Parsimonious Classification using Higher Order Markov Chains

A project report submitted by

Aniruddha Pathak (181013)

Project Supervisor : Prof. Subhajit Dutta

Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
June, 2020

Abstract

Classification problems with categorical feature variables can be found in various fields of study like biology, computer science, finance, etc. Our goal is to discuss some parsimonious methodologies that can be used in this context. Here we discuss about three different classifiers. The first one uses maximum likelihood estimates. The second one uses a mixture probability model for the conditional probabilities. And, the last one uses LASSO regularized logistic regression on counting statistics. We prove the weak consistency of the classifier based on maximum likelihood estimates. We apply these classifiers on simulated datasets to assess the performance by their misclassification rates and time efficiency.

Acknowledgments

I would like to thank all the people who have helped and supported me for this project. First and foremost, I would like to thank my project supervisor Prof. Subhajit Dutta for his constant support and guidance at every step of my project. I would also like to thank my classmates, who have many times patiently heard and suggested me regarding many points in this project.

Contents

1 Introduction

In many real world applications, we frequently come across categorical data sequences. For example in sales demand prediction, products can be classified into several states like high, normal and low, with respect to the sales volume. In DNA sequences analysis, the observations are represented by large categorical sequences with categories {A, T, C, G}. Fields of study like biology, computer science, finance many times deal with classification problems where the feature variables are categorical data sequences.

Let us consider a two-class classification problem with class probabilities π_0 and π_1 respectively, where $\pi_0+\pi_1 = 1, \pi_0 > 0$ and the features $\mathbf{x} \in \{0,1\}^p$. For this case, the Bayes classifier with minimum misclassification probability is given,

$$\delta_B(\mathbf{x}) = \begin{cases} 0 & \text{if } \pi_0\, p_0(\mathbf{x}) > \pi_1\, p_1(\mathbf{x}) \\ 1 & \text{otherwise,} \end{cases} \tag{1}$$

where p_j is the probability mass function under the j^{th} class, $j \in \{0,1\}$.

1.1 What is the Problem?

In order to find the Bayes classifier, it is fundamental to calculate $p_j(\mathbf{x})\ \forall\ \mathbf{x}$ for $j \in \{0,1\}$. As $|\{0,1\}^p| = 2^p$, we have to estimate a total of $2^{p+1} - 2$ parameters (probabilities).

When we are considering a classification problem we estimate the unknown parameters (probabilities) in $p_j,\ j \in \{0,1\}$ from the training data. Now, if the number of observations are less than $2^p - 1$ for any of the classes, then we may not be able to estimate all the parameters from the training data. In such a case, we may have some observations absent in the training set, but present in the test set. Thus we cannot calculate $\hat{p}_j$'s for those $\mathbf{x}$'s. Hence, we cannot use the Bayes classifier without a large training data. Also the number of parameters is $O(2^{p+1})$, which makes the classifier inefficient for computational purpose when p is high.

1.2 Why Higher Order Markov Chains?

It is well-known that the feature variables in the classification problem may not be independent and can be modeled using a higher-order Markov chain

(see Dutta et al. 2014). Let $(X_n)_{n\geq 0}$ follow a k^{th} order Markov Chain, i.e.

$$X_n|X_{n-1},\ldots,X_1 \stackrel{d}{=} X_n|X_{n-1},\ldots,X_{n-k}, \forall n \geq k, k \in \mathbb{N} \tag{2}$$

For an observation from the j^{th} class $\mathbf{x}_{ji},\ i \in \{1,\ldots,n_j\},\ j \in \{0,1\}$ which follows a higher order Markov chain of order k_j, we can write

$$\begin{aligned} &p_j(x_{ji1}, x_{ji2}, \ldots, x_{jip}) \\ &\quad = p_j(x_{ji1},\ldots,x_{jik_j}) \cdot p_j(x_{ji(k_j+1)}|x_{jik_j}\ldots,x_{ji2},x_{ji1})\ldots \\ &\ldots p_j(x_{ji(p-1)}|x_{ji(p-2)}\ldots,x_{ji2},x_{ji1}) \cdot p_j(x_{jip}|x_{ji(p-1)}\ldots,x_{ji2},x_{ji1}) \\ &= p_j(x_{ji1},\ldots,x_{jik_j}) \prod_{l=k_j+1}^{p} p_j(x_{jil}|x_{ji(l-1)},\ldots,x_{ji(l-k_j)}) \ \forall\ j \in \{0,1\} \end{aligned} \tag{3}$$

In the above model (3), the $p_j(\mathbf{x})$, we have to calculate $p_j(x_{ji1},\ldots,x_{jik_j})$ and $p_j(x_{jil}|x_{ji(l-1)},\ldots,x_{ji(l-k_j)})$ which requires to estimate $2^{k_j}-1$ and 2^{k_j}, parameters respectively for each class, i.e. for a two-class problem we try to estimate a total of $2^{k_0+1}+2^{k_1+1}-2$ parameters. Hence we can attain a reduction in dimension with respect to the parameters concerned, which may lead to more efficiency in our computation.

Higher order Markov chains have also been used in a range of applications including the analysis of wind speed and direction, DNA sequences, social behaviour and financial series (see Berchtold and Raftery 2002).

In the literature, there are many methodologies proposed for modelling the higher order Markov chains. Raftery (1984) proposed a mixture probability model for this purpose. While Ching et al. (2004) generalized Raftery's model. Machler and Buhlmann (2004) presented a new computational tool named variable length Markov chains (VLMC). On the other hand, Yang and Dunson (2016) proposed a methodology for a categorical response and high-dimensional categorical predictors based on Bayesian conditional tensor factorization. Dutta et al. (2014) proposed methodologies based on the occurrences of words (specific sequences) in the feature variables.

In this project, we will try to use the methodologies proposed by Dutta et al. (2014) and, Ching et al. (2004) for the purpose of modelling the higher order Markov chains. We will propose a new methodology in this context. At last we will also try to compare the efficiencies of these methodologies.

2 Methodology

2.1 Classifier based on Maximum Likelihood Estimates

For a stochastic sequence $\mathbf{x}_{ji} = (x_{ji1}, \ldots, x_{jip})^T$ generated from the probability distribution G_j belonging to the j^{th} class, the Bayes rule $\delta_B(\mathbf{x}_{ji})$ with minimum misclassification probability is given by,

$$\delta_B(\mathbf{x}_{ji}) = \begin{cases} 0 & \text{if } \pi_0\ p_{G_0}(\mathbf{x}_{ji}) > \pi_1\ p_{G_1}(\mathbf{x}_{ji}) \\ 1 & \text{otherwise,} \end{cases} \tag{4}$$

where p_j is the probability mass function under the j^{th} class, for $j \in \{0, 1\}$.

Now for any integer $k > 0$, we refer to the elements of $\{0, 1\}^k$ as k-words. For a fixed k and a fixed k-word $(m_1, \ldots, m_k) \in \{0, 1\}^k$, if $I(\cdot)$ is the indicator variable, then

$$f_{\mathbf{x}_{ji}}(m_1, \ldots, m_k) = \sum_{l=1}^{p-k+1} I(x_{jil} = m_1, x_{ji(l+1)} = m_2, \ldots, x_{ji(l+k-1)} = m_k)$$

is defined as the frequency for the k-word $(m_1, \ldots, m_k)$ in $\mathbf{x}_{ji}$. The frequencies of different k-words in $\mathbf{x}_{ji}$ can be considered as features of $\mathbf{x}_{ji}$, for $i = \{1, 2, \ldots, n_i\},\ j \in \{0, 1\}$

Now let us assume that, for $j \in \{0, 1\}$, the probability distribution G_j corresponding j^{th} class is Markov with order k_j. Let, $\theta_j(k_j)$ be the vector of model parameters, $j \in \{0, 1\}$. Also, let $\mathbf{k} = (k_0, k_1)$ and $\phi(\mathbf{k}) = (\theta_0(k_0), \theta_1(k_1))$. Then, we can rewrite the decomposition of the likelihood in (3) under the Markov model as,

$$\begin{aligned} \log\ p_{G_j}(\mathbf{x}) = & \sum_{(m_1,\ldots,m_k)\in\{0,1\}^k} I(x_1 = m_1, \ldots, x_k = m_k)\ \log\ q_j(m_1, \ldots, m_k) \\ + & \sum_{(m_1,\ldots,m_{k+1})\in\{0,1\}^{k+1}} f_x(m_1, \ldots, m_{k+1})\ \log\ p_j(m_{k+1} \mid m_k, \ldots, m_1), \end{aligned} \tag{5}$$

and the Bayes rule based on such Markov likelihoods become

$$\delta(\mathbf{x}, \phi(\mathbf{k}), \mathbf{k}) = \begin{cases} 0 & \text{if } \log\ \pi_0 + \log\ p_{G_0}(\mathbf{x}) > log\ \pi_1 + log\ p_{G_1}(\mathbf{x}) \\ 1 & \text{otherwise.} \end{cases} \tag{6}$$

Now, the Bayes rule consists of unknown parameters which can be estimated from the training set. For a set of random samples $\mathbf{x}_{j1}, \mathbf{x}_{j2}, \ldots, \mathbf{x}_{jn_j}$ from the j^{th} class in the training set, the maximum likelihood estimates of model parameters based on the observations are given by,

$$\hat{q}_j(m_1, \ldots, m_{k_j}) = \frac{1}{n_j} \sum_{i=1}^{n_i} I(x_{ji1} = m_1, \ldots, x_{jik_j} = m_{k_j}) \tag{7}$$

and

$$\hat{p}_j(m_{k_j+1} \mid m_{k_j}, \ldots, m_1) = \frac{\sum_{i=1}^{n_j} f_{\mathbf{x_{ji}}}(m_1, \ldots, m_{k_j}, m_{k_j+1})}{\sum_{i=1}^{n_j} \sum_{m=\{0,1\}} f_{\mathbf{x_{ji}}}(m_1, \ldots, m_{k_j}, m)}, \tag{8}$$

where $m_l \in \{0, 1\}\ \forall\ l \in \{1, 2, \ldots, k+1\},\ j \in \{0, 1\}$.

Let $\hat{\theta}_j(k_j)$ be the maximum likelihood estimates of $\theta_j(k_j)$. It can be proved that if the assumptions regarding higher order Markov chain of order k_j and its stationarity are true, and $p > k_j$ $\hat{\theta}_j(k_j)$ converges to in probability to $\theta_j(k_j)$ as $\min_j n_j \to \infty\ \forall j \in \{0, 1\}$. The proof of the convergence can be found in the appendix.

For the purpose of choosing the hyper-parameter $\mathbf{k} = (k_0, k_1)$, we conduct V-fold cross-validation. We choose $\mathbf{k}$ such that the cross-validation error $\hat{\Delta}_{CV}(\mathbf{k})$ is minimum and denote it by $\tilde{\mathbf{k}}$ (see Dutta et al. 2014). We fit the model with $(\phi(\tilde{\mathbf{k}}),\ \tilde{\mathbf{k}})$ and can use the Bayes classifier given in (6) subsequently to calculate the misclassifcation rate from the test set. Then it can be proved that with assumptions same as above this classifier has weak consistency (see Devroye et al. 1996).

2.2 Classifier based on Mixture Probability Model

2.2.1 The Model

Similar as the previous method, our goal here is to estimate the likelihood functions, so that we can use them in the Bayes classifier given in (1). Let us consider the decomposition of the likelihood given in (3). Given the training data, for the class j, we estimate $p_j(x_1, \ \ldots \ , x_k)$ from the sample probability mass function of the first k variables $(x_1, \ \ldots \ , x_k)$ similarly as given in (7).

For the purpose of modelling the higher order Markov chains, Raftery (1984) proposed a simple mixture probability model given by,

$$p_j(X_n = t_0 \mid X_{n-1} = t_1, \ldots, X_{n-k} = t_k) = \sum_{l=1}^{k_j} \lambda_l^{[j]} q_{t_0 t_l}^{[j]} \ , \tag{9}$$

where $t_0, t_l \in \{0,1\}, l = 1, \ldots, k_j, \ j \in \{0,1\} \ \sum_{l=1}^{k_j} \lambda_l^{[j]} = 1$ and $\boldsymbol{Q}^{[j]} = \left((q_{st}^{[j]})\right)$ is a transition matrix with column sums are equal to one such that

$$0 \leq \sum_{l=1}^{k_j} \lambda_l^{[j]} q_{t_0 t_l}^{[j]} \leq 1, t_0, t_l \in \{0,1\}, \ l = 1, \ldots, k_j, \forall j \in \{0,1\}.$$

The model in (9) can also be written as

$$\boldsymbol{\chi}_{n+k_j+1}^{[j]} = \sum_{l=1}^{k_j} \lambda_l^{[j]} \boldsymbol{Q}^{[j]} \boldsymbol{\chi}_{n+k_j+1-l}^{[j]} \ , \ \forall j \in \{0,1\}, \tag{10}$$

where $\boldsymbol{\chi}_{n+k_j+1-l}^{[j]}$ is the probability distribution of states at time $(n+k_j+1-l)$. Ching et al. (2004) generalized (10) as follows:

$$\boldsymbol{\chi}_{n+k_j+1}^{[j]} = \sum_{l=1}^{k_j} \lambda_l^{[j]} \boldsymbol{Q}_{\boldsymbol{l}}^{[j]} \boldsymbol{\chi}_{n+k_j+1-l}^{[j]} \ , \ \forall j \in \{0,1\}. \tag{11}$$

2.2.2 Estimation

Primary Estimation : In (11), Ching et al. (2004) assumed that $\boldsymbol{\chi}_{n+k_j+1}^{[j]}$ depends on $\boldsymbol{\chi}_{n+k_j+1-l}^{[j]}$, through the matrix $\boldsymbol{Q}_{\boldsymbol{l}}^{[j]}$ and the weights $\lambda_l^{[j]} \ \forall \ l \in$

$\{1, \ldots, k_j\}$ and $j \in \{0, 1\}$. The authors related $\boldsymbol{Q}_l^{[j]}$ to the l^{th} step transition matrix of the process for the j^{th} class.

Given a sequence from the training set, they estimate $\boldsymbol{Q}_l^{[j]}$ by the l^{th} step transition matrix of the observations belonging to class j and $\lambda_l^{[j]}$ from the minimization problem :

$$\min_{\lambda^{[j]}} \left\| \sum_{l=1}^{k_j} \lambda_l^{[j]} \boldsymbol{Q}_l^{[j]} \hat{\boldsymbol{\chi}}^{[j]} - \hat{\boldsymbol{\chi}}^{[j]} \right\|, \tag{12}$$

subject to $\sum_{l=1}^{k} \lambda_l^{[j]} = 1$ and $\lambda_l^{[j]} \geq 0,\ j \in \{0, 1\}\ \forall\ l$, where $\hat{\boldsymbol{\chi}}^{[j]}$ is vector of the proportions of the occurrence of each state in the given sequence. In the above minimization problem, if we take l_1 norm then the problem reduces to a linear programming problem and $\lambda_l^{[j]}$ can be estimated from the optimization problem.

Combination of Estimates : Now, $\hat{\boldsymbol{\lambda}}^{[j](i)}$ and $\hat{\boldsymbol{Q}}_l^{[j](i)}$ be the estimates of $\boldsymbol{\lambda}^{[j]}$ and $\boldsymbol{Q}_l^{[j]}$ from the i^{th} sequence $\mathbf{x}_{ji}$ from the training set of class j. To have an overall model from the training set for the class j, we propose the overall estimates for $\boldsymbol{\lambda}^{[j]}$ and $\boldsymbol{Q}_l^{[j]}$ as :

$$\bar{\boldsymbol{\lambda}}^{[j]} = \frac{1}{n_j} \sum_{i=1}^{n_j} \boldsymbol{\lambda}^{[j](i)} \text{ and } \bar{\boldsymbol{Q}}_l^{[j]} = \frac{1}{n_j} \sum_{i=1}^{n_j} \boldsymbol{Q}_l^{[i](j)},\ l = 1, \ldots, k_j,\ j \in \{0, 1\}.$$

Thus our model for the j^{th} class becomes

$$\hat{p}_j(X_n = t_0 | X_{n-1} = t_1, \ldots, X_{n-k_j} = t_{k_j}) = \sum_{l=1}^{n_j} \bar{\lambda}_l^{[j]} \left[\bar{\boldsymbol{Q}}_l^{[j]} \right]_{t_0 t_l}, \tag{13}$$

where $\bar{\boldsymbol{\lambda}}^{[j]} = \left(\bar{\lambda}_1^{[j]}, \ldots, \bar{\lambda}_{k_j}^{[j]} \right)^T$ and $\left[\bar{\boldsymbol{Q}}_l^{[j]} \right]_{t_0 t_l}$ is the $(t_0, t_l)^{th}$ element of $\bar{\boldsymbol{Q}}_l^{[j]}$.

We can easily see that $\bar{\boldsymbol{\lambda}}^{[j]}$ and $\bar{\boldsymbol{Q}}_l^{[j]},\ l = 1, \ldots, k_j$ also satisfies the conditions which were satisfied by, $\boldsymbol{\lambda}^{[j]}$ and $\boldsymbol{Q}_l^{[j]}$'s. Hence by (3), we can estimate $p_j(x_1,\ \ldots\ , x_n)$ and use the estimated Bayes classifier for the purpose of classification.

Choosing the order of the higher order Markov chains : For the purpose of choosing the hyper-parameter related to the order of the higher order Markov chains, we conduct V-fold cross-validation. Let $\mathbf{k} = (k_0, k_1)$ be the true orders of the higher order Markov chains corresponding class 0 and 1, respectively. We choose $\mathbf{k}$ such that the cross-validation error $\hat{\Delta}_{CV}(\mathbf{k})$ is minimum and denote it by $\tilde{\mathbf{k}}$. We fit higher order Markov models with the given estimates the model parameters $\tilde{\mathbf{k}}$, and use the Bayes classifier subsequently to calculate misclassification rate.

2.3 LASSO Regularized Logistic Regression Classifier based on Counting Statistics

From (5), we can get the sufficient statistic in this model given the training set as

$$T(\mathbf{x}) = \Bigg(I(x_1 = 0, \ldots, x_k = 0), \ldots, I(x_1 = 1, \ldots, x_k = 1), \\ f_{\mathbf{x}}(0, \ldots, 0, 0), \ldots, f_{\mathbf{x}}(1, \ldots, 1, 1) \Bigg)^T,$$

where the last 2^{k+1} entries are the set of the frequencies of all the possible $(k+1)$-words. Now, it can be easily seen that the first 2^k entries of $T(\mathbf{x})$ are indicator functions and hence sparse (actually only one of them will be 1 and others will be 0). So, if we ignore the first 2^k entries, the loss of information would not be very significant.

We define our new statistics as follows

$$\tilde{T}(\mathbf{x}) = \Big(f_{\mathbf{x}}(0, \ldots, 0, 0), \ldots, f_{\mathbf{x}}(1, \ldots, 1, 1) \Big)^T_{2^{k+1} \times 1}$$

In general we use the set of the frequencies of all the possible k-words in $\mathbf{x}$ and denoted it as $\tilde{T}_k$. It may be noted that here we are considering a mapping $\tilde{T}_k : \{0,1\}^p \to \mathbb{Z}_{\geq}^{2^k}$, where $\mathbb{Z}_{\geq}$ is the set of all non-negative integers. Thus, we go from dimension p to 2^k, if we have k such that $p > 2^k$ then we can have a dimension reduction in the data. We must remember this fact while choosing k.

Now, we will consider a classification rule on $\mathbb{Z}_{\geq}^{2^k}$ based on the training set. So, $\tilde{T}_k$ is supposed to be sparse, as the occurrence of all the k-words in a single observation $\mathbf{x}$ is less likely, in case k is not very small. Then, we must implement a classification rule which would perform well in such a scenario. We propose the use of elastic net regularization on the logistic regression model(see Hastie et al. 2008). The elastic net regularization uses a convex combination of l_1 and l_2 penalties on the log-likelihood function. Here we only use the LASSO regularization.

Let $\tilde{T}_k(\mathbf{x}) = \mathbf{z}$ be the transformed variable. Then the l_1 penalized log-likelihood for the parameters of the logistic regression $\beta = (\beta_0, \ldots, \beta_q)^T$

where $q = 2^k$ is given by,

$$\log L(\beta|\mathcal{X}) = \sum_{j=0}^{1}\sum_{i=1}^{n_j}\left(y_{ji}\cdot\beta^T\mathbf{z_{ji}} - \log(1+e^{\beta^T\mathbf{z_{ji}}})\right) + \lambda\|\beta\|_1$$

For a given choice of (k, λ), we can get the logistic classifier by maximizing $\log L(\beta|\mathcal{X})$. For the purpose of choosing the hyper-parameters (k, λ), we conduct V-fold cross-validation on the training set. We choose (k, λ) such that the cross-validation error $\hat{\Delta}_{CV}(k, \lambda)$ is minimum and denote it by $(\tilde{k}, \tilde{\lambda})$. For the purpose of convenience, we propose a two-stage minimization approach for $\hat{\Delta}_{CV}(k, \lambda)$, and $(\tilde{k}, \tilde{\lambda})$ is defined as

$$(\tilde{k}, \tilde{\lambda}) = \arg\min_k \min_\lambda \hat{\Delta}_{CV}(k, \lambda).$$

Hence, we fit the model with $(\tilde{k}, \tilde{\lambda})$. Let k_0 and k_1 be the true orders of the higher order Markov chains corresponding the two classes. In this method, we only consider a single parameter k regarding the orders. But it should be noted that, this method do not not violate the Markov assumption as long as $k \geq \max\{k_0, k_1\}$.

3 Numerical Results

For the purpose of comparison of performance of the three different classifiers that we have described earlier, we conduct simulation studies. We consider equal class probabilities. To generate the dataset we use the transition probabilities,

Class 0 :

$$p_0(0|0,0) = 0.4, p_0(0|0,1) = 0.6,$$

$$p_0(0|1,0) = 0.6, p_0(0|1,1) = 0.4.$$

Class 1 :

$$p_1(0|0,0) = 0.6,\ p_1(0|0,1) = 0.4,$$

$$p_1(0|1,0) = 0.4,\ p_1(0|1,1) = 0.6.$$

We generated sequences of length 100, and formed the training and test samples with 50 and 150 observations for each class, respectively. We conduct 100 Monte Carlo simulations (repetitions of the procedure) to get a better idea about the misclassification rates. We pre-initialize each sequence with (0,1). We have taken a reducible, transient Markov chain which has a stationary distribution. So, the sequences will not be affected by the pre-initialization.

The average misclassification rates with their standard errors for the classifiers based on maximum likelihood estimates (MLE), mixture probability model and LASSO-logistic regression are tabulated in the Table 1. The details regrading the empirical probability distributions of the hyper-parameters are tabulated in Tables 2-6. The average time taken by the classifiers for each replication with their standard deviation are tabulated in Table 7.

In all the classifiers described earlier, we have to estimate the hyper-parameters through V-fold cross-validation. We conduct 2-fold cross-validation with 10 random splits for each case. Keeping in the mind the fact about dimension reduction we only consider $k_0, k_1 \in \{1, 2, 3\}$ for the first two classifiers and $k \in \{1, 2, 3, 4, 5\}$, $\lambda \in \{0, 0.01, \ldots, 0.34, 0.35\}$ for the LASSO-logistic regression classifier.

Table 1: Misclassification rates with standard error (within parentheses).

Methods	Avg. Misclassification Rate
MLE	0.0227 (0.00851)
Mixture Model	0.391 (0.03532279)
LASSO-Logistic	0.0342 (0.01320)

Table 2: Joint empirical probability distribution of $(\hat{k}_0^{CV},\ \hat{k}_1^{CV})$ for the classifier based on maximum likelihood estimates.

$\hat{k}_0^{CV}$ \ $\hat{k}_1^{CV}$	1	2	3
1	0	0	0
2	0	0.95	0.04
3	0	0.01	0

Table 3: Joint empirical probability distribution of $(\hat{k}_0^{CV},\ \hat{k}_1^{CV})$ for the classifier based on mixture probability model.

$\hat{k}_0^{CV}$ \ $\hat{k}_1^{CV}$	1	2	3
1	0.14	0.05	0.09
2	0.10	0.09	0.11
3	0.09	0.15	0.18

Table 4: Joint empirical probability distribution of ($\hat{k}^{CV}$, $\hat{\lambda}^{CV}$) for the LASSO-Logistic classifier.

$\hat{\lambda}^{CV}$ \ $\hat{k}^{CV}$	1	2	3	4	5
0	0	0	0.20	0.05	0.01
0.01	0	0	0.13	0.07	0.01
0.02	0	0	0.08	0.08	0
0.03	0	0	0.11	0.01	0.01
0.04	0	0	0.07	0.04	0
0.05	0	0	0.02	0.04	0.01
0.06	0	0	0.03	0	0
0.07	0	0	0.02	0	0
0.09	0	0	0.01	0	0
otherwise	0	0	0	0	0

Table 5: Marginal empirical probability distribution of $\hat{k}^{CV}$

$\hat{k}^{CV}$	1	2	3	4	5
Probability	0	0	0.67	0.29	0.04

Table 6: Marginal empirical probability distribution of $\hat{\lambda}^{CV}$

$\hat{\lambda}^{CV}$	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.09
Probability	0.26	0.21	0.16	0.13	0.11	0.07	0.03	0.02	0.01

Table 7: Average time taken (with standard deviation within parentheses) by different classifiers per replication of a simulated data set.

Methods	Avg. time taken (in seconds)
MLE	25.90008 (0.1429632)
Mixture Model	78.21576 (3.150723)
LASSO-Logistic	6.794869 (0.2085855)

4 Conclusion

We have discussed three different classifiers in this context of a classification problem with categorical feature variables. We have applied the classifiers to simulated datasets and have obtained some numerical results regarding the misclassification rates, time efficiency, and the hyperparameters.

In Table 1, we can see the average misclassification rates of the different classifiers that were obtained for the simulated datasets. The classifier based on the mixture probability model has very high misclassification rates when compared with others. On the other hand, the classifiers based on maximum likelihood estimates and counting statistics performed quite similarly well.

An important task of all the classifiers is to detect the true order of the higher order Markov chains involved. From the Tables 2-6, we can see how the classifiers work regarding this context. We already stated earlier that we do not violate the Markov assumptions as long as the estimated orders are greater than or equal to the true orders (in our case, 2 for both the classes). From Tables 2 and 5, we can see that the classifiers based on maximum likelihood estimates and counting statistics always estimated the orders correctly. But for the classifier based on the mixture probability model, the assumptions are violated for 47% of the replications, where in 14% of the replications, both of the hyperparameters were not correct. This may be contributed as a reason for the high misclassification rate. May be by taking a higher number of folds for cross-validation with more replications, we can get better results from this classifier. We want to add that from Table 6, the LASSO-logistic classifier tends to choose small LASSO penalization rather than taking the large ones.

Table 7 shows that the LASSO-Logistic classifier based on counting statistics is the most time-efficient between the three concerned. Whereas the classifiers based on maximum likelihood estimates and mixture probability model take about four times and eleven times, respectively, of what the LASSO-Logistic classifier takes for a single replication data.

From the simulation studies, we can observe that the classifier based on maximum likelihood estimates and LASSO-logistic classifier performs quite well if we consider the misclassification rate. On the other hand, the LASSO-logistic classifier performs most efficiently concerning the execution time. In contrast, the classifier based on the mixture probability model does not perform well when we consider either the misclassification rate or time efficiency.

5 Appendix : Proof of Weak Consistency for Section 2.1

Proposition 5.1. *If G_j is an irreducible, stationary Markov of order k_j then $\hat{q}_j(m_1, \ldots, m_{k_j})$ converges in probability to $q_j(m_1, \ldots, m_{k_j})$ as $\min_j n_j \to \infty \; \forall \; j \in \{0, 1\}$*

Proof. Now,

$$E_{G_j}\Bigg(I(x_{ji1} = m_1, \ldots, x_{jik_j} = m_{k_j}) \Bigg) \\ = P_{G_j}(X_1 = m_1, \ldots, X_{k_j} = m_{k_j}) \\ = q_j(m_1, \ldots, m_{k_j}) \; \forall \; j \in \{0, 1\}.$$

Similarly,

$$V_{G_j}\Bigg(I(x_{ji1} = m_1, \ldots, x_{jik_j} = m_{k_j}) \Bigg) \\ = P_{G_j}(X_1 = m_1, \ldots, X_{k_j} = m_{k_j})(1 - P_{G_j}(X_1 = m_1, \ldots, X_{k_j} = m_{k_j})) \\ = q_j(m_1, \ldots, m_{k_j})(1 - q_j(m_1, \ldots, m_{k_j})) \; \forall \; j \in \{0, 1\}.$$

Hence,

$$E_{G_j}\Bigg(\hat{q}_j(m_1, \ldots, m_{k_j}))\Bigg) = q_j(m_1, \ldots, m_{k_j}) \; \forall \; j \in \{0, 1\}.$$

$$V_{G_j}\Bigg(\hat{q}_j(m_1, \ldots, m_{k_j})\Bigg) \\ = \frac{1}{n_j^2} \sum_{i=1}^{n_j} \mathrm{V}_{G_j}\Bigg(I(x_{ji1} = m_1, \ldots, x_{jik_j} = m_{k_j})\Bigg) \\ = \frac{1}{n_j} q_j(m_1, \ldots, m_{k_j})\Bigg(1 - q_j(m_1, \ldots, m_{k_j})\Bigg) \; \forall \; j \in \{0, 1\}.$$

Note that, the covariances are $0 \; \forall \; i \neq i'$, as $(x_{ji1}, \ldots, x_{jik_j})$ are first k_j entries of the i.i.d. observations from the Markov chain G_j. So, $V_{G_j}(\hat{q}_j(m_1, \ldots, m_{k_j})$ converges to 0 as $\min_j n_j \to \infty \; \forall \; j. \in \{0, 1\}$. Thus,

$$\hat{q}_j(m_1, \ldots, m_{k_j}) \xrightarrow{P} q_j(m_1, \ldots, m_{k_j}) \text{ as } \min_j n_j \to \infty \; \forall \; j \in \{0, 1\}. \; \blacksquare$$

Proposition 5.2. *If G_j is an irreducible, stationary Markov of order k_j and $p > k_j \ \forall j$, then $\hat{p}_j(m_{k_j+1}|\ m_1, \ldots, m_{k_j})$ converges in probability to $p_j(m_{k_j+1}|\ m_1, \ldots, m_{k_j})$ as $\min_j n_j \to \infty \ \forall\ j \in \{0, 1\}$, where p is the length of each of the feature data sequences.*

Proof. Consider

$$\hat{p}_j(m_{k_j+1}|\ m_1, \ldots, m_{k_j}) = \frac{A_{n_j}}{B_{n_j}},$$

where

$$A_{n_j} = \frac{1}{n_j \cdot (p - k_j)} \sum_{i=1}^{n_j} f_{\mathbf{x}_{ji}}(m_1, \ldots, m_{k_j}, m_{k_{j+1}}), \text{ and}$$

and

$$B_{n_j} = \frac{1}{n_j \cdot (p - k_j)} \sum_{i=1}^{n_j} \sum_{m \in S} f_{\mathbf{x}_{ji}}(m_1, \ldots, m_{k_j}, m) \ \forall\ j \in \{0, 1\}.$$

Now,

$$\begin{aligned}
&E\Big(f_{\mathbf{x}_{ji}}(m_1, \ldots, m_{k_j}, m_{k_{j+1}})\Big) \\
&\quad = \sum_{l=1}^{p-k_j} E\Big(I(x_{jil} = m_1, \ldots, x_{ji(l+k_j-1)} = m_{k_j}, x_{ji(l+k_j)} = m_{k_j+1})\Big) \\
&\quad\quad = \sum_{l=1}^{p-k_j} P_{G_j}(X_l = m_1, \ldots, X_{l+k_j} = m_{k_j+1}) \\
&= \sum_{l=1}^{p-k_j} P_{G_j}(X_n = m_1, \ldots, X_{n+k_j} = m_{k_j+1}) \ \Big[\text{due to stationarity of } (X_{jil})_{l \geq 0}\Big] \\
&= (p - k_j) \cdot P_{G_j}(X_n = m_1, \ldots, X_{n+k_j-1} = m_{k_j}, X_{n+k_j} = m_{k_j+1}) \ \forall\ j \in \{0, 1\}.
\end{aligned}$$

Then,

$$\begin{aligned}
E(A_{n_j}) &= \frac{1}{n_j \cdot (p - k_j)} \sum_{i=1}^{n_j} E\Big(f_{\mathbf{x}_{ji}}(m_1, \ldots, m_{k_j}, m_{k_{j+1}})\Big) \\
&= P_{G_j}(X_n = m_1, \ldots, X_{n+k_j-1} = m_{k_j}, X_{n+k_j} = m_{k_j+1}),
\end{aligned}$$

$$E(B_{n_j}) = \frac{1}{n_j \cdot (p-k_j)} \sum_{i=1}^{n_j} \sum_{m \in S} E\Bigg(f_{\mathbf{x}_{ji}}(m_1, \ldots, m_{k_j}, m) \Bigg)$$

$$= \sum_{m \in S} \frac{1}{n_j \cdot (p-k_j)} \sum_{i=1}^{n_j} E\Bigg(f_{\mathbf{x}_{ji}}(m_1, \ldots, m_{k_j}, m) \Bigg)$$

$$= \sum_{m \in S} P_{G_j}(X_n = m_1, \ldots, X_{n+k_j-1} = m_{k_j}, X_{n+k_j} = m) \Bigg[\text{similar as above}\Bigg]$$

$$= P_{G_j}(X_n = m_1, \ldots, X_{n+k_j-1} = m_{k_j}) \; \forall \; j \in \{0,1\}.$$

Again,

$$A_{n_j} = \frac{1}{n_j \cdot (p-k_j)} \sum_{i=1}^{n_j} \sum_{l=1}^{p-k_j} I(x_{jil} = m_1, \ldots, x_{ji(l+k_j)} = m_{k_j+1}) \; \forall \; j \in \{0,1\}.$$

Let, the random variable Y_{jil} be defined such that, $\forall \; \omega \in \Omega$,

$$\{\omega : Y_{jil}^{-1}(\omega) = f(m_1, \ldots, m_{k_j+1})\}$$

$$= \{\omega : (X_{jil} \ldots, X_{ji(l+k_j)})^{-1}(\omega) = (m_1, \ldots, m_{k_j+1})\},$$

where f can be defined as an one-one function from $\{0,1\}^{k_j+1}$ to a finite set $\mathcal{A}$. If $(X_{jil})_{l \geq 0}$'s follows a higher order Markov chain order k_j, then $(Y_{jil})_{l \geq 0}$ follows a higher order Markov chain order $2k_j$ which has a state space $\mathcal{A}$. Then A_{n_j} can be written as follows:

$$A_{n_j} = \frac{1}{n_j \cdot (p-k_j)} \sum_{i=1}^{n_j} \sum_{l=1}^{p-k_j} I\Bigg(y_{jil} = f(m_1, \ldots, m_{k_j+1}) \Bigg) \; \forall \; j \in \{0,1\}.$$

Then we can write by applying the Ergodic theorem that (see Shalizi 2009)

$$A_{n_j} \xrightarrow{P} P_{G_j}(X_n = m_1, \ldots, X_{n+k_j-1} = m_{k_j}) \; \text{as} \min_j n_j \to \infty \; \forall \; j \in \{0,1\},$$

and similarly applying the Ergodic theorem we can prove that

$$B_{n_j} \xrightarrow{P} P_{G_j}(X_n = m_1, \ldots, X_{n+k_j-1} = m_{k_j}) \; \text{as} \min_j n_j \to \infty \forall \; j \in \{0,1\}.$$

By Slutsky's theorem,

$$\begin{aligned}
&\hat{p}_j(m_{k_j+1}|m_1,\ldots,m_{k_j}) \\
&\quad = \frac{A_{n_j}}{B_{n_j}} \xrightarrow{P} \frac{P_{G_j}(X_n = m_1,\ldots,X_{n+k_j-1} = m_{k_j}, X_{n+k_j} = m_{k_j+1})}{P_{G_j}(X_n = m_1,\ldots,X_{n+k_j-1} = m_{k_j})} \\
&\quad\quad = P_{G_j}(X_{n+k_j} = m_{k_j+1}|X_n = m_1,\ldots,X_{n+k_j-1} = m_{k_j}) \\
&\quad\quad\quad = p_j(m_{k_j+1}|m_1,\ldots,m_{k_j}),\ \text{as} \min_j n_j \to \infty\ \forall\ j \in \{0,1\}.
\end{aligned}$$

Thus, $\hat{p}_j(m_{k_j+1}|\ m_1,\ldots,m_{k_j}) \xrightarrow{P} p_j(m_{k_j+1}|\ m_1,\ldots,m_{k_j})$,
as $\min_j n_j \to \infty\ \forall\ j \in \{0,1\}$. ■

Proposition 5.3. *If G_j is an irreducible, stationary Markov of order k_j and $p > k_j\ \forall j$, then the classifier based on the maximum likelihood estimators of the model parameters has weakly consistency as $\min_j n_j \to \infty\ \ \forall\ j$, where p is the length of each of the feature data sequences.*

Proof. Let $\mathbf{x} \in \{0,1\}^p$ be an observation from the test set and y be its true class. Then, the Bayes classifier Δ_B is defined by,

$$\delta_B(\mathbf{x}) = \arg\max_j \pi_j\ p_{G_j}(\mathbf{x}),$$

where p_{G_j} is the probability mass function under the j^{th} class.

Then, the Bayes risk Δ_B is defined as, $\Delta_B = P(\delta_B(\mathbf{X}) = Y)$. It can be noted that Δ_B is a function of the model parameters. The estimated Bayes classifier based on the maximum likelihood estimates is given by

$$\delta_n(\mathbf{x}) = \arg\max_j \hat{\pi}_j\ \hat{p}_{G_j}(\mathbf{x}),$$

where $\hat{p}_{G_j}$ is the estimated probability mass function under the j^{th} class, estimated with maximum likelihood estimates of the model parameters and $n = \sum_j n_j$.

Let $\mathbf{x}_1,\ldots,\mathbf{x}_N \in \{0,1\}^p$ be sample observations from the test set and $y_1,\ldots,y_N$ be their true classes, respectively. Then, we define the estimated Bayes risk as, $\Delta_n = \frac{1}{N}\sum_{i=1}^{N} I(\delta_n(\mathbf{x}_i) = y_i)$. Now, we have already proved

that the maximum likelihood estimators of the model parameters converge in probability to the true model parameters as $\min_j n_j \to \infty$.

So, by the continuous mapping theorem we can say that, $\Delta_n \xrightarrow{P} \Delta_B$ as $\min_j n_j \to \infty$ (see Devroye et al. 1996). Hence, we have the weak consistency for δ_n. ■

References

Andre Berchtold and Adrian E. Raftery. The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science*, 17(3):328–356, 2002.

Wai Ki Ching, Eric S Fung, and Michael K Ng. Higher-order Markov chain models for categorical data sequences. *Naval Research Logistics (NRL)*, 51 (4):557–574, 2004.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

Subhajit Dutta, Probal Chaudhuri, and Anil Kumar Ghosh. Linear discriminant analysis of character sequences using occurrences of words. *Statistica Sinica*, 24(1):493–514, 2014.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer, 2008.

Martin Machler and Peter Buhlmann. Variable length Markov chains: Methodology, computing and software. *Journal of Computational and Graphical Statistics*, 13(2):435–455, 2004.

Adrian E. Raftery. A model for high-order Markov chains. *Journal of the Royal Statistical Society Series B (Methodological)*, 47(3):528–539, 1984.

Cosma Shalizi. Lecture notes on chaos, complexity, and inference. *Carnegie Mellon University*, 2009.

Yun Yang and David B. Dunson. Bayesian conditional tensor factorizations for high-dimensional classification. *Journal of American Statistical Association*, 111(514):656–669, 2016.