AM207 Project Proposal

Avery Faller, Fanny Heneine, Crystal Lim

April 21, 2016

Expectation Maximization

Using Expectation Maximization, we will attempt to discover groups of animals that frequently appear together. For this problem we will use gbif.org to examine occurrences of mammals in the United States over the last few years. We will break down the United States into a series of lat-long grid squares that we will use to create groupings of species based on the locations of the observed sightings. We will create a topic model and perform EM on this model, using the log-likelihood as an objective function that we will attempt to maximize.

We can start off by making the simplifying assumption that each grid-square only contains a single topic of animals. We can define each 'document' to be all of the sightings within a particular lat-long grid-square. We will label each of these grids g. Each grid has a set of sightings, labeled $\boldsymbol{s}_g$. Each gird has N_g sightings in total, where $N_g = \sum_a s_{g,a}$, where A is the set of all animals.

Given this notation and the simplification mentioned above, we can write the log-likelihood as follows:

$$p(\{z_g, \boldsymbol{s}_g\}_{g=1}^{G} | \boldsymbol{\theta}, \{\beta_k\}_{k=1}^{K}) = \sum_{g=1}^{G} \sum_{k=1}^{K} z_{gk} \left(\ln \theta_k + \sum_{a=1}^{A} s_{g,a} \ln \beta_{k,a} \right)$$

where z_g is a vector that represents which 'topic' the observations in this specific grid-square belong to. β is a matrix where each row represents a different 'topic' and each column a different animal in our set of all animals A. So the values of β, i.e. $\beta_{k,a}$, represent the percentages of a topic that are made up of a particular animal.

If we have time at the end, we can extend this model to allow for multiple topics per lat-long grid-square. It would be nice to have your help working through the math of the multi-topic model. Another way to extend this component of the project would be to run the model on a larger variety of animals, or across multiple countries.

Hidden Markov Chains

Using Hidden Markov Chains, we will attempt to estimate the number of animals from a specific species in the United states, namely deer. To solve this problem, we will use gbif.org to examine occurrences of observed dears every year between 1980 and 2010, and combine this data with the actual values of (estimated) deer in the United States. As this is a continuous state case, we will need to discretize the state space and we would then be able to approximate the continuous state Markov Decision Process via a discrete one.

We will use a forward message pass algorithm to produce the most likely state for any year of the MDP, which would be the predicted number of deer in the US, based on observation data. We would use a combination of the actual number of deer and the observations to produce emission probability, while transition probability would be calculated using our actual number data set.

In the first pass, the algorithm computes a set of forward probabilities which provide, for all $k \in \{1, \dots, t\}$, the probability of ending up with a given number of deer (an interval of numbers) given the first k observations in the sequence, i.e. $P(X_k \mid Y_{1:k})$. In the second pass, the algorithm computes a set of backward probabilities which provide the probability of observing the remaining observations given any starting point k, i.e. $P(Y_{k+1:t} \mid X_k)$. These two sets of probability distributions can then be combined to obtain the distribution over states at any specific point in time given the entire observation sequence:

$$P(X_k \mid Y_{1:t}) = P(X_k \mid Y_{1:k}, Y_{k+1:t}) \propto P(Y_{k+1:t} \mid X_k) P(X_k \mid Y_{1:k})$$

However, as the discretization has downsides as it assumes that the value function is takes a constant value over each of the discretization intervals.

Therefore, we will extend the problem to the Kalman filtering. We will use the observations occurrences over time, which might contain inaccuracies, and produce estimates of our unknown variables.The Kalman filter model assumes that at time k an observation (or measurement) $\mathbf{z}_k$ of the true state $\mathbf{x}_k$ is made according to:

$$\mathbf{z}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k$$

where $\mathbf{H}_k$ is the observation model which maps the true state space into the observed space and $\mathbf{v}_k$ is the noise which is drawn from a normal distribution.

We will then be able to compare both methods, and compare the prediction with the actual number to evaluate the performance of our algorithms in both cases.

Bayesian Hierarchical Model

Using a Bayesian Hierarchical Model, we will examine which areas have higher populations of a given animal. In our case, we will be observing the squirrel subfamily *Sciuridae Hemprich* as recorded in 1820. Reported sightings, our data points, will be correlated based on distance such that a large number of sightings implies a large number of squirrels in the surrounding regions as well. We will be focusing on the United States and break up the map into grid squares of an appropriate size. We will also use a slice sampler after creating our model so that we can visualize highly populated regions.

We will be representing the number of sightings, our indicator of population, as Y for its respective location X. Each Y_i will be drawn from a Poisson distribution characterized by our chosen grid square size w and λ, our population intensity. Under the Log-Gaussian Cox process we represent λ by $Exp[\alpha + Z_i]$, where α (base population count) is drawn from a normal distribution and Z is drawn from a multivariate normal distribution with covariance depending on a square exponential kernel whose value is determined by distance between points where map size is adjusted for. A covariance matrix, for α and Z is then created.

Our log-likelihood for M grid squares:

$$P(Y|\lambda) = \sum_{i=1}^{M} w^2 \lambda_i$$

Also, for $Z \sim MVN(0, \Sigma)$:

$$\Sigma = Exp[-||X_i - X_j||^2/\phi]$$

We will choose ϕ so that the large distances between sightings in the US are accounted for.

To extend the model further we can add multiple animal species and have points be correlated not only by position, but also by level of biodiversity. We would then be modeling the locations with great biodiversity with the assumption that highly biodiverse areas are also surrounded by biodiverse areas.