Bridging the two cultures: Latent variable statistical modeling with boosted regression trees

Thomas G. Dietterich and Rebecca Hutchinson

Oregon State University
Corvallis, Oregon, USA
A Species Distribution Modeling Problem:

- eBird data
  - 12 bird species
  - 3 synthetic species
  - 3124 observations from New York State, May-July 2006-2008
  - 23 covariates
Two Cultures

- Probabilistic Graphical Models
  - MacKenzie, et al., 2002

- Flexible Nonparametric Models
  - Boosted Regression Trees
    - Friedman, 2001
    - Elith et al, 2006
    - Elith, Leathwick & Hastie, 2008
Occupancy-Detection Model

Occupancy features (e.g. elevation, vegetation)

Probability of occupancy (function of $X_i$)

True (latent) presence/absence
$Z_i \sim \text{Bern}(o_i)$

Observed presence/absence
$Y_{it} \mid Z_i \sim \text{Bern}(Z_i d_{it})$

Probability of detection (function of $W_{it}$)

Detection features (e.g. time of day, effort)

Sites
$\text{Sites} = \{i = 1, \ldots, M\}$

Visits
$\text{Visits} = \{t = 1, \ldots, T\}$

$X_i \rightarrow O_i \rightarrow Z_i \rightarrow Y_{it} \rightarrow W_{it}$
Parameterizing the model

\[ Z_i \sim P(Z_i | X_i) : \text{Species Distribution Model} \]
\[ P(Z_i = 1 | X_i) = o_i = F(X_i) \quad \text{“occupancy probability”} \]

\[ y_{it} \sim P(y_{it} | z_i, w_{it}) : \text{Observation model} \]
\[ P(Y_{it} = 1 | Z_i, W_{it}) = Z_i d_{it} \]
\[ d_{it} = G(W_{it}) \quad \text{“detection probability”} \]
Standard Approach: Log Linear (logistic regression) models

- \[ \log \frac{F(X_i)}{1-F(X_i)} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_J X_{iJ} \]
- \[ \log \frac{G(W_{it})}{1-G(W_{it})} = \alpha_0 + \alpha_1 W_{it1} + \cdots + \alpha_K W_{itK} \]
- Fit via maximum likelihood
- Can apply hypothesis tests to assess importance of covariates
  - \( H_0: \beta_1 = 0 \)
  - \( H_a: \beta_1 > 0 \)
Results on Synthetic Species with Nonlinear Interactions

- Predictions exhibit high variance because model cannot fit the nonlinearities well
A Flexible Predictive Model

- Predict the observation $y_{it}$ from the combination of occupancy covariates $x_i$ and detection covariates $w_{it}$

- Boosted Regression trees
  \[
  \log \frac{P(Y_{it}=1|X_i, W_{it})}{P(Y_{it}=0|X_i, W_{it})} = \beta_1 \text{tree}_1(X_i, W_{it}) + \cdots + \beta_L \text{tree}_L(X_i, W_{it})
  \]

- Fitted via functional gradient descent

- Model complexity is tuned to the complexity of the data
  - Number of trees
  - Depth of each tree
Results

- Systematically biased because it does not capture the latent occupancy
  - Underestimates occupancy at occupied sites to fit detection failures
- Much lower variance than the Occupancy-Detection model, because it can handle the non-linearities
Two Cultures: Summary

Probabilistic Graphical Models

- Advantages
  - Supports latent variables
  - Supports hypothesis tests on meaningful parameters

- Disadvantages
  - Model must be carefully designed (interactions? non-linearities?)
  - Data must be transformed to match modeling assumptions (linearity, Gaussianity)
  - Model has fixed complexity so either under-fits or over-fits

Flexible Nonparametric Models

- Advantages
  - Model complexity adapts to data complexity
  - Easy to use “off-the-shelf”

- Disadvantages
  - Cannot support latent variables
  - Cannot provide parametric hypothesis tests
The Dream

Probabilistic Graphical Models

Flexible Nonparametric Models

Flexible Nonparametric Probabilistic Models
A Simple Idea:
Parameterize $F$ and $G$ as boosted trees

- $\log \frac{F(X)}{1-F(X)} = f^0(X) + \rho_1 f^1(X) + \ldots + \rho_L f^L(X)$
- $\log \frac{G(W)}{1-G(W)} = g^0(W) + \eta_1 g^1(W) + \ldots + \eta_L g^L(W)$
- Perform functional gradient descent in $F$ and $G$
Results: OD-BRT

- Occupancy probabilities are predicted very well
Interpreting Non-Parametric Models: Partial Dependence Plots

- Simulate manipulating one variable (e.g., Distance of Survey)
- Visualize the predicted response
Partial Dependence Plot
Synthetic Species 3

- OD-BRT correctly captures the bimodal detection probability
Partial Dependence Plot
Blue Jay vs. Time of Day
Partial Dependence Plot
Blue Jay vs. Duration of Observation
Summary: We can have our cake (latent variables, interpretable submodels) and eat it too (have flexible, easy-to-use modeling tools)

- Easier to use
- More accurate
Concluding Remarks

- With limited data, the most accurate predictive model is much simpler than the “true model”
- Predictive accuracy on a single data set is not a sufficient criterion for a scientific model
Acknowledgements

- Liping Liu: Boosted Regression Trees in OD models
- Steve Kelling and colleagues at the Cornell Lab of Ornithology
- National Science Foundation Grants 0083292, 0307592, 0832804, and 0905885
Supporting Materials
Regression Trees

- Interactions are captured by the if-then-else structure of the tree
- Nonlinearities are approximated by piecewise constant functions
- Tree can be flattened into a linear model:

\[
Y_1 = -5 \cdot I(X_1 \geq 3, X_2 \geq 0) + 3 \cdot I(X_1 \geq 3, X_2 < 0) + 8 \cdot I(x_1 < 3, X_2 \geq 0) + 1 \cdot I(X_1 < 3, X_2 < 0)
\]
Friedman (2000), Mason et al. (NIPS 1999), Breiman (1996)

Fit a logistic regression model as a weighted sum of regression trees:

\[
\log \frac{P(Y = 1)}{P(Y = 0)} = tree^0(X) + \eta_1 tree^1(X) + \cdots + \eta_L tree^L(X)
\]

When “flattened” this gives a log linear model with complex interaction terms
L2-Tree Boosting Algorithm

- Let $F^0(X) = f^0(X) = 0$ be the zero function
- For $\ell = 1, \ldots, L$ do
  - Construct a training set $S^\ell = \{(X^i, \tilde{Y}^i)\}_{i=1}^N$
    - where $\tilde{Y}$ is computed as
      $\tilde{Y}^i = \frac{\partial L(L(F))}{\partial F} \bigg|_{F=F^\ell-1(X^i)}$ “how we wish $F$ would change at $X^i$”
  - Let $f^\ell = \text{regression tree fit to } S^\ell$
  - $F^\ell := F^{\ell-1} + \eta^\ell f^\ell$
- The step sizes $\eta^\ell$ are the weights computed in boosting
- This provides a general recipe for learning a conditional probability distribution for a Bernoulli or multinomial random variable
Alternating Functional Gradient Descent

- Loss function $L(F, G, y)$
- $F^0 = G^0 = f^0 = g^0 = 0$
- For $\ell = 1, \ldots, L$
  - For each site $i$ compute
    $$\tilde{z}_i = \frac{\partial L(F^{\ell-1}(x_i), G^{\ell-1}, y_i)}{\partial F^{\ell-1}(x_i)}$$
  - Fit regression tree $f^\ell$ to $\{(x_i, \tilde{z}_i)\}_{i=1}^M$
  - Let $F^\ell = F^{\ell-1} + \rho_\ell f^\ell$
  - For each visit $t$ to site $i$, compute
    $$\tilde{y}_{it} = \frac{\partial L(F^\ell(x_i), G^{\ell-1}(w_{it}), y_{it})}{\partial G^{\ell-1}(w_{it})}$$
  - Fit regression tree $g^\ell$ to $\{(w_{it}, \tilde{y}_{it})\}_{i=1,t=1}^{M,T_i}$
  - Let $G^\ell = G^{\ell-1} + \eta_\ell g^\ell$
## Multiple Visit Data

<table>
<thead>
<tr>
<th>Site</th>
<th>True occupancy (latent)</th>
<th>Visit 1 (rainy day, 12pm)</th>
<th>Visit 2 (clear day, 6am)</th>
<th>Visit 3 (clear day, 9am)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (forest, elev=400m)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B (forest, elev=500m)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C (forest, elev=300m)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D (grassland, elev=200m)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Covariates

<table>
<thead>
<tr>
<th>$X^{(1)}$</th>
<th>Human population per sq. mile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X^{(2)}$</td>
<td>Number of housing units per sq. mile</td>
</tr>
<tr>
<td>$X^{(3)}$</td>
<td>Percentage of housing units vacant</td>
</tr>
<tr>
<td>$X^{(4)}$</td>
<td>Elevation</td>
</tr>
<tr>
<td>$X^{(5)} \ldots X^{(19)}$</td>
<td>Percent of surrounding 22,500 hectares in each of 15 habitat classes from the National Land Cover Dataset</td>
</tr>
<tr>
<td>$W^{(1)}$</td>
<td>Time of day</td>
</tr>
<tr>
<td>$W^{(2)}$</td>
<td>Observation duration</td>
</tr>
<tr>
<td>$W^{(3)}$</td>
<td>Distance traveled during observation</td>
</tr>
<tr>
<td>$W^{(4)}$</td>
<td>Day of year</td>
</tr>
</tbody>
</table>
Synthetic Species 2

- $F$ and $G$ nonlinear

\[
\log \frac{o_i}{1 - o_i} = -2 \left[ x_i^{(1)} \right]^2 + 3 \left[ x_i^{(2)} \right]^2 - 2x_i^{(3)}
\]

\[
\log \frac{d_{it}}{1 - d_{it}} = \exp(-0.5w_{it}^{(4)}) + \sin(1.25w_{it}^{(1)} + 5)
\]
Predicting Occupancy

Synthetic Species 2
Open Problems

- Sometimes the OD model finds trivial solutions
  - Detection probability = 0 at many sites, which allows the Occupancy model complete freedom at those sites
  - Occupancy probability constant (0.2)

- Log likelihood for latent variable models suffers from local minima
  - Proper initialization?
  - Proper regularization?
  - Posterior regularization?

- How much data do we need to fit this model?
  - Can we detect when the model has failed?