Correlated Bernoulli Process using De Bruijn Graphs

LOUISE KIMPTON

EMAIL: LMK212@EXETER.AC.UK
Introduction

- Want to create chains of 0’s and 1’s that cluster or stick together

- Put structure into a Bernoulli distribution to make a correlated Bernoulli process

- We do this using de Bruijn graphs
Motivation

- Improvement on logistic regression and classification
- Need to include correlation between points
- Look for a clean boundary with no drop-outs
De Bruijn Graphs

- Directed graphs where nodes consist of all possible length \( m \) sequences (words) given a set of symbols
- \( m \) is the word length which controls how spread the correlation is (how many points the current point is dependent on)
- A probability is associated with each arc of the graph – gives the probability of transitioning from word to word

\[ P_{ij} \] – probability of transitioning from word \( i \) to word \( j \)

- Use symbols 0 and 1 to correspond to regions
Markov Properties

- Can write the transition probabilities in matrix form
- Then can use this to generate chains of 0s and 1s
- Can create stickiness in the chains by choosing specific transition probabilities
- Marginal probabilities stay the same over time but 0s and 1s are grouped together

\[ T = \begin{pmatrix} 1 - p_{00}^{01} & p_{00}^{01} & 0 & 0 \\ 0 & 1 - p_{01}^{01} & 0 & 0 \\ 1 - p_{10}^{01} & p_{10}^{01} & 0 & 0 \\ 0 & 1 - p_{11}^{01} & p_{11}^{01} & 0 \end{pmatrix} \]
Examples
Examples
Run Length Distribution

Word Length

\[ m = 2 \]

\[ P(\text{run length} = n) = \begin{cases} 
  p_{01}^{10} & \text{for } n = 1 \\
  p_{01}^{11} (p_{11}^{11})^{n-2} p_{11}^{10} & \text{for } n \geq 2, 
\end{cases} \]

\[ E[\text{run length}] = p_{01}^{10} + \frac{p_{01}^{11} (1 - (p_{11}^{10})^2)}{p_{11}^{11} p_{11}^{10}}. \]
Run Length Distribution

Word Length
\( m \geq 3 \)

\[
P(\text{Run Length} = n) = \begin{cases} 
\sum_{i=0}^{2^{m-2}-1} \pi(i) \frac{2^3(i \mod 2^{m-3})+2}{p_{4i+1}} \\
\sum_{i=0}^{2^{m-2}-1} \pi(i) \frac{2^3(i \mod 2^{m-3})+3}{p_{4i+1}} \\
\quad \times \left[ \prod_{j=1}^{n-1} p_{2^{j+2}(i \mod 2^{m-3})+(2^{j+1}-1)} \right] \\
\quad \times \left[ \prod_{j=1}^{m-2} p_{2^{j+2}(i \mod 2^{m-3})+(2^{j+1}-1)} \right] \\
\quad \times \left[ \left( \frac{p_{2^m-1}}{p_{2^m-2}} \right)^{n-m} \right] \\
\end{cases} 
\]

for \( n = 1 \)

for \( n \geq m \).

where,

\[
\pi(i) = \sum_{j=0}^{2^{m-1}-1} \prod_{k=0}^{m-3} p_{2^k(j \mod 2^{m-k})+\sum_{s=1}^{k+1} 2^{k-s}[(\frac{1}{2^{m-s-2}}(i-(i \mod 2^{m-2})) \mod 2)]} \pi(j) 
\]
Transition Likelihood

\[ p_{ij} \] – transition probability for the word \( ij \)

\[ n_{ij} \] – number of words, \( ij \), in the sequence

\[ m = 2 \]

\[ \mathcal{L} = (p_{00}^{00})^{n_{00}^{00}} (p_{00}^{01})^{n_{00}^{01}} (p_{01}^{10})^{n_{01}^{10}} (p_{01}^{11})^{n_{01}^{11}} (p_{10}^{00})^{n_{10}^{00}} (p_{10}^{01})^{n_{10}^{01}} (p_{11}^{10})^{n_{11}^{10}} (p_{11}^{11})^{n_{11}^{11}} (1 - p_{00}^{01})^{n_{00}^{01}} (1 - p_{01}^{11})^{n_{01}^{11}} (1 - p_{10}^{01})^{n_{10}^{01}} (1 - p_{11}^{11})^{n_{11}^{11}} \]

\[ m \geq 3 \]

\[ \mathcal{L} = \prod_{i=0}^{2^m - 1} \left( p_{i} \right)^{\frac{i \pmod{2^m+1}}{2^m+1}} \frac{1}{2} \frac{2^m+1}{2} [-2(-1)^{i+1}(i+1) - 3(-1)^{i+1} - 1] \]

\[ = \prod_{i=0}^{2^m - 1} \left( 1 - p_{i}^{(2i+1) \pmod{2^m}} \right)^{n_{i}^{(2i+1) \pmod{2^m}} - 1} \left( p_{i}^{(2i+1) \pmod{2^m}} \right)^{n_{i}^{(2i+1) \pmod{2^m}}} \]
Conjugate Prior

Likelihood (m=2)

\[ \mathcal{L} = (p_{00}^{00})^{n_{00}^{00}} (p_{01}^{01})^{n_{00}^{01}} (p_{01}^{00})^{n_{01}^{01}} (p_{01}^{10})^{n_{01}^{11}} (p_{01}^{10})^{n_{01}^{11}} (p_{10}^{00})^{n_{10}^{00}} (p_{10}^{01})^{n_{10}^{01}} (p_{11}^{00})^{n_{10}^{10}} (p_{11}^{01})^{n_{10}^{10}} (p_{11}^{11})^{n_{11}^{11}} \]

Posterior with beta prior

\[ P \propto (p_{00}^{00})^{n_{00}^{00}} (p_{00}^{01})^{n_{00}^{01}} (p_{01}^{00})^{\beta_1 - 1} (p_{00}^{01})^{\alpha_1 - 1} \times \]
\[ (p_{01}^{10})^{n_{01}^{10}} (p_{10}^{01})^{n_{01}^{11}} (p_{01}^{10})^{\beta_2 - 1} (p_{01}^{10})^{\alpha_2 - 1} \times \]
\[ (p_{10}^{10})^{n_{10}^{10}} (p_{10}^{11})^{n_{10}^{11}} (p_{01}^{10})^{\beta_3 - 1} (p_{01}^{11})^{\alpha_3 - 1} \times \]
\[ (p_{11}^{10})^{n_{11}^{10}} (p_{11}^{11})^{n_{11}^{11}} (p_{01}^{11})^{\beta_4 - 1} (p_{01}^{11})^{\alpha_4 - 1} \]

\[ = (p_{00}^{00})^{n_{00}^{00} + \beta_1 - 1} (p_{00}^{01})^{n_{00}^{01} + \alpha_1 - 1} (p_{01}^{00})^{n_{01}^{01} + \beta_2 - 1} (p_{01}^{10})^{n_{01}^{11} + \alpha_2 - 1} (p_{10}^{01})^{n_{10}^{10} + \beta_3 - 1} (p_{10}^{11})^{n_{11}^{11} + \alpha_3 - 1} (p_{11}^{10})^{n_{11}^{10} + \beta_4 - 1} (p_{11}^{11})^{n_{11}^{11} + \alpha_4 - 1} \]

pdf of beta distribution:

\[ P(x; \alpha, \beta) \propto x^{\alpha - 1} (1 - x)^{\beta - 1} \]
Inference for word length m

The posterior is now a product of beta densities. With the conjugate relationship, we can state the following:

\[ P(seq) = \int P(seq|p)P(p)dp = \frac{\Gamma(n_{00}^0 + \beta_1)\Gamma(n_{01}^0 + \alpha_1)}{\Gamma(n_{00}^0 + n_{01}^0 + \beta_1 + \alpha_1)} \times \frac{\Gamma(n_{01}^{10} + \beta_2)\Gamma(n_{01}^{11} + \alpha_2)}{\Gamma(n_{01}^{10} + n_{01}^{11} + \beta_2 + \alpha_2)} \times \frac{\Gamma(n_{10}^{00} + n_{10}^{01} + \beta_3 + \alpha_3)\Gamma(n_{11}^{00} + n_{11}^{01} + \beta_4 + \alpha_4)}{\Gamma(n_{10}^{00} + n_{10}^{01} + n_{10}^{10} + n_{10}^{11} + n_{11}^{00} + n_{11}^{01} + n_{11}^{10} + n_{11}^{11} + \beta_3 + \alpha_3 + \beta_4 + \alpha_4)} \]

For \( m \geq 3 \), this becomes:

\[ \int P(seq|p)P(p)dp = \prod_{i=0}^{2^{m}-1} \frac{\Gamma(n_i^{(2i+1) \mod 2^m} - 1 + \beta_i + 1)\Gamma(n_i^{(2i+1) \mod 2^m} + \alpha_i)}{\Gamma(n_i^{(2i+1) \mod 2^m})}\quad \]

Bayes factors are calculated for each model with word lengths \( m=1,\ldots,10 \), so that the word length that best represents the given sequence is chosen.
2d De Bruijn Graph

- Similar to the 1d version, but with a different word structure
- Words are formed by including all points that are a certain number of points away moving only upwards and to the right
- Can find the 1d equivalence for each 2d word so that we can apply the same theory
- Should be extendable to n dimensions
Non-directional de Bruijn process

- Direction does not make logical sense in a spatial grid
- Attempt to remove the direction, but keep the de Bruijn structure
- Change the form of the word, but inference remains the same
Conclusion

- Create chains of 0’s and 1’s with correlation using de Bruijn graphs
- Developed a run length distribution and inference
- Working on the 2d version – with hope to eventually take out the directionality
- Apply the method to applications with classification problems