DEALING WITH CATEGORICAL DATA

- Generalizing SOM for categorical data [1]
- Self-organizing map (SOM), proposed by Kohonen, is an unsupervised neural network which projects high-dimensional data onto a low-dimensional grid
- Typical approach to processing categorical values to a pre-process such as binary encoding
- Concept of Distance Hierarchies
CATEGORICAL DATA ANALYSIS - TOWARDS LOGIT MIXED MODELS [2]

- Commonly used - ANOVAs
- This paper suggests that Logit Models are better
- Logit Model - Logistic Regression
  - Regression to proportions of categorical outcome variables
  - Real binomially distributed data
  - Odds(p) = p/(1-p)
  - Logit = ln(p/(1-p))
    - Log odds
    - Centered around 0 => p=0.5
    - Range (-∞, ∞)
  - Logit Model - linear regression in log odds space
- Better suited to binomial categorical variables
SIMILARITY MEASURES FOR CATEGORICAL DATA [3]

- Similarity = $1/(1+\text{distance})$
- Classification of measures:
  - Diagonal
  - Off-diagonal
  - Diagonal + Off-diagonal

![Similarity Matrix](image_url)

Figure 1: Similarity Matrix for a Single Categorical Attribute
<table>
<thead>
<tr>
<th>Measure</th>
<th>$S_k(X_k, Y_k)$</th>
<th>$w_k, k = 1...d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Overlap</td>
<td>$\begin{cases} 1 &amp; \text{if } X_k = Y_k \ 0 &amp; \text{otherwise} \end{cases}$</td>
<td>$\frac{1}{d}$</td>
</tr>
<tr>
<td>2. Eskin</td>
<td>$\begin{cases} \frac{1}{2} \frac{\sum_k n_k^2}{n_k^2 + 2} &amp; \text{if } X_k = Y_k \ \frac{1}{2} &amp; \text{otherwise} \end{cases}$</td>
<td>$\frac{1}{d}$</td>
</tr>
<tr>
<td>3. IOF</td>
<td>$\begin{cases} 1 &amp; \text{if } X_k = Y_k \ \frac{1}{1 + \log f_k(X_k) \times \log f_k(Y_k)} &amp; \text{otherwise} \end{cases}$</td>
<td>$\frac{1}{d}$</td>
</tr>
<tr>
<td>4. OF</td>
<td>$\begin{cases} 1 &amp; \text{if } X_k = Y_k \ \frac{1}{1 + \log f_k(X_k) \times \log f_k(Y_k)} &amp; \text{otherwise} \end{cases}$</td>
<td>$\frac{1}{d}$</td>
</tr>
<tr>
<td>5. Lin</td>
<td>$\begin{cases} 2 \log \frac{\hat{p}_k(X_k)}{2 \log (\hat{p}_k(X_k) + \hat{p}_k(Y_k))} + \hat{p}<em>k(Y_k)} &amp; \text{if } X_k = Y_k \ \frac{1}{\sum</em>{i=1}^d \log p_i(X_i) + \log p_i(Y_i)} &amp; \text{otherwise} \end{cases}$</td>
<td>$\frac{1}{d}$</td>
</tr>
<tr>
<td>6. Lin1</td>
<td>$\begin{cases} \frac{1}{2} \sum_{q \in Q} \log \hat{p}<em>k(q) &amp; \text{if } X_k = Y_k \ 2 \log \sum</em>{q \in Q} \hat{p}_k(q) &amp; \text{otherwise} \end{cases}$</td>
<td>$\frac{1}{d}$</td>
</tr>
<tr>
<td>7. Goodall1</td>
<td>$\begin{cases} 1 - \frac{1}{2} \sum_{q \in Q} p_k^2(q) &amp; \text{if } X_k = Y_k \ 0 &amp; \text{otherwise} \end{cases}$</td>
<td>$\frac{1}{d}$</td>
</tr>
<tr>
<td>8. Goodall2</td>
<td>$\begin{cases} 1 - \frac{1}{2} \sum_{q \in Q} p_k^2(q) &amp; \text{if } X_k = Y_k \ 0 &amp; \text{otherwise} \end{cases}$</td>
<td>$\frac{1}{d}$</td>
</tr>
<tr>
<td>9. Goodall3</td>
<td>$\begin{cases} 1 - p_k^2(X_k) &amp; \text{if } X_k = Y_k \ 0 &amp; \text{otherwise} \end{cases}$</td>
<td>$\frac{1}{d}$</td>
</tr>
<tr>
<td>10. Goodall4</td>
<td>$\begin{cases} p_k^2(X_k) &amp; \text{if } X_k = Y_k \ 0 &amp; \text{otherwise} \end{cases}$</td>
<td>$\frac{1}{d}$</td>
</tr>
</tbody>
</table>
DEALING WITH MISSING VALUES [4]

- **MCAR** - Missing Completely at Random
  - Very rare

- **MAR** - What MAR means is missing, but conditional on some other 'X-variable' observed in the data set, although not on the 'Y-variable' of interest

- **NMAR** - Not Missing at Random, (or informatively missing, as it is often known) occurs when the missingness mechanism depends on the actual value of the missing data.
  - Difficult to model
The principal options for dealing with missing data are:
1. Analysing only the available data (i.e. ignoring the missing data)
2. Imputing the missing data with replacement values, and treating these as if they were observed (e.g. last observation carried forward, imputing an assumed outcome such as assuming all were poor outcomes, imputing the mean, imputing based on predicted values from a regression analysis)
3. Imputing the missing data and accounting for the fact that these were imputed with uncertainty (e.g. multiple imputation, simple imputation methods (as point 2) with adjustment to the standard error)
4. Using statistical models to allow for missing data, making assumptions about their relationships with the available data.
Four general recommendations for dealing with missing data in Cochrane reviews are as follows:

- Whenever possible, contact the original investigators to request missing data.
- Make explicit the assumptions of any methods used to cope with missing data: for example, that the data are assumed missing at random, or that missing values were assumed to have a particular value such as a poor outcome.
- Perform sensitivity analyses to assess how sensitive results are to reasonable changes in the assumptions that are made.
- Address the potential impact of missing data on the findings of the review in the Discussion section.
DEALING WITH MISSING VALUES

- Ignorability
  - MCAR, MAR are ignorable. Not NMAR
  - Percentage of missingness should not be too high

How to deal with missing data?

Case Deletion

Mean/Mode substitution

Single Imputation Methods

Dummy Variable method

Single regression

Model Based Methods

Max likelihood

Multiple imputation
SINGLE IMPUTATION METHODS

- **Mean/Mode substitution**
  - Reduces variability
  - Weakens covariance and correlation estimates in data

- **Dummy variable method**
  - Create indicator for missing value: 1 = value missing, 0 = value observed
  - Impute missing values to a constant
  - Disadvantage - results in biased estimates, not theoretically driven

- **Single regression**
  - Replace with predicted score from regression equation
  - Disadvantage
    - overestimates model fit and correlation estimates
    - Weakens variance
MODEL BASED METHODS

- **Max Likelihood**
  - Identifies set of parameter values that produces highest log-likelihood
  - **Advantage**
    - uses full information
    - Unbiased parameter estimates with MCAR or MAR data

- **Multiple imputation**
  - Repeat regression multiple times. Analyse. Pool results into 1 estimate
  - **Advantage**
    - Variability more accurate with multiple imputations for each missing value
  - **Disadvantage**
    - Room for error when specifying models
    - Cumbersome coding
TECHNIQUES FOR DEALING WITH MISSING VALUES IN CLASSIFICATION [6]

- Decision tree based inductive learning
- Missing values in training set
  - “Predictor” Method
  - Estimate value - deals with MNAR case!
- Missing values in test set
  - Explore all branches below current node
  - Surrogate split
  - Dynamic Path Generation
  - Lazy decision tree method
FEW OTHERS

- Principal Component Analysis for Large Scale Problems with Lots of Missing Values [7]
  - High dimensional data and majority values are missing
  - MAR data

- Gower Distance [8]
  - Dissimilarity between two rows is weighted mean of contributions of each variable
  - Weight = 0 if 1 or both values are missing
  - Else weight is 1
REFERENCES


