Multivariate Prediction of Binary Response Variables

An Applied Lecture

Robert McDonald,
Graduate School of Design
Talk objectives

• Discuss statistical goal
  – Multivariate predictors, not multivariate response
• Discuss my example datasets
• Descriptive analysis techniques
• Three methodologies:
  – Discriminant analysis
  – Logistic regression
  – Classification and regression trees
• Tuning/ROC
Our statistical goal...

- Probability of occurrence $d$ at place $i$ varies based on $X_i$
- Binary examples:
  - Habitat vs. not…
- Multiple categories:
  - Cover classes…
- $X_i$ can include info on neighboring pixels or spatial location

$$P(d)_i = f(x_i) + \varepsilon$$
Deforestation dataset

- Classified Thematic Mapper Imagery 1991, 2001
- \( P(\text{deforestation}) \) is a function of:
  - Driving time
  - Distance to road
  - Municipality
  - Protected status
  - Distance to Stream
  - TCI
Logging dataset
Explanatory variables

- Physical controls: slope, soil type, value of forest timber
- Socioeconomic controls: distance to road, housing density, value of land…
An aside: presence-only data

- Many datasets missing null-case
- Examples: health data, species occurrence data...
- Severely limits the statistical interpretation possible
Presence-only data, the solution

• Solution 1: Create a plausible null-set
  – For the health data, we might…

• Solution 2: Calculate whether the presence data is different from a random grab of the whole landscape
  – Implicit null set is everywhere, equally!
Scale and the MAUP

• For all classification models, scale matters:
  – Extent
  – Grain
  – Coverage
• Often, averages (rates) are calculated for discrete spatial regions (e.g., townships, counties)
• Spatial patterns are very different with different units of analysis: Modifiable Areal Unit Problem
Averaging over discrete regions- Zonal statistics
Buffering and kernel density estimation - Focal statistics
Box-plots by response variable

- Sample one explanatory variable for each level of response variable
- Graph box-plots
- Univariate statistical tests possible…
Determining spatial regions of controlling regimes

• It’s important to note if there are different areas following different rules!
• In effect, if the function $f$ is not constant in form over the study area, you get *spatial averaging*
Discriminant analysis—
an overview

• Oldest technique, still widely used
• Related mathematically to an ANOVA:
  – ANOVA~ seeing if the mean of a continuous response variable is a function of discrete treatment groups
  – DA~ seeing if the value of a continuous predictor predicts a discrete response variable
• DA gives something called "discriminant function": a linear combination of explanatory variables that maximally separates the two groups
• Aside: Similar to Principal Components Analysis
Discriminant analysis - some issues

• Multivariate normal distribution of variables
• Homogeneity of variances/covariances
• Correlations between means and variances
• Discriminant functions somewhat difficult to interpret
Logistic regression and GLM

- Generalized linear models are extensions of the ANOVA/regression framework to other error distributions
- Logistic regression is a type of GLM with a logit link function and a binomial error distribution
- The predicted value is a probability, scaled from 0-1

\[
\begin{align*}
z &= \alpha + \beta X_1 + \beta X_2 + \ldots \\
P(d) &= \frac{e^z}{1 + e^z}
\end{align*}
\]
A data table...

Table 2
Regression coefficients from the GLM model, as well as the sensitivity of each coefficient

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Sensitivity (%)</th>
<th>d.f.</th>
<th>Deviance</th>
<th>Residual deviance</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.7160</td>
<td>-12.6</td>
<td>1</td>
<td>NA</td>
<td>1250.8</td>
<td>NA</td>
</tr>
<tr>
<td>Municipality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apex</td>
<td>1.0178</td>
<td>97.5</td>
<td>13</td>
<td>57.7</td>
<td>1193.1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Butner</td>
<td>0.6770</td>
<td>61.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carrboro</td>
<td>0.1119</td>
<td>8.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cary</td>
<td>0.0804</td>
<td>6.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapel Hill</td>
<td>-0.2916</td>
<td>-20.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Durham</td>
<td>-0.0846</td>
<td>-6.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Garner</td>
<td>-0.0987</td>
<td>-7.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hillsborough</td>
<td>0.0754</td>
<td>5.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morrisville</td>
<td>0.0776</td>
<td>6.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raleigh</td>
<td>-0.0721</td>
<td>-5.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wake Forest</td>
<td>-0.1183</td>
<td>-8.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fuquay-Varina</td>
<td>-0.0464</td>
<td>-3.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Holly Springs</td>
<td>-0.0636</td>
<td>-4.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Driving time-distance</td>
<td>0.0375</td>
<td>8.0</td>
<td>1</td>
<td>29.7</td>
<td>1163.4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SV within 30 m</td>
<td>-2.3382</td>
<td>-1.1</td>
<td>1</td>
<td>14.1</td>
<td>1149.3</td>
<td>0.0002</td>
</tr>
<tr>
<td>Slope</td>
<td>-0.3999</td>
<td>-3.2</td>
<td>1</td>
<td>12.3</td>
<td>1137.0</td>
<td>0.0003</td>
</tr>
<tr>
<td>Protected</td>
<td>-0.5177</td>
<td>-34.4</td>
<td>1</td>
<td>9.5</td>
<td>1127.8</td>
<td>0.0023</td>
</tr>
<tr>
<td>Development within 120 m</td>
<td>-1.5056</td>
<td>-0.5</td>
<td>1</td>
<td>7.9</td>
<td>1120.0</td>
<td>0.0030</td>
</tr>
<tr>
<td>Development within 30 m</td>
<td>-3.4191</td>
<td>-0.4</td>
<td>1</td>
<td>4.0</td>
<td>1115.9</td>
<td>0.0463</td>
</tr>
<tr>
<td>Distance to road</td>
<td>0.0007</td>
<td>1.2</td>
<td>1</td>
<td>2.6</td>
<td>1113.3</td>
<td>0.1052</td>
</tr>
</tbody>
</table>

See text for details.
Mapping predicted probability
Classification and regression trees

• Can have response variables that are continuous (RT) or discrete (CT)
• At each node in a tree, finds the optimal split on one of the explanatory variables -> hierarchical decision tree
• Trees are usually pruned using cross-validation to avoid overfitting
• Advantages: Simple to interpret, method is nonparametric and nonlinear
• Aside: CT similar to hierarchical cluster analysis
A CART tree...

In Apex, Butner, Carrboro, Cary, Hillsborough, Knightdale, or Wake Forest?

- Yes (17%)
- No

Distance to stream < 63m?

- Yes (51%)
  - Yes
  - No (61%)

- No

Driving time to UNC < 36 min?

- Yes (15%)
  - Yes
  - No

- No

Distance to road < 87m

- Yes (10%)
  - Yes (4%)
  - No (12%)

- No

In Durham or Morrisville?

- Yes (23%)
  - Yes (67%)
  - No (41%)

- No

TCI < 6.8

- Yes (21%)
  - Yes (17%)
  - No

Prop. of SV within 30m < 0.3

- Yes (36%)
  - Yes (41%)
  - No (0%)
Mapping “purity” at each node

- Can be mapped with a series of IF, THEN statements in Arc
- This is very tedious, however, for big trees
- Essential for understanding what the splits in the tree are doing!
Other classification techniques

• Neural networks: a very flexible non-linear extension of multiple logistic regression
• Nearest-neighbor rule: a simple non-parametric rule used in remote sensing
How to map residuals?

• For GLM (and CART), prediction continuous probability
How to map residuals?

• For GLM (and CART), prediction continuous probability
• Events are discrete
How to map residuals?

- For GLM (and CART), prediction continuous probability
- Events are discrete
- Solution: calculate proportion deforested
Model tuning

- Often, models are applied to new data: classification
- From a continuous probability (or purity) to a discrete prediction
- The optimal decision rule is often not the intuitive 0.5
Conclusions

• There are many techniques for classification—choose one whose assumptions match your data!
• Always do a thorough graphical examination of your data before statistical analysis
• Always graph residuals spatially to see where your model is failing
• A model’s failure often points toward new explanatory variables…
Questions?