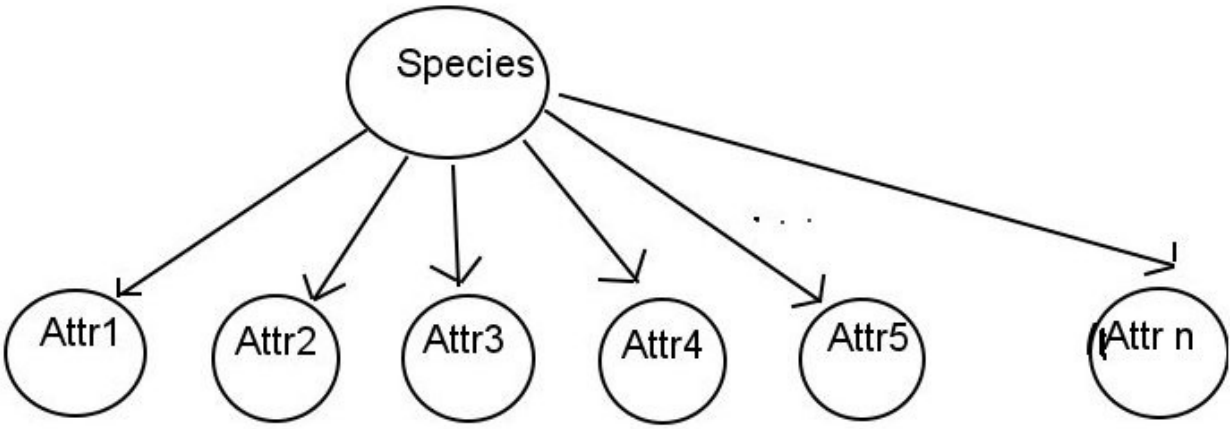
Avian Asker Team 10
Sonal Gupta
Jonathan Krause
Manuel Lagang

Final Report

Overview

The overall approach is model a probability distribution over birds and attributes using a bayesian network. The distribution on the birds is updated by conditioning on the values of the attribute queries, and then a decision is made to finally guess a bird when the value of guessing a bird is more than that of querying another attribute (more details provided in later section).

The structure of the bayesian network is that of a naïve Baye's classifier, the attribute queries are assumed to be conditionally independent given the species. Below is an illustration:



Species is a discrete random variable taking values from 1-200 and the Attr variables are discrete random variables that take values in [present, not present] x [confident, probably, not sure]. This Bayesian Network was chosen because it is intuitive (knowing the species gives all information about attributes) and it is also easy to learn and perform conditioning on the attributes.

Learning the probability distribution

There are 2 classes of probability distributions to learn, the probability distribution of species and the conditional distribution of an attribute given a species. The probability distribution of species is assumed to be uniform. The conditional distributions are then estimated from the training data using counts.

The regularization method used to prevent overfitting on the conditional distributions is based on “diffusion”. Once the initial probability estimates are obtained by counts, the probability estimates are diffused toward the uniform distribution by moving probabilities to neighbors in the confidence spectrum (for example, some of the probability in the [present, confident] answer is transferred to the [present, probably] answer). This hopefully prevents conditional probabilities from being 0 and accounts for the observations being noisy in the sense that asking the same question again can yield a different answer.

Deciding what question to ask

The next question is chosen greedily based on the expected entropy of the species distribution after asking the question. For attribute queries, this is calculable by subtracting the mutual information of that attribute and the species from the currenty entropy. For guessing a particular bird species, the entropy will go down to 0 if the species is correct, and or will otherwise retain the entropy of the remaining species if incorrect.

This greedy decision seems to lead to overguessing where perhaps querying more attributes may have been overall better, which infrequently results in a large amount of questions asked. To combat this, after a threshold of questions (currently 20), if a species guess is the greedy choice, then with some probability the best attribute query is chosen instead. It is not clear how much this threshold improves performance since this tail of large amounts of questions asked still exists. However, other attempts to mitigate this have not improved performance.