_p) \times \mathbf{X}_p \times \eta_p$$
where d_p is the desired output for pattern \mathbf{X}_p and o_p is the actual output of the perceptron (threshold logic unit) for the pattern \mathbf{X}_p and $0 < \eta_{Min} \leq \eta_p \leq \eta_{Max} < \infty$ where η_{Max} and η_{Min} are (respectively) upper and lower bounds on the training sample specific learning rate η_p . *Prove* that the perceptron algorithm, given a training set $\{(\mathbf{X}_p, d_p)\}$, is guaranteed to find a weight vector \mathbf{W}^* (if one exists) such that for all examples in the training set, $\mathbf{W}^* \cdot \mathbf{X}_p \geq 0$ if $d_p = 1$ and $\mathbf{W}^* \cdot \mathbf{X}_p < 0$ if $d_p = -1$.
2. 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 whenever a training example (\mathbf{x}_p, d_p) is misclassified ($y_p \neq d_p$): $w_i^+ \leftarrow w_i^+ \beta^{-y_p x_{ip}}$ and $w_i^- \leftarrow w_i^- \beta^{y_p x_{ip}}$ where $0 < \beta < 1$ is a learning rate. Prove that this algorithm is guaranteed to converge to a pair of weight vectors $(\mathbf{w}_*^+, \mathbf{w}_*^-)$ that correctly classify the training data whenever such weight vectors exist. Comment on the potential advantages of such a multiplicative weight update algorithm over its additive counterpart.
3. **(25 pts.)** Consider a binary Naive Bayes classifier with m nominal input variables, and $N - m$ continuous valued input variables. Suppose further that each of the nominal variables have multinomial distribution and the continuous variables have Gaussian distributions (each with its own variance). Explore whether the posterior probability of the class variable can be expressed in the form of a logistic function.
4. **(25 pts.)** Consider 2-class binary Naive Bayes classifier with N multinomially distributed input variables. Recall that the discriminative counterpart of such a classifier is a logistic regression classifier trained using maximum likelihood estimation. Suppose instead, we assume that the parameters estimated by logistic regression have Gaussian distributions with means corresponding to the respective Maximum Likelihood estimates and some assumed variances (which can be estimated through a procedure akin to crossvalidation). Derive the gradient based learning algorithm for this setting.