A Review & Comparison of Classification Methods

TEAM 17

Agenda

- Bayes Classifier
- Logistic Regression
- Linear Discriminant Analysis
- QDA Vs. KNN

What is the Bayes Classifier?

Bayes Classifier is the general name of a class of classification algorithms, which are based on Bayes theorem, so they are collectively called Bayes classifier.

It is possible to show that the test error rate given by the expression is minimized, on average, by a very simple classifier that assigns each observation to the "most likely" class, given its predictor values.

Easy way →

we should simply assign an observation with predictor vector x_0 to the class j for which the following conditional probability is largest:

$$p(Y=j|X=x\_0)$$

This classifier is called the **Bayes Classifier**

Basic concepts of Bayes' Theorem

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

$$P(c|x)=\frac{P(x,c)}{P(x)}=\frac{P(c)P(x|c)}{P(x)}$$

A, B	events
P(A|B)	probability of A given B is true
P(B|A)	probability of B given A is true
P(A), P(B)	the independent probabilities of A and B

Bayesian formula is based on conditional probability to find the cause of the occurrence of events

For each feature x, the samples we want to know under this feature x belongs to which category, which seeks a Posterior probability P (c | x) the largest class tag.

P(category | feature) =P(c | x)

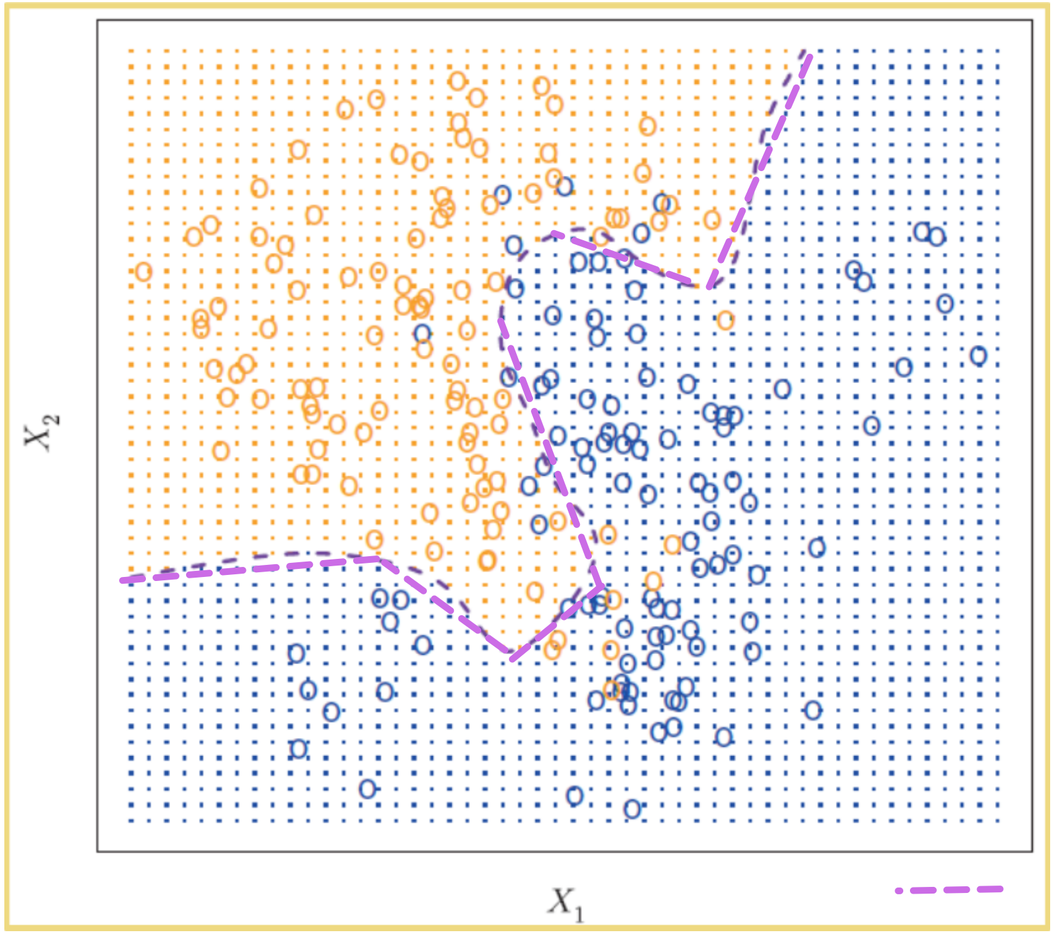
P(feature | category) = P(x | c)

P(feature) =P(x)

P(category) = P(c)

- **Prior probability: refers to the probability obtained based on previous experience and analysis.**
- **Posterior probability: The probability that something has happened and that it happened because of some factor.**
- **Posterior probability is a conditional probability**

Graph of Bayes Classifier



The Graph consists of a simulated data set of 100 observations consisting of predictors X_1 and X_2.

- Each observation falls into one of two groups orange and blue.
- For each value of X_1 and X_2, there is a different probability in the orange (or blue) group.

Since this is simulated data, we know how the data were generated and we can calculate the Bayes classifier's conditional probabilities for each value of X_1 and X_2.

The purple dashed line is called the Bayes decision boundary.

The Bayes classifier produces the lowest possible test error rate, called the Bayes error rate.

Problems We Meet !

- **The problem is that we never know how to compute the actual values of these conditional probabilities –in a real-world setting, computing the Bayes Classifier isn't possible.**

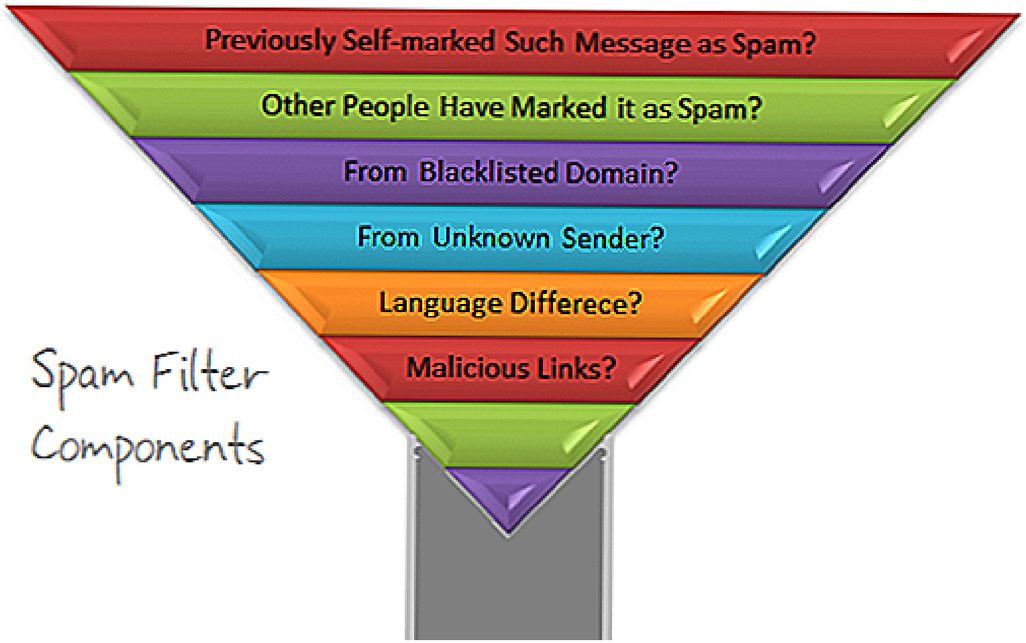
- **The main difficulty in estimating the posterior probability P(c|x) with the Bayesian formula is that P(x|c) is the joint probability on all attributes (x represents multiple attributes) , which is difficult to estimate directly from limited training samples.**

- **Therefore, the Bayes Classifier serves as an unattainable gold standard against which to compare other methods.**

Naïve Bayes Classifier

What is the Naive Bayes Classifier ?

- Naive Bayes classifier is the simplest and most common classification method in Bayesian classifier.
- It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.



Examples of Naïve Bayes :

- spam filtration
- classifying articles
- Sentimental analysis

Naïve Bayes Classifier

Example:

Lab/Assignment/Practice
Features(1) Features(2) Features(3)

- Prior probability: P (good grade) =6/10=0.6
- Conditional Probability: P (Lab | Good) =4/6=2/3
- Conditional Probability: P (Assignment | Good) =4/6=2/3
- Conditional Probability: P (Practice | Good) =2/6=1/3

A good Grade
or
A bad Grade

P (Good) x P (Lab | Good) x P (Assignment | Good) x P (Practice | Good) = 0.6 × (2 / 3) × (2 / 3) × (1 / 3) = 4 / 45

P(Bad) × P(Lab | Bad) × P(Assignment | Bad) × P(Practice | Bad) = 0.4 × 0.25 × 0.25 × 0.25 = 1 / 160

Features(1) Features(2) Features(3)
Lab/Assignment/Practice

- Prior probability: P (bad grade) =4/10=0.4
- Conditional Probability: P (Lab | Bad) =1/4=0.25
- Conditional Probability: P (Assignment | Bad) =1/4=0.25
- Conditional Probability: P (Practice | Bad) =1/4=0.25

Compare the posterior probabilities in the good and bad grades categories: 4/45 >1/160 So the Lab/Assignment/Practice will think as Good Grade.

Logistic Regression

About

Logistic regression is a type of classification, and the main distinction from linear regression is that the dependent variable is binary categorical instead of continuous.
Logistic regression could solve problems like detecting SPAM, or detecting Cancer.

The Logistic Regression Function

- The conditional probability distribution= p(X)=Pr(Y=1|X)
- Y is a binary response variable and we are using the generic 0/1 coding for its two classes.

$$\hat{p} = \frac{\exp(b_0 + b_1 x)}{1 + \exp(b_0 + b_1 x)}$$

- Where b_0 and b_1 are the estimates of the population parameters β_(0)and β_1 and p̂ is the predicted probability of success.
- exp is the exponent constant [Euler's number] ≈2.71828

LDA

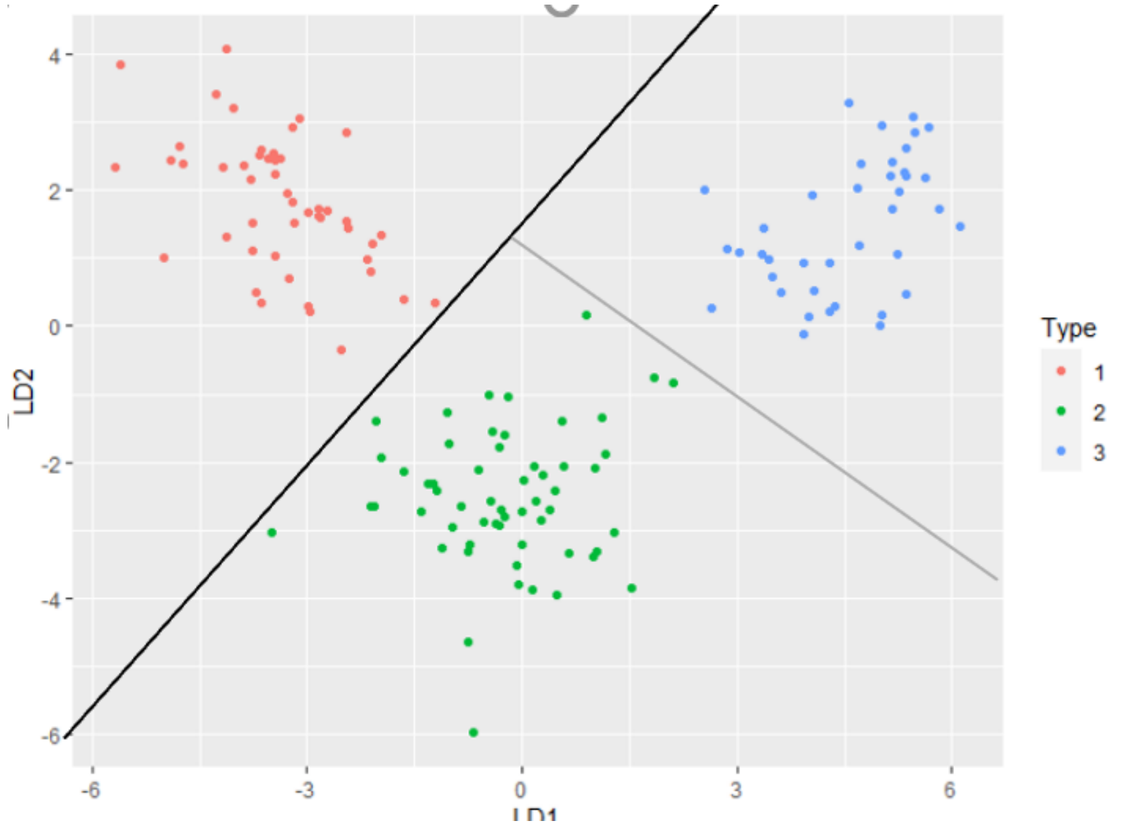
- LDA is a classification method for classifying a set of observations into predefined classes. The purpose is to determine the class of an observation based on a set of predictor variables.
    a) Uses predefined classes based on a set of linear discriminant functions of the predictor variables.

- normal (Gaussian) distributions for each class

Why discriminant analysis?

- When the classes are **well-separated**, the parameter estimates for logistic regression model are surprisingly unstable. Linear discriminant analysis does not suffer from this problem.
- If n is small and the distribution of the predictors X is approximately normal in each of the classes, the LDA model is again **more stable** than the logistic regression model.
- LDA is **popular** when we have more **than 2 response classes**, because it also provides low-dimensional views of the data.

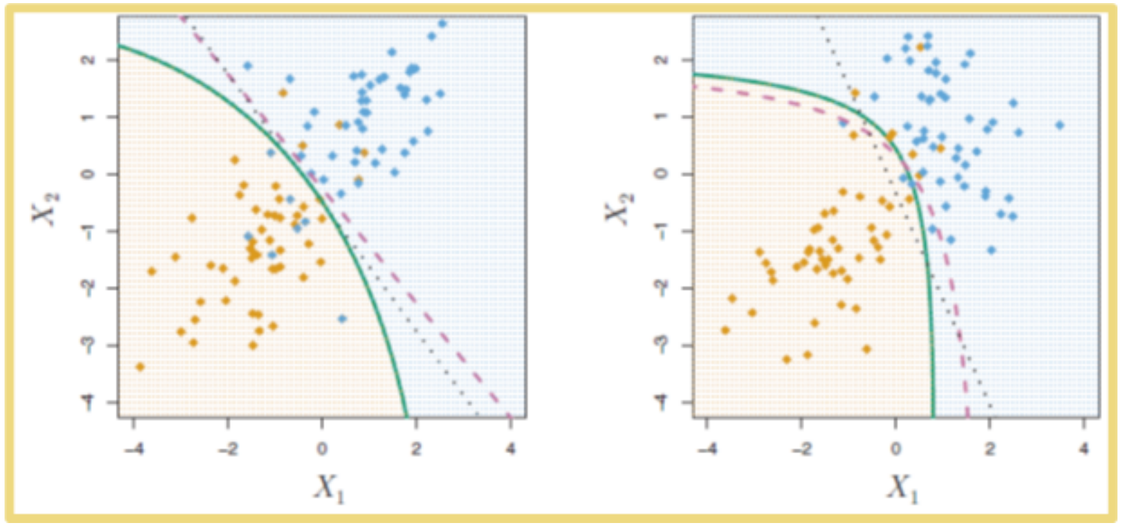
Fisher's Discriminant Plot



- When there are K classes, linear discriminant analysis can be viewed exactly in a **K-1** dimensional plot, because it essentially classifies to the closest centroid, and they span a K-1 dimensional plane.

LDA vs QDA

- The model on the left is both an LDA and QDA with normal Gaussian distribution .
- As a result, the Bayes decision boundary is linear and is accurately approximated by the LDA decision boundary.
- The QDA decision boundary is inferior, because it suffers from higher variance without a corresponding decrease in bias.



- In contrast, the model on the right displays a situation in which the orange class has a correlation of 0.7 between the variables and the blue class has a correlation of −0.7.
- Now the Bayes decision boundary is quadratic, and so QDA more accurately approximates this boundary than does LDA.

KNN

- KNN is a completely non-parametric approach: no assumptions are made about the shape of the decision boundary. Therefore, we can expect this approach to dominate LDA and logistic regression when the decision boundary is highly non-linear.
- KNN does not tell us which predictors are important
    - it is very unreliable in high dimensions.

QDA

- QDA serves as a compromise between the non-parametric KNN method and the linear LDA and logistic regression approaches.
- Since QDA assumes a quadratic decision boundary, it can accurately model a wider range of problems than can the linear methods.
- Though not as flexible as KNN, QDA can perform better in the presence of a limited number of training observations because it does make some assumptions about the form of the decision boundary

Review of Code in R