Probabilistic Foundations of Artificial Intelligence

Problem Set 7

Dec 5, 2014

1. Learning Bayesian Networks for classification of bird into species

In this problem, you will implement an algorithm for learning a Naive Bayes Model for classification of birds into species, given a set of attributes for the bird. In this task, the answer to all the attributes is given and goal is to predict the bird specie correctly. You will do this by learning a Naive Bayes model. You will also consider a very useful extension called Tree Augmented Naive Bayes to improve the prediction accuracy.

The training data set is given in a file called `trainingData.txt`, available on the course webpage. There are two species in data denoted by class variable C and we are given attribute data for 200 birds. Given a sample for a bird (corresponding to one row in the file), the $1st$ column is the class variable (specie of the bird) C , and the $2nd$ to the $6th$ columns are the bird attributes A_1, A_2, A_3, A_4, A_5. The testing data set is given in a file called `testingData.txt`. There are 100 testing samples, with the same format for each sample.

(i) **Learning a Naive Bayes model**. You are asked to learn a naive Bayesian network based on a given training data set. The structure of the naive Bayes Network is given as follows:

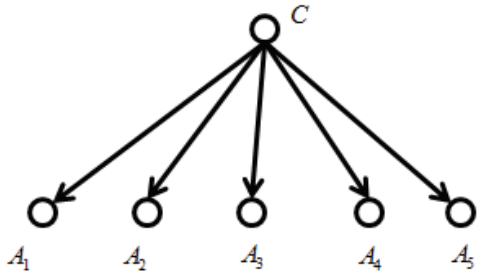

Figure 1: Naive Bayes network.

Estimate the parameters for the conditional probability distributions in the network using MLE on the training data. Based on the constructed naive Bayesian network you can classify samples by applying Bayes rule to compute conditional class probabilities $P(C|A_1, A_2, A_3, A_4, A_5)$, and predicting the label with the highest probability.

Please write down the parameters θ_C and $\theta_{A_1|C}$, and the percentage of classification error on the testing data set.

(ii) **Learning a Tree Augmented Naive Bayes (TAN) model**. Tree augmented naive Bayes models are formed by adding directional edges between attributes. After removing the class variable, the attributes should form a tree structure (no V-structures). See Fig. 2 as an example.

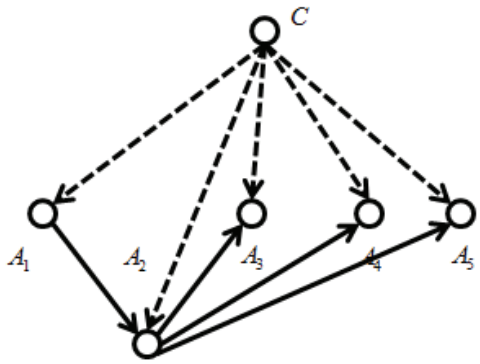

Figure 2: An example of a tree augmented naive Bayes network.

Use the following procedure to learn the tree augmented naive Bayes model for the training data, then draw the structure of the obtained model.

(a) Compute $I_{\widehat{P}_D}(A_i; A_j|C)$ between each pair of attributes, $i \neq j$, where $I_{\widehat{P}_D}(A_i; A_j|C)$ is the conditional mutual information (with respect to the empirical distribution $\widehat{P}_D$ on the training data) between A_i, A_j given the class variable.

$$I_{\widehat{P}_D}(X; Y|C) = \sum_{x,y,c} \widehat{P}_D(x, y, c) \log \frac{\widehat{P}_D(x, y|c)}{\widehat{P}_D(x|c)\widehat{P}_D(y|c)}$$

(b) Build a complete undirected graph in which the vertices are the attributes A_1, A_2, A_3, A_4, A_5. Annotate the weight of an edge connecting A_i and A_j by $I_{\widehat{P}_D}(A_i; A_j|C)$.

(c) Build a maximum weighted spanning tree.

(d) Transform the resulting undirected tree to a directed one by choosing a root variable and setting the direction of all edges to be outward from it.

(e) Construct a tree augmented naive Bayes model by adding a vertex labeled by C and adding an directional edge from C to each A_i.

(iii) **TAN for classification**. Based on the structure above, you are asked to estimate the parameters for conditional probability distributions using the training data set (using MLE). Then you can classify the testing data set by computing the highest probability $P(C|A_1, A_2, A_3, A_4, A_5)$.

What is the percentage of classification error for the testing data set? Please compare this result with that using the Naive Bayes model.