Bayesian Sequence Prediction with Mult-step modeling

Longhai Li

Department of Statistics

University of Toronto

23 March 2006

Joint work with Radford Neal

Work still in progress...

Question and Motivation

- We want to predict the next state of a sequence given its previous states. Taking binary sequences as example, we want to know the **transition probability** given a **history pattern**:

$$p_{x_1,\cdots,x_{n-1}} = P(x_n = 1 \mid x_1, \cdots, x_{n-1})$$

- A simple way of learning $p_{x_1,\cdots,x_{n-1}}$ from the data is by calculating the frequency of $x_n = 1$ following $x_1, \cdots, x_{n-1}$.
- The number of all patterns of length $n-1$ is 2^{n-1}. It is often the case that the number of available sequences is much less than this number, as presents difficulty for statistical learning. For instance, some patterns in test sequences may have never been expressed in training sequences or expressed by very few cases, the estimation of their parameters with above simple method is impossible or inaccurate.
- Therefore people usually decide a small number of steps to look back, i.e. assuming $P(x_n \mid x_1, \cdots, x_{n-1}) = P(x_n \mid x_{n-k}, \cdots, x_{n-1})$. But it is hard to determine k and more importantly it takes the risk of losing information.

Tree Modeling of Binary Sequences

- Transition probability given a pattern is modeled as follows:

  $P(x_n = 1 \mid x_1, \cdots, x_{n-1}) = \text{logistic}(\beta_0+\beta_{(x_{n-1})}+\beta_{(x_{n-2},x_{n-1})}+\cdots+\beta_{(x_1,\cdots,x_{n-1})})$,

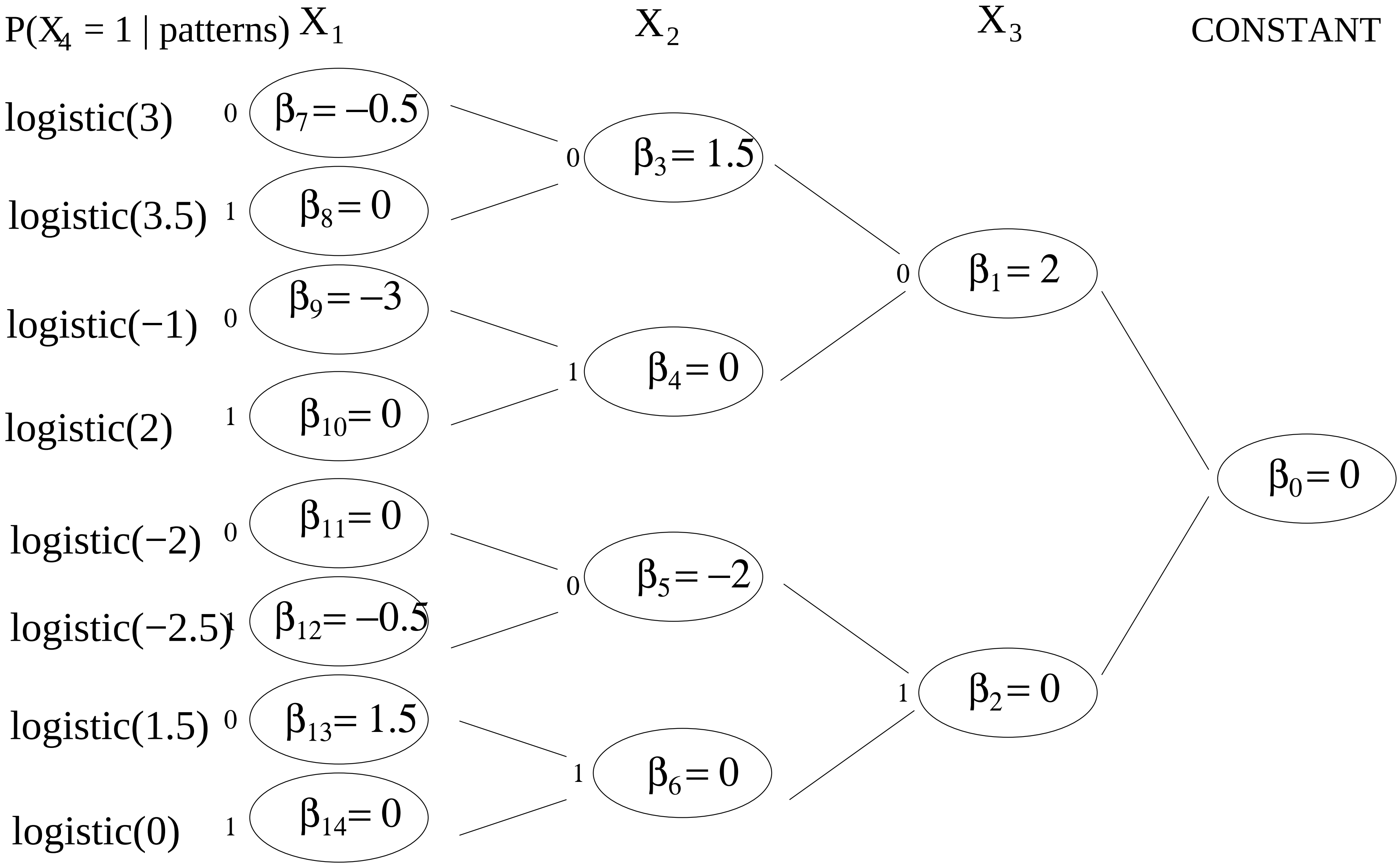
  where $\text{logistic}(x) = \frac{1}{1+\exp(-x)}$

- This is a logistic regression model by using the indicators over all patterns of all lengths as predictors.

- Prior: $\beta_{ptn} \mid \sigma_1, r$ id $\sim$ Symmetric Stable$(\alpha, 0, \sigma_1 r^{l-1})$, where l=length(ptn) and σ_1 and $r \sim f(\cdot)$

- When there are more than two possibilities for each state, logistic model is replaced by softmax model.

An example of binary sequences of length 3



Remarks on the Model

- It does not truncate the sequence arbitrarily. Instead, we let the influence of sequences decreases gradually with its length through the prior.
- The lengths of sequences could be different acrossing the cases.
- We don't need to express all β's explicitly. For example, when the patterns in case i starting from step j are expressed only in case i, only the sum of these associated β's needs to be expressed explicitly. The prior of the sum of these β's is still symmetric stable distribution with some width. This reduces the number of model parameters dramatically (greatly less than 2^{n-1}) when the number of sequences is small.
- We expect that the heavy-tailed symmetric stable distribution would allow the β's of some long history patterns turn to large values if the data favour this situation even though most of their "peers" are very small.

Searching All Explicit β's in Sequences

Searching all explicit β's in all available sequences turns out to be very simple, as shown by the graph below:

Data			Indice of β's		
0	0	0	7	3	1
1	0	0	8	3	1
0	1	0	4	4	1
0	1	0	4	4	1
0	0	1	9	5	2
1	0	1	10	5	2
0	1	1	stop	6	2

A Simulation Study

- Generating the dataset: 3 more levels are added to the graph shown in previous slide. The β's of the first three steps are shown on the graph. Then the previous graph is appended to each of the 8 nodes in the third level again. 1000 sequences were generated from this tree of 6 levels, 200 of which were used as training cases and the remaining 800 as test cases.

- Specifying the priors: The prior of σ_1 is $Unif(0.01, 0.3)$ for Cauchy distribution and $Unif(0.1, 03)$ for Gaussian distribution. The prior of r is $Unif(0.1, 0.9)$.

- Training the model: We used Gibbs sampling with each univariate distribution sampled with slice sampling to train the model. The σ_1 and r are sampled with Metropolis-Hasting method with simple proposal. 500 iterations were run and the final 450 ones were used to make prediction.

A Simulation Study (Cont'd)

- Making comparison: Two measures are used to evaluate the methods. Suppose $y_1, \cdots, y_{n_{ts}}$ are the true values in test cases, $\hat{p}_1, \cdots, \hat{p}_{n_{ts}}$ are the predictive probabilities, and using 0.5 as the threshold, $\hat{y}_1, \cdots, \hat{y}_{n_{ts}}$ are the predicted values, then "error.fit" and "error.rate" can be defined as:

$$\text{error.fit} = -(\sum_{i=1}^{n_{ts}} (\log(\hat{p}_i) y_i + \log(1-\hat{p}_i)(1-y_i))/n_{ts})$$

$$\text{error.rate} = \sum_{i=1}^{n_{ts}} I(y_i \neq \hat{y}_i)/n_{ts}$$

No. of back steps	1	2	3	4	5	6
error.fit by Cauchy	0.64	0.60	0.60	0.55	0.51	0.48
error.fit by Gaussian	0.65	0.59	0.60	0.53	0.48	0.46
error.rate by Cauchy	0.40	0.28	0.28	0.28	0.28	0.25
error.rate by Gaussian	0.40	0.28	0.32	0.28	0.24	0.23

An Example of Phased Haplotype Data

I chose a phased haplotype dataset from hapmap which contains 120 patients and 1294 SNPs. 100 patients were chosen randomly as training cases. The 400th SNP was chosen to predict, which has mean 0.55. The prior of σ_1 of Cauchy is $Unif(0.01, 0.3)$ and of Gaussian is $Unif(0.1, 3)$. The prior of r is $Unif(0.1, 0.9)$. 500 Gibbs sampling iterations were run and the final 450 were used to make prediction.

No. of back steps	1	5	10	15	100	399
No. of patterns expressed	3	11	15	21	85	167
Total training time	56	122	166	233	1997	4307
error.fit by Cauchy	0.51	0.38	0.28	0.22	0.22	0.23
error.fit by Gaussian	0.51	0.39	0.32	0.21	0.33	0.33
error.rate by Cauchy	0.15	0.15	0.15	0.10	0.10	0.10
error.rate by Gaussian	0.15	0.15	0.15	0.15	0.15	0.15

Conclusion and future work

- We've proposed a Bayesian tree modeling method for the discrete sequences that allows us to use the full or long history.
- The simulation results show that when the long predictive patterns exists in the dataset the performance of our proposed methods increases with length of history taken into consideration.
- In an example of phased haplotype dataset, the predicting performance is improved by using long history. The time increase due to using long history is acceptable. In this example, Cauchy prior is seen to work better than Gaussian prior in term of predicting error, especially when long history is used. Therefore it suggests that we use long history, like history longer than 20, to make prediction in haplotype inference.
- We will apply the method to text prediction.