Correlated Bernoulli Processes Using De Bruijn Graphs

L. Kimpton
University of Exeter

Supervisor(s): Prof. P. Challenor (University or Exeter) and Dr. D. Williamson (University of Exeter)

Ph.D. expected duration: Sep. 2016 - Mar. 2020

Adress: Room 601
Laver Building
University of Exeter Streatham Campus
North Park Road Exeter
EX4 4QE

Email: lmk212@exeter.ac.uk

Abstract:

The aim of my current PhD work is to produce a correlated Bernoulli process using a concept known as de Bruijn graphs [1, 2]. The motivation for this work came from studying numerical models which have two distinct regions in output space where a classification method is required. For example, we may have a computer model that fails to complete for specific input regions, and we would like to predict where to avoid running the model or incorrectly running an emulator.

A widely used method for classification is logistic regression, which produces a distribution for the predictive class membership of being in one of the two regions. When sampling from this to make predictions, the current practice is to draw from an independent Bernoulli distribution. Drawing a 0 would represent one of the regions and drawing a 1 would represent the other. However, drawing marginally means that any correlation that was being considered between data is lost and this can result in large numbers of misclassifications. This is especially true near the boundaries of the regions where we have very high uncertainty.

Although we can take many independent Bernoulli draws and average them to create a smooth boundary, this is still not a full solution as we would have to assign a threshold to find the two regions. Therefore, the aim of my work is to produce an equivalence to a Bernoulli process where correlation is incorporated when making draws or samples. This novel process should have a high correlation between points that are close together and a low correlation for points that are far apart. In a one dimensional scenario, this would correspond to being able to produce sequences of 0s and 1s such that like symbols cluster together instead of appearing fairly random. We will then hopefully see a clean cut boundary between regions when making classification predictions, instead of having frequent misclassifications.

We use the structure from de Bruijn Graphs. A de Bruijn graph is a directed graph, where given a set of 'letters', V, and a 'word' length, m, the nodes of the graph consist of all possible sequences of V of length m. Edges are drawn between node pairs in such a way that the connected nodes have overlaps of $m-1$ nodes. An edge is created by removing the first symbol and adding a new symbol to the end from V. So from each vertex, $(v_1, ..., v_m) \in V^m$, there is an edge to vertex $(v_2, ..., v_m, v) \in V$ for every $v \in V$. The word length controls the number of states that each individual state is dependent on, increasing correlation over a wider area. On each directed edge of the de Bruijn graph, we are able to assign a probability of transitioning from the previous node to the next. We can thus immediately see that there is a connection to Markov chains. However, we make a key definition here that de Bruijn graphs have a Markov property on the de Bruijn

word and not the letter. I.e. the current word is dependent on only the previous word in the sequence and no other. This means that we can create far more structure than if it were simply the letters that were Markov.

To link de Bruijn graphs with this idea of a correlated Bernoulli process, we will be dealing with the set of $s = 2$ letters, $V = \{0, 1\}$, and we can make the following definition:

Definition (de Bruijn Process): The de Bruijn process is a process to produce sequences of 'letters' from the set, $V = \{0, 1\}$, where correlation is included through a de Bruijn graph structure with length m 'words'. There is defined to be a Markov property on the de Bruijn words but not on the letters such that for time step, t:

$$P(X_t = i_t | X_{t-1} = i_{t-1}, X_{t-2} = i_{t-2}, ..., X_0 = i_0) = P(X_t = i_t | X_{t-1} = i_{t-1}) = p_{i_{t-1}}^{i_t}, \quad (1)$$

for random variables, X, and where $p_{i_{t-1}}^{i_t}$ is the probability of transitioning from the word i_{t-1} to word i_t.

We can then define several properties for the de Bruijn process. The most interesting of these is a run length distribution which specifies the probability of observing each length of runs of 1s (or 0s) in a sequence for a given de Bruijn process. The run length gives an idea of the number of consecutive 1s (or 0s) in a row, giving a measure of how 'sticky' a sequence generated from a specific de Bruijn process is likely to be. From this we can calculate the run length expectation, variance and generating functions.

We have also developed a method of inference for the de Bruijn process so that given a sequence of 0s and 1s, we can estimate the de Bruijn process that was used to create it. This involves estimating both the word length, m, and the transition probabilities, p, that determine the correlated structure of the corresponding Markov chains.

Following on from the 1d de Bruijn process, we have also looked into ways to expand the method to higher dimensions, and whether we can remove the unnatural direction that is associated with de Bruijn graphs.

References

[1] N G De Bruijn. A combinatorial problem. *Communicated by Prof. W Van Der Woude*, 1946.

[2] I J Good. Normal Recurring Decimals. *Journal of the London Mathematical Society*, 21(3), 1946.

Short biography – I have been a PhD student in Mathematics at the University of Exeter for the past three years, and hope to graduate this year. Previously to this I studied for an MMath, also at Exeter. My PhD thesis is entitled "Uncertainty Quantification for Models with Two Regions of Solution", which I am completing under the supervision of Prof. Peter Challenor, funded by EPSRC.