Classification of Iris dataset using Bayesian Models

Shruti Sharma[1], Khyati Chopra[2]
{sml@yahoo.com[1], khyatichopra134@gmail.com[2]}

IIT Delhi[1], Jamia Hamdard[2]

Abstract: Classification of Iris flower dataset is the best known problem to be found in the pattern recognition literature. It contains four attributes of the flowers belonging to three different species. The objective is to design a model which can differentiate the species based on the attributes of the flowers. In this paper, we have used three different Bayesian models to perform this classification task viz. mixture model based method, hierarchical modeling based method and classical multinomial logistic regression. It was found out that mixture model based method was able to classify the data upto 83.33% accuracy without exploiting any prior knowledge except the number of components of mixtures (which is 3). Bayesian hierarchical model did use the knowledge of the membership of data but main advantage of using this technique was that we obtained the posterior predictive density which is not possible if we use logistic regression or mixture models. Logistic Regression significantly performed better as compared to mixture models with classification accuracy upto 98.33%.

Keywords: classification, Iris dataset, Bayesian model.

1 Introduction

Classification of Iris flower dataset is the best known problem to be found in the pattern recognition literature [1,2]. It contains four attributes of the flowers belonging to three different species. The objective of paper is to classify the multivariate Iris dataset (with three classes) using various Bayesian models. We have specifically used three models, viz. Bayesian Hierarchical Model, Mixture Model and Multinomial' Logistic Regression based Model.

ICIDSSD 2022, May 24-25, New Delhi, India

DOI 10.4108/eai.24-5-2022.2319063

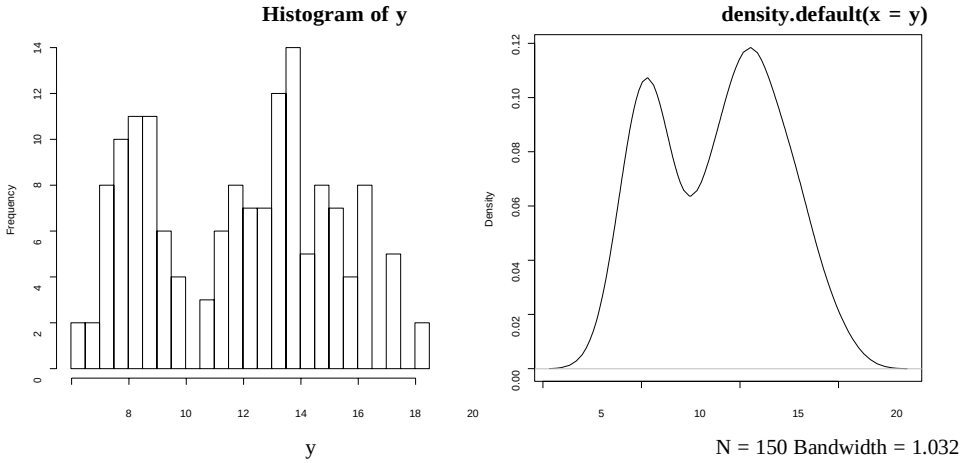

Fig 1: y = sepal.length + sepal.width + petal.length + petal.width is used to predict the underlying species of the flower. Histogram and density plot of y.

2 Dataset

The Iris flower dataset [2] has been used throughout the paper. The dataset consists of three species, namely, 'setosa', 'virginica' and 'versicoloar', each having 50 samples. Four attributes, namely, sepal length, sepal width, petal length and petal width of the flowers are used to design a classifier which can distinguish the species from each other. In particular, sum of all attributes is used as an observation variable.

Histogram and density plots for the observation variable are shown in Fig. 1. It can be seen that there are two peaks in the density plot which shows that by appropriately modeling the density, we can at least identify two species in the dataset1. This is the motivation behind using mixtures model and hierarchical model to model the dataset. For comparison, we will use multinomial logistic regression (generalization of logistic regression) to classify three classes present in the data.[1]

[1]It can be shown that one of the classes (viz. Setosa) is linearly separable from the other two, while the remaining two are not linearly separable.

3 Model

Gaussian Mixtures Model

When we fit a mixture model to data, we usually have the y values and do not know which 'population' they belong to. In mixture models, latent variables are used to provide this information regarding membership of the observation point to a particular population [1].
To solve the problem, we have used Gaussian Mixture Model (GMM). The motivation behind using GMM is that GMM can approximate any density (multimodal) if we choose proper number of populations with ap- propriate memberships. The GMM model looks like:

$$y_i | z_i, \mu_z, \sigma^2 \sim_i \mathsf{N}_{z_i}(\mu_z, \sigma^2), \qquad i = 1, \ldots, n \tag{1}$$

where zi is the latent variable used to identify the population which an observation yi belongs to . Here, z[i] was assumed to follow categorical distribution where each probability ωi of z[i] followed dirichlet prior. The choice of this prior was to exploit the conjugacy of the model.
From Fig 1, it apears that we have two populations, where observations are coming from. However, since we know that we have to identify three classes, we will assume three normal distributions (again owing to conjugacy of normal distribution for mean) with variance 1 and different (and unknown means). All the three means are further assumed to come from normal distribution with parameters:

$$\mu_1 \sim \mathsf{N}(-1, 0.01), \qquad \mu_2 \sim \mathsf{N}(0, 0.01), \qquad \mu_3 \sim \mathsf{N}(1, 0.01)$$

It was found out (empirically) that all μ's did not depend on the choice of means and result obtained were same irrespective of these means.

Checking the model and Results

We ran Monte-carlo markov chain (MCMC) convergence tests with 3 chains with 1000 iterations as burn in period [3]. Three means (with standard deviations) corresponding to three classes were obtained as:

$$\mu_1 = 10.355 \pm 0.179, \qquad \mu_2 = 14.93 \pm 0.278, \qquad \mu_3 = 17.75 \pm 0.377.$$

z latent variable show the membership of the observation to the unknown population.
The confusion matrix is given by Table 1.
It can be seen that around 83.33% predictions have been made correct by the model exploing GMM. Specifically, there was a lot of ambiguity while classifying the species 'Virginica' and 'versicoloar'. This can be attributed the fact that the 'Setosa' is linearly separable from the other two classes, while the remaining two are not linearly separable.
For this model, we could not obtain the Gelman And Rubin's Conver- gence Diagnostic. Effective size of the variables were found to be 4152.926 for μ1, 1103.172

for μ2 and 1168.982 for μ3. Deviance Information Criterion (DIC) for the model was obtained as 549.8.

Table 1: Confusion matrix for GMM

True (↓) Predicted(→)	Setosa	Virginica	versicoloar
Setosa	50	0	0
Virginica	5	45	0
Versicoloar	0	16	34

Bayesian Hierarchical Model

In Bayesian hierarchical modeling [1] of data, we shall again assume normal likelihood function:

$$y_{ij}|\mu_j, \sigma \sim \mathsf{N}(\mu_j, \sigma^2) \qquad i = 1, \ldots, n, j = 1, 2, 3 \tag{2}$$

where j subscript denotes three classes of data. This model implies that observation yij is assumed to follow normal distribution with mean μji and σ2. One of the key difference between Eq. (2) and Eq. (1) is that in the former we are trying to estimate the membership of the observation variable using latent variable zi whereas in the later case μji clearly utilizes the knowledge of the membership of observation to the population.

Checking the model and Results

We ran MCMC convergence tests with 3 chains with 1000 iterations as burn in period. Three means (with standard deviations) corresponding to three classes were obtained as:

$$\mu_1 = 10.143 \pm 0.171, \ \mu_2 = 14.296 \pm 0.173, \ \mu_3 = 17.135 \pm 0.172.$$

It is noted that there was not a very huge difference between the means obtained using mixture model and means obtained using hierarchical model. However, we again emphasize that Eq. (1) does not exploit any information related to the membership of observation to the population it belongs to.

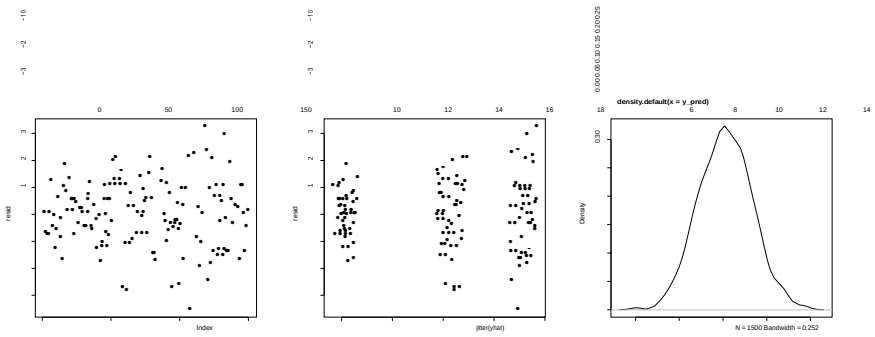

Fig 2: (a) and (b)Residual Analysis associated with the posterior means of the parameters. (c) Predictive posterior density corresponding to Species 'Setosa'

For this model, the Gelman And Rubin's Convergence Diagnostic was obtained to be 1 for all the variables. Effective size of the variables were found to be 9867.433 for μ1, 8814.902 for μ2 and 15763.297 for μ3. These numbers are significantly larger than the one obtained by mixture model. Deviance Information Criterion (DIC) for the model was obtained as 488.1 which is smaller than the one obtained using mixture model. This is quite natural because mixture models have additional latent variables which leads to greater DIC value.
To check the fit via residuals, we looked at the residuals associcated with the posterior means of the parameters as shown in Fig. 2.

Multinomial Logistic Regression
For binary discrete variables, data is fit into linear regression model, which then be acted upon by a logistic function (where 'logit' is link function) predicting the target categorical dependent variable [1]. This is the heuristic behind 'binary' logistic variable. When there are several possible categories that the dependent variable can fall into, we have multinomial logistic regression. If y ∈ {1, 2, . . . , k}, model the data using multinomial likelihood function where k parameters φ1, . . . , φk−1 are used to specify the probability of each of the outcomes with $\Sigma_{k-1\varphi} = 1_{i=1} \varphi$.

Therefore ik

$$\varphi_i = p(y = i; \varphi), \qquad \text{with } \varphi_k = p(y = k; \varphi) = 1 - \sum_{i=1}^{k-1} \varphi_i.$$

Table 2: Confusion matrix for Multinomial Logistic Regression with threshold = 0.7.

True (↓) Predicted(→)	Setosa	Virginica	versicoloar
Setosa	50	0	0
Virginica	2	48	0
versicoloar	0	2	48

It can be shown that

$$\varphi_i = p(y = i|x;\theta) = \frac{e^{\theta_i^T x}}{\sum_{j=1}^{k} e^{\theta_j^T x}}$$

where x is a vector of covariates and θj is the parameter vector (weights) corresponding to class j. This is also called softmax regression.

By using the log link function and Eq. (3), we classify the dataset and obtain the confusion matrix shown in Table 2.

It can be seen that using this model, we have obtained 98.33% classification accuracy.

4 Conclusions

In this paper, we tried to classify Iris flower dataset using three models, viz. Gaussian mixture model, Bayesian hierarchical model and multinomial logistic regression. While GMM does not use any a priori information re- garding the membership of data to a specific population (except number of mixtures which in this case is 3), Bayesian hierarchical modeling and multinomial regression do use this information. It was seen that the results obtained using GMM and hierarchical model were based on the approxima- tion of the underlying density of the data and they were almost equivalent. Whereas logistic regression based modeling was based on firstly fitting data onto linear regression model and then using an 'appropriate' link function to predict the target categorical variable. For reproducibility of the results and more detailed analysis of the data, R notebook has been uploaded on the github page https://github.com/shruti51/Iris Classification.

References

[1] Christopher M. Bishop, "Pattern Recognition and Machine Learning" (Information Science and Statistics), Springer-Verlag, Berlin, Heidel-berg, 2006.

[2] Fisher,R.A. "The use of multiple measurements in taxonomic prob-lems" Annual Eugenics, 7, Part II, 179-188, 1936.

[3] https://www.r-project.org/