Maximum entropy distribution modeling

A: The maximum entropy approach to distribution modeling

Distribution modeling

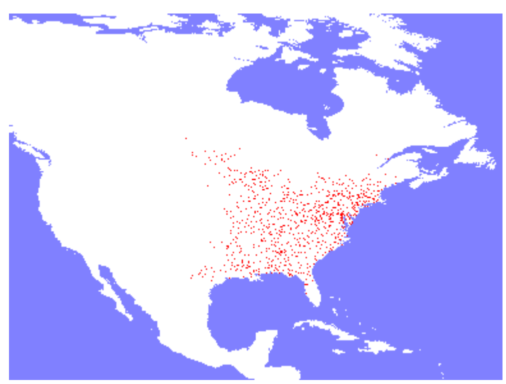
So far we've mentioned a few canonical distributions:
Normal, Poisson, binomial, beta, gamma, ...

- Is there any commonality to these?
- What do we do in situations where these models are not appropriate?

Maximum entropy: a broad framework for distribution modeling.

Modeling the geographical distribution of a species

Example: the yellow-throated vireo.



1611 sightings

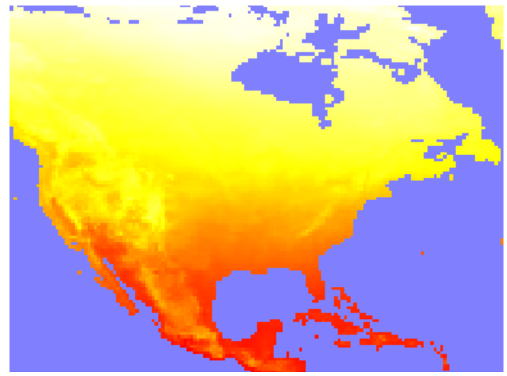
- Estimate the distribution of this species of bird
- Desired resolution: 386×286 grid ($= 110{,}396$ pixels)

Environmental features

- Want a distribution over $S = \{\text{locations}\}$
- Represent each pixel $x \in S$ by environmental features $T(x) = (T_1(x), \ldots, T_k(x))$

Features:

- Annual precipitation, number of wet days
- Average daily temperature, temperature range
- Elevation, aspect, slope



Annual average temperature

Use data to estimate $\mathbb{E}\,T_i(x)$ for $1 \leq i \leq k$.

The maximum entropy approach

1. Suppose we had no sightings at all.
   The most reasonable model, in the absence of any information, might just be the uniform distribution over S.

2. But we have some sightings, yielding estimates $\mathbb{E}\,T_i(x) = b_i$, $1 \leq i \leq k$.
   Pick a distribution p over S that respects these constraints, i.e., for all i,

$$\sum_{x \in S} p(x)\,T_i(x) = b_i,$$

   but is otherwise **as random as possible**.

B: Entropy

Entropy

The **entropy** of distribution p over finite set S is

$$H(p) = \sum_{x \in S} p(x) \log \frac{1}{p(x)}.$$

- Fair coin.
  - Specify S and p.
  - What is $H(p)$?
- Coin with bias $3/4$.
- Coin with bias 0.99.

Entropy: more examples

- Two fair coins.

- Uniform distribution over k outcomes.

Justifying entropy: Appealing properties

(1) Expansibility. If X has distribution $(p_1, \ldots, p_n)$ and Y has distribution $(p_1, \ldots, p_n, 0)$ then $H(X) = H(Y)$.

(2) Symmetry. Distribution $(p, 1-p)$ has the same entropy as $(1-p, p)$.

(3) Additivity. If X, Y are independent then $H(X, Y) = H(X) + H(Y)$.

(4) Subadditivity. $H(X, Y) \leq H(X) + H(Y)$.

(5) Normalization. A fair coin has entropy 1.

(6) "Small for small probability". The entropy of a coin of bias p goes to 0 as $p \downarrow 0$.

Aczel-Forte-Ng (1975): Entropy is the **only** measure that satisfies these 6 properties.

Additivity

If X, Y are independent then $H(X, Y) = H(X) + H(Y)$.

Subadditivity

$H(X, Y) \leq H(X) + H(Y)$.

Relation to KL divergence

- Canonical distance function between probability distributions: KL divergence.

$$K(p, q) = \sum_{x \in S} p(x) \log \frac{p(x)}{q(x)}.$$

- Can show that $K(p, q) \geq 0$, with equality iff $p = q$.

Lemma. If p is a distribution on S and u is uniform on S then $H(p) = \log |S| - K(p, u)$.

C: Distribution modeling by convex programming

Back to maximum entropy distribution modeling

- **Domain**: finite set S, e.g. geographical locations
- **Features** $T : S \to \mathbb{R}^k$, e.g. environmental features
- **Observed constraints**: $\mathbb{E}T_i(x) = b_i$, for $i = 1, \ldots, k$

Find the distribution p on S that has maximum entropy subject to the constraints.

$$\max \sum_{x \in S} p_x \ln \frac{1}{p_x}$$

$$\sum_{x \in S} p_x T_i(x) = b_i, \quad 1 \leq i \leq k$$

$$p_x \geq 0, \quad x \in S$$

$$\sum_{x \in S} p_x = 1$$

This is a convex optimization problem!

A slight generalization

Suppose we have a prior distribution π on S (e.g. distribution of a broader class of birds).

$$\min K(p, \pi)$$

$$\sum_{x \in S} p_x T_i(x) = b_i, \quad 1 \leq i \leq k$$

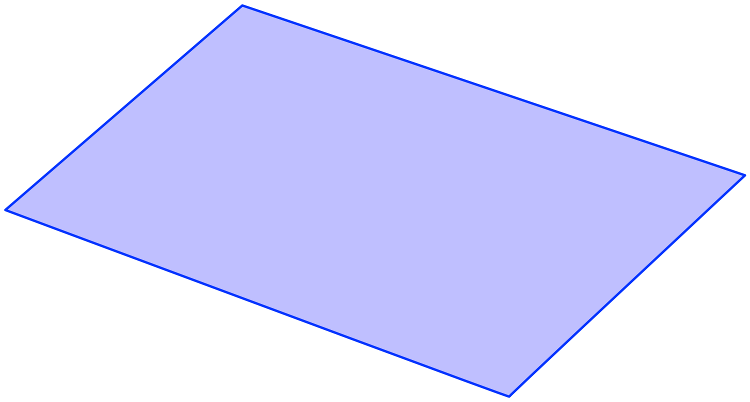
$$p_x \geq 0, \quad x \in S$$

$$\sum_{x \in S} p_x = 1$$

- Why is this a generalization?
- Why not $K(\pi, p)$?

Information projection

Think of this page as the probability simplex $\Delta = \{p \in \mathbb{R}^{|S|} : p_x \geq 0, \sum_x p_x = 1\}$



- Find the point $p \in L$ that is closest to π in KL divergence.
- We say p is the *I*-**projection** of π onto affine subspace L.

Solution by Lagrange multipliers

$$\min_{p \geq 0} \sum_x p_x \ln \frac{p_x}{\pi_x}$$

$$\sum_{x \in S} p_x T_i(x) = b_i, \quad 1 \leq i \leq k$$

$$\sum_{x \in S} p_x = 1$$

Form of solution

The solution has a specific functional form:

$$p(x) \;=\; \frac{1}{Z}\exp\left(\sum_{i=1}^{k}\theta_i T_i(x)\right)\pi(x) \;=\; \frac{1}{Z}e^{\theta\cdot T(x)}\pi(x)$$

where $T(x) = (T_1(x), \ldots, T_k(x))$ and $\theta = (\theta_1, \ldots, \theta_k)$.

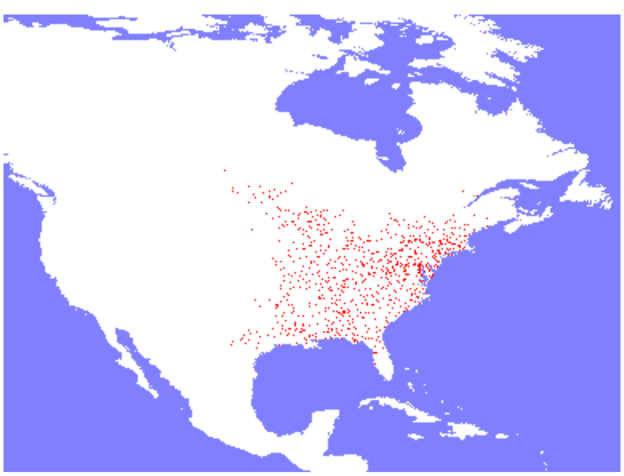
Return to example:

$$p(x) \propto \pi(x)\exp\left(\theta_1 \cdot \mathrm{avgtemp}(x) + \theta_2 \cdot \mathrm{elevation}(x) + \cdots\right)$$

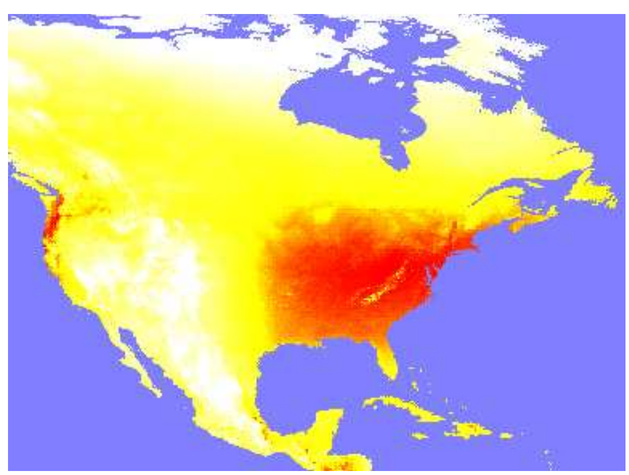
E.g. $\theta_2 = 0.81 \implies$ each additional unit of elevation multiplies the probability by $e^{0.81} \approx 2.25$

Yellow-throated vireo

Data: 1611 sightings



Maximum entropy distribution



In generality

Given any:

- Set of outcomes S
- Features $T(x) = (T_1(x), \ldots, T_k(x))$
- Base measure π on S
- Constraints $\mathbb{E}T_i(x) = b_i, 1 \leq k$

the I-projection of π onto the constraints is of the form

$$p(x) = \frac{1}{Z} e^{\theta \cdot T(x)} \pi(x)$$

where $\theta = (\theta_1, \ldots, \theta_k)$.

By varying (S, T, π), we get Gaussians, Poissons, binomials, Markov random fields...
These are all **exponential families**.

D: Exponential families of distributions

Exponential families: basics

Start with:

- **Outcome space** $S \subset \mathbb{R}^r$.
- **Base measure** $\pi : \mathbb{R}^r \to \mathbb{R}$.
- **Features** $T_1, \ldots, T_k : S \to \mathbb{R}$. Write $T(x) = (T_1(x), \ldots, T_k(x)) \in \mathbb{R}^k$.

The **exponential family generated by** (S, π, T) consists of distributions

$$p_\theta(x) = \frac{1}{Z_\theta} e^{\theta \cdot T(x)} \pi(x), \qquad \theta \in \mathbb{R}^k,$$

where **partition function** $Z_\theta = \sum_{x \in S} e^{\theta \cdot T(x)} \pi(x)$ or $\int_S e^{\theta \cdot T(x)} \pi(x) dx$.

Conventional form: Write $G(\theta) = \ln Z_\theta$, the **log partition function**. Then

$$p_\theta(x) = \exp(\theta \cdot T(x) - G(\theta)) \pi(x), \quad \theta \in \Theta$$

where $\Theta = \{\theta \in \mathbb{R}^k : G(\theta) < \infty\}$ is the **natural parameter space**.

The Bernoulli (coin flip) distribution

- How to express a coin of bias q in exponential family form?

- What are S, π, T in this case?

Poisson distribution

Recall: Poisson(λ) is a distribution over non-negative integers with $\Pr(k) = e^{-\lambda}\lambda^k/k!$.

- How to express in exponential family form?

- What are S, π, T in this case?

Normal distribution

- How to express $N(\mu, \sigma^2)$ in exponential family form?

- What are S, π, T in this case?

Fitting exponential family distributions

Pick an exponential family $\{p_\theta : \theta \in \Theta\}$ with

- Outcome space $S \subset \mathbb{R}^r$.
- Base measure $\pi : \mathbb{R}^r \to \mathbb{R}$.
- Features $T_1, \ldots, T_k : S \to \mathbb{R}$. Write $T(x) = (T_1(x), \ldots, T_k(x)) \in \mathbb{R}^k$.

Given data $x_1, \ldots, x_n \in S$, want to choose a model p_θ.

1. Maximum-likelihood coincides with the method of moments.
   The maximum-likelihood solution is the θ for which
   $$\mathbb{E}_{X \sim p_\theta}[T(X)] = \frac{1}{n}\sum_{i=1}^{n} T(x_i).$$
   Hence $T(x)$ is a **sufficient statistic** for estimating θ.
2. We can find this θ by solving a convex program.