COS 424: Interacting with Data

Lecturer: Rob Schapire
Scribe: Jingyuan Wu

Lecture #19
April 17, 2007

1 A Density Estimation Problem: Modelling the Habitat of Plant and Animal Species

Conservation biologists are often concerned with modeling the population distribution of plants and animals. In particular, our problem is concerned with modeling the population distribution of a particular species across a grid map.

1.1 Description of Data

In this problem, two types of data will be available. The first type of data we will have is called **presence records**. Presence records are pixels on the grid map where the species of concern was observed. The same pixel may be present multiple times if the species was observed more than one time within that pixel. The second type of data we will have is called **environmental variables**. Each environmental variable will contain information such as average rain fall on a particular pixel. Environmental variable data are available for each pixel on the grid map.

This problem is not simply one of classifying each point on the map as a habitat or a non-habitat, because we have only positive examples and no negative examples. Just because the biologist did not observe the species at a particular location does not enable us to label that location as non-habitat.

1.2 Formal Definition of Variables

We will define the following set of variables:

- X is defined as the set of all pixels or locations on the grid map.
- $|X|$ is the size of cardinality of the set X. This value is generally very large, ranging from tens of thousands to millions.
- $x_1, ..., x_m \in X$ are the set of pixels that are included as the presence records. Note that the x_is are not necessarily distinct.
- $f_1, ..., f_n$ is defined as the set of *features*. Each f_j is defined for all pixels on the grid map, i.e. $f_j : X \to \Re$.
- π is defined as the true distribution of the species. In other words, $\pi(x)$ is the fraction of the population living at pixel $x \in X$.

The set of features includes the environmental variables, but it may also contain additional functions derived from the environmental variables such as the average rainfall squared.

1.3 Assumptions

We assume that the set of presence records $x_1, ..., x_n$ is chosen i.i.d. according to π. This means that $\pi(x)$, the probability that x will be chosen as a presence record, is proportional to the population living at x

Unfortunately, the assumptions may affect our model's ability to approximate the real world for a number of reasons. First, we have assumed implicitly that the true distribution π does not change with time. This may not be realistic since π may in fact change with day and night cycles and with seasonal cycles. We have also assumed that there is no sample bias. This assumption means that the biologist sampled all the points in X with equal diligence. This, however, may not be the case since some locations are definitely harder to access than others.

It may also be noted that even the assumption that the presence records x_i are independent may be suspect since it may be that a biologist is more likely to sample a nearby location after having observed a butterfly in the present location.

2 Approach One: Maximum Likelihood Estimation

To solve this problem, our goal is to create an estimate of π, call it $\widehat{\pi}$. One method of attack is to use *Maximum Likelihood Estimation*. To proceed, we need to first express $\widehat{\pi}$ in a parametric form. We may begin by choosing a linear parametric form:

$$\widehat{\pi}(x) = \sum_{j=1}^{n} \lambda_j f_j(x).$$

However, there are several problems with this simple formulation. First, the values $\widehat{\pi}(x)$ may not lie in $[0,])$. Since $\widehat{\pi}(x)$ represent proportions, it would make little sense for them to take on negative values or values greater than one. Also, $\sum_{x \in X} \widehat{\pi}(x)$ may not equal one. Again, this equality is required because $\widehat{\pi}(x)$ represent proportions.

As a result, we may choose to transform the simple linear form to an exponential form and set $\widehat{\pi}(x)$ equal to:

$$q_\lambda(x) = \frac{\exp\left(\sum_{j=1}^{n} \lambda_j f_j(x)\right)}{Z_\lambda}.$$

Here Z_λ is chosen so that $\sum_x q_\lambda(x) = 1$. This form has the advantage that it is strictly positive and lies $[0, 1]$. To maximize the likelihood function, then, we would choose $\widehat{\pi}$ to be equal to $q_{\widehat{\lambda}}$ where:

$$\begin{aligned} \widehat{\lambda} &= \arg\max_{\widehat{\lambda}} \prod_{i=1}^{m} q_{\widehat{\lambda}}(x_i) \\ &= \arg\max_{\widehat{\lambda}} \sum_{i=1}^{m} \ln q_{\widehat{\lambda}}(x_i). \end{aligned}$$

This last equation turns out to be concave in λ, which means there exist efficient methods for solving it.

There are a few problems with this maximum likelihood estimation. One problem is that this estimation technique is prone to overfitting especially if the number of features is large. Another problem is that the transformation of the linear model seems somewhat arbitrary. Therefore, we will next explore a different approach to this problem.

	altitude	July temp.	···
record #1	1410m	16°C	···
record #2	1217m	22°C	···
record #3	1351m	17°C	···
⋮	⋮	⋮	
Average	1327m	17.2°C	···
Stand. Dev.	117m	3.2°C	···

Figure 1: average of features over presence records

3 Approach Two: Entropy Maximization

Before we proceed, we make the following definitions. We define the sample average of the features as the average of a feature over all presence records. (see figure 1 above). Mathematically, we define

$$\widehat{E}[f_j] = \frac{1}{m}\sum_{i=1}^{m} f_j(x_i).$$

Also, we define the true expection of the features as

$$E_\pi[f_j] = \sum_{x\in X} \pi(x) f_j(x).$$

In general, we would expect $E_\pi[f_j] \approx \widehat{E}[f_j]$ for all j. We may then have as our constraints for estimating π that

$$E_{\widehat{\pi}}[f_j] \quad = \quad \widehat{E}[f_j]\ \forall j.$$

These constraints, in general, do not reduce the possible choices for $\widehat{\pi}$ to 1. As a result, we need other constraints to narrow down our choices.

For example, if we were to estimate π with no information nor data of any kind, then the most intuitive estimate for π is a uniform distribution. Therefore, we can choose to have as our goal the selection of $\widehat{\pi}$ that is as close to the uniform distribution as possible while still satisfying a set of constraints.

Closeness to the uniform distribution may be measured by entropy H:

$$H(\widehat{\pi}) \quad = \quad -\sum_{x\in X} \widehat{\pi}(x)\ln(\widehat{\pi}(x)).$$

It can be shown that H is never negative, and that it is maximized when $\widehat{\pi}$ is uniform. This maximization follows the *principle of maximum entropy*. Namely, when modelling a distribution, we should maximize the entropy subject to constraints representing what we know about the distribution. So our problem now is to find $\widehat{\pi}$ that maximizes $H(\widehat{\pi})$ and such that

$$\begin{aligned} E_{\widehat{\pi}}[f_j] &= \widehat{E}[f_j]\ \forall j \\ \widehat{\pi}(x) &\geq 0 \\ \sum_{x\in X} \widehat{\pi}(x) &= 1. \end{aligned}$$

Note that here, $\widehat{\pi}$ is not parametrized but is instead manipulated directly.

The entropy function turns out to be concave and the constraints are linear. As a result, we know that there is a local maximum which is also the global maximum.

We may solve this maximization problem using lagrange multipliers:

$$L = \sum_{x \in X} \widehat{\pi}(x) \ln \widehat{\pi}(x) - \sum_{j=1}^{n} \lambda_j \left(\sum_{x \in X} \widehat{\pi}(x) f_j(x) - \widehat{E}[f_j] \right) + \gamma \left(\sum_{x \in X} \widehat{\pi}(x) - 1 \right).$$

Solving this equation by setting $\frac{\partial L}{\partial \widehat{\pi}(x)} = 0$ for each $x \in X$, we get

$$\widehat{\pi}(x) = \frac{\exp\left(\sum_{j=1}^{n} \lambda_j f_j(x)\right)}{Z_\lambda}.$$

But this is the same estimate that we obtained by maximum likelihood estimate!

Furthermore, if we plug $\widehat{\pi}(x)$ back in to the expression for L we get

$$L = \frac{1}{m} \sum_{i=1}^{m} \ln q_\lambda(x_i).$$

So L is the same as the objective for the maximum likelihood estimate! The two methods are identical and give the same estimate.

Unfortunately, as a result, this method also suffers from the problem of overfitting. It also seems odd to require that $E_{\widehat{\pi}}[f_j] = \widehat{E}[f_j]$ given that $E_{\widehat{\pi}}[f_j]$ is only approximately equal to $\widehat{E}[f_j]$.

3.1 Relaxing the constraints

We can overcome both of these difficulties by relaxing the constraints used in the maximum entropy formulation. Through VC theory and some other techniques, we can in general obtain bounds of the form

$$|E_\pi[f_j] - \widehat{E}[f_j]| \leq \beta_j.$$

We may then use this inequality to replace the previous equality constraint. When we solve the lagrangian equation now, we will still obtain an estimate of π of the same form q_λ, but our objective will now have the form:

$$L = \frac{1}{m} \sum_{i=1}^{m} \ln q_\lambda(x_i) - \sum_{j=1}^{n} \beta_j |\lambda_j|.$$

The second term is a penalty term known as a lasso or L_1-regularization. As discussed previously, L_1-regularization is well suited to decrease overfitting when there are many features and some features are irrelevant for the classification task at hand.

4 Application to Conservation Biology

Modeling the distribution of a species is important for many reasons such as selecting an appropriate environment for a reserve that protects endangered species. The model as discussed above will also be useful for predicting how environmental changes such as climate changes will impact a particular species. Finally, the model above has been used to actually discover new species of animals.

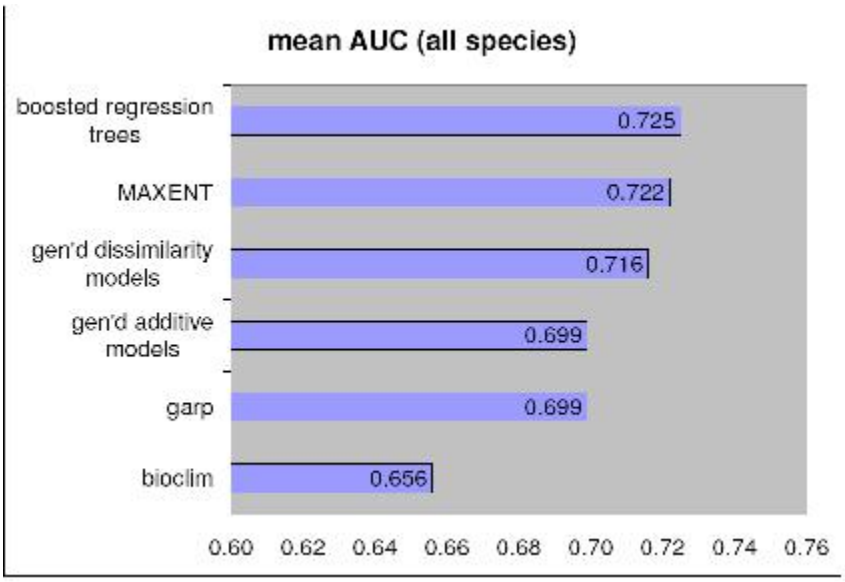

Figure 2: effectiveness of max entropy v.s other techniques

4.1 Modeling bake-off

Figure 2 shows the effectiveness of the max entropy approach with relaxed constraints in modeling the population distribution of various species, and compares it to other effectiveness of other approaches. The results for the graph were obtained during a modeling bake off. The effectiveness of the various approaches were measured by AUC, with 1 being a best. As we can see, the max entropy approach was a fairly effective model.

All the algorithms performed poorly, however, for the modeling population distribution in Canada, as can be seen in figure 3. The reason for this poor performance is severe sample bias. Biologists tended to sample warmer areas much more than colder areas of Canada. Removing this bias increased the effectiveness of max entropy greatly (see figure 4).

4.2 Discovering New Species

The biologist Raxworthy and his colleagues used the max entropy algorithm to model the population distribution of several species of gekkos and chameleons on Madagascar. The algorithm predicted the presence of these species in certain areas of the island that had not been closely investigated. When biologists investigated these areas, they discovered many new species.

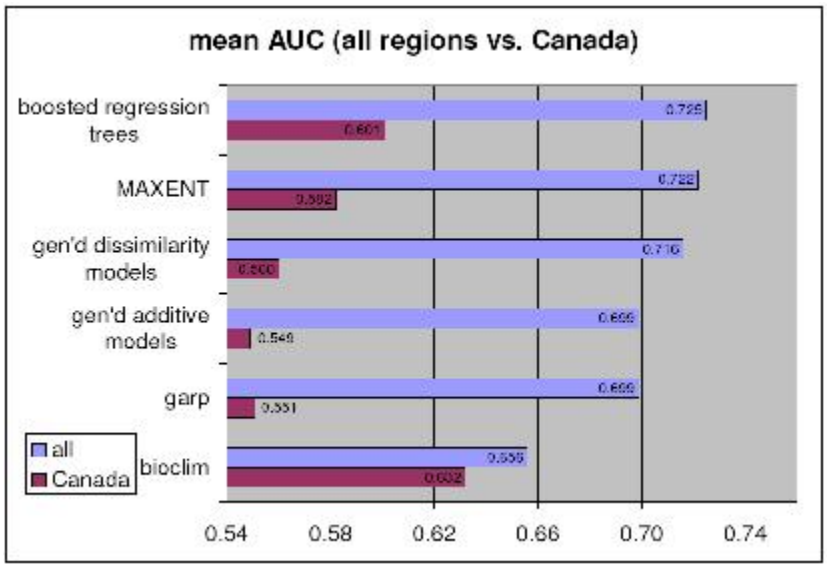

Figure 3: effectiveness of max entropy and other techniques for the canadian data

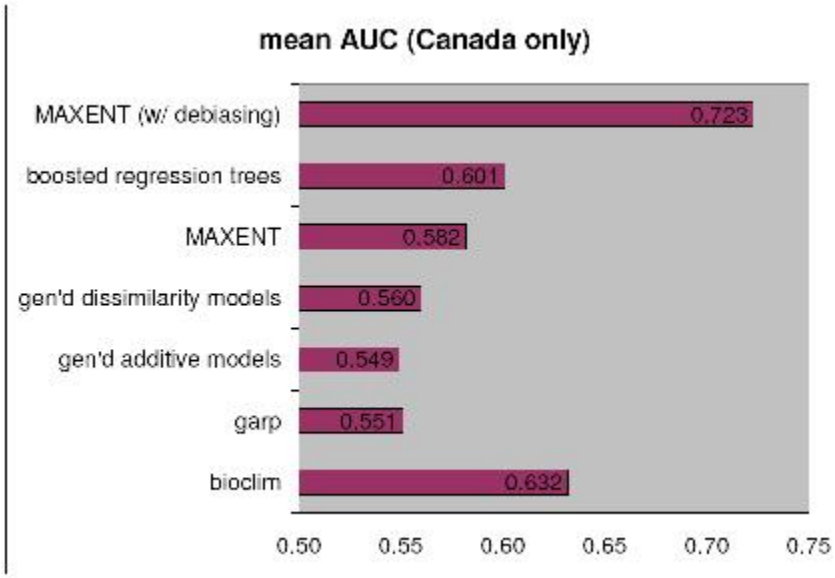

Figure 4: effect of removing biase on the max entropy algorithm