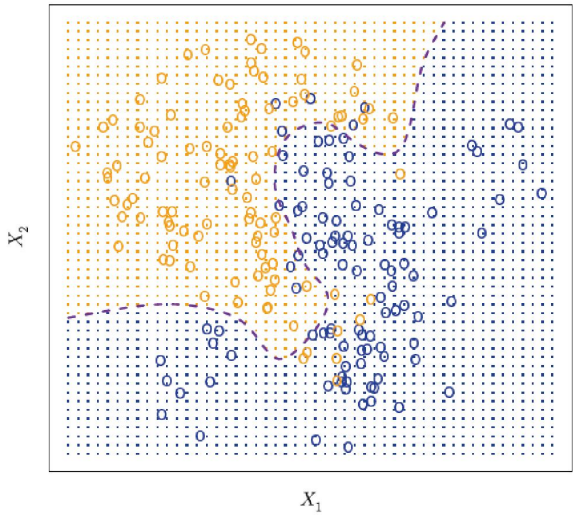
CLASSIFICATION

What is classification?

- In linear regression, we assume that the response variable Y is quantitative.
- But in classification the response variable Y is qualitative.
- Qualitative variables are referred as categorical.
- Widely used classifiers are
- K- nearest neighbors
- Logistic regression
- Linear discriminant analysis

Bayes' classification

- This method assigns each observation to the most likely class, given its predictor values. The test error rate Ave (I(y0 ≠ ˆy0)) is minimized.
- Here, we simply assign a test observation with predictor vector *x*0 to the class *j* for which Pr(Y = j|X = x0) is largest.
- It is the conditional probability that *Y* = *j*, given the observed predictor vector *x*0.
- If Pr(Y = j|X = x0) > 0.5 it belongs to class 1or else it belongs to class 2
- The dashed line in the figure represents bayes' decision boundary



Why not Bayes' classification?

- For real data, we do not know the conditional distribution of *Y* given *X*, and so computing the Bayes classifier is impossible.
- Therefore, classification using Bayes' method is not possible.

- Many approaches attempt to estimate the conditional distribution of *Y* given *X*, and then classify a given observation to the class with highest *estimated* probability.
- One such method is the *K-nearest neighbors* (KNN) classifier.

K-Nearest neighbors

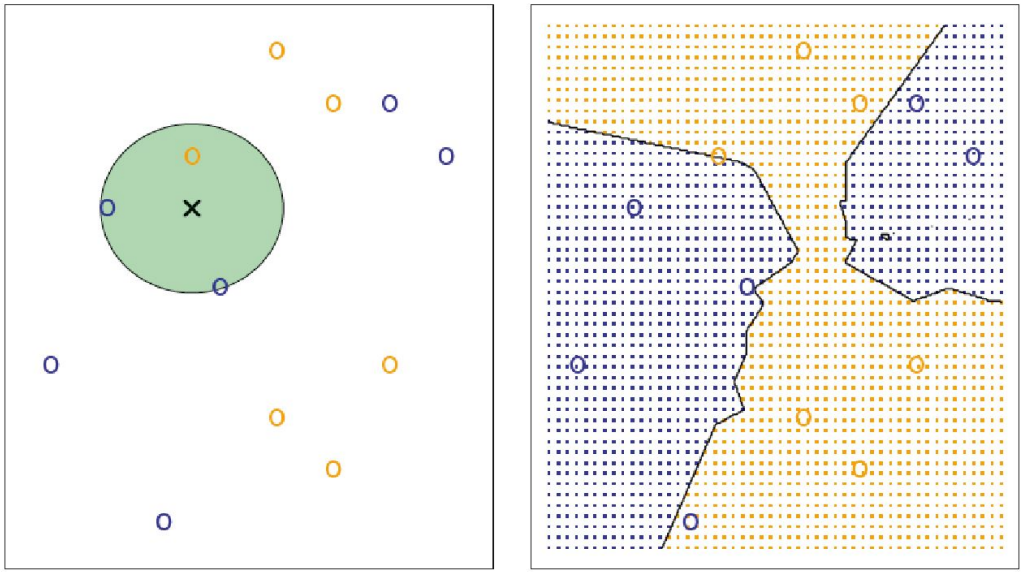
- KNN classifier first identifies the neighbors *K* points in the training data that are closest to *x*0, represented by *N*0.
- It then estimates the conditional probability for class *j* as the fraction of points in *N*0 whose response values equal *j*:

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j).$$

- Finally, KNN applies Bayes rule and classifies the test observation *x*0 to the class with the largest probability.

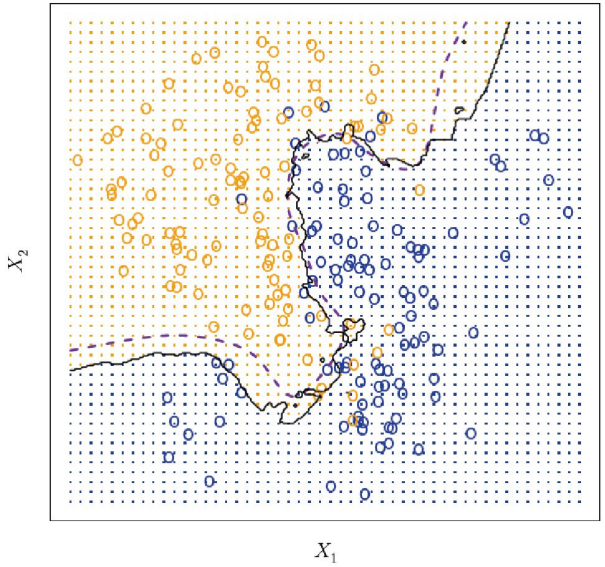
KNN approach with different K values

K=3 (by considering 3 nearest neighbors for each data point.



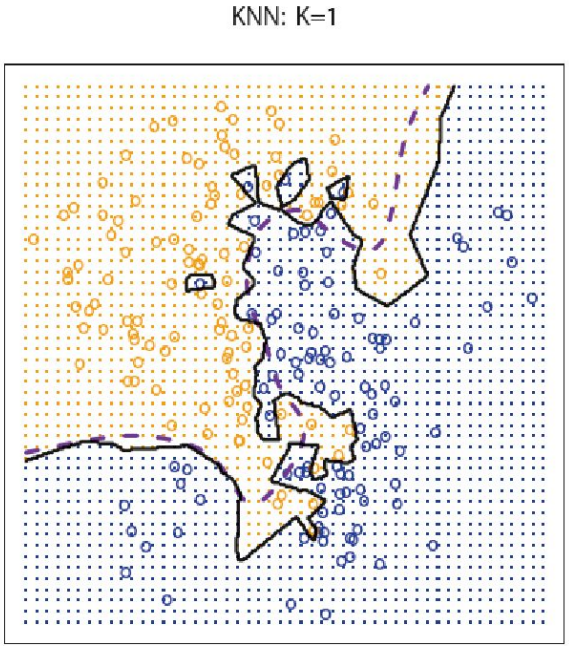
KNN approach with different K values

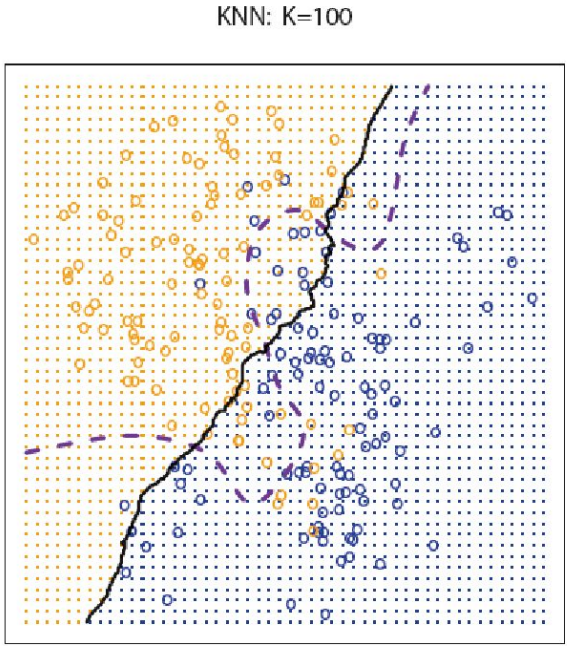
K=10 (by considering 10 nearest neighbors for each data point.



KNN approach with different K values

K = 1 and K = 100



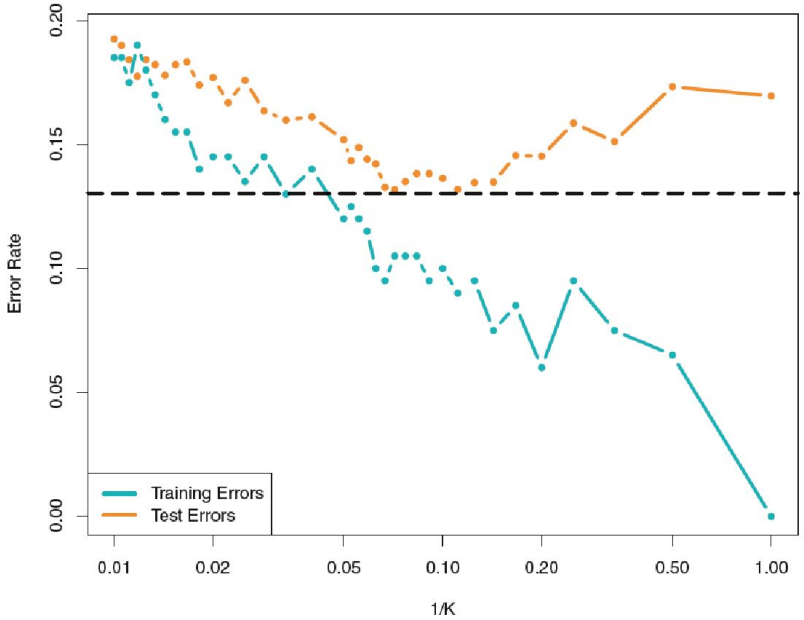


Optimal K value

- We see that the boundaries vary largely with the choice of K values we use.
- When *K* = 1, the decision boundary is overly flexible and finds patterns in the data that don't correspond to the Bayes decision boundary. This corresponds to a classifier that has low bias but very high variance.
- As *K* grows, the method becomes less flexible and produces a decision boundary that is close to linear. This corresponds to a low-variance but high-bias classifier.

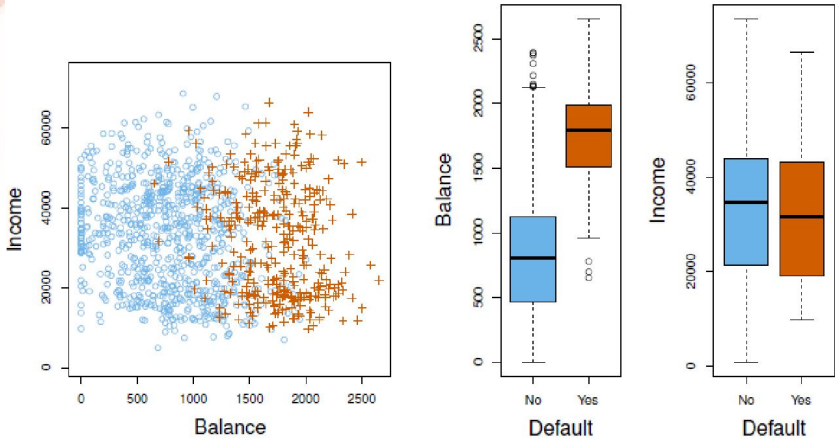
Optimal K value

This is how we choose the K value. The one which has the lowest error rate should be chosen.



Logistic regression

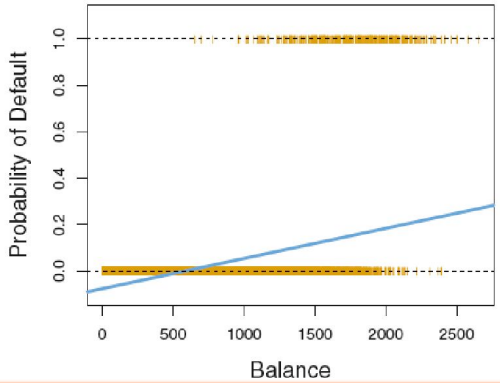
- Here we consider an example where we are classify the individuals who will default credit card payments from who does not.
- The first graph shows the relation between Income and Balance with The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.
- On the right side, these box plots is plotted for balance and income as a function default respectively.

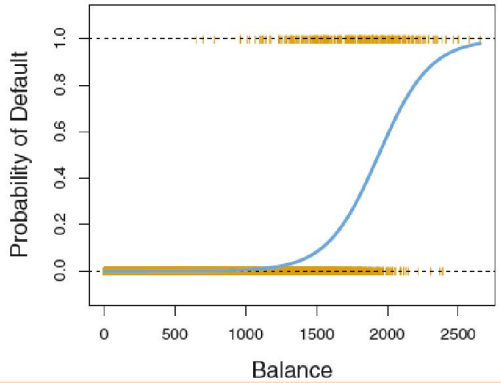


Why Not Linear Regression?

- Suppose for the Default classification task that we code $$Y = \begin{cases} 0 & \text{if No} \\ 1 & \text{if Yes.} \end{cases}$$

- Can we perform a linear regression of Y on X and classify as Yes if Y > 0.5?
- If we use linear regression, some of our estimates might be outside the [0, 1] interval making them hard to interpret as probabilities!





Logistic regression

- Rather than modeling this response *Y* directly, logistic regression models the probability that *Y* belongs to a particular category.
- The probability of default given balance can be written as

  Pr(default = Yes|balance).
- We must model $p(X)$ using a function that gives outputs between 0 and 1 for all values of X.
- In logistic regression we use logistic model

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

- And this model is fit using a maximum likelihood method

Logistic model

- The logistic function will always produce an *S-shaped* curve of this form, and so regardless of the value of *X*, we will obtain a sensible prediction.

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}.$$

- This quantity is called odds and it take the value between 0 and ∞

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

Z- statistics

- Z – statistics is same as t- statistics which is used in linear regression
- The *z*-statistic associated with $\beta 1$ is equal to $\hat{\beta}1/SE(\hat{\beta}1)$, and so a large (absolute) value of the *z*-statistic indicates evidence against the null hypothesis $H0 : \beta_1 = 0$
- It is calculated by taking the ratio of coefficient and the standard error.
- For the above example, the given table give the relationship between probability of default and balance

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

Making predictions

- To make predictions of probability of default for given balance, We use this function.

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}}$$

- This function gives the probability of default for give X value. β_0 and β_1 values are used from the one which is predicted using the maximum likelihood method.

Multiple logistic regression

- For binary response with multiple predictors, $$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

  where X = (X1, . . .,Xp) are p predictors
- For finding the probability this equation is written as, $$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

- And we use the maximum likelihood method to estimate $\beta 0, \beta 1,\ldots,\beta p$.

Multiple logistic regression

- When we consider balance, income and student/non-student to find whether the individual will default of not we calculate the following.

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

- The negative coefficient for student in the multiple logistic regression indicates that *for a fixed value of* balance *and* income, a student is less likely to default than a non-student.

Problems in multiple linear regression

- When we calculate the relation between the probability of default with the person being a student of non- student we get

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

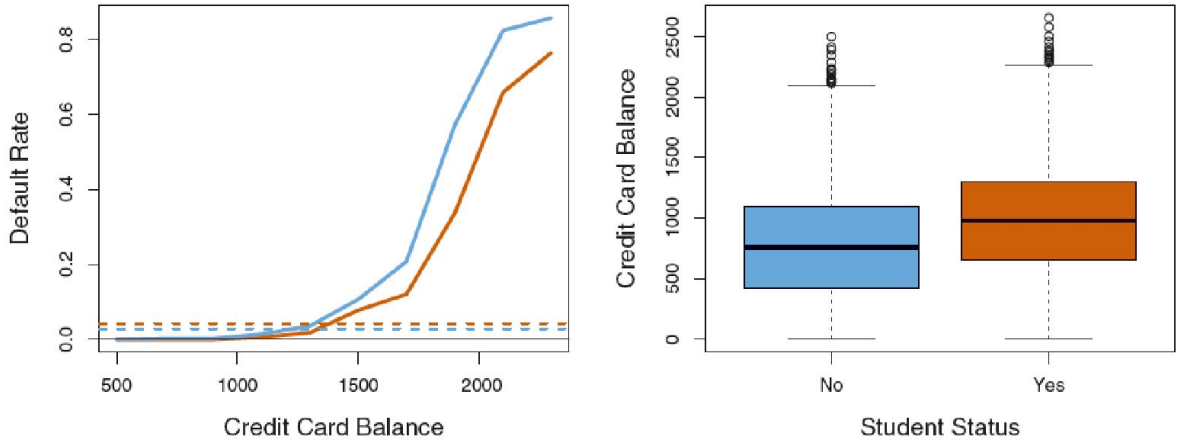
Where the coefficient is positive.

- But when we take multiple logistic regression, we get its coefficient to be negative.

Confounding

We observe from the left side graph that the student default rate is at or below that of the non-student default rate for every value of balance.

But the horizontal broken lines near the base of the plot shows that default rates for students and non-students averaged over all values of balance and income, suggest the reverse. This phenomenon is called confounding.



Reason for confounding

- The variables student and balance are correlated.
- Students tend to hold higher levels of debt, which is in turn associated with higher probability of default.
- Students are more likely to have large credit card balances, which tend to be associated with high default rates.
- Thus, even though an individual student with a given credit card balance will tend to have a lower probability of default than a non-student with the same credit card balance, the fact that students on the whole tend to have higher credit card balances means that overall, students tend to default at a higher rate than non-students.
- However, that student is less risky than a non-student *with the same credit card balance.*

Linear discriminant analysis

- Here, we model the distribution of the predictors X separately in each of the response classes (i.e. given Y), and then use Bayes' theorem to flip these around into estimates for $\Pr(Y = k|X = x)$.

- When these distributions are assumed to be normal, it turns out that the model is very similar in form to logistic regression.

Why not logistic regression?

- When the classes are well-separated, the parameter estimates for the 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 linear discriminant model is again more stable than the logistic regression model.

- Linear discriminant analysis is popular when we have more than two response classes.

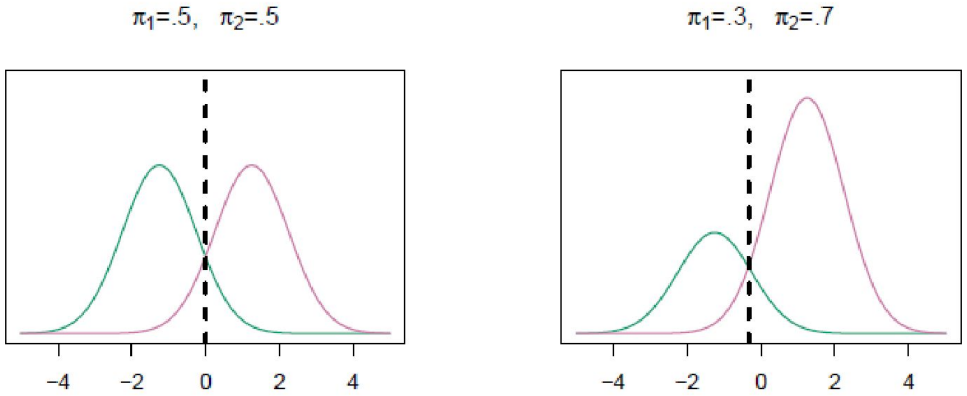
Using Bayes' Theorem for Classification

- Bayes' theorem states that

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

- We can simply plug in estimates of πk and $fk(X)$.
- πk denotes the prior probability of an observation belongs to the kth class
- We refer to pk(x) as the *posterior* probability that an observation $X = x$ belongs to the kth class. That is, it is the probability that the observation belongs to the kth class, *given* the predictor value for that observation.

Classify to the highest density



- We classify a new point according to which density is highest.
- When the priors are different, we take them into account as well, and compare $\prod$kfk(x).
- On the right, we favor the pink class - the decision boundary has shifted to the left.

Linear Discriminant Analysis for one predictor

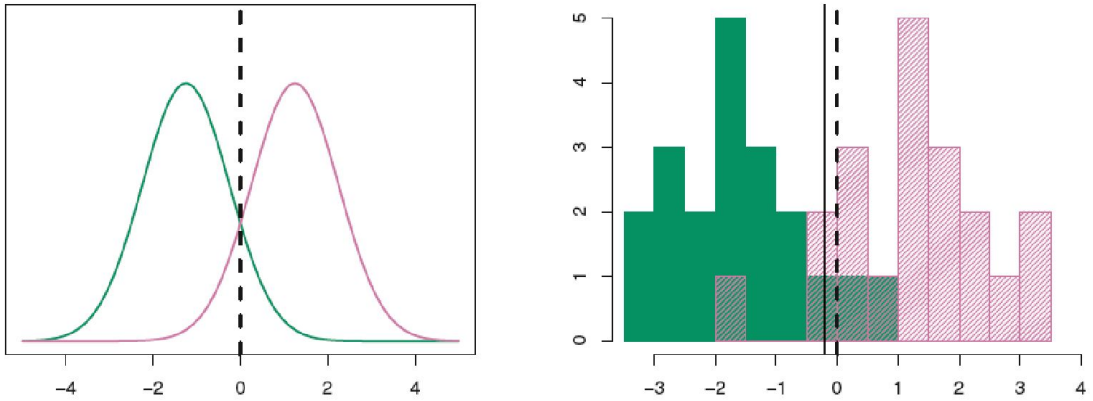
- We assume that *fk*(*x*) is *normal* or *Gaussian*. $f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right),$
-
- where μk and σk^2 are the mean and variance for the *k*th class.
- When we assu $\sigma_1^2 = \ldots = \sigma_K^2$ We get

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}.$$

- Taking the log rearranging the terms will give $\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$
- For K = 2 classes, $\pi_1 = \pi_2$ = then one can see that the decision boundary is at $x = \frac{\mu_1 + \mu_2}{2}.$

Discriminant functions

- The dashed vertical line represents the Bayes decision boundary.
- The solid vertical line represents the LDA decision boundary estimated from the training data.



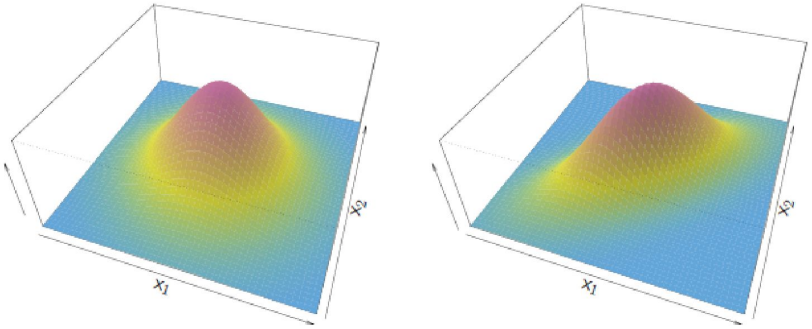
- We have to estimate the parameters to plug them into the rule.

Linear discriminant analysis for p>1

We assume that $X = (X1, X2, \ldots, Xp)$ is drawn from a *multivariate Gaussian* distribution.

$$\text{Density: } f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)}$$

$$\text{Discriminant function: } \delta_k(x) = x^T\mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}^{-1}\mu_k + \log \pi_k$$

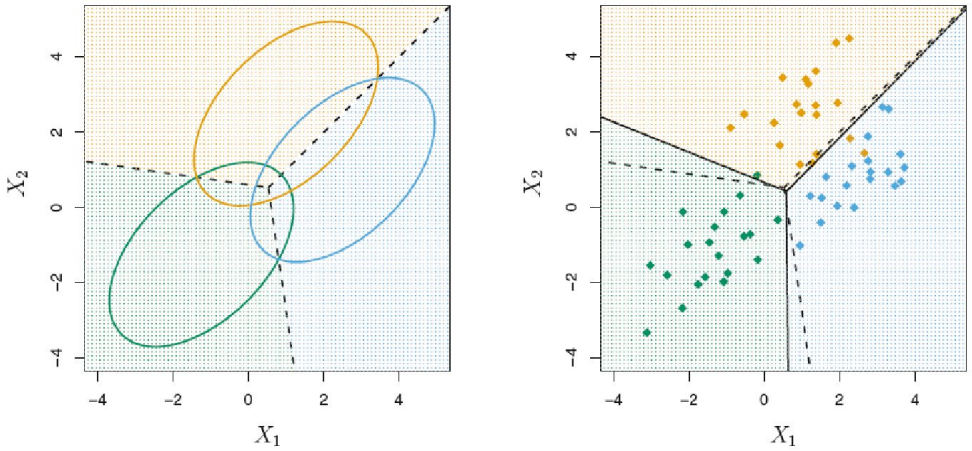


P = 2 and K = 3

The dashed lines are known as the Bayes decision boundaries.

They will yield the fewest misclassification errors, among all possible classifiers.

Here $\pi_1 = \pi_2 = \pi_3 = 1/3$.



Quadratic discriminant analysis

LDA assumes that the observations within each class are drawn from a multivariate Gaussian distribution with a class specific mean vector and a covariance matrix that is common to all K classes.

Unlike LDA, QDA assumes that each class has its own covariance matrix.

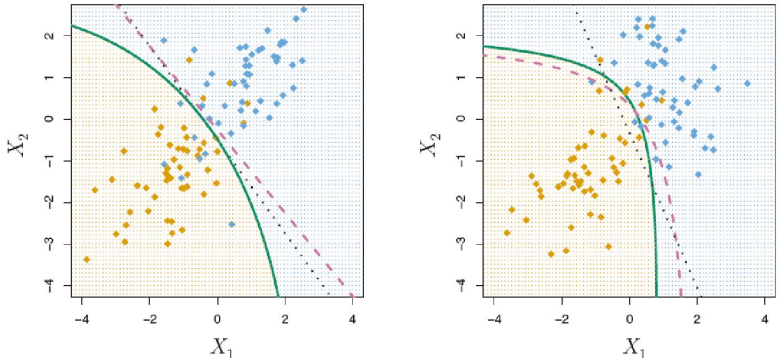
That is, it assumes that an observation from the kth class is of the form $X \sim N(\mu k, \Sigma k)$, where Σk is a covariance matrix for the kth class.

The Bayes classifier assigns an observation $X = x$ to the class for which

is the largest.

LDA and QDA

- The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with $\Sigma 1 = \Sigma 2$.
- The shading indicates the QDA decision rule. In the first one, since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA.
- Details are as given in the left-hand panel, except that $\Sigma 1 = \Sigma 2$. Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA.



Summary

We studied about different classification methods they are :

- Bayes' theorem
- K nearest neighbor approach
- Linear regression
- Logistic regression
- Linear discriminant analysis
- Quadratic discriminant analysis
- LDA vs QDA

Thank you