Maximum Entropy and Species Distribution Modeling

Rob Schapire
Steven Phillips
Miro Dudík

Also including work by or with: Rob Anderson, Jane Elith, Catherine Graham, Chris Raxworthy, NCEAS Working Group, ...
The Problem: Species Habitat Modeling

- goal: model distribution of plant or animal species
The Problem: Species Habitat Modeling

- **goal**: model distribution of plant or animal species
- **given**: presence records
The Problem: Species Habitat Modeling

- **goal**: model distribution of plant or animal species
- **given**: presence records

- **given**: environmental variables

  - precipitation
  - wet days
  - avg. temp.
  - ...
The Problem: Species Habitat Modeling

- **goal**: model distribution of plant or animal species
- **given**: presence records

- **given**: environmental variables
  - precipitation
  - wet days
  - avg. temp.

- **desired output**: map of range
**Biological Importance**

- **fundamental** question: what are survival requirements (niche) of given species?
- core problem for conservation of species
- first step for many applications:
  - reserve design
  - impact of climate change
  - discovery of new species
  - clarification of taxonomic boundaries
A Challenge for Machine Learning

- no negative examples
- very limited data
  - often, only 20-100 presence records
  - usually, not systematically collected
  - may be museum records years or decades old
- sample bias
  - (toward most accessible locations)
Our Approach

- assume presence records come from probability distribution $\pi$
- try to estimate $\pi$
- apply maximum entropy approach
Maxent Approach

- presence records for species X:

<table>
<thead>
<tr>
<th></th>
<th>altitude</th>
<th>July temp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>record #1</td>
<td>1410m</td>
<td>16°C</td>
</tr>
<tr>
<td>record #2</td>
<td>1217m</td>
<td>22°C</td>
</tr>
<tr>
<td>record #3</td>
<td>1351m</td>
<td>17°C</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...
### Maxent Approach

- presence records for species X:

<table>
<thead>
<tr>
<th></th>
<th>altitude</th>
<th>July temp.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>record #1</strong></td>
<td>1410m</td>
<td>16°C</td>
</tr>
<tr>
<td><strong>record #2</strong></td>
<td>1217m</td>
<td>22°C</td>
</tr>
<tr>
<td><strong>record #3</strong></td>
<td>1351m</td>
<td>17°C</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>1327m</td>
<td>17.2°C</td>
</tr>
<tr>
<td><strong>Stand. Dev.</strong></td>
<td>117m</td>
<td>3.2°C</td>
</tr>
</tbody>
</table>
Maxent Approach (cont.)

- data allow us to infer many facts, e.g.:
  - average altitude of species X’s habitat $\approx 1327$ m
  - average July temperature of species X’s habitat $\approx 17.2^\circ$C
  - stand. dev. of altitude of species X’s habitat $\approx 117$ m
Maxent Approach (cont.)

- data allow us to infer many facts, e.g.:
  - average altitude of species X’s habitat ≈ 1327m
  - average July temperature of species X’s habitat ≈ 17.2°C
  - stand. dev. of altitude of species X’s habitat ≈ 117m
  - probability species X lives above 1100m ≈ 0.78
  - probability species X lives above 1200m ≈ 0.62
Maxent Approach (cont.)

- data allow us to infer many facts, e.g.:
  - average altitude of species X’s habitat ≈ 1327m
  - average July temperature of species X’s habitat ≈ 17.2°C
  - stand. dev. of altitude of species X’s habitat ≈ 117m
  - probability species X lives above 1100m ≈ 0.78
  - probability species X lives above 1200m ≈ 0.62
  - each tells us something about true distribution
  - idea: find distribution satisfying all constraints
  - among these, choose distribution closest to uniform (i.e., of highest entropy)
This Talk

- theory
  - maxent with relaxed constraints
  - new performance guarantees for maxent
    - useful even with very large number of features (or constraints)
  - algorithm and convergence
- experiments and applications
The Abstract Framework

- $\pi \equiv \text{(unknown) true distribution}$

[Della Pietra, Della Pietra & Lafferty]
The Abstract Framework

- $\pi = \text{(unknown) true distribution}$
- given:
  - samples $x_1, \ldots, x_m \in X$
    - $x_i \sim \pi$
  - features $f_1, \ldots, f_n$
    - $f_j : X \rightarrow [0, 1]$

[Della Pietra, Della Pietra & Lafferty]
The Abstract Framework

- $\pi$ = (unknown) true distribution
- given:
  - samples $x_1, \ldots, x_m \in X$
    - $x_j \sim \pi$
  - features $f_1, \ldots, f_n$
    - $f_j : X \rightarrow [0, 1]$
- goal: find $\hat{\pi} = \text{estimate of } \pi$

[Della Pietra, Della Pietra & Lafferty]
Maxent and Habitat Modeling

- $x_i =$ presence record
- $X =$ all localities (discretized)
- $\pi =$ distribution of localities inhabited by species
  - ignores sample bias and dependence between samples
Maxent and Habitat Modeling

- $x_i =$ presence record
- $X =$ all localities (discretized)
- $\pi =$ distribution of localities inhabited by species
  - ignores sample bias and dependence between samples
- features: use raw environmental variables $v_k$
  - linear: $f_j = v_k$
**Maxent and Habitat Modeling**

- \( x_i = \) presence record
- \( X = \) all localities (discretized)
- \( \pi = \) distribution of localities inhabited by species
  - ignores sample bias and dependence between samples
- **features**: use raw environmental variables \( v_k \) or derived functions
  - linear: \( f_j = v_k \)
  - quadratic: \( f_j = v_k^2 \)
Maxent and Habitat Modeling

- $x_i =$ presence record
- $X =$ all localities (discretized)
- $\pi =$ distribution of localities inhabited by species
  - ignores sample bias and dependence between samples
- features: use raw environmental variables $v_k$ or derived functions
  - linear: $f_j = v_k$
  - quadratic: $f_j = v_k^2$
  - product: $f_j = v_k v_\ell$
Maxent and Habitat Modeling

- \( x_i \) = presence record
- \( X \) = all localities (discretized)
- \( \pi \) = distribution of localities inhabited by species
  - ignores sample bias and dependence between samples
- features: use raw environmental variables \( v_k \) or derived functions
  - linear: \( f_j = v_k \)
  - quadratic: \( f_j = v_k^2 \)
  - product: \( f_j = v_k v_\ell \)
  - threshold: \( f_j = \begin{cases} 1 & \text{if } v_k \geq a \\ 0 & \text{else} \end{cases} \)
Maxent and Habitat Modeling

- $x_i$ = presence record
- $X =$ all localities (discretized)
- $\pi =$ distribution of localities inhabited by species
  - ignores sample bias and dependence between samples
- features: use raw environmental variables $v_k$ or derived functions
  - linear: $f_j = v_k$
  - quadratic: $f_j = v_k^2$
  - product: $f_j = v_k v_\ell$
  - threshold: $f_j = \begin{cases} 1 & \text{if } v_k \geq a \\ 0 & \text{else} \end{cases}$
- # features can become very large (even infinite)
• $\tilde{\pi} = \text{empirical distribution}$
  • i.e., $\tilde{\pi}(x) = \#\{i : x_i = x\}/m$
• $\pi[f] = \text{expectation of } f \text{ with respect to } \pi$
  (so $\tilde{\pi}[f] = \text{empirical average of } f$)
• $\tilde{\pi}$ typically very poor estimate of $\pi$
Maxent

- $\tilde{\pi}$ typically very poor estimate of $\pi$
- but: $\tilde{\pi}[f_j]$ likely to be reasonable estimate of $\pi[f_j]$
• $\tilde{\pi}$ typically very poor estimate of $\pi$

• but: $\tilde{\pi}[f_j]$ likely to be reasonable estimate of $\pi[f_j]$

• so: choose distribution $\hat{\pi}$ such that

$$\hat{\pi}[f_j] = \tilde{\pi}[f_j]$$

for all features $f_j$
• \( \hat{\pi} \) typically very poor estimate of \( \pi \)
• but: \( \tilde{\pi}[f_j] \) likely to be reasonable estimate of \( \pi[f_j] \)
• so: choose distribution \( \hat{\pi} \) such that

\[
\hat{\pi}[f_j] = \tilde{\pi}[f_j]
\]

for all features \( f_j \)
• among these, choose one closest to uniform, i.e., of maximum entropy

[Jaynes]
• $\tilde{\pi}$ typically very poor estimate of $\pi$
• but: $\tilde{\pi}[f_j]$ likely to be reasonable estimate of $\pi[f_j]$
• so: choose distribution $\hat{\pi}$ such that

$$\hat{\pi}[f_j] = \tilde{\pi}[f_j]$$

for all features $f_j$

• among these, choose one closest to uniform, i.e., of maximum entropy [Jaynes]
• problem: can badly overfit, especially with a large number of features
A More Relaxed Version

- generally, only expect $\tilde{\pi}[f_j] \approx \pi[f_j]$
A More Relaxed Version

- generally, only expect $\tilde{\pi}[f_j] \approx \pi[f_j]$
- usually, can estimate upper bound on $|\tilde{\pi}[f_j] - \pi[f_j]|$
• generally, only expect $\tilde{\pi}[f_j] \approx \pi[f_j]$

• usually, can estimate upper bound on

$$|\tilde{\pi}[f_j] - \pi[f_j]|$$

• so: compute $\hat{\pi}$ to maximize $H(\hat{\pi})$ ( = entropy)
subject to

$$\forall j : |\tilde{\pi}[f_j] - \hat{\pi}[f_j]| \leq \beta_j$$

where $\beta_j =$ known upper bound

[Kazama & Tsujii]
Duality

- Can show solution must be Gibbs distribution:

\[ \hat{\pi}(x) = q \lambda(x) \propto \exp \left( \sum_j \lambda_j f_j(x) \right) \]
can show solution must be Gibbs distribution:

\[
\hat{\pi}(x) = q_\lambda(x) \propto \exp \left( \sum_{j} \lambda_j f_j(x) \right)
\]

in unrelaxed case, solution is Gibbs distribution that maximizes likelihood, i.e., minimizes:

\[
-\frac{1}{m} \sum_i \ln q_\lambda(x_i)
\]

negative log likelihood
Duality

• can show solution must be Gibbs distribution:

\[ \hat{\pi}(x) = q\lambda(x) \propto \exp \left( \sum_j \lambda_j f_j(x) \right) \]

• in unrelaxed case, solution is Gibbs distribution that maximizes likelihood

• in relaxed case, solution is Gibbs distribution that minimizes:

\[ \frac{1}{m} \sum_i \ln q\lambda(x_i) + \sum_j \beta_j |\lambda_j| \]

\( \text{negative log likelihood} \quad \text{“regularization”} \)
Equivalent Motivations

- maxent with relaxed constraints
- log loss with regularization
- MAP estimate with Laplace prior on weights $\lambda$
How Good Is Maxent Estimate?

- want to bound distance between $\hat{\pi}$ and $\pi$
  (measure with relative entropy)

$$\text{RE}(\pi \mid \mid \hat{\pi}) \leq$$
How Good Is Maxent Estimate?

- want to bound distance between $\hat{\pi}$ and $\pi$ (measure with relative entropy)
- can never beat “best” Gibbs distribution $\pi^*$

$$\text{RE}(\pi \parallel \hat{\pi}) \leq \text{RE}(\pi \parallel \pi^*) +$$
How Good Is Maxent Estimate?

- want to bound distance between $\hat{\pi}$ and $\pi$ (measure with relative entropy)
- can never beat “best” Gibbs distribution $\pi^*$
- additional term
  - $\to 0$ as $m \to \infty$
- depend on
  - number or complexity of features
  - “smoothness” of $\pi^*$

$$\text{RE}(\pi \mid \hat{\pi}) \leq \text{RE}(\pi \mid \pi^*) + \text{additional term}$$
with high probability, for all $\lambda^*$

$$RE(\pi \| \hat{\pi}) \leq RE(\pi \| \pi^*) + O \left(\|\lambda^*\|_1 \sqrt{\frac{\ln n}{m}}\right)$$

(for choice of $\beta_j$ based only on $n$ and $m$)

- $\pi^* = q\lambda^* = “best”$ Gibbs distribution
- $\|\lambda^*\|_1$ measures “smoothness” of $\pi^*$
- very moderate in number of features
Bounds for Infinite Binary Feature Classes

- assume binary features with VC-dimension $d$
- then with high probability, for all $\lambda^*$:

$$RE(\pi \parallel \hat{\pi}) \leq RE(\pi \parallel \pi^*) + \tilde{O}\left(\|\lambda^*\|_1 \sqrt{\frac{d}{m}}\right)$$

(for choice of $\beta_j$ based only on $d$ and $m$)
- e.g., infinitely many threshold features, but very low VC-dimension
Main Theorem

- both bounds follow from main theorem:
  - assume $\forall j : |\pi[f_j] - \hat{\pi}[f_j]| \leq \beta_j$
  - then
    \[
    \text{RE}(\pi \mid \hat{\pi}) \leq \text{RE}(\pi \mid \pi^*) + 2 \sum_j \beta_j |\lambda_j^*|
    \]

- preceding results are simple corollaries using standard uniform convergence results

- in practice, theorem tells us how to set $\beta_j$ parameters:
  use tightest bound available on $|\pi[f_j] - \hat{\pi}[f_j]|$
Finding an Algorithm

- want to minimize

\[ L(\lambda) = -\frac{1}{m} \sum_i \ln q\lambda(x_i) + \sum_j \beta_j |\lambda_j| \]

- no analytical solution
- instead, iteratively compute \( \lambda_1, \lambda_2, \ldots \) so that \( L(\lambda_t) \) converges to minimum
- most algorithms for maxent update all weights \( \lambda_j \) simultaneously
- less practical when very large number of features
Sequential-update Algorithm

- instead update just one weight at a time
- leads to sparser solution
- sometimes can search for best weight to update very efficiently
- analogous to boosting
  - weak learner acts as oracle for choosing function (weak classifier) from large space
- can prove convergence to minimum of $L$
Experiments and Applications

- broad comparison of algorithms
  - improvements by handling sample bias
- case study
- discovering new species
- clarification of taxonomic boundaries
• species distribution modeling “bake-off” comparing 16 methods
• 226 plant and animal species from 6 world regions
• mostly 10’s to 100’s of presence records per species
  • min = 2, max = 5822, average = 241.1, median = 58.5
• design:
  • training data:
    • incidental, non-systematic, presence-only
    • mainly from museums, herbaria, etc.
  • test data:
    • presence and absence data
    • collected in systematic surveys
Results

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Mean AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>boosted regression trees</td>
<td>0.725</td>
</tr>
<tr>
<td>MAXENT</td>
<td>0.722</td>
</tr>
<tr>
<td>gen'd dissimilarity models</td>
<td>0.716</td>
</tr>
<tr>
<td>gen'd additive models</td>
<td>0.699</td>
</tr>
<tr>
<td>garp</td>
<td>0.699</td>
</tr>
<tr>
<td>bioclim</td>
<td>0.656</td>
</tr>
</tbody>
</table>
Results

mean AUC (all species)

- boosted regression trees: 0.725
- MAXENT: 0.722
- gen'd dissimilarity models: 0.716
- gen'd additive models: 0.699
- garp: 0.699
- bioclim: 0.656

• newer statistical/machine learning methods (including maxent) performed better than more established methods
• reasonable to use presence-only incidental data
Maxent versus Boosted Regression Trees

- very similar, both mathematically and algorithmically, as methods for combining simpler features
- differences:
  - maxent is generative; boosting is discriminative
  - as implemented, boosting uses complex features; maxent uses simple features
- open: which is more important?
The Problem with Canada

- results for Canada are by far the weakest:

- apparent problem: very bad sample bias
  - sampling much heavier in (warm) south than (cold) north
can modify maxent to handle sample bias
  - use sampling distribution (assume known) as “default” distribution (instead of uniform)
  - then factor bias out from final model
problem: where to get sampling distribution
Sample Bias

- can modify maxent to handle sample bias
  - use sampling distribution (assume known) as “default” distribution (instead of uniform)
  - then factor bias out from final model
- problem: where to get sampling distribution
- typically, modeling many species at once
- so, for sampling distribution, use all locations where any species observed
## Results of Sample Debiasing

### Mean AUC (Canada only)

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAXENT (w/ debiasing)</td>
<td>0.723</td>
</tr>
<tr>
<td>boosted regression trees</td>
<td>0.601</td>
</tr>
<tr>
<td>MAXENT</td>
<td>0.582</td>
</tr>
<tr>
<td>gen'd dissimilarity models</td>
<td>0.560</td>
</tr>
<tr>
<td>gen'd additive models</td>
<td>0.549</td>
</tr>
<tr>
<td>garp</td>
<td>0.551</td>
</tr>
<tr>
<td>bioclim</td>
<td>0.632</td>
</tr>
</tbody>
</table>

- huge improvements possible using debiasing
Results of Sample Debiasing

![Bar chart showing mean AUC (all species) for different models]

- MAXENT (w/ debiasing): 0.752
- boosted regression trees: 0.725
- MAXENT: 0.722
- gen'd dissimilarity models: 0.716
- gen'd additive models: 0.699
- garp: 0.699
- bioclim: 0.656

- huge improvements possible using debiasing
Interpreting Maxent Models

- recall $q_\lambda(x) \propto \exp(\sum_j \lambda_j f_j(x))$
- for each environmental variable, can plot total contribution to exponent

value of environmental variable

threshold, $\beta=1.0$  
threshold, $\beta=0.01$  
linear+quadratic, $\beta=0.1$
Case Study: Microryzomys minutus

[with Phillips & Anderson]
Case Study: Microryzomys minutus

[with Phillips & Anderson]
Case Study: Microryzomys minutus
[with Phillips & Anderson]
• accurately captures realized range
• did not predict other wet montane forest areas where could live, but doesn’t
  • examined predictions of maxent on six of these (chosen by biologist)
  • found all had characteristics well outside typical range for actual presence records
    • e.g., four sites had July precipitation $\geq$ 5 standard deviations above mean
Finding New Species

- build models of several gekkos and chameleons of Madagascar (≈ 10-20 presence records each)
- identify isolated regions where predicted but not found
Finding New Species (cont.)

- combine all identified regions into single map
Finding New Species (cont.)

- combine all identified regions into single map
- many regions already well known areas of local endemism
Finding New Species (cont.)

- combine all identified regions into single map
- many regions already well known areas of local endemism
- survey regions not previously studied
New Species

- result: discovery of many new species (possibly 15-30)

Thanks to Chris Raxworthy for all maps and photos!
Clarifying Taxonomic Boundaries

- “cryptic species”: classified as *single* species, but suspected *mixture* of $\geq 2$ species
- sometimes, maxent model is wildly wrong if trained on all records
- but model is much more reasonable if trained on each sub-population *separately*
- gives strong evidence actually dealing with multiple species
- can then follow up with morphological or genetic study
Phelsuma madagascariensis subspecies

One species

Three species

Phelsuma mad. grandis
Phelsuma mad. kochi
Phelsuma mad. madagascariensis

Observed localities:
- Phelsuma grandis
- Phelsuma madagascariensis
- Phelsuma kochi

Modeled distributions (part B):
- Phelsuma grandis
- Phelsuma madagascariensis
- Phelsuma kochi
- Model overlap

This slide courtesy of Chris Raxworthy.
Summary

- maxent provides clean and effective fit to habitat modeling problem
- works with positive-only examples
- seems to perform well with limited number of examples
- theoretical guarantees indicate can be used even with a very large number of features
- other nice properties:
  - easy to interpret by human expert
  - can be extended to handle sample bias
- many biological applications