Category Selection for Multinomial Data

Rebecca Baker

Department of Mathematical Sciences, University of Durham

The multinomial NPI model

Model for learning from multinomial data

- inferences about a future observation
- in form of a probability interval
- based entirely on past observations

Have observed $Y_1, ..., Y_n$, want to find out about Y_{n+1}

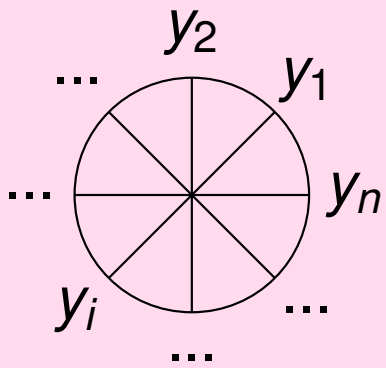
- each observation belongs to a particular category
- K categories in total
- we have already observed $c_1, ..., c_k$
- n_j observations in category c_j

Event of interest is ($Y_{n+1} \in E$) where E is a subset of the K categories

The probability wheel representation

Represent data on a **probability wheel**

- Y_{n+1} has probability $\frac{1}{n}$ of being in each slice

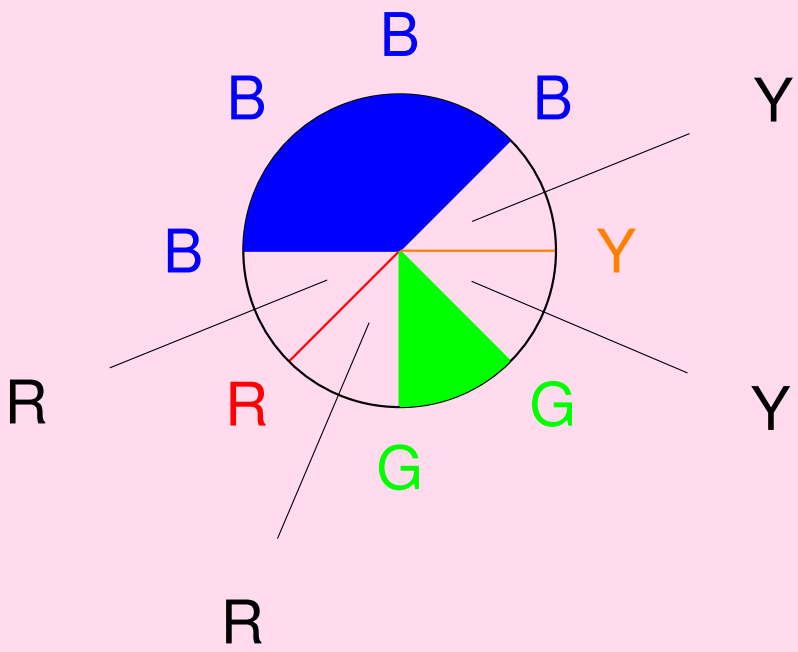


- Slice bordered by two observations in the same category is assigned to this category
- Slice bordered by two observations in different categories may be assigned to any available category

Note: Each category may only be represented by a single segment of the wheel.

Deriving lower probabilities

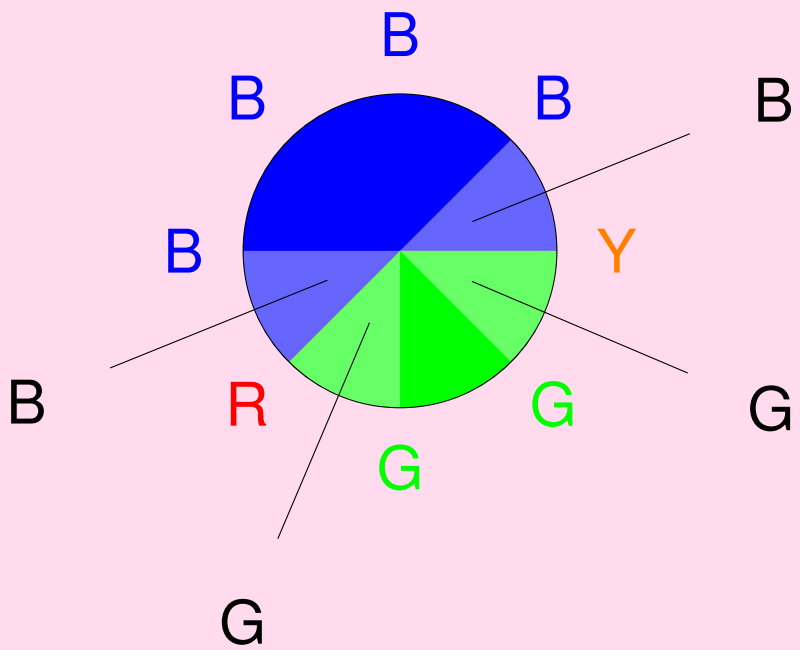
- Possible categories are blue (B), green (G), red (R), yellow (Y), pink (P) and orange (O)
- Event $E = \{B, G, P\}$



- $\underline{P}(Y_{n+1} \in E) = \frac{4}{8}$

Deriving upper probabilities

- Possible categories are blue (B), green (G), red (R), yellow (Y), pink (P) and orange (O)
- Event $E = \{B, G, P\}$



- $\overline{P}(Y_{n+1} \in E) = 1$

Research topics

NPI with subcategories

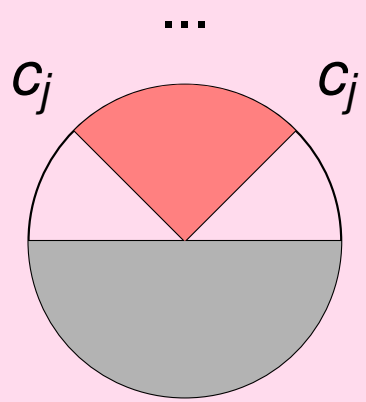
- A generalised NPI model to deal with data described at subcategory level
- Enables consistent inferences at different levels of detail

Category selection

- A generalised NPI model which uses inferences about multiple future observations
- Selection of an optimal category or subset of categories which meets some specified probability criterion
  - What are the relevant lower and upper probabilities?
  - How large does the subset need to be?

NPI for multiple future observations

We derive new NPI lower and upper probabilities using m future observations



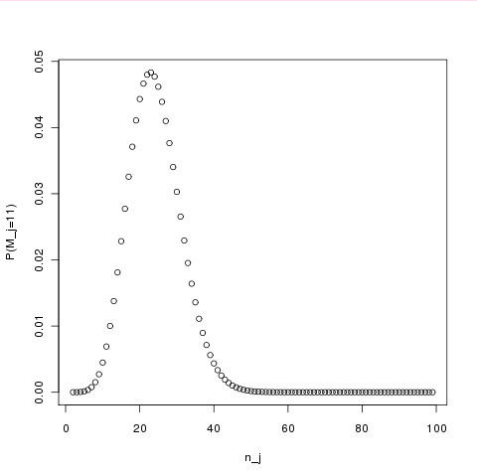
- There are $\binom{n+m-1}{m}$ arrangements of m future observations amongst the n slices of the wheel
- There are $\binom{(s-1)+f}{f}$ arrangements of f future observations within a segment made up of s slices

Selecting a single category to maximise $P(M_j = m_j)$

Problem: Select the category which maximises $\underline{P}(M_j = 11)$

- We have observed 20 B, 27 G, 25 R, 28 Y

By theorem, $n_j = 23$ will maximise this probability



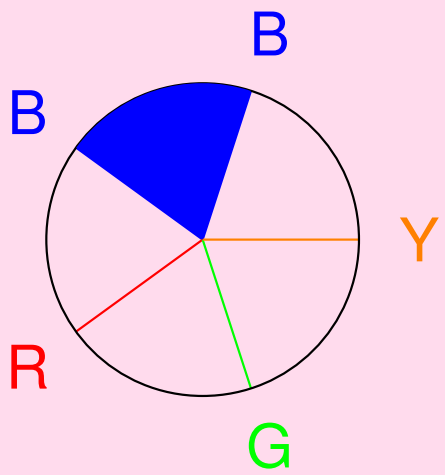
- Closest values are $n_B = 20$ and $n_R = 25$
- $\underline{P}(M_B = 11) = 0.0443$
- $\underline{P}(M_R = 11) = 0.0462$

The category we should select is R.

Selecting a single optimal category: $P(M_j \geq m_j)$

Problem: Select the optimal category for the event $M_j \geq \frac{m}{3}$

- We have observed 2 B, 1 G, 1 R, 1 Y



Take $m = 3$

- $P(M_B \geq 1) = [\frac{15}{35}, \frac{31}{35}]$
- $P(M_G \geq 1) = P(M_R \geq 1) = P(M_Y \geq 1) = [0, \frac{25}{35}]$

The category we should select is B.

Selecting an optimal subset of categories: $P(M_S \geq m_s)$

Problem: Select the optimal subset such that $\underline{P}(M_{S_i} \geq 1) \geq 0.8$

Category	A	B	C	D	E	F	G	H
Observations	25	20	18	13	10	9	5	0

i	S_i	$P(M_{S_i} \geq 1)$	$P(M_{S_i} \geq 2)$
1	A	$[0.4206, 0.4505]$	$[0.0594, 0.0695]$
2	A,B	$[0.6727, 0.7166]$	$[0.1873, 0.2234]$
3	A-C	$[0.8376, 0.8822]$	$[0.3624, 0.4378]$
4	A-D	$[0.9196, 0.9543]$	$[0.5204, 0.6257]$
5	A-E	$[0.9697, 0.9846]$	$[0.6903, 0.7754]$
6	A-F	$[0.9945, 0.9980]$	$[0.8655, 0.9220]$
7	A-G	$[0.9998, 1.0000]$	$[0.9802, 1.0000]$
8	A-H	$[1.0000, 1.0000]$	$[1.0000, 1.0000]$

The subset we should select is $S_3 = \{A, B, C\}$.

Future research

Classification

- Classification trees with NPI probabilities
- Investigating naive classification with NPI

NPI in finance

References

- Augustin, T. and Coolen, F.P.A. (2004) Nonparametric predictive inference and interval probability *Journal of Statistical Planning and Inference*, **124**, 251-272.
- Coolen, F.P.A. and Augustin, T. (2005) Learning from multinomial data: a nonparametric predictive alternative to the Imprecise Dirichlet Model *ISIPTA '05*, 125-134.
- Coolen, F.P.A. (2006) On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, **15**, 21-47.
- Coolen, F.P.A. and Augustin, T. (2009) A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories *International Journal of Approximate Reasoning*, **50**, 217-230.