U. PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Statistical Inference Project

Generative Models for Classification

Miguel Bourgin Gonçalves
nº202107127

Index

1 Generative Models 3

1.1 Linear Discriminant Analysis (LDA) . . . 3

1.2 Quadratic Discriminant Analysis (QDA) . . . 3

1.3 Naive Bayes . . . 3

2 Implications of Assumptions 4

3 Data Pre-Processing 5

4 Model Fitting 5

4.1 Strategy . . . 5

4.2 Implementation . . . 5

5 Results 5

6 Conclusion 6

Abstract

This report presents a project conducted within the Statistical Inference course at the Faculty of Sciences of the University of Oporto. It explores the application of Naive Bayes, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA) models to a dataset containing various credit related variables. The goal of this project is to compare the performance of these classification techniques in predicting income based on factors such as gender, education, etc [1]. This report is based on 'An Introduction to Statistical Learning' [2].

1 Generative Models

Logistic regression models $P(Y = k|X = x)$ directly using the logistic function, focusing on the conditional distribution of the response Y given the predictors X. Alternatively, generative models estimate these probabilities by modeling the distribution of the predictors X separately for each response class $Y = k$, and then applying Bayes' theorem to compute:

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \tag{1.1}$$

Here 1.1, π_k represents the prior probability of class k, while $f_k(x)$ is the density function of X for class k. Specifically, $f_k(x) = P(X|Y = k)$ is relatively large when an observation in class k has $X \approx x$, and small otherwise. The posterior probability $P(Y = k|X = x)$, often abbreviated as $p_k(x)$, is the probability that an observation with predictors $X = x$ belongs to class k. By estimating π_k and $f_k(x)$, we can approximate the Bayes classifier, which assigns an observation to the class with the highest posterior probability and achieves the lowest possible error rate, provided that π_k and $f_k(x)$ are correctly specified.

Generative models have several advantages over logistic regression. First, when there is substantial separation between the two classes, the parameter estimates for the logistic regression model can become unstable, whereas generative models are not affected by this issue. Second, if the distribution of the predictors X is approximately normal within each class, generative models may yield more accurate results, particularly for small sample sizes. Lastly, generative models naturally extend to multiclass classification problems, while logistic regression requires extensions such as multinomial logistic regression.

To implement generative models, it is necessary to estimate π_k and $f_k(x)$. Estimating π_k is relatively straightforward, as it can be calculated as the proportion of training observations in each class. However, estimating $f_k(x)$ is more challenging and typically requires simplifying assumptions about the distribution of the predictors. For example, it is often assumed that $f_k(x)$ follows a multivariate Gaussian distribution.

Several classifiers are derived from generative modeling principles, differing mainly in their assumptions about the distribution of $f_k(x)$. The most common methods are:

1.1 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis assumes that the predictors X follow a multivariate Gaussian distribution within each class, with a shared covariance matrix across all classes. This assumption simplifies the estimation of $f_k(x)$, leading to linear decision boundaries between classes.

1.2 Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis relaxes the shared covariance matrix assumption made by LDA and allows each class to have its own covariance matrix. This flexibility enables QDA to model more complex, non-linear decision boundaries, but it requires more parameters to be estimated, which may lead to overfitting for small datasets.

1.3 Naive Bayes

Naive Bayes assumes that the predictors are conditionally independent given the class label Y. This strong independence assumption simplifies the estimation of $f_k(x)$, making it computationally efficient, even when the number of predictors is large. Naive Bayes is particularly useful in high-dimensional settings but may perform poorly if the independence assumption does not hold.

These generative approaches approximate the Bayes classifier and can often outperform logistic regression under the right conditions, particularly when the assumptions about the predictor distributions are valid.

2 Implications of Assumptions

After selecting a model, we estimate the parameters of the within-class distributions to determine the likelihood of our test observation, and obtain the final conditional probability we use to classify it.

The different models result in a different number of parameters being estimated. Reminder: we have p predictors and K total classes. For all models we need to estimate means of the Gaussian distribution of the predictors, that can be different in each class. This results in a base, $p \times K$ parameters to be estimated for all methods.

Additionally, if we pick LDA we estimate the variances for all p predictors and covariances for each pair of predictors, resulting in

$$p + \frac{p!}{2!(p-2)!} \tag{2.1}$$

parameters. These are constant across classes.

For QDA, since they differ in each class, we multiply the number of parameters for LDA times K, resulting in the following equation for the estimated number of parameters:

$$[p + \frac{p!}{2!(p-2)!}] \times K. \tag{2.2}$$

For Naive Bayes, we only need the prior probabilities for each class and the probability distributions for each feature (predictor) conditioned on the class. These distributions can vary depending on the nature of the features (e.g., Gaussian, multinomial, etc.), but in the general case, we assume conditional independence of the features given the class. For a given number of features p and classes K, the model's complexity is determined by the number of parameters to estimate, which involves the priors for each class and the parameters of the feature distributions for each class. It is easy to see the advantage of using Naive Bayes for large values of p and/or K, especially in high-dimensional spaces. For the frequently occurring problem of binary classification (i.e., when K=2), this is how the model complexity evolves for increasing p for Naive Bayes compared to other algorithms.

In terms of complexity, GNB (Gaussian Naive Bayes) and NB are similar as long as feature independence is assumed. The only difference is that GNB assumes continuous features with a Gaussian distribution, while general NB (e.g., Multinomial or Bernoulli) is suited for categorical or discrete data.

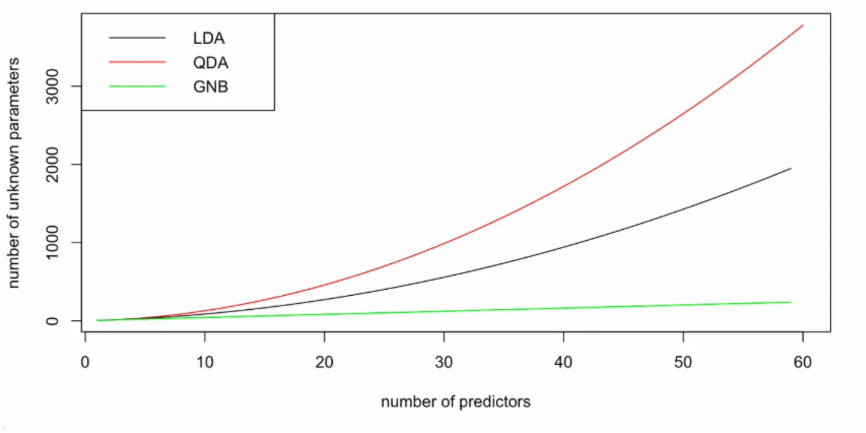

Figure 1: Model Complexity

Figure 1 compares the complexity of three classification models, LDA, QDA, and NB (since it is equal to GNB), based on the number of predictors and unknown parameters. LDA (black line) shows a linear increase in parameters, as it assumes a shared covariance matrix across classes. QDA (red line) exhibits a quadratic growth in parameters, as it allows each class its own covariance matrix. NB (green line) maintains a constant number of parameters, assuming feature independence.

The differences highlight trade-offs: NB is simplest and works well with high-dimensional data but assumes independence, LDA balances complexity and robustness, and QDA, while flexible, risks overfitting in high-dimensional spaces due to its high parameter count. This emphasizes the importance of matching the complexity of the model with the size and structure of the dataset.

3 Data Pre-Processing

The dataset used in this project is based on the **Credit** dataset, consisting of demographic, financial, and behavioral variables. The objective is to predict income categories using statistical models. Several preprocessing steps were performed to prepare the data for modeling.

The dataset was imported into the R environment, from **ISLR** package and columns all columns besides `Id` were selected for analysis. Categorical variables, such as `Gender`, `Ethnicity`, `Married`, and `Student`, were converted into numeric factors for correlation analysis. A correlation matrix was computed, revealing strong correlations among `Rating`, `Limit`, and `Balance`. Both keeping and removing some variables is going to be done and we will see how each model performs. The target variable, `Income`, was categorized into three classes: Low, Medium, and High, using quantile-based bins.

The dataset was split into training (80%) and test (20%) sets to facilitate model evaluation. These preprocessing steps ensured the data was clean and ready for models like Naive Bayes, LDA, and QDA.

These preprocessing steps ensured the dataset was clean, consistent, and ready for implementing the generative models used in this project, including Naive Bayes (NB), Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA).

4 Model Fitting

4.1 Strategy

The primary objective is to predict income categories (Low, Medium, High) using features like gender, marital status, student status, etc. Naive Bayes is going to be implemented as a probabilistic generative model and its effectiveness will be compared with LDA and QDA.

4.2 Implementation

- **Naive Bayes:** This model classifies observations based on Bayes' theorem, assuming Gaussian distributions for continuous predictors. The implementation was carried out using the `naiveBayes()` function from the `e1071` package.

- **Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA):** Both models were implemented using the `lda()` and `qda()` functions from the `MASS` package. LDA assumes a shared covariance matrix across classes, resulting in linear decision boundaries, while QDA allows for distinct covariance matrices per class, enabling the modeling of non-linear decision boundaries.

For all models, performance metrics such as accuracy, precision, and recall are to be computed. The results will be discussed in the following section.

5 Results

After fitting the Naive Bayes, LDA, and QDA models, we evaluate their performance on the test data.

Model Setup	Naive Bayes Accuracy	LDA Accuracy	QDA Accuracy
All Variables	0.514	0.654	0.615
Without Rating	0.500	0.654	0.615
Without Limit	0.474	0.628	0.590
Without Balance	0.519	0.500	0.539
Only Limit	0.539	0.487	0.526
Only Rating	0.500	0.500	0.513
Only Balance	0.385	0.385	0.359

Table 1: Model Accuracy Comparison All Numerical Variables

The table presents the accuracy of three models (Naive Bayes, LDA, and QDA) across different subsets of variables. Notably, the **Naive Bayes** model performs best with the full set of variables except `Balance` and `Rating` (0.539 accuracy), with accuracy dropping slightly in other cases. Removing `Limit` appears to have the most significant impact, with the accuracy falling to 0.474. The **LDA** model achieves its highest accuracy (0.654) with all variables, but this accuracy is maintained even when `Rating` or `Limit` is removed. **QDA** tends to perform worse than LDA overall, with the lowest accuracy (0.359) observed when using only `Balance`. These results suggest that while LDA is relatively robust to variable exclusion, Naive Bayes and QDA are more sensitive to the choice of variables, particularly the presence of `Limit` and `Balance`.

Surprisingly, Naive Bayes is performing worse than expected, and we will explore possible strategies to improve its accuracy.

Model	**Accuracy (Reduced)**	**Accuracy (Scaled)**	**Accuracy (Final)**
Naive Bayes	0.462	0.513	0.962
LDA	0.705	0.692	0.731
QDA	0.667	0.590	0.885

Table 2: Model Accuracies for Reduced, Scaled, and Final Models

The table presents the accuracy of four models (Naive Bayes, LDA and QDA) across three different setups: Reduced, Scaled, and Final models. The **Reduced** and **Final** models include only `Cards`, `Limit`, `Balance`, and `Education`, while the **Scaled** model uses all variables. These variables were selected through backward elimination, as they were always statistically significant. The **Scaled** and **Final** models are scaled using the function **scale()** from **R**, while the **Reduced** is not scaled.

The Naive Bayes model shows a significant improvement in accuracy from 0.462 and 0.513 in the Reduced and Scaled models to 0.962 in the Final model, indicating that scaling and using only some more relevant variables have a substantial positive impact. LDA improves moderately from 0.705 (Reduced) to 0.731 (Final), while QDA shows a bigger improvement, going from 0.667 to 0.885.

Overall, the results highlight the importance of variable selection and scaling in improving model performance. The Naive Bayes model, initially performing poorly with the Reduced and Scaled setups, demonstrates a remarkable improvement in accuracy when both scaling and relevant variables are used in the Final model. This suggests that careful feature selection and normalization are crucial for optimizing Naive Bayes. LDA and QDA also show improvements, with QDA experiencing a large accuracy boost, because of the covariance assumptions. Overall, scaling and including more relevant features significantly enhance model performance, especially for models like Naive Bayes and QDA, which are sensitive to feature set and data scaling. These findings suggest that feature selection and preprocessing steps, such as scaling, play an essential role in improving model accuracy.

6 Conclusion

In this study, we compared three classification methods—Naive Bayes, LDA, and QDA—on a dataset of personal credit. Naive Bayes (Final) emerged as the most effective model, with the highest accuracy, precision, and recall. Despite its strong assumption of conditional independence between features, Naive Bayes was able to make reliable predictions on the Income Category after the alterations. Suggesting that feature independence were a critical limitation in this context, but once we dealt with that the model, as expected was the best one.

LDA and QDA, while still useful, did not outperform Naive Bayes in this case. LDA's linear assumptions and QDA's additional complexity with covariance matrices were less suited to the nature of the dataset, where relationships between features may not be strictly linear.

The significant improvement in the Naive Bayes model, as well as QDA and LDA, can be attributed to the fact that scaling and feature selection help address issues related to feature correlation and differing variable magnitudes. In the Reduced and Scaled models, Naive Bayes may have struggled due to irrelevant or highly correlated variables, leading to inaccurate probability estimates. By focusing on the most relevant features (Cards, Limit, Balance, and Education), the model becomes more robust, as scaling normalizes the data and reduces the influence of outliers. Similarly, QDA and LDA benefit from the scaling, as these models are sensitive to the scale of the data, which can distort the decision boundaries if not properly addressed. In essence, scaling and careful feature selection allow the models to better capture the underlying patterns in the data, resulting in improved accuracy.

References

[1] Credit, package-ISLR, from R.

[2] An Introduction to Statistical Learning. Available: https://www.statlearning.com/

[3] Differences between LDA, QDA and Gaussian Naive Bayes classifiers [Online]. Available: https://towardsdatascience.com/differences-of-lda-qda-and-gaussian-naive-bayes-classifiers-eaa4

[4] LDA, QDA, Naive Bayes Generative Classification Models. Marek Petrik. [Online]. Available: https://www.cs.unh.edu/~mpetrik/teaching/intro_ml_17/intro_ml_17_files/class5.pdf