4.1 Introduction to Classification Problems

https://youtu.be/sqq21-VIa1c

Classification

- Qualitative variables take values in an unordered set $\mathcal{C}$, such as:
  `eye color` $\in \{$`brown`, `blue`, `green`$\}$
  `email` $\in \{$`spam`, `ham`$\}$.
- Given a feature vector X and a qualitative response Y taking values in the set $\mathcal{C}$, the classification task is to build a function $C(X)$ that takes as input the feature vector X and predicts its value for Y; i.e. $C(X) \in \mathcal{C}$.
- Often we are more interested in estimating the *probabilities* that X belongs to each category in $\mathcal{C}$.

Classification

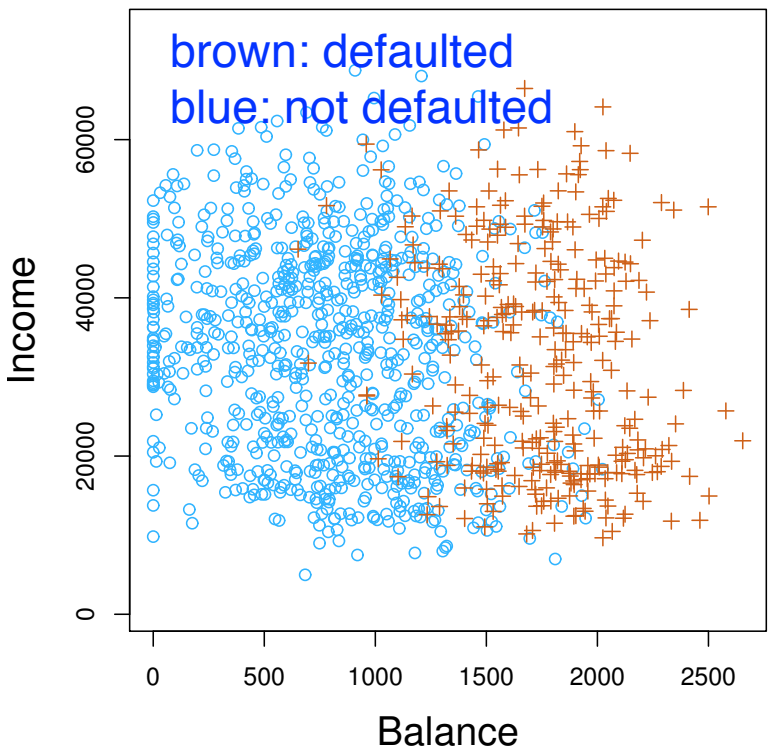
- Qualitative variables take values in an unordered set $\mathcal{C}$, such as:
  `eye color` $\in \{$`brown`, `blue`, `green`$\}$
  `email` $\in \{$`spam`, `ham`$\}$.
- Given a feature vector X and a qualitative response Y taking values in the set $\mathcal{C}$, the classification task is to build a function $C(X)$ that takes as input the feature vector X and predicts its value for Y; i.e. $C(X) \in \mathcal{C}$.
- Often we are more interested in estimating the *probabilities* that X belongs to each category in $\mathcal{C}$.

  For example, it is more valuable to have an estimate of the probability that an insurance claim is fraudulent, than a classification fraudulent or not.

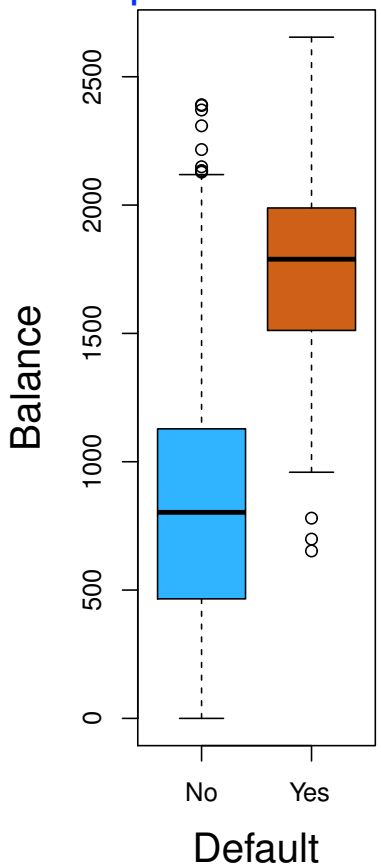
Almost one of the first things you should do when you get some data to analyze is do some scatter plots and create some box plots.

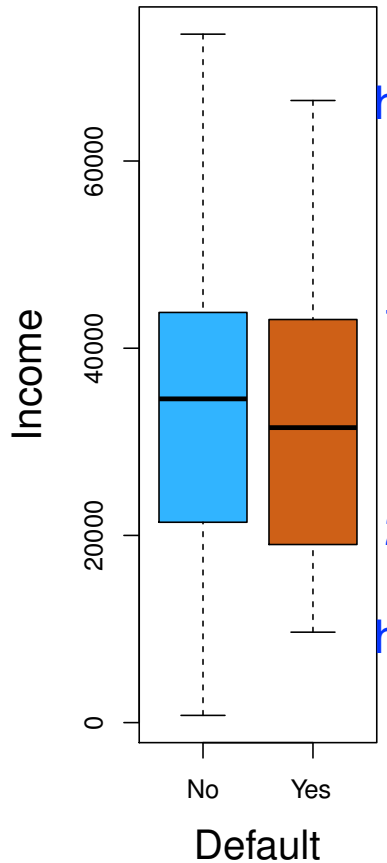
Example: Credit Card Default

2 variables: balance, income



boxplot:
black line = median
top/bottom of box = 75% and 25% quartile





hinge

75%

median

25%

hinge

in this plot, it looks like balance is the important variable. Notice that there's a big separation between the blues and the browns

And if data points fall outside the hinges, they're considered to be outliers.

Can we use Linear Regression?

Suppose for the `Default` classification task that we code

$$Y = \begin{cases} 0 & \text{if } \texttt{No} \\ 1 & \text{if } \texttt{Yes}. \end{cases}$$

Can we simply perform a linear regression of Y on X and classify as `Yes` if $\hat{Y} > 0.5$?

Can we use Linear Regression?

Suppose for the `Default` classification task that we code

$$Y = \begin{cases} 0 & \text{if } \texttt{No} \\ 1 & \text{if } \texttt{Yes}. \end{cases}$$

Can we simply perform a linear regression of Y on X and classify as `Yes` if $\hat{Y} > 0.5$?

https://stats.stackexchange.com/questions/300267/linear-regression-for-binary-response-same-classifications-as-lda

- In this case of a binary outcome, linear regression does a good job as a classifier, and is equivalent to *linear discriminant analysis* which we discuss later.
- Since in the population $E(Y|X = x) = \Pr(Y = 1|X = x)$, we might think that regression is perfect for this task.

Can we use Linear Regression?

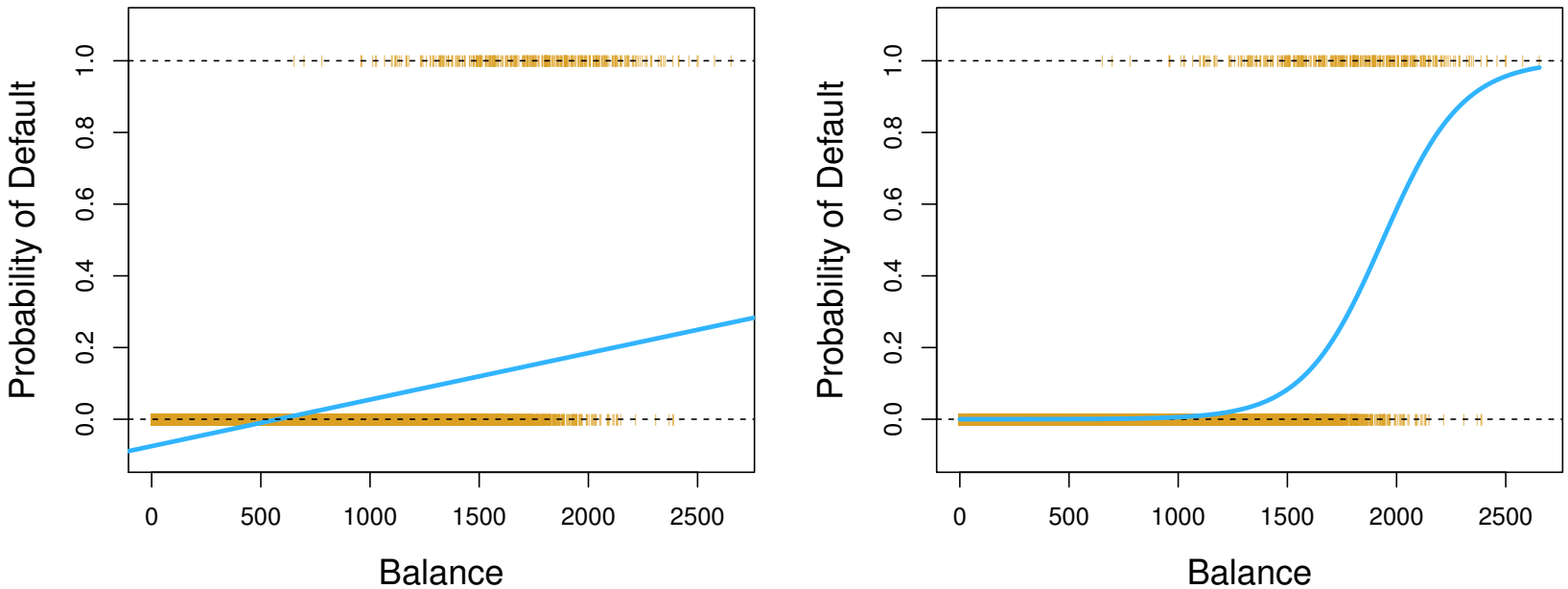
Suppose for the `Default` classification task that we code

$$Y = \begin{cases} 0 & \text{if } \texttt{No} \\ 1 & \text{if } \texttt{Yes}. \end{cases}$$

Can we simply perform a linear regression of Y on X and classify as `Yes` if $\hat{Y} > 0.5$?

- In this case of a binary outcome, linear regression does a good job as a classifier, and is equivalent to *linear discriminant analysis* which we discuss later.
- Since in the population $E(Y|X = x) = \Pr(Y = 1|X = x)$, we might think that regression is perfect for this task.
- However, *linear* regression might produce probabilities less than zero or bigger than one. *Logistic regression* is more appropriate.

Linear versus Logistic Regression



The orange marks indicate the response Y, either 0 or 1. Linear regression does not estimate $\Pr(Y = 1|X)$ well. Logistic regression seems well suited to the task.

Linear Regression continued

Now suppose we have a response variable with three possible values. A patient presents at the emergency room, and we must classify them according to their symptoms.

$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases}$$

in fact there's not necessarily an ordering here at all.

This coding suggests an ordering, and in fact implies that the difference between `stroke` and `drug overdose` is the same as between `drug overdose` and `epileptic seizure`.

Linear Regression continued

Now suppose we have a response variable with three possible values. A patient presents at the emergency room, and we must classify them according to their symptoms.

$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases}$$

This coding suggests an ordering, and in fact implies that the difference between `stroke` and `drug overdose` is the same as between `drug overdose` and `epileptic seizure`.

Linear regression is not appropriate here.
Multiclass Logistic Regression or *Discriminant Analysis* are more appropriate.

4.2 Logistic Regression

https://youtu.be/31Q5FGRnxt4

Logistic Regression

Let's write $p(X) = \Pr(Y = 1|X)$ for short and consider using `balance` to predict `default`. Logistic regression uses the form

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

($e \approx 2.71828$ is a mathematical constant [Euler's number.])
It is easy to see that no matter what values β_0, β_1 or X take, $p(X)$ will have values between 0 and 1.

So this is a transformation of a linear model to guarantee that what we get out is a probability.

Logistic Regression

Let's write $p(X) = \Pr(Y = 1|X)$ for short and consider using `balance` to predict `default`. Logistic regression uses the form

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

($e \approx 2.71828$ is a mathematical constant [Euler's number.])
It is easy to see that no matter what values β_0, β_1 or X take, $p(X)$ will have values between 0 and 1.

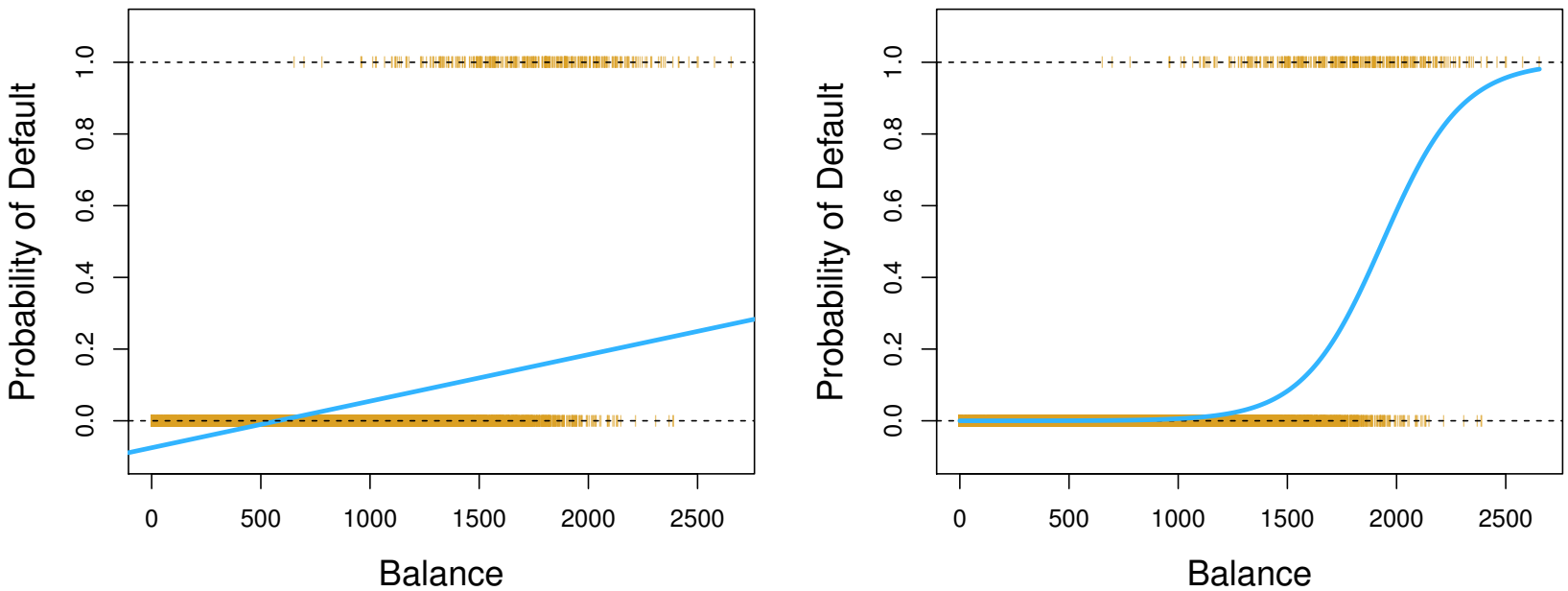
A bit of rearrangement gives

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

This monotone transformation is called the *log odds* or *logit* transformation of $p(X)$. (by log we mean *natural log*: ln.)

To summarize, we got a linear model still. But it's modeling the probabilities on a non-linear scale.

Linear versus Logistic Regression



Logistic regression ensures that our estimate for $p(X)$ lies between 0 and 1.

Maximum Likelihood

We use maximum likelihood to estimate the parameters.

$$\ell(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)).$$

This *likelihood* gives the probability of the observed zeros and ones in the data. We pick β_0 and β_1 to maximize the likelihood of the observed data.

https://www.youtube.com/watch?v=XepXtl9YKwc
The goal of maximum likelihood is to find the optimal way to fit a distribution to the data
ex: guess a distribution, find mean and standard deviation to best fit the distribution to the data (see how it maximize the likelihood)

https://www.quora.com/How-do-you-explain-maximum-likelihood-estimation-intuitively
Maximum likelihood estimation is to provide the parameters of a statistical distribution which would most likely produce data similar to what is given.

https://www.quora.com/What-is-Maximum-Likelihood
Lets say you have a biased coin. Let the probability of head be p. You toss the coin 10 times and you get 7 Heads and 3 Tails. The likelihood here is simply the probability of getting this result (data).
Likelihood ~ p^7 * (1-p)^3.
The idea is that the best estimate of p is the one which will maximize the likelihood. To make it easier take the log of this
Log Likelihood = 7*log(p) + 3*log(1-p). Taking the derivative and equating it to zero we get the value of p = 7/10 which is an estimate of p calculated using maximum likelihood estimation.

https://www.quora.com/What-are-the-differences-between-maximum-likelihood-and-cross-entropy-as-a-loss-function
…as you can see, maximizing the (log) likelihood is equivalent to minimizing the binary cross entropy….

my interpretation:
suppose we have 10 samples: (x1,y1) (x2,y2) …..(x10,y10), among which y1 ~ y7 are classified as “1”, and y8~y10 are “0”

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
y1	y2	y3	y4	y5	y6	y7	y8	y9	y10
1.	1.	1.	1.	1.	1.	1.	0.	0.	0

what is the distribution/parameters i can use to describe this data set ?

for each xi, assume that there is a transformation p to transform xi into a “0” or “1” , where p(x) = p(y=1 | x)
if this transformation is perfect, then we should get

p(x1)=1 p(x2)=1 p(x3)=1 p(x4)=1 p(x5)=1 p(x6)=1 p(x7)=1 p(x8)=0 p(x9)=0 p(x10)=0

the likelihood = p(x1) * p(x2) * p(x3) * p(x4) * p(x5) * p(x6) * p(x7) * [1-p(x8)] * [1-p(x9)] * [1-p(x10)]

since p = exp(b0 + b1*xi) / [1 + exp(b0 + b1*xi)], what is b0 and b1 that maximize the likelihood, so that
p(x1)~1 p(x2)~1 p(x3)~1 p(x4)~1 p(x5)~1 p(x6)~1 p(x7)~1 p(x8)~0 p(x9)~0 p(x10)~0

Maximum Likelihood

We use maximum likelihood to estimate the parameters.

$$\ell(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)).$$

This *likelihood* gives the probability of the observed zeros and ones in the data. We pick β_0 and β_1 to maximize the likelihood of the observed data.

Most statistical packages can fit linear logistic regression models by maximum likelihood. In `R` we use the `glm` function.

	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

We usually don't care so much about the P-value for the intercept. The intercept's largely to do with the preponderance of 0's and 1's in the data set. And so that's of less importance. That's just inherent in the data set. It's the slope that's really important.
ref: 4_sigmoid_function_Kaggle.pdf

Making Predictions

What is our estimated probability of `default` for someone with a balance of \$1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

Making Predictions

What is our estimated probability of `default` for someone with a balance of \$1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

With a balance of \$2000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586$$

Lets do it again, using `student` as the predictor.

	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

Lets do it again, using `student` as the predictor.

	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

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

4.3 Multivariate Logistic Regression

https://youtu.be/MpX8rVv_u4E

Logistic regression with several variables

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

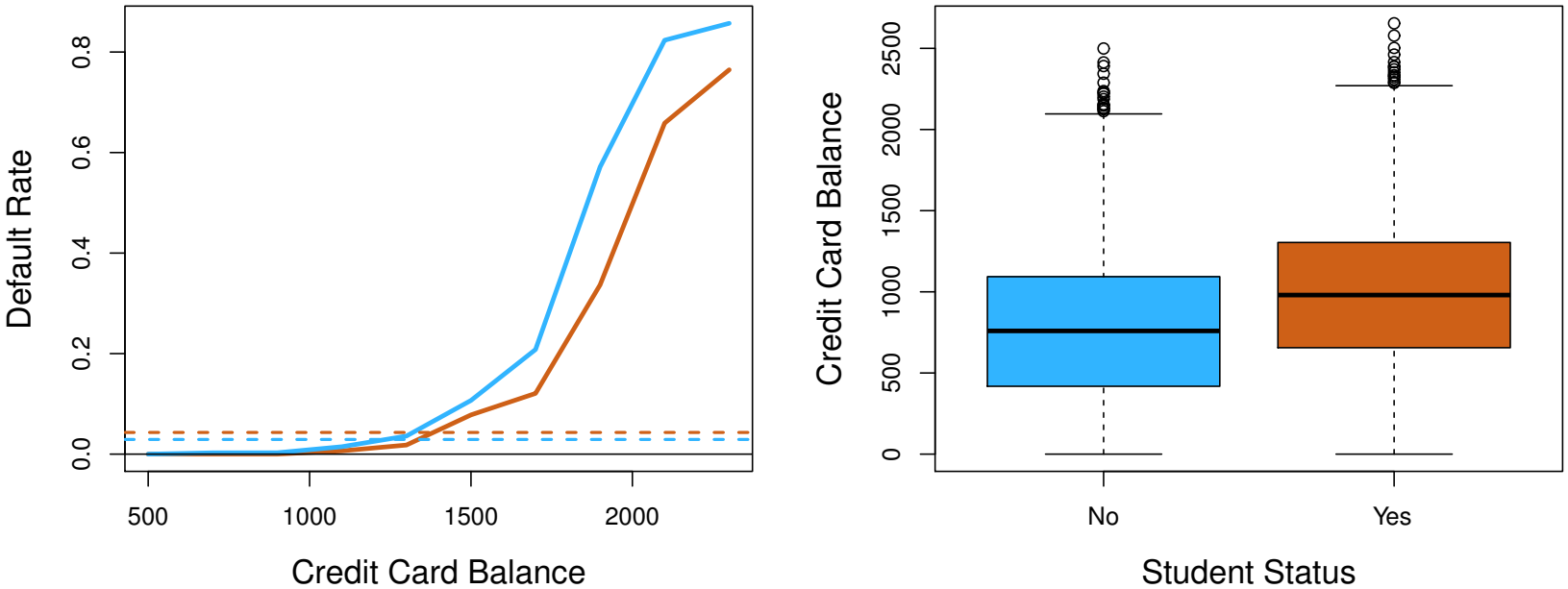
balance and student are significant, Income is not significant.

	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

Why is coefficient for `student` negative, while it was positive before?

last time we talked about in regression models how difficult it is to interpret coefficients in a multiple regression model, because the correlations between the variables can affect the signs. so we are going to see the role of correlations in the variables.

Confounding



- Students tend to have higher balances than non-students, so their marginal default rate is higher than for non-students.
- But for each level of balance, students default less than non-students.
- Multiple logistic regression can tease this out.

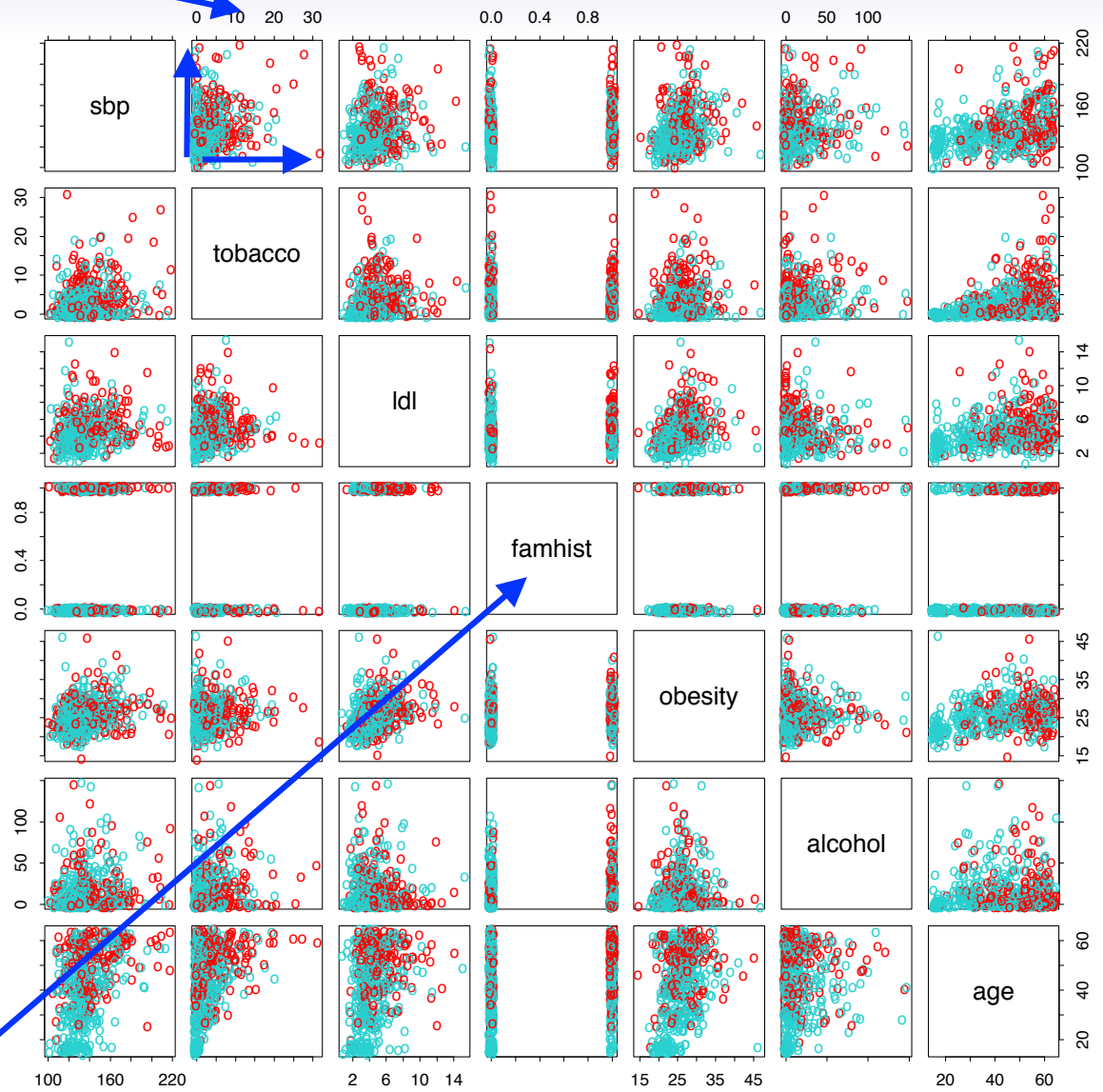
Example: South African Heart Disease

A study that compares patients who have a disease or outcome of interest (cases) with patients who do not have the disease or outcome (controls)

heart attach

- 160 cases of MI (myocardial infarction) and 302 controls (all male in age range 15-64), from Western Cape, South Africa in early 80s.
- Overall prevalence very high in this region: 5.1%.
- Measurements on seven predictors (risk factors), shown in scatterplot matrix.
- Goal is to identify relative strengths and directions of risk factors.
- This was part of an intervention study aimed at educating the public on healthier diets.

look at the top plot, for example, if you were high in tobacco usage and your systolic blood pressure is high, you tend to be a brown point (who tended to have heart attacks.)



Scatterplot matrix of the *South African Heart Disease* data. The response is color coded — The cases (MI) are red, the controls turquoise. `famhist` is a binary variable, with 1 indicating family history of MI.

famhist=family history =categorical variable.
It turns out to be an important risk factor,
If you have a family history of heart
disease, the risk is high.

response is CHD, which is the name of the response variable.

dot means all the other variables in the data frame (in this case = heart)

~: modeled as

the family is binomial, which just tells to fit the logistic regression model.

```
> heartfit<-glm(chd~.,data=heart,family=binomial)
> summary(heartfit)

Call:
glm(formula = chd ~ ., family = binomial, data = heart)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -4.1295997  0.9641558  -4.283 1.84e-05 ***
sbp              0.0057607  0.0056326   1.023  0.30643
tobacco          0.0795256  0.0262150   3.034  0.00242 **
ldl              0.1847793  0.0574115   3.219  0.00129 **
famhistPresent   0.9391855  0.2248691   4.177 2.96e-05 ***
obesity         -0.0345434  0.0291053  -1.187  0.23529
alcohol          0.0006065  0.0044550   0.136  0.89171
age              0.0425412  0.0101749   4.181 2.90e-05 ***

(Dispersion parameter for binomial family taken to be 1)
```

Tobacco usage is significant.

Low density lipoproteins (ldl), that's a bad cholesterol measure, that's significant.

Family history, very significant.

age is significant. We know the risk of heart disease goes up with age.

large P value doesn't mean its insignificant for multiple variables! correlations play a role!

```
    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 483.17  on 454  degrees of freedom
AIC: 499.17
```

obesity and alcohol usage are not significant ??? this is a case of having correlated variables. If we look in the previous plot, you see that there is a lot of correlation between variables. So age and tobacco usage are correlated. Alcohol usage and LDL seem to be negatively correlated.So there's lots of correlations. And so, for example, we've got LDL is significant in the model. And once LDL is in the model, perhaps alcohol usage is not needed anymore.

4.4 Logistic Regression - Case-Control Sampling and Multiclass

https://youtu.be/GavRXXEHGqU

what this case-control sampling is ? The most obvious way to study the risk factors for hear disease would be to take a large group of people, maybe 1,000 or 100,000 people, follow them for maybe 20 years, record their risk factors, and see who gets heart disease and who doesn't after 20 years. Now that actually is a good way to do things, except it's very expensive and it takes a long time. You have to get a lot of people, and you have to wait for many years.
Case-control sampling is a lot more attractive. Because what you do is rather than taking people and following them forward in time, you sample people who you know have heart disease. You also get a comparison sample of people who do not have heart disease, the controls. and then you record their risk factors. So it's much cheaper, and it's much quicker to do. And that's why case-control sampling is a very commonly used technique in epidemiology.

case-control sampling is one of the favorite tools in epidemiology. Especially when you have a rare disease, you take all the cases you can find, and then you can just sample from the controls.

Case-control sampling and logistic regression

- In South African data, there are 160 cases, 302 controls — $\tilde{\pi} = 0.35$ are cases. Yet the prevalence of MI in this region is $\pi = 0.05$.

risk ~ 160/(160+302)~0.346

- With case-control samples, we can estimate the regression parameters β_j accurately (if our model is correct); the constant term β_0 is incorrect.
- We can correct the estimated intercept by a simple transformation

here's the logit transformation of the true probablity

$$\hat{\beta}_0^* = \hat{\beta}_0 + \log \frac{\pi}{1-\pi} - \log \frac{\tilde{\pi}}{1-\tilde{\pi}}$$

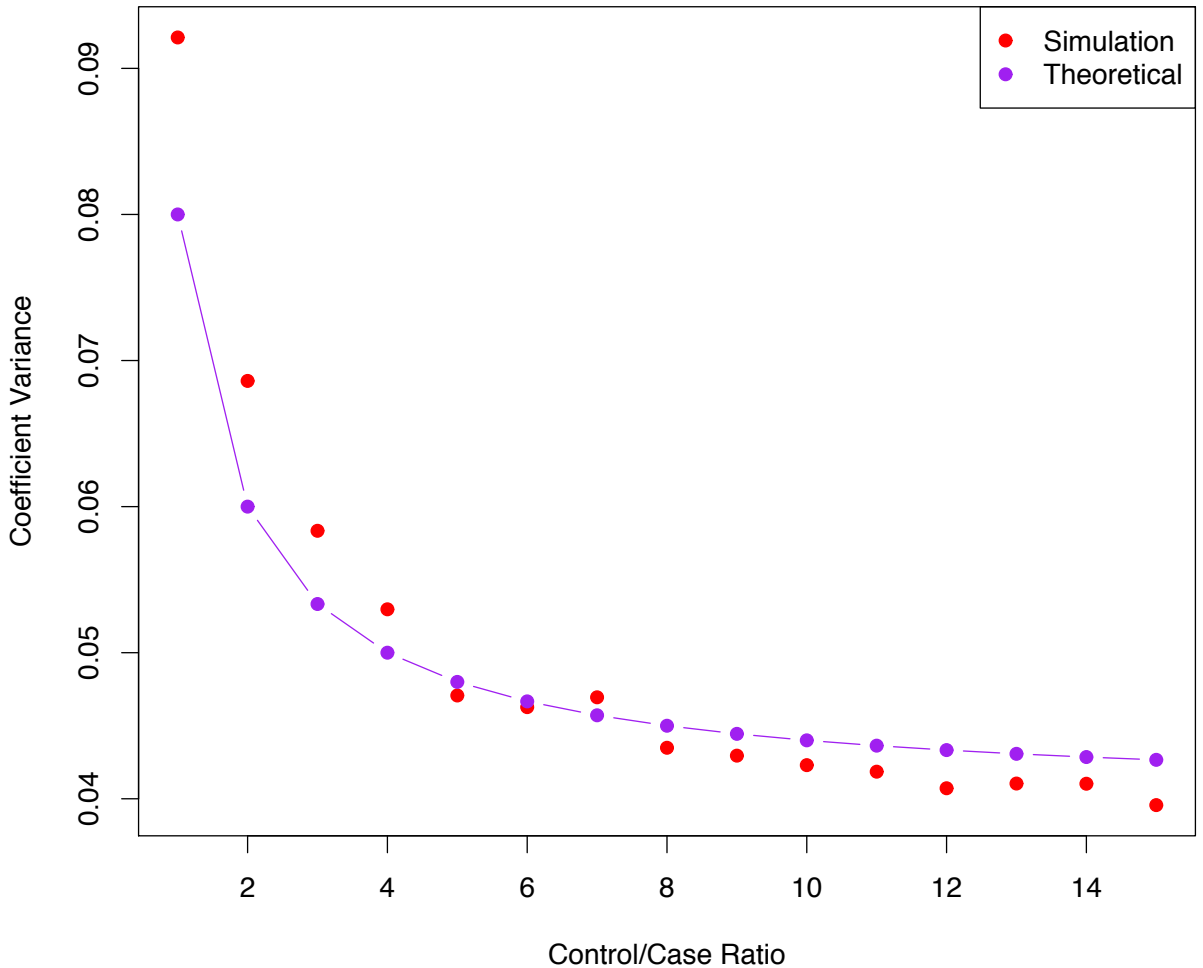
logit transformation of the prior probability or the prior apparent probability.

- Often cases are rare and we take them all; up to five times that number of controls is sufficient. See next frame

for the logistic regression model, you can estimate the regression parameters of interest (coefficients of the x's) accurately if the model's correct. But the constant term will be incorrect. Then you can just go ahead and correct the constant by a simple transformation.

in many modern data sets, we'll have very imbalanced situations, for example, if you're modelling the click-through rate on an ad on a web page, the probability of someone clicking is less than 1%, which means if you just take a random sample of subjects who've been exposed to ads, you're going to get very few 1's and a huge amount of 0's. So do we need to use all of that 0, 1 data to fit the models? Well, no! You can take a sample of the controls. This picture over here just gives an indication of the trade-off.

Diminishing returns in unbalanced binary data



Sampling more controls than cases reduces the variance of the parameter estimates. But after a ratio of about 5 to 1 the variance reduction flattens out.

So if you've got a very sparse situation, sample about 5 or 6 controls for every case, and now you can work with a much more manageable data set.

Logistic regression with more than two classes

So far we have discussed logistic regression with two classes. It is easily generalized to more than two classes. One version (used in the R package `glmnet`) has the symmetric form

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+...+\beta_{pk}X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_{1\ell}X_1+...+\beta_{p\ell}X_p}}$$

Here there is a linear function for *each* class.

K classes (K>2)

each class has its own linear model

Logistic regression with more than two classes

So far we have discussed logistic regression with two classes. It is easily generalized to more than two classes. One version (used in the R package `glmnet`) has the symmetric form

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\ldots+\beta_{pk}X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_{1\ell}X_1+\ldots+\beta_{p\ell}X_p}}$$

Here there is a linear function for *each* class.
(The *mathier* students will recognize that some cancellation is possible, and only $K-1$ linear functions are needed as in 2-class logistic regression.)

Logistic regression with more than two classes

So far we have discussed logistic regression with two classes. It is easily generalized to more than two classes. One version (used in the R package `glmnet`) has the symmetric form

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\ldots+\beta_{pk}X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_{1\ell}X_1+\ldots+\beta_{p\ell}X_p}}$$

Here there is a linear function for *each* class.
(The *mathier* students will recognize that some cancellation is possible, and only $K-1$ linear functions are needed as in 2-class logistic regression.)

Multiclass logistic regression is also referred to as *multinomial regression*.

4.5 Discriminant Analysis

https://youtu.be/RfrGiG1Hm3M

https://www.youtube.com/watch?v=azXCzI57Yfc
https://statquest.org/2016/07/10/statquest-linear-discriminant-analysis-lda-clearly-explained/
Linear discrimination analysis (LDA) is like PCA, but it focus on maximizing the seperatiblity among know categories.

Discriminant Analysis

Here the approach is to model the distribution of X in each of the classes separately, and then use *Bayes theorem* to flip things around and obtain $\Pr(Y|X)$.

When we use normal (Gaussian) distributions for each class, this leads to linear or quadratic discriminant analysis.

However, this approach is quite general, and other distributions can be used as well. We will focus on normal distributions.

Bayes theorem for classification

Thomas Bayes was a famous mathematician whose name represents a big subfield of statistical and probabilistic modeling. Here we focus on a simple result, known as Bayes theorem:

probability of hypothesis conditional on some evidence

probability of evidence conditional on hypothesis

prior probability of hypothesis

$$\Pr(Y = k|X = x) = \frac{\Pr(X = x|Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$

probability of y= K given x

prior probability of evidence

https://www.khanacademy.org/partner-content/wi-phi/wiphi-critical-thinking/wiphi-fundamentals/v/bayes-theorem

ex: you suspect that you get sick k, because of the symptom of x you have, the probability of getting sick k given symptom x = ?
=>
(1) probability of getting symptom x if you are sick k = 95 %
(2) probability of getting sick k = 0.01%
(3) probability of getting symptom x = 1 %

the probability of getting sick Y given symptom x = P(klx) = 0.95 * 0.0001 / 0.01 = 0.0095

Bayes theorem for classification

Thomas Bayes was a famous mathematician whose name represents a big subfield of statistical and probabilistic modeling. Here we focus on a simple result, known as Bayes theorem:

$$\Pr(Y = k|X = x) = \frac{\Pr(X = x|Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$

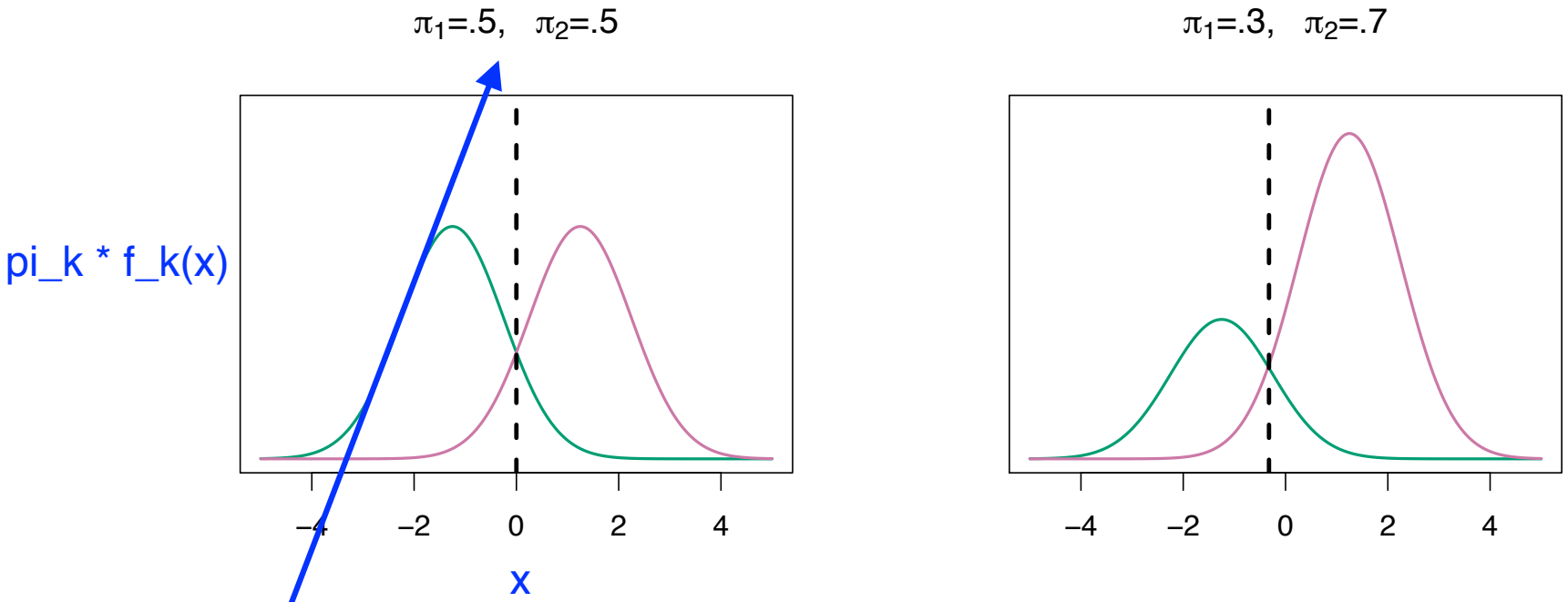
One writes this slightly differently for discriminant analysis:

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

the marginal probability f of x is this, summing over all classes.

- $f_k(x) = \Pr(X = x|Y = k)$ is the *density* for X in class k. Here we will use normal densities for these, separately in each class.
- $\pi_k = \Pr(Y = k)$ is the marginal or *prior* probability for class k.

Classify to the highest density

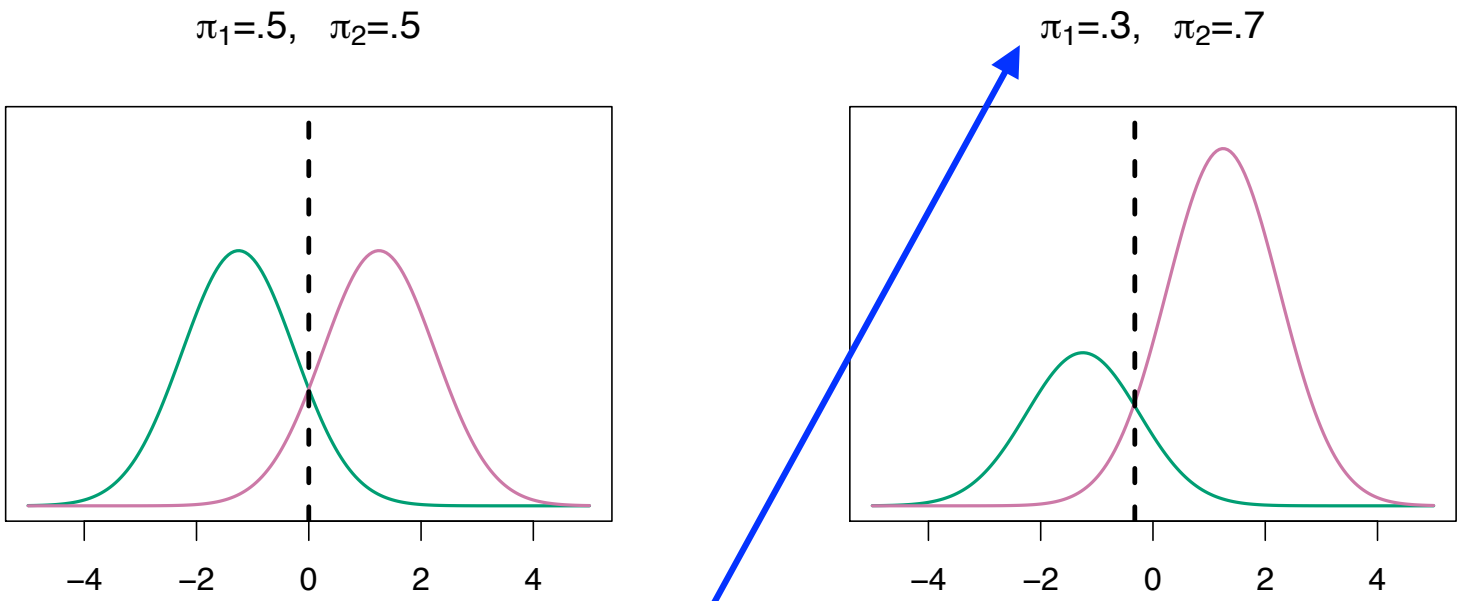


remember in the previous slide, the probability was essentially proportional to pi_k *f_k(x), in this case, the pi's are the same for both. So it's really to do with which density is the highest.

We classify a new point according to which density is highest.

And you can see that the decision boundary, or the vertical dash line, is at zero. And so anything to the left of zero we classify as green. And anything to the right we'd classify as purple.

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 $\pi_k f_k(x)$. On the right, we favor the pink class — the decision boundary has shifted to the left.

Why discriminant analysis?

In fact, if you've go a feature that separates the classes perfectly, the coefficients go off to infinity. So it really doesn't do well there. Logistic aggression was developed in largely the biological and medical fields where you never found such strong predictors. Now, you can do things to make logistic regression better behave. But it turns out linear discriminant analysis doesn't suffer from this problem and is better behaved in those situations.

- 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 sample size is small
  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. ref: 4_sigmoid_function_Kaggle.pdf
- Linear discriminant analysis is popular when we have more than two response classes, because it also provides low-dimensional views of the data.

https://stats.stackexchange.com/questions/254124/why-does-logistic-regression-become-unstable-when-classes-are-well-separated/254205
….. It corresponds to an intercept of negative infinity and a slope of infinity. …..
(ref: 4_sigmoid_function_Kaggle.pdf)

https://stats.stackexchange.com/questions/224863/understanding-complete-separation-for-logistic-regression/224864#224864

https://stats.stackexchange.com/questions/239928/is-there-any-intuitive-explanation-of-why-logistic-regression-will-not-work-for logistic regression with/without regularization

4.6 Gaussian Discriminant Analysis - One Variable

https://youtu.be/QG0pVJXT6EU

Linear Discriminant Analysis when $p = 1$

The Gaussian density has the form

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

Here μ_k is the mean, and σ_k^2 the variance (in class k). We will assume that all the $\sigma_k = \sigma$ are the same.

assume sigma_k, is sigma, the same in each of the classes.
(It turns out this is an important convenience. And it's going to determine whether the discriminant function that we get, the discriminant analysis, gives us linear functions or quadratic functions.)

Linear Discriminant Analysis when $p = 1$

The Gaussian density has the form

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

Here μ_k is the mean, and σ_k^2 the variance (in class k). We will assume that all the $\sigma_k = \sigma$ are the same.

Plugging this into Bayes formula, we get a rather complex expression for $p_k(x) = \Pr(Y = k|X = x)$:

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

Happily, there are simplifications and cancellations.

whenever you see exponentials the first thing you want to do is take the logs.

Discriminant functions

To classify at the value $X = x$, we need to see which of the $p_k(x)$ is largest. Taking logs, and discarding terms that do not depend on k, we see that this is equivalent to assigning x to the class with the largest *discriminant score*:

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Note that $\delta_k(x)$ is a *linear* function of x.

ln(p1)= ln(pi1) - 1/2((x-mu1)/sig)^2 - ln((..)+(..)+..)
ln(p2)= ln(pi2) - 1/2((x-mu2)/sig)^2 - ln((..)+(..)+..)
=> compare p1 or p2, which one is larger ?
=> equivalent to compare this 3 terms, all the other terms can be ignored because they are the same in ln(p1) and ln(p2)

Discriminant functions

To classify at the value $X = x$, we need to see which of the $p_k(x)$ is largest. Taking logs, and discarding terms that do not depend on k, we see that this is equivalent to assigning x to the class with the largest *discriminant score*:

the idea of the discriminant function is, you compute one of these for each of the classes, and then you classify it to the class for which it's largest.

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

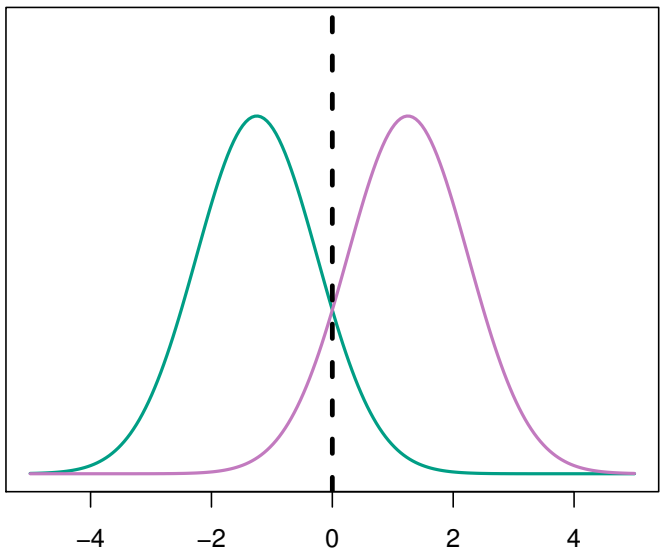
Note that $\delta_k(x)$ is a *linear* function of x.

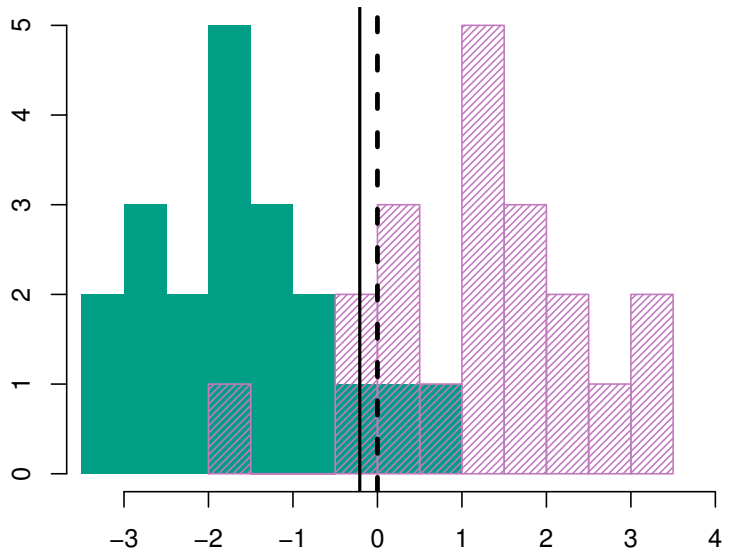
If there are $K = 2$ classes and $\pi_1 = \pi_2 = 0.5$, then one can see that the *decision boundary* is at

delta1=delta2 => x =

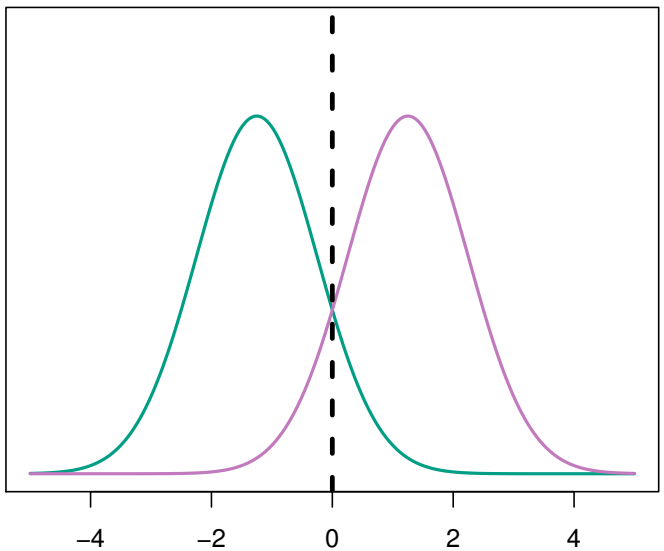
$$x = \frac{\mu_1 + \mu_2}{2}.$$

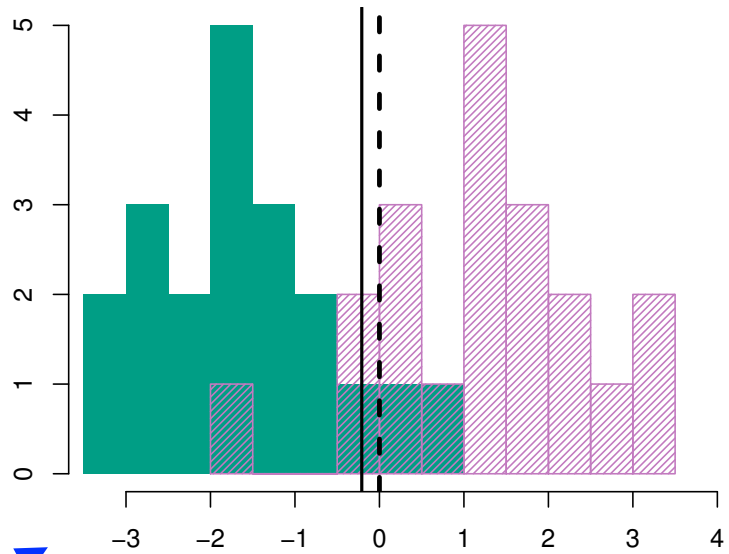
(See if you can show this)





Example with $\mu_1 = -1.5$, $\mu_2 = 1.5$, $\pi_1 = \pi_2 = 0.5$, and $\sigma^2 = 1$.





Example with $\mu_1 = -1.5$, $\mu_2 = 1.5$, $\pi_1 = \pi_2 = 0.5$, and $\sigma^2 = 1$.

Typically we don't know these parameters; we just have the training data. In that case we simply estimate the parameters and plug them into the rule.

Estimating the parameters

$$\hat{\pi}_k = \frac{n_k}{n}$$

n_k=number in class k

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:\, y_i = k} x_i$$

So this will just sum those xi's that are in class k.

pooled variance estimate.

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:\, y_i = k} (x_i - \hat{\mu}_k)^2$$

So this is just like a weight on each of those variances.

$$\text{or} = \sum_{k=1}^{K} \frac{n_k - 1}{n - K} \cdot \hat{\sigma}_k^2$$

we average them using this formula over here.

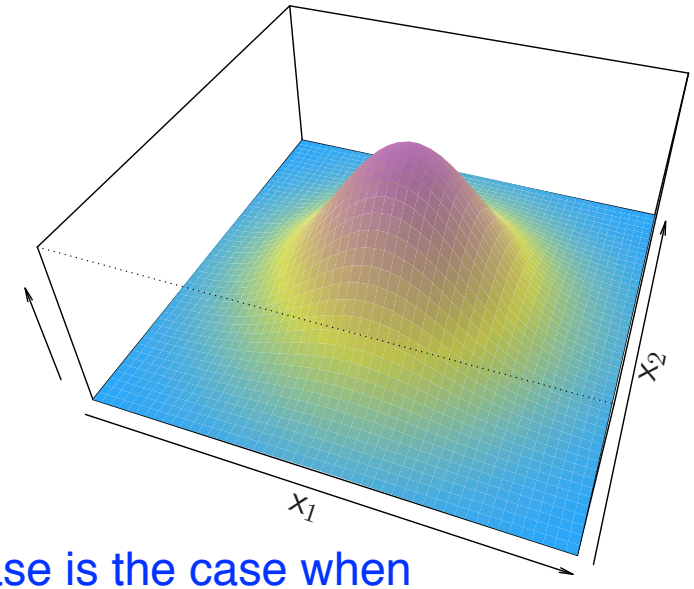
estimate the sample variance separately in each of the classes.

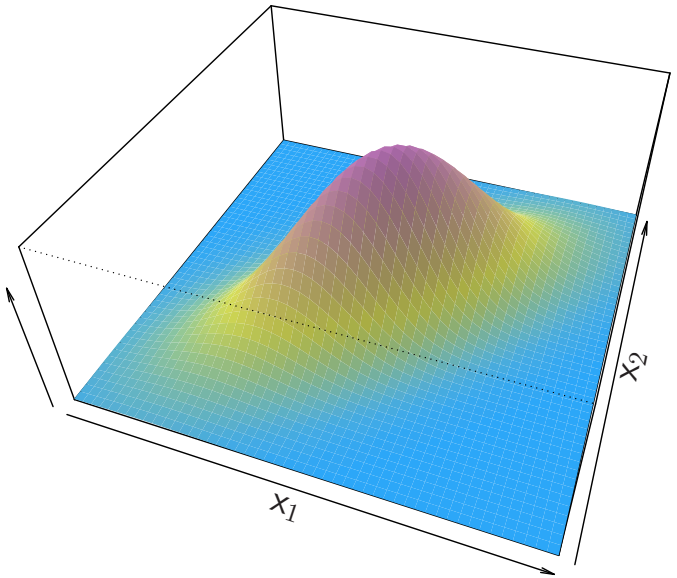
where $\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i:\, y_i = k} (x_i - \hat{\mu}_k)^2$ is the usual formula for the estimated variance in the kth class.

4.7 Gaussian Discriminant Analysis - Many Variables

https://youtu.be/X4VDZDp2vqw

Linear Discriminant Analysis when $p > 1$



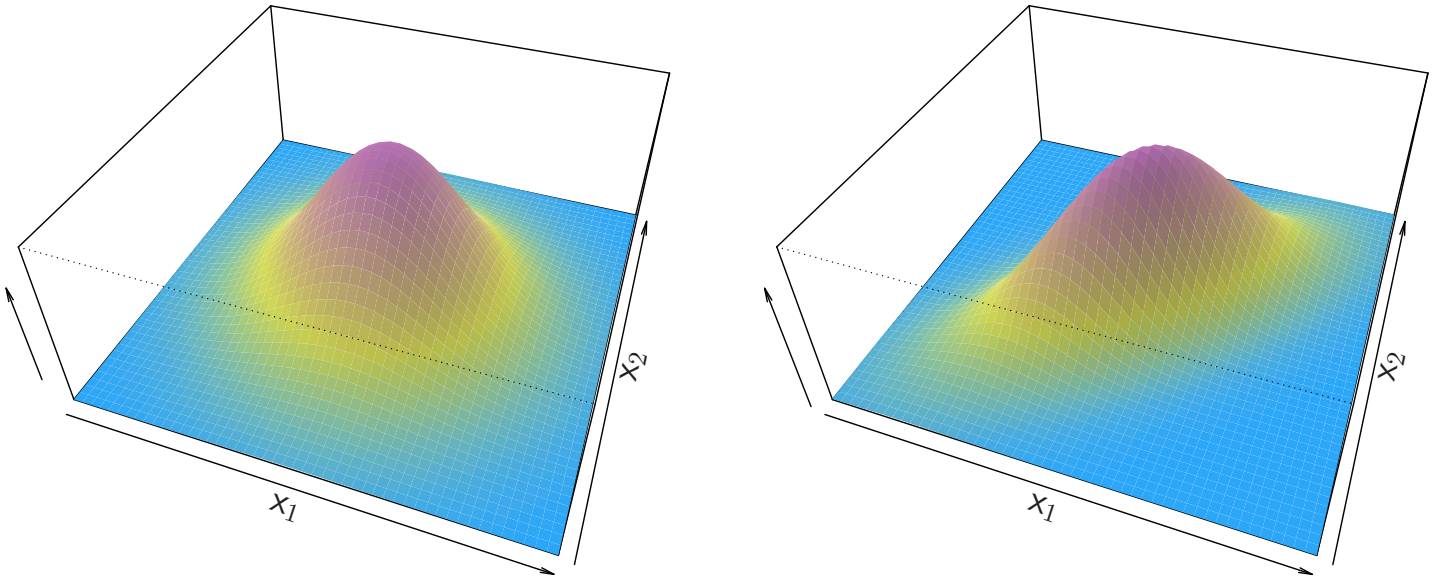


left hand case is the case when two variables are uncorrelated, so it's just really like a bell.

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

And the right hand case is when there's correlation, so it's like a stretched bell.

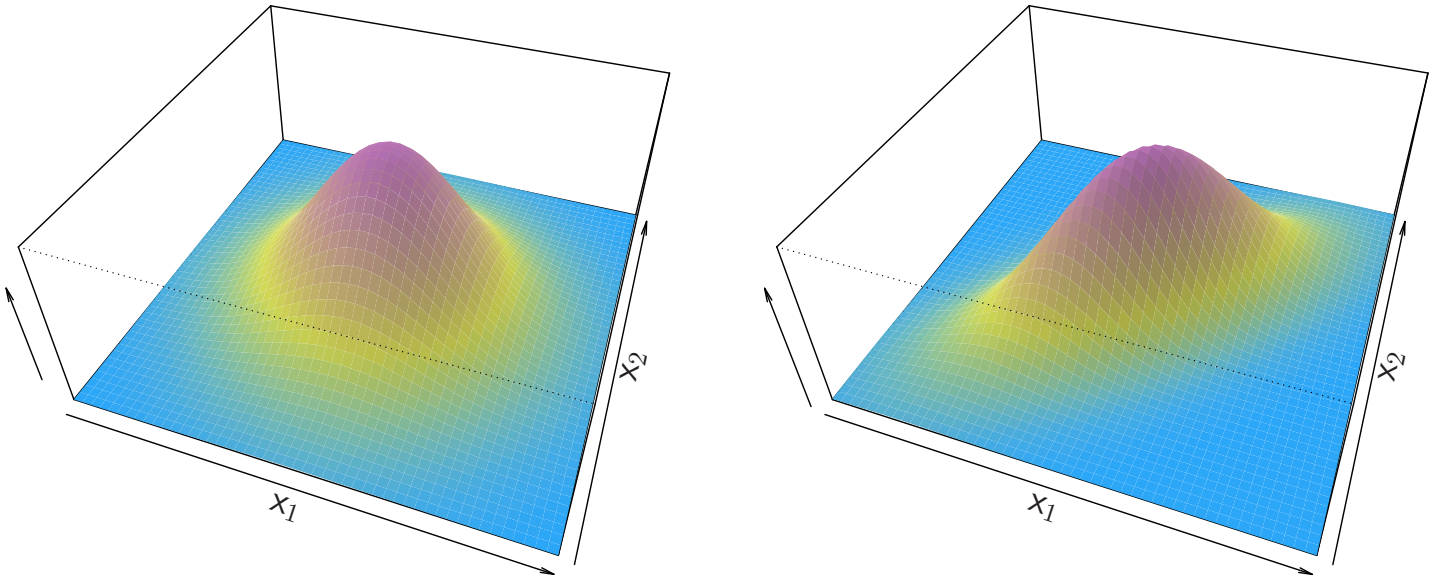
Linear Discriminant Analysis when $p > 1$



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

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

Linear Discriminant Analysis when $p > 1$



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

covariance matrix

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

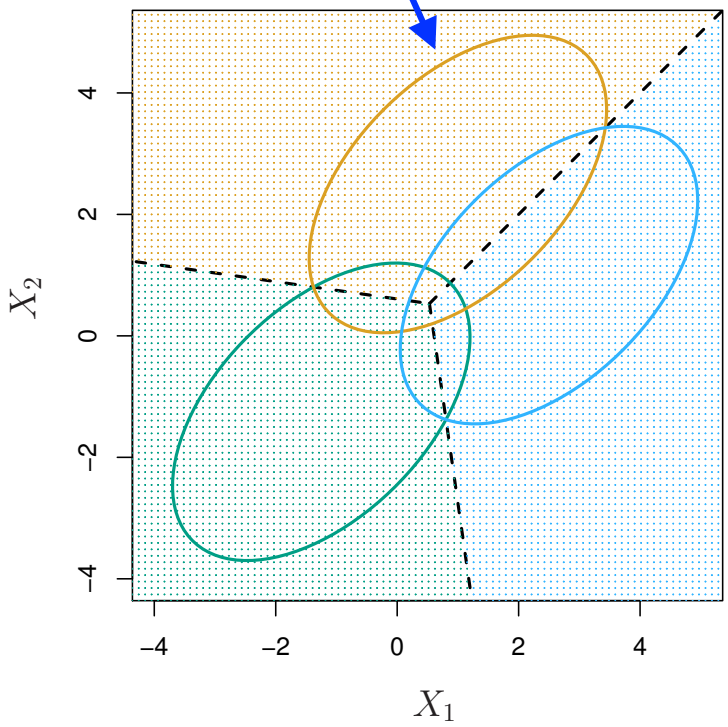
the idea of the discriminant function is, you compute one of these for each of the classes, and then you classify it to the class for which it's largest.

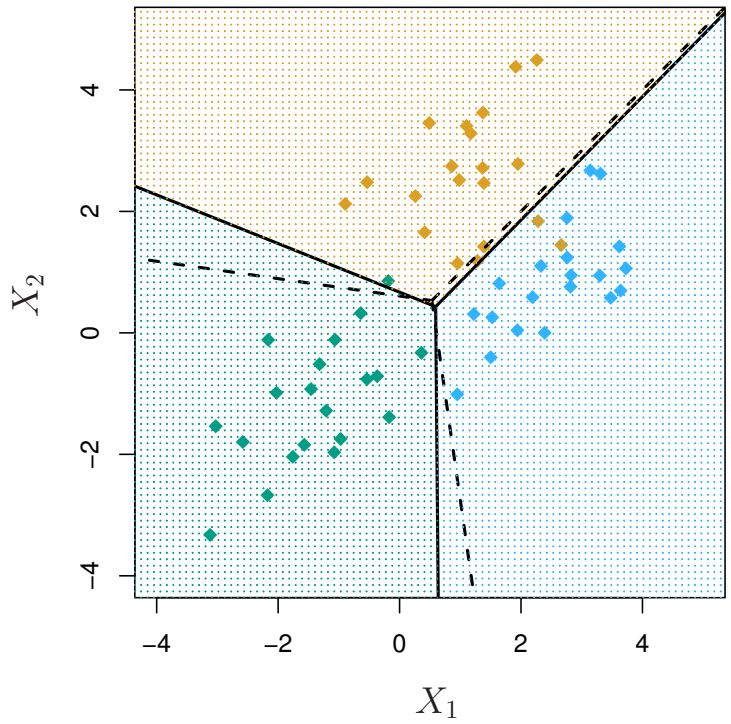
Despite its complex form,

$\delta_k(x) = c_{k0} + c_{k1}x_1 + c_{k2}x_2 + \ldots + c_{kp}x_p$ — a linear function.

Illustration: $p = 2$ and $K = 3$ classes

And we show the contour of a particular level of probability





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

The dashed lines are known as the *Bayes decision boundaries*.

true decision boundaries

Were they known, they would yield the fewest misclassification errors, among all possible classifiers.

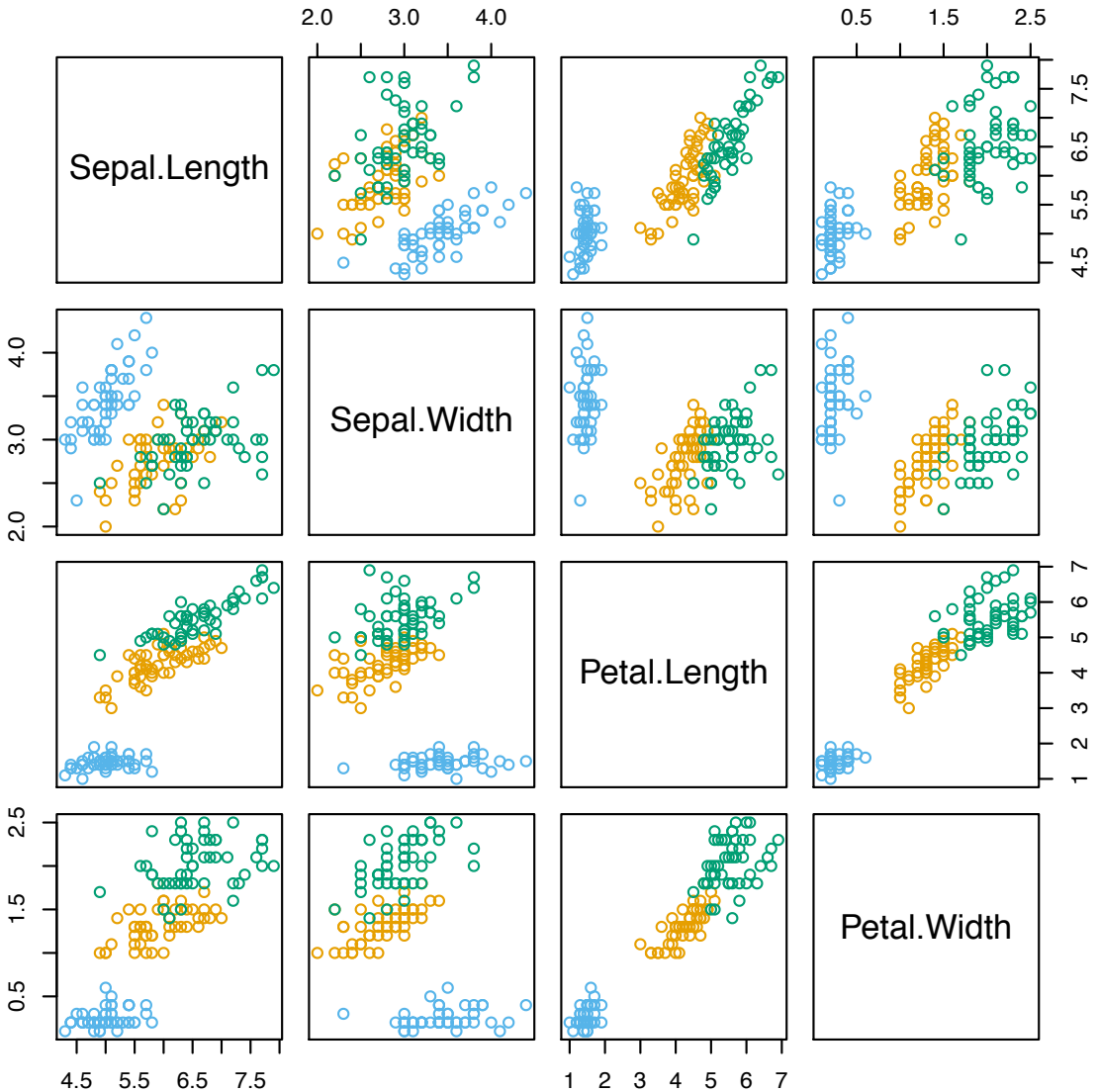
if you knew the true densities (in this case, the Gaussian) you get the exact decision boundaries (Bayes decision boundaries, dashed lines), which is the true decision boundaries, but we don't. We estimate the parameters using the formula and we get the solid lines, it is very close to the true boundary (dashed lines).

Fisher's Iris Data

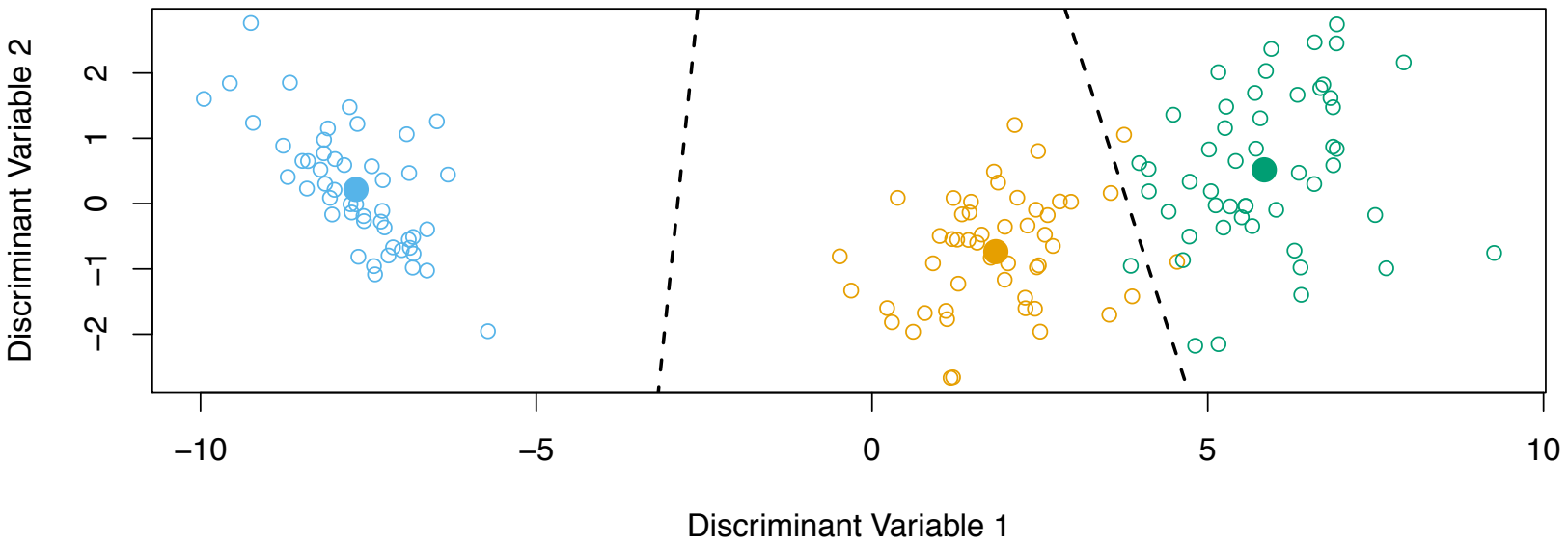
4 variables
3 species
50 samples/class

- Setosa
- Versicolor
- Virginica

LDA classifies all but 3 of the 150 training samples correctly.



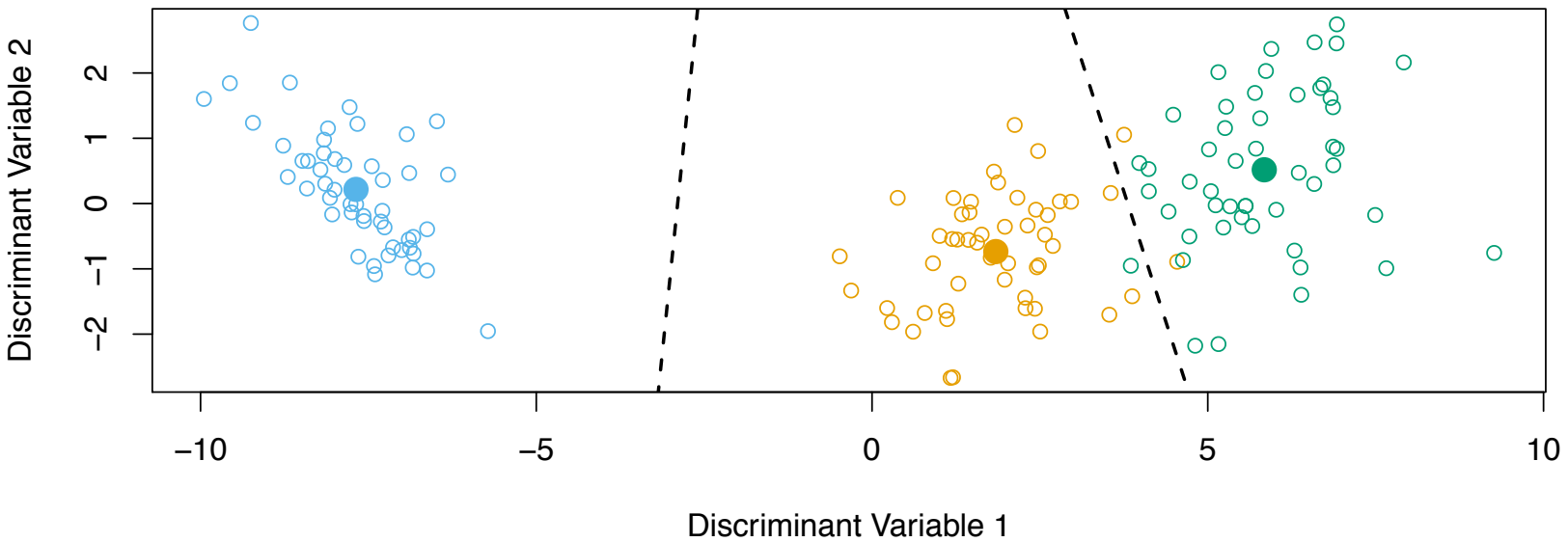
Fisher's Discriminant Plot



When there are K classes, linear discriminant analysis can be viewed exactly in a $K-1$ dimensional plot.
Why?

And it turns out these are linear combinations of the original variables. But they're special linear combinations. And when you plot the variables against these two, you see really good separation.

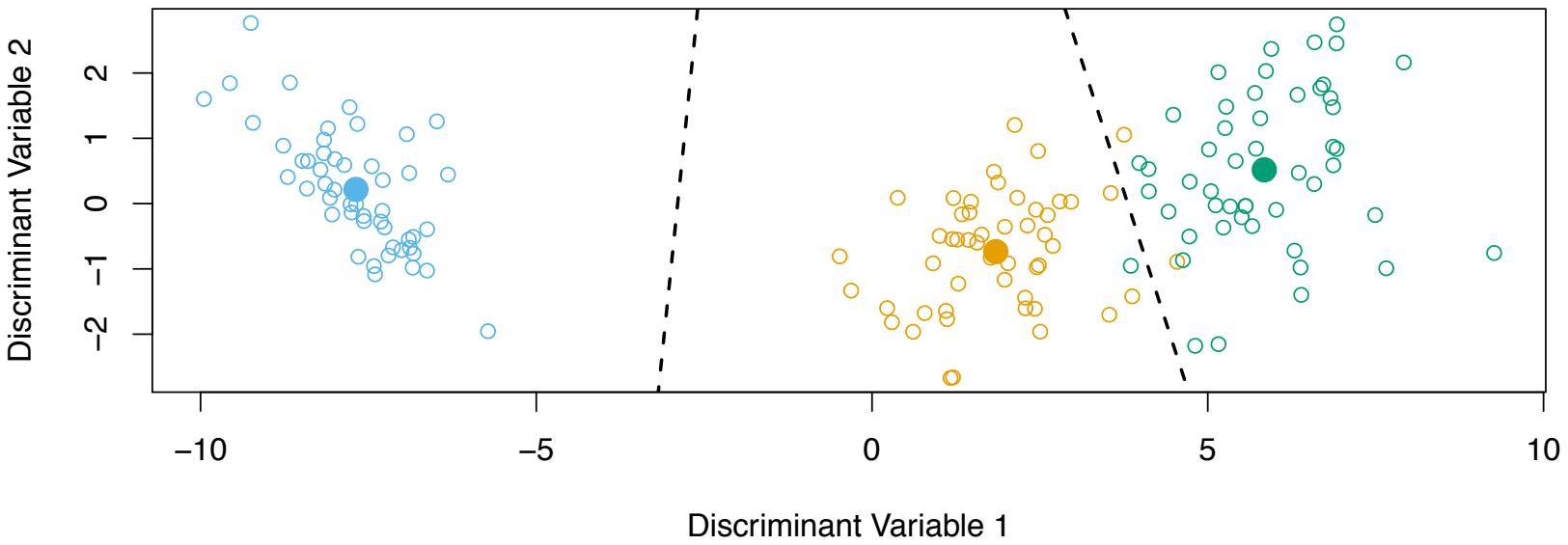
Fisher's Discriminant Plot



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

Discriminant analysis was a very attractive method, But imagine we had 4,000 features. Then we had to plug in an estimate of the covariance matrix of size 4,000 by 4,000. And we just can't carry out discriminant analysis without other modifications if the number of variables are very large.

Fisher's Discriminant Plot



When there are K classes, linear discriminant analysis can be viewed exactly in a $K-1$ dimensional plot.

Why? Because it essentially classifies to the closest centroid, and they span a $K-1$ dimensional plane.

Even when $K>3$, we can find the "best" 2-dimensional plane for visualizing the discriminant rule.

And when we have more than three classes, we can still find two-dimensional plots. But in that case, it doesn't capture all the information in the two-dimensional plot, but you can find the base two-dimensional plot for visualizing the discriminant rule. And that's another important reason why linear discriminant analysis is very popular for multi-class classification

From $\delta_k(x)$ to probabilities

Once we have estimates $\hat{\delta}_k(x)$, we can turn these into estimates for class probabilities:

$$\widehat{\Pr}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}.$$

So classifying to the largest $\hat{\delta}_k(x)$ amounts to classifying to the class for which $\widehat{\Pr}(Y = k|X = x)$ is largest.

From $\delta_k(x)$ to probabilities

Once we have estimates $\hat{\delta}_k(x)$, we can turn these into estimates for class probabilities:

$$\widehat{\Pr}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}.$$

So classifying to the largest $\hat{\delta}_k(x)$ amounts to classifying to the class for which $\widehat{\Pr}(Y = k|X = x)$ is largest.

When $K = 2$, we classify to class 2 if $\widehat{\Pr}(Y = 2|X = x) \geq 0.5$, else to class 1.

LDA on Credit Data

confusion matrix

		True Default Status		
		No	Yes	Total
Predicted	No	9644	252	9896
Default Status	Yes	23	81	104
	Total	9667	333	10000

$(23 + 252)/10000$ errors — a 2.75% misclassification rate!

Some caveats:

- This is *training* error, and we may be overfitting.

LDA on Credit Data

		True Default Status		
		No	Yes	Total
Predicted	No	9644	252	9896
Default Status	Yes	23	81	104
	Total	9667	333	10000

$(23 + 252)/10000$ errors — a 2.75% misclassification rate!

Some caveats:

- This is *training* error, and we may be overfitting. Not a big concern here since $n = 10000$ and $p = 2$!

LDA on Credit Data

		True Default Status		
		No	Yes	Total
Predicted	No	9644	252	9896
Default Status	Yes	23	81	104
	Total	9667	333	10000

$(23 + 252)/10000$ errors — a 2.75% misclassification rate!

Some caveats:

- This is *training* error, and we may be overfitting. Not a big concern here since $n = 10000$ and $p = 2$!
- If we classified to the prior — always to class `No` in this case — we would make 333/10000 errors, or only 3.33%.

	No	Yes	
No	9667	333	
Yes	0	0	
	9667	333	

its" Null rate"
And so you always bear in mind the null rate when getting excited about a misclassification error rate.

LDA on Credit Data

		True Default Status		
		No	Yes	Total
Predicted	No	9644	252	9896
Default Status	Yes	23	81	104
	Total	9667	333	10000

$(23 + 252)/10000$ errors — a 2.75% misclassification rate!

Some caveats:

- This is *training* error, and we may be overfitting. Not a big concern here since $n = 10000$ and $p = 2$!
- If we classified to the prior — always to class `No` in this case — we would make 333/10000 errors, or only 3.33%.
- Of the true `No`'s, we make $23/9667 = 0.2\%$ errors; of the true `Yes`'s, we make $252/333 = 75.7\%$ errors!

Types of errors

False positive rate: The fraction of negative examples that are classified as positive — 0.2% in example.

23/9667 = 0.23 %

False negative rate: The fraction of positive examples that are classified as negative — 75.7% in example.

252/333 = 75.6 %

We produced this table by classifying to class `Yes` if

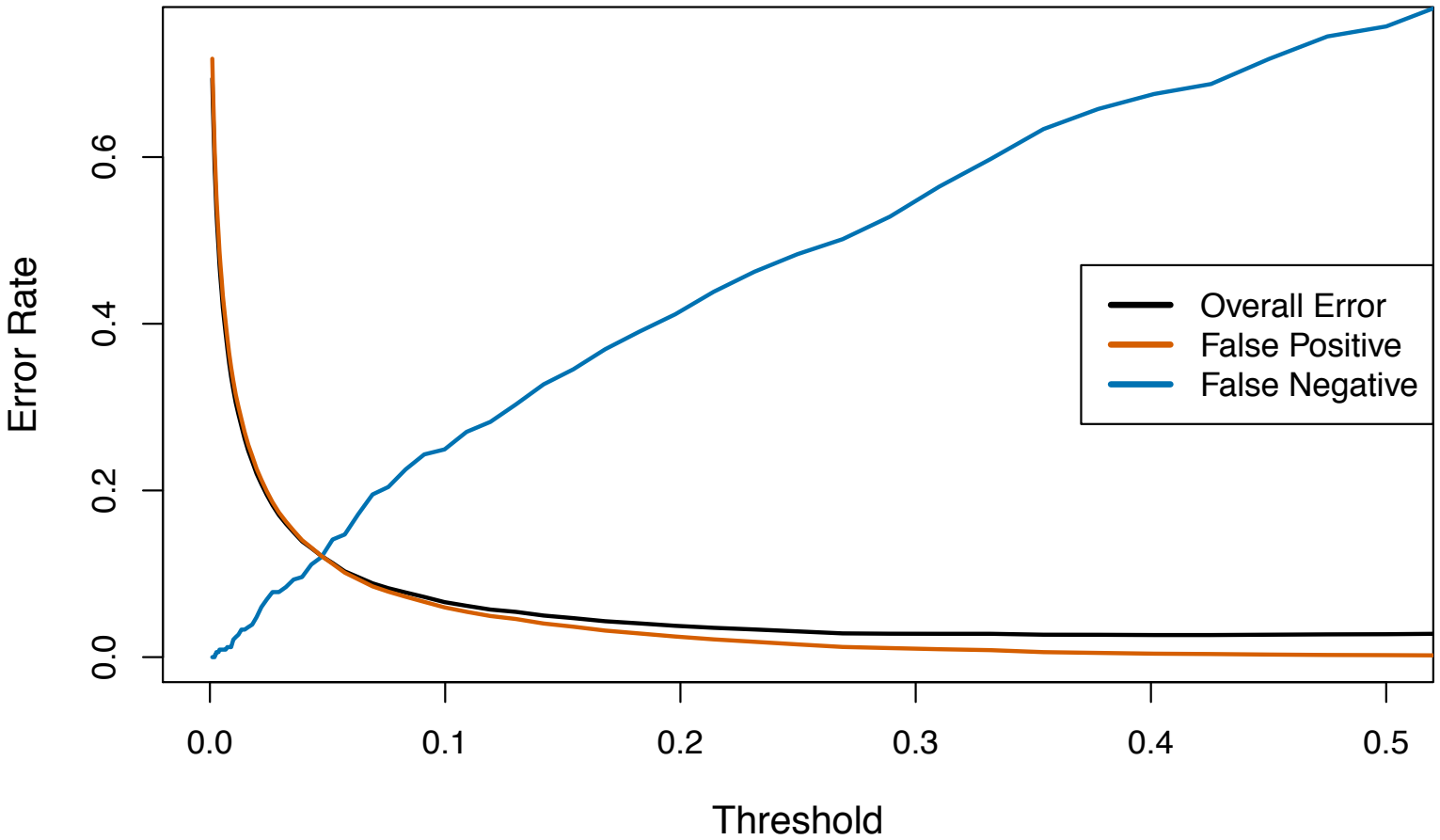
$$\widehat{\Pr}(\texttt{Default} = \texttt{Yes} | \texttt{Balance}, \texttt{Student}) \geq 0.5$$

We can change the two error rates by changing the threshold from 0.5 to some other value in $[0, 1]$:

$$\widehat{\Pr}(\texttt{Default} = \texttt{Yes} | \texttt{Balance}, \texttt{Student}) \geq \textit{threshold},$$

and vary *threshold*.

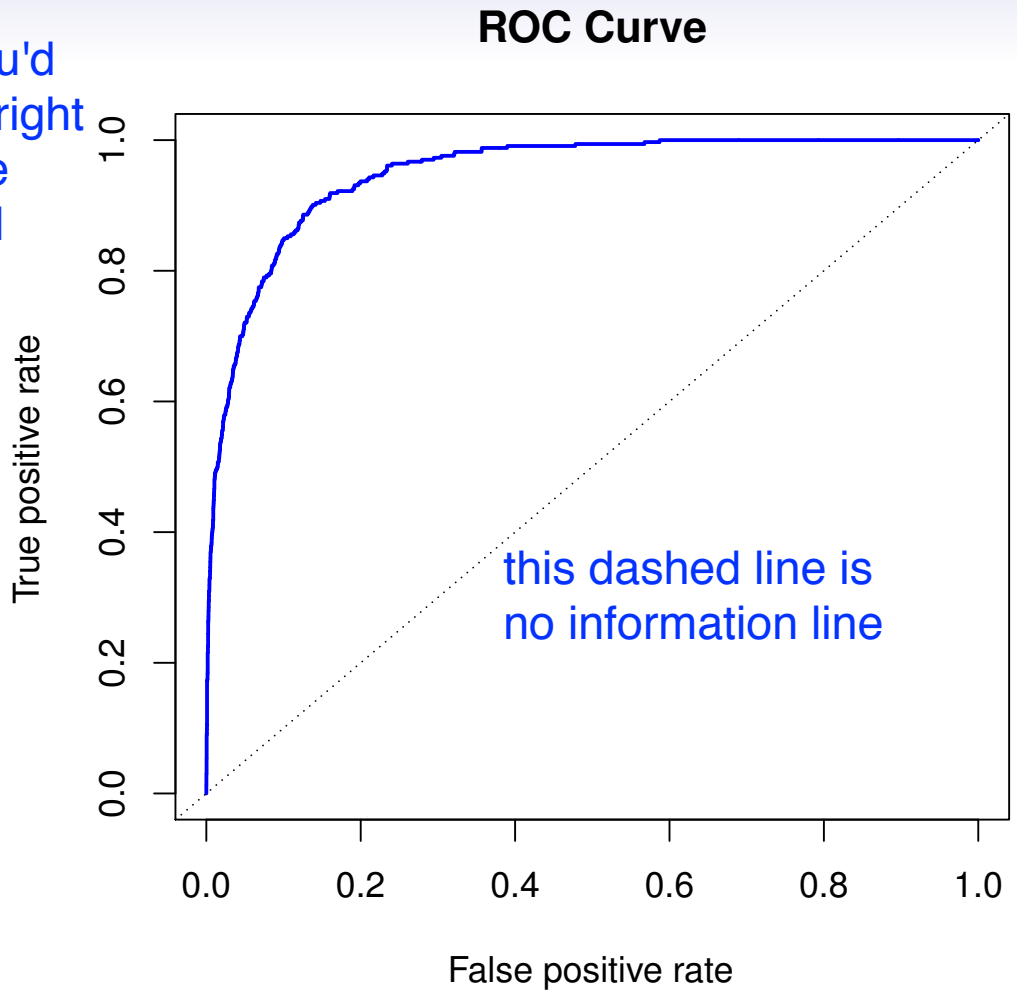
Varying the *threshold*



In order to reduce the false negative rate, we may want to reduce the threshold to 0.1 or less.

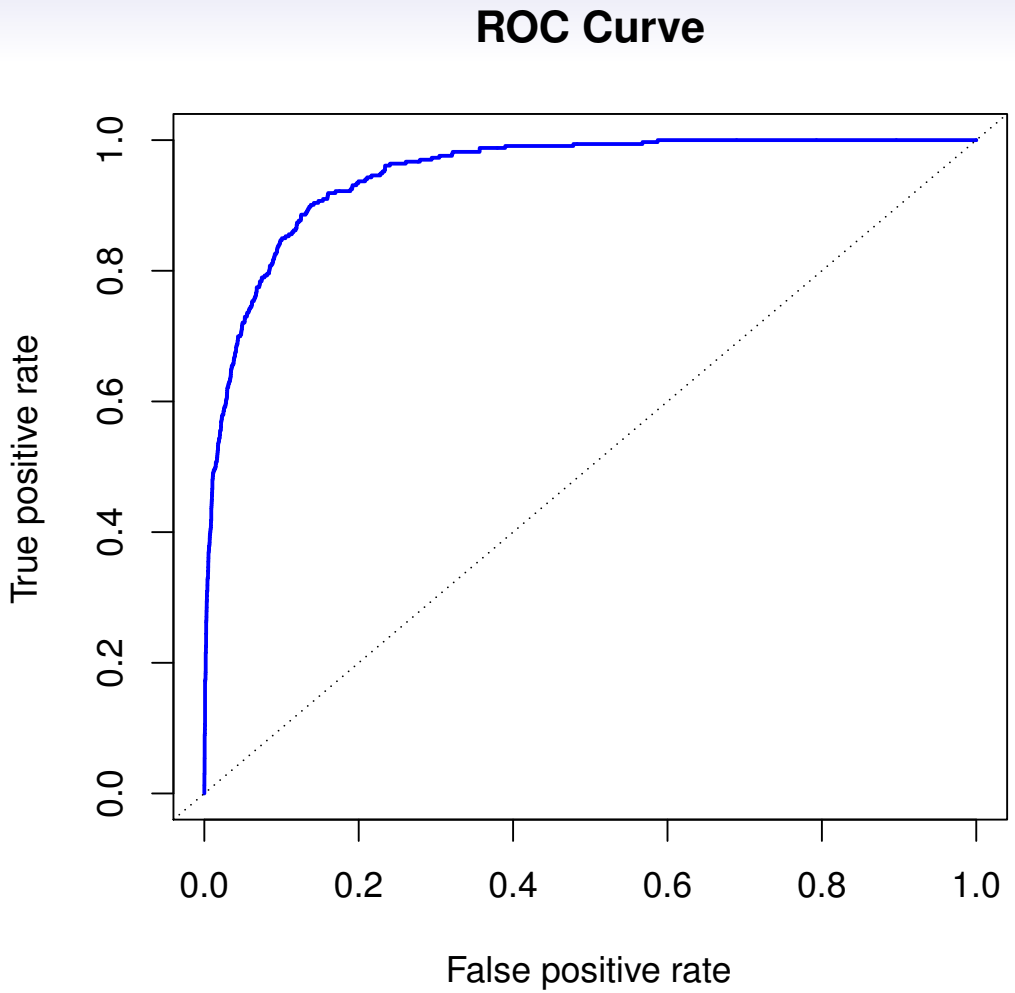
you can capture that change in threshold in what's known as an ROC curve.
And you can compare different classifiers by comparing the ROC curves.

what you'd like is you'd like this curve to be right up as far as possible into the top left hand corner



So what this shows is the two error rates, in this case, false positive rate and true positive rate, as we change the threshold.

The *ROC plot* displays both simultaneously.



The *ROC plot* displays both simultaneously.

Sometimes we use the *AUC* or *area under the curve* to summarize the overall performance. Higher *AUC* is good.

4.8 Quadratic Discriminant Analysis and Naive Bayes

https://youtu.be/6FiNGTYAOAA

Other forms of Discriminant Analysis

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

When $f_k(x)$ are Gaussian densities, with the same covariance matrix $\mathbf{\Sigma}$ in each class, this leads to linear discriminant analysis. By altering the forms for $f_k(x)$, we get different classifiers.

- With Gaussians but different $\mathbf{\Sigma}_k$ in each class, we get *quadratic discriminant analysis*.

Well, when the variances are different in each class, the quadratic terms don't cancel. And so now your discriminant functions are going to be quadratic functions of X.

Other forms of Discriminant Analysis

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

When $f_k(x)$ are Gaussian densities, with the same covariance matrix $\mathbf{\Sigma}$ in each class, this leads to linear discriminant analysis. By altering the forms for $f_k(x)$, we get different classifiers.

- With Gaussians but different $\mathbf{\Sigma}_k$ in each class, we get *quadratic discriminant analysis*.
- With $f_k(x) = \prod_{j=1}^{p} f_{jk}(x_j)$ (conditional independence model) in each class we get *naive Bayes*. For Gaussian this means the $\mathbf{\Sigma}_k$ are diagonal.

you can assume that in each class the density factors into a product of densities.

For linear discriminant analysis, this means that the covariances, sigma_k are diagonal. And instead of estimating the covariance matrix, if you've got P variables it's got P squared parameters. But if you assume that it's diagonal, then you need to estimate P parameters again.

LDS:
sigma_k = sigma
fk = Gaussian
(cpmpute cov matrix)

QDS:
sigma_k != sigma (different sigma_k)
fk = Gaussian
(cpmpute cov matrix)

Other forms of Discriminant Analysis

naive Bayes with Gaussian f:
sigma_k != sigma (different sigma_k)
fk = Gaussian + conditional independence => diagonal variance
(compute cov matrix, easier to compute)

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

When $f_k(x)$ are Gaussian densities, with the same covariance matrix $\mathbf{\Sigma}$ in each class, this leads to linear discriminant analysis. By altering the forms for $f_k(x)$, we get different classifiers.

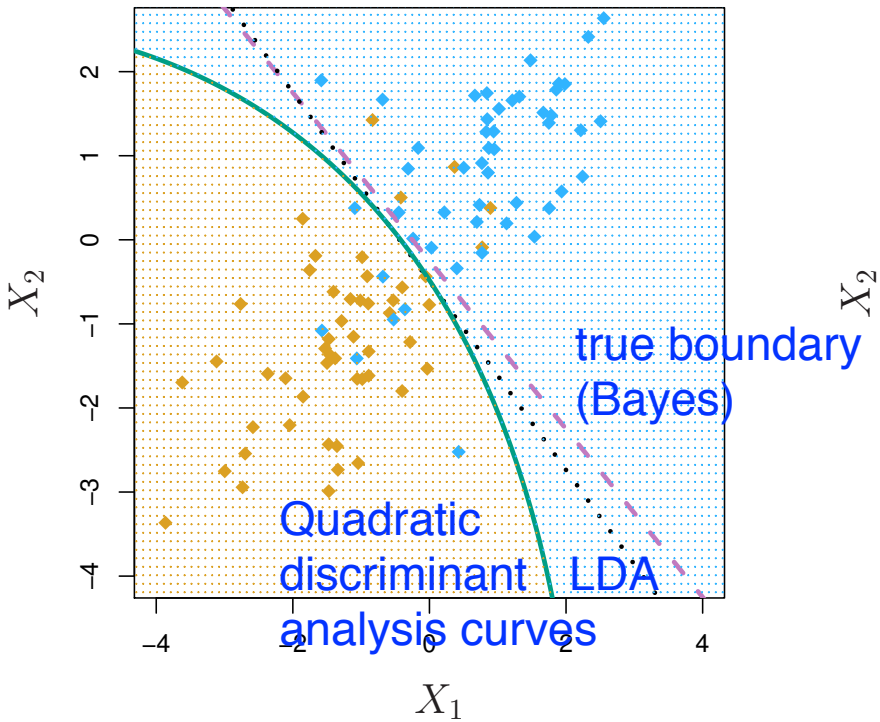
- With Gaussians but different $\mathbf{\Sigma}_k$ in each class, we get *quadratic discriminant analysis*.
- With $f_k(x) = \prod_{j=1}^{p} f_{jk}(x_j)$ (conditional independence model) in each class we get *naive Bayes*. For Gaussian this means the $\mathbf{\Sigma}_k$ are diagonal.
- Many other forms, by proposing specific density models for $f_k(x)$, including nonparametric approaches.

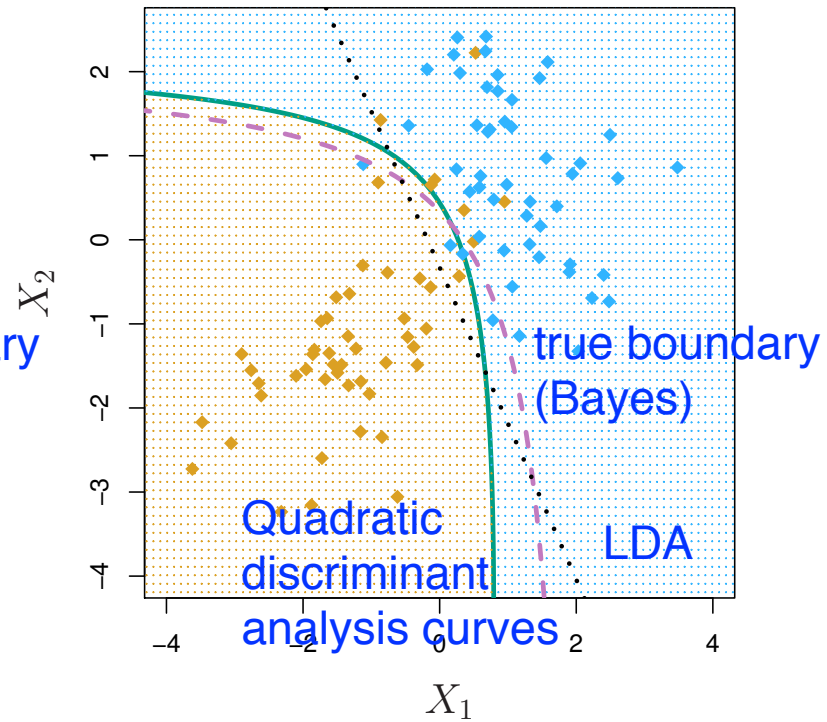
the assumption is wrong-- this naive Bayes classifier is actually very useful in high-dimensional problems. And it's one actually we'll return to later in different forms.

So here we have it.
Quadratic discriminant analysis uses a different covariance matrix for each class.

Quadratic Discriminant Analysis

(QDA)





$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}_k^{-1}(x-\mu_k) + \log \pi_k - \frac{1}{2}\log|\Sigma_k|$$

Because the $\mathbf{\Sigma}_k$ are different, the quadratic terms matter.

Quadratic discriminant analysis is attractive if the number of variables is small. When the number of variables or features is large, you've got to estimate these big covariance matrices, and things can break down. And even for LDA it can break down. Here's where naive Bayes becomes attractive.

Naive Bayes

Assumes features are independent in each class.

Useful when p is large, and so multivariate methods like QDA and even LDA break down.

- Gaussian naive Bayes assumes each $\mathbf{\Sigma}_k$ is diagonal:

It makes a much stronger assumption.It assumes that the covariance in each of the classes, although different, are diagonal. And so that's much fewer parameters.

$$\begin{aligned} \delta_k(x) &\propto \log\left[\pi_k \prod_{j=1}^{p} f_{kj}(x_j)\right] \\ &= -\frac{1}{2}\sum_{j=1}^{p}\left[\frac{(x_j-\mu_{kj})^2}{\sigma_{kj}^2} + \log\sigma_{kj}^2\right] + \log\pi_k \end{aligned}$$

- can use for *mixed* feature vectors (qualitative and quantitative). If X_j is qualitative, replace $f_{kj}(x_j)$ with probability mass function (histogram) over discrete categories.

Despite strong assumptions, naive Bayes often produces good classification results.

even though it has strong assumptions, it often produces good classification results. Because, in classification, we're mainly concerned about which class has the highest probability, and not whether we got the probabilities exactly right.

Logistic Regression versus LDA

For a two-class problem, one can show that for LDA

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x_1 + \ldots + c_p x_p$$

So it has the same form as logistic regression.

The difference is in how the parameters are estimated.

Logistic Regression versus LDA

For a two-class problem, one can show that for LDA

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x_1 + \ldots + c_p x_p$$

So it has the same form as logistic regression.

The difference is in how the parameters are estimated.

- Logistic regression uses the conditional likelihood based on $\Pr(Y|X)$ (known as *discriminative learning*).

Logistic regression uses the conditional likelihood based on probability of Y given X. Remember, it was using the probabilities of a 1 or a 0 given X in each of the classes. And in machine learning, this is known as discriminative learning using the conditional distribution of Y given X.

Logistic Regression versus LDA

For a two-class problem, one can show that for LDA

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x_1 + \ldots + c_p x_p$$

So it has the same form as logistic regression.

The difference is in how the parameters are estimated.

- Logistic regression uses the conditional likelihood based on $\Pr(Y|X)$ (known as *discriminative learning*).
- LDA uses the full likelihood based on $\Pr(X, Y)$ (known as *generative learning*).

Discriminant analysis is estimating these parameters using the full likelihood. Because it's using the distribution of X's and Y's (whereas logistic regression was only using the distribution of Y's), And in that case, it's known as generative learning. Remember, we modelled the means and variances of X in each of the classes, and we modeled the prior probability. So that can be seen as modelling the joint distribution of X and Y.

Logistic Regression versus LDA

For a two-class problem, one can show that for LDA

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x_1 + \ldots + c_p x_p$$

So it has the same form as logistic regression.

The difference is in how the parameters are estimated.

- Logistic regression uses the conditional likelihood based on $\Pr(Y|X)$ (known as *discriminative learning*).
- LDA uses the full likelihood based on $\Pr(X, Y)$ (known as *generative learning*).
- Despite these differences, in practice the results are often very similar.

Logistic Regression versus LDA

For a two-class problem, one can show that for LDA

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x_1 + \ldots + c_p x_p$$

So it has the same form as logistic regression.

The difference is in how the parameters are estimated.

- Logistic regression uses the conditional likelihood based on $\Pr(Y|X)$ (known as *discriminative learning*).
- LDA uses the full likelihood based on $\Pr(X, Y)$ (known as *generative learning*).
- Despite these differences, in practice the results are often very similar.

Footnote: logistic regression can also fit quadratic boundaries like QDA, by explicitly including quadratic terms in the model.

Just like we did in linear regression, in logistic regression we can put in X^2 and Xi * Xj and terms like that and just explicitly get a quadratic boundary.

Summary

- Logistic regression is very popular for classification, especially when $K = 2$.
- LDA is useful when n is small, or the classes are well separated, and Gaussian assumptions are reasonable. Also when $K > 2$.
- Naive Bayes is useful when p is very large.
- See Section 4.5 for some comparisons of logistic regression, LDA and KNN.