DS 310 Machine Learning
Vasant Honavar
Fall 2018

Problemset 2: Decision Trees, Perceptrons, Naive Bayes

Available: October 10, 2018
Due: October 17, 2018

1. (25 pts.) Consider a version of the perceptron learning algorithm in which the learning rate $\eta(t)$ can vary with each weight change step t (e.g., it might be different at different times of the day or it may be a function of the weather, the learner's mood, the amount of coffee consumed, etc.). Prove from first principles that the resulting variant of the perceptron learning algorithm is guaranteed to terminate with a weight vector that defines a separating hyperplane whenever the training data are linearly separable, so long as $0 < A \leq \eta(t) \leq B$ where A and B are fixed lower and upper bounds respectively.

2. (25 pts.) Use the decision tree learning algorithm to construct a decision tree classifier from the following training set and show all your calculations.

Outlook	Temperature	Humidity	Wind	PlayTennis
sunny	hot	high	weak	No
sunny	hot	high	strong	No
overcast	hot	high	weak	Yes
rain	mild	high	weak	Yes
rain	cool	normal	weak	Yes
rain	cool	normal	strong	No
overcast	cool	normal	strong	Yes
sunny	mild	high	weak	No
sunny	cool	normal	weak	Yes
rain	mild	normal	weak	Yes
sunny	mild	normal	strong	Yes
overcast	mild	high	strong	Yes
overcast	hot	normal	weak	Yes
rain	mild	high	strong	No

3. (25 pts.) Recall the use of *bag of words* representation for text classification. Note that in a *bag of words* representation each document is encoded using a *bag of words*. Consider a more general scenario in which it might be beneficial to encode a data sample to be classified using multiple bags of features that correspond to perhaps different

modalities (e.g., textual features or words, visual features, etc. in the case of a document that contains both words and pictures). We can think of each modality as being associated with its own *vocabulary* (e.g., a set of words, a set of visual features, etc.). Precisely formulate the problem of learning classifiers in a setting wherein each data sample is represented using multiple (say M) bags of features, one for each modality. Outline a learning algorithm (at sufficient detail needed for implementation). Hint: You may consider variations or adaptations of the learning algorithms that you have studied so far.

4. (25 pts.) Suppose you have been hired by an AI consulting firm. Your clients are in a position to acquire software for data driven knowledge acquisition using one or more of the following: perceptron, decision tree, random forest, naive bayes. Indicate which algorithm you would choose in each of the following applications. 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 features for different diseases.

   (b) Your client has a web-based information system for a large organization. She would like to enhance its functionality to support customization of information retrieved and presented to different users. He is able to record each users's actions when presented with specific documents in a given context. She would like to use such a database of records for designing a proactive information assistant for each user that goes out and retrieves documents that might be of interest to each user.

   (c) Your client is a financial organization which has managed to gather a large database of credit related information on its customers. Experts in the organization are convinced that they can automate the decision to approve or deny a loan based on a simple rule that checks whether one or more (usually a small number) of a set of possible conditions are satisfied. You are told that it is important that the decision-making process be transparent – that is, it should be easy to understand why a loan was approved or denied in each case.