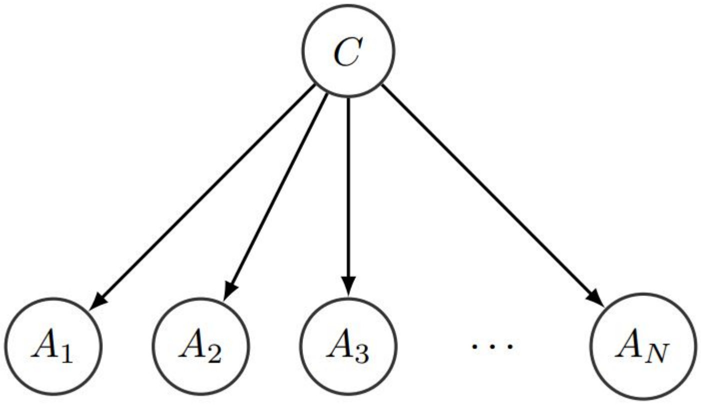
Programming Assignment 1

Naive Bayes (NB) classifiers are often competitive classifiers even though their strong independence assumptions may be unrealistic. If C denotes the class variable and $(A_1, ..., A_N)$ the attributes, then a NB model can be represented as a directed graph with these variables as nodes and edges $\{(C, A_i) : 1 \leq i \leq n\}$. The following figure illustrates the graph structure.



In this assignment, you will implement the parameter learning for NB classifiers.You will apply these classifiers to predict the party affiliation of either Democrat or Republican of US Congressmen (the class variable) based on their votes for 16 different measures (the attribute variables) shown in Table 1. Not all congressmen voted on all 16 measures, so sometimes entries in this dataset will have missing attributes; however, we will still be able to utilize our Bayes Network to accurately classify these examples. To keep things simple, the class and attribute variables are all binary with 0, 1 corresponding to a no and yes vote respectively.

When training the models, some of the parameters may not have enough examples for accurate estimation. To mitigate this, we will use a Beta$(0.1, 0.1)$ prior on the parameters of the vote distributions.

In order to evaluate the performance of our classifiers on the dataset, we will use 10-fold cross-validation. Under 10-fold cross-validation, the dataset is first partitioned into 10 equally sized partitions. Of these 10 partitions, one of them is used for the test set while the rest of the data are used as the training set to compute test error on this partition. This process is repeated for the other nine partitions, and we can take the average of the resulting test errors to obtain the 10-fold CV test error. We have implemented this procedure for you in the function `evaluate`.

Attribute	Name	Incomplete Entry 1
A_1	handicapped infants	1
A_2	water project cost sharing	1
A_3	adoption of the budget resolution	0
A_4	physician fee freeze	0
A_5	El Salvador aid	0
A_6	religious groups in schools	0
A_7	anti satellite test ban	0
A_8	aid to Nicaraguan Contras	?
A_9	mx missile	?
A_{10}	immigration	0
A_{11}	synfuels corporation cutback	0
A_{12}	education spending	?
A_{13}	superfund right to sue	?
A_{14}	crime	?
A_{15}	duty free exports	0
A_{16}	export administration act south Africa	0

Table 1: Attribute names for Congressional Voting Records together with an incomplete example that has some voting records missing for a particular Congressman.

We have provided a starter code for this assignment [1]. You can download it here.

What to submit

Please submit the following two files to CourSYS:

- nb.py – Your completed implementation. Do not change the signature of the functions that you were supposed to implement.
- report.pdf – A pdf file answering all the questions in this assignment.

(8 points) Question 1

Implement a NB classifier that both learns the parameters from the training data and can use these parameters to score and classify examples in the training data. What is your test error rate using 10-fold cross-validation?

Note: you can use the ***evaluate*** *function in the starter code, but leave the optional argument train subset to its default value until question 3.*

[1] This assignment is adapted from Stanford CS 228.

(6 points) Question 2

In general, working with data where the values of attributes and labels are missing is difficult when learning model parameters. However, we can still use our generative model from a fully trained Bayes Network to classify examples in which some of the attributes may be unobserved. Suppose A_i is unobserved. We can still compute $P(C|A_1, ..., A_{i1}, A_i, A_{i+1}, ..., A_N)$ by computing $P(C|A_1, ..., A_{i-1}, A_{i+1}, ..., A_N)$ and marginalizing over A_i . Update your NB implementation to handle the case where some attributes may have missing values and use this new implementation to classify Incomplete Entry 1 in Table 11 . Given the observed attributes, what is the marginal probability this Congressman is Democrat (C = 1) given the votes we did observe? Can you predict how this Congressman voted on education spending (A12)?

Note the power of a generative model: it can easily handle missing data and can be used to answer all sorts of probabilistic queries. In contrast, this would not be possible with a discriminative model, e.g., if you trained a logistic regression or a random forest classifier to directly predict the affiliation of a Congressman (C), because these models would only provide you with the conditional distribution $P(C|A_1, ..., A_N)$.

*Note: You should train your classifier on the full dataset. You can use the function **evaluate_incomplete_entry**, which both trains on the full dataset and loads Incomplete Entry 1 for classification.*

(4 points) Question 3

Set the arguments *train_subset=True* when calling **evaluate** so that the classifiers are trained on data size of 16 (instead of 208 in question 1). What is the test error when you train NB classifiers on a smaller subset of the training data?

Explain why the test error may not strictly be worse than the test error in question 1.

Question 4

a) Give one short piece of feedback about the course so far. What have you found most interesting? Is there a topic that you had trouble understanding? Are there any changes that could improve the value of the course to you?

b) How many hours did you spend on this assignment?

Please provide your answers in your report.