Chapter_4_Classification

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:

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

  You can easily see that $p(X)$ will always take values between 0 and 1.
- A bit of rearrangement gives

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

  This monotone transformation is called the **log odds** or **logit** transformation of **p(X)**. The quantity whose logarithmic is being taken is called **odds** .

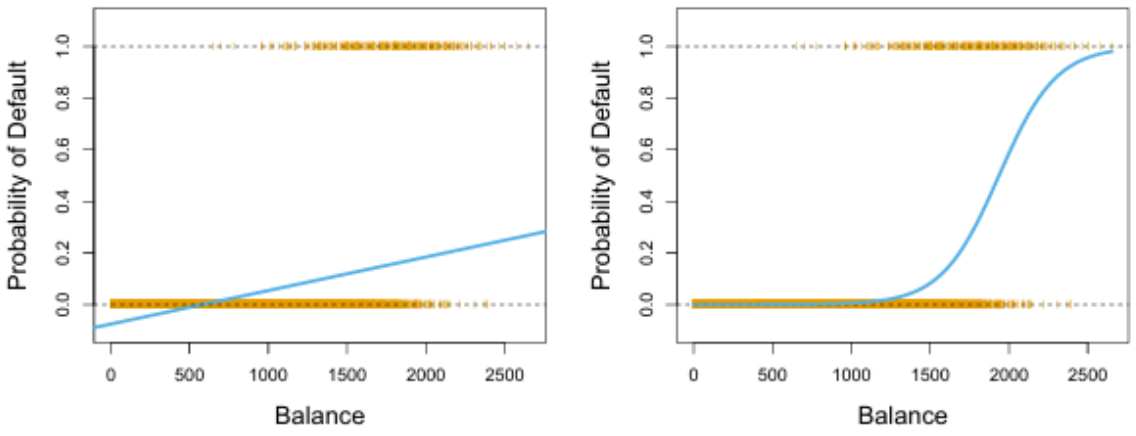


FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default` *(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

Maximum Likelihood

- We use maximum likelihood to estimate the parameters

$$l(\beta_0, \beta_1) = \Pi_{i:y_i=1} p(x_i) \Pi_{i:y_i=0} (1 - p(x_i))$$

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 a linear logistical model by maximum likelihood. In **R** , **glm** function is used.

Making Predictions

We get the following:

	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

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?

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

Let's do this again using **student** as 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

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

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

Multivariate Logistic Regression

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

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

	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?

Confounding

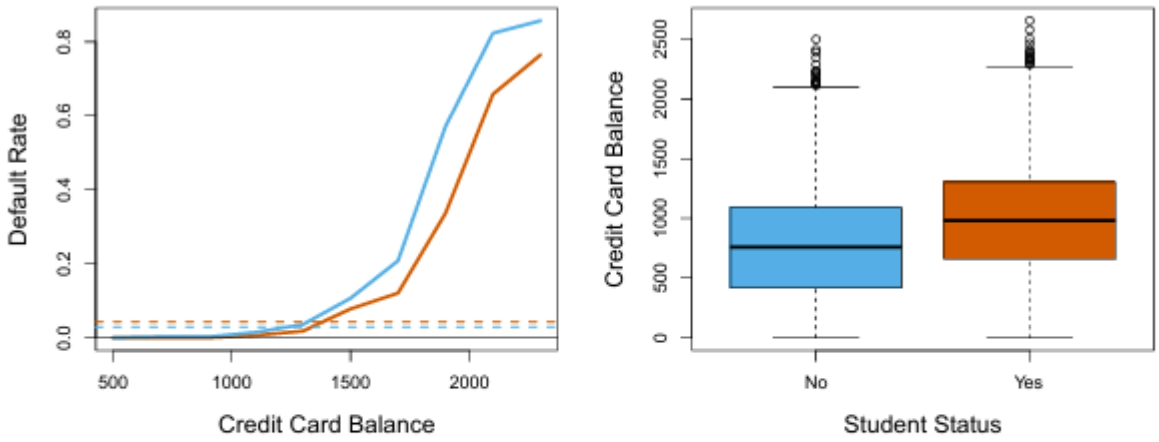
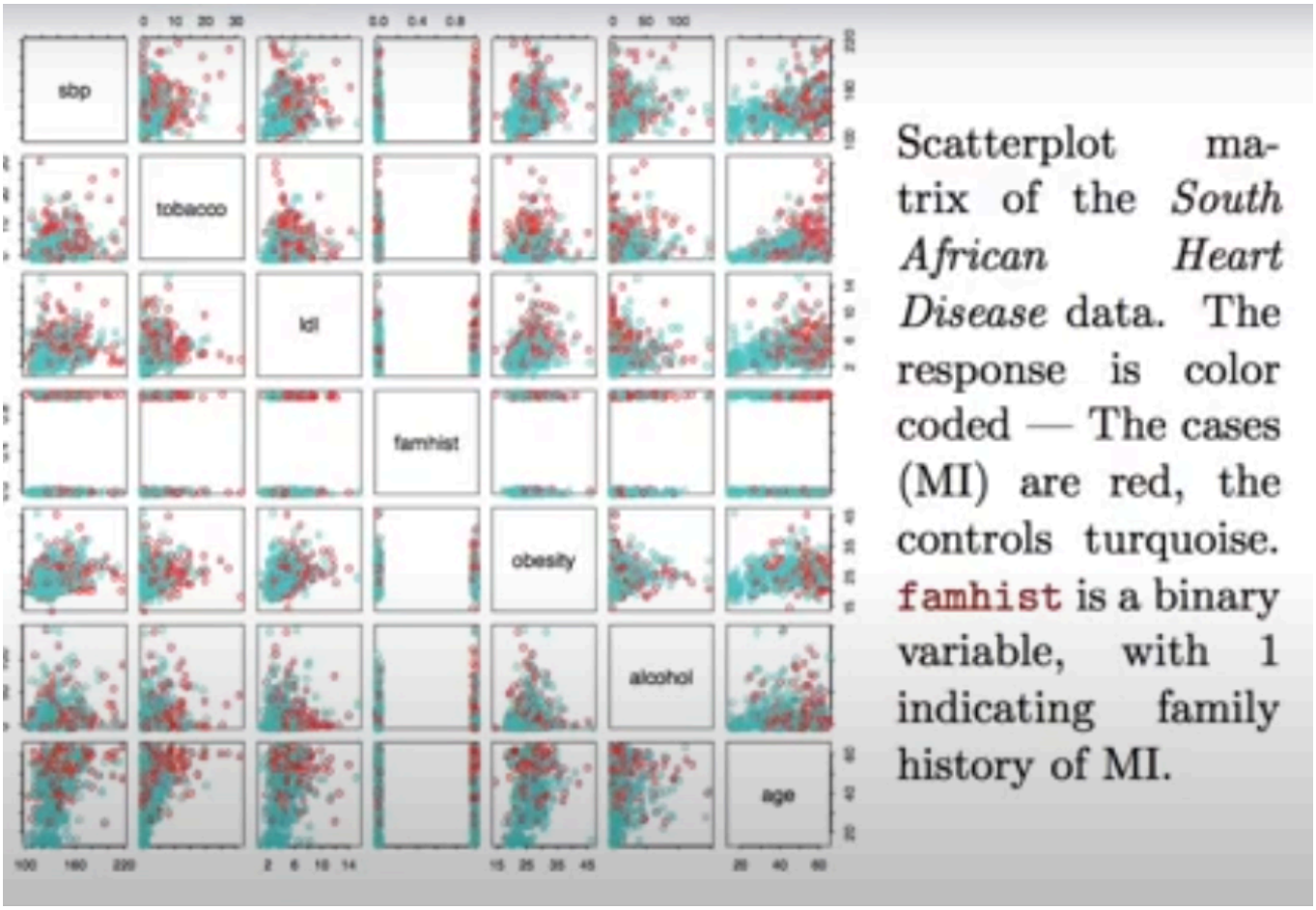


FIGURE 4.3. *Confounding in the* `Default` *data.* Left: *Default rates are shown for students (orange) and non-students (blue). The solid lines display default rate as a function of* `balance`*, while the horizontal broken lines display the overall default rates.* Right: *Boxplots of* `balance` *for students (orange) and non-students (blue) are shown.*

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

Example: South African Heart Disease

- 160 cases of MI(myocardial infarction) and 302 controls(all male in the age range of 15-64), from Western Cape ,South Africa in early 80s.
- Overall prevalence very high in this region : 5.1%
- Measurements on seven predictors, 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.



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.

```
> 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)

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

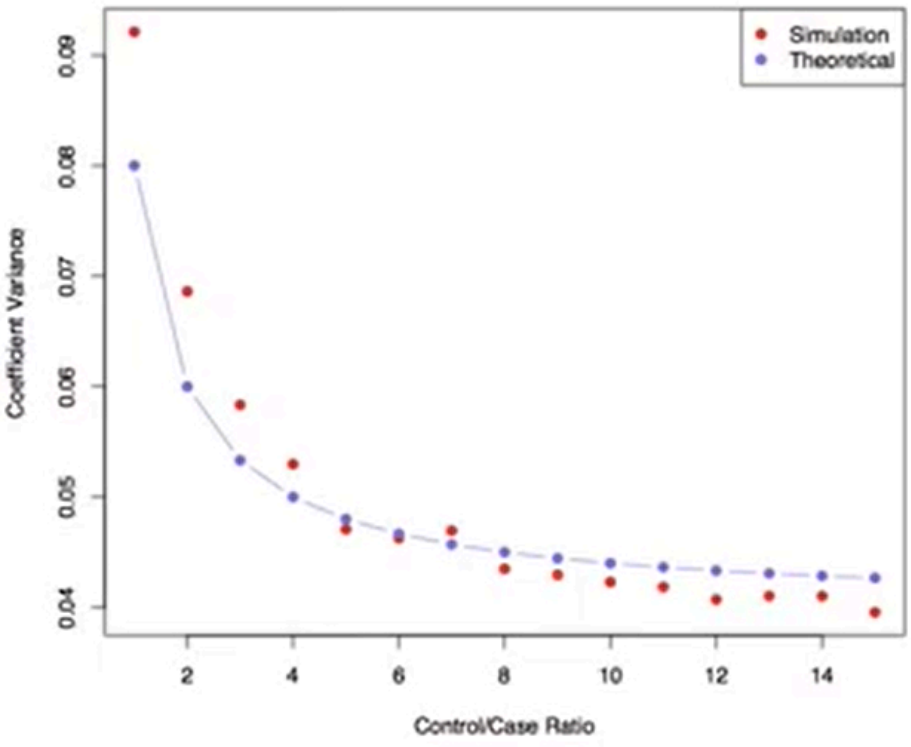
Case Control Sampling and Logistic Regression

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

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

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

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.

Logistic Regression with more than two classes

So far we have discussed logistic regression with two classes. Now for **K** classes :

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

Here there is a linear function for each class. The above function is also known as **softmax function** . Notice that after some cancellation only **K-1** linear functions are required. Also multivariate logistic regression is known as **multinomial logistic regression** .

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 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 distribution though.

Bayes Theorem for Classification

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

What is written above is known as **Bayes Theorem** .
One writes this slightly different for discriminant analysis:

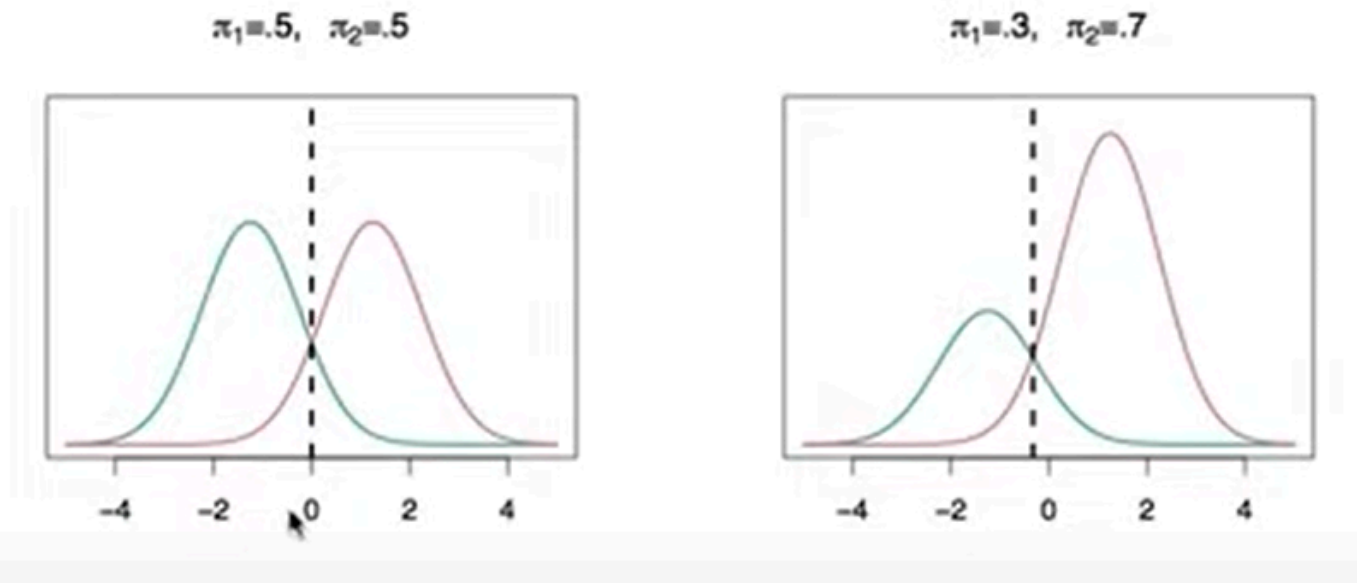
$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)},$$

where

- $f_k(x) = Pr(X = x|Y = k)$ 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

Now assuming the priors are same, the classification only depends on the densities:



- We classify a new point according to which density is highest. Like in figure one with same densities, the decision boundary is at zero. So any positive *X* belongs to *pink* class.
- 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 - decision boundary has shifted rightwards.

WHY DISCRIMINANT ANALYSIS?

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

Linear Discriminant Analysis when $p = 1$

- Gaussian density has the form

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

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

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

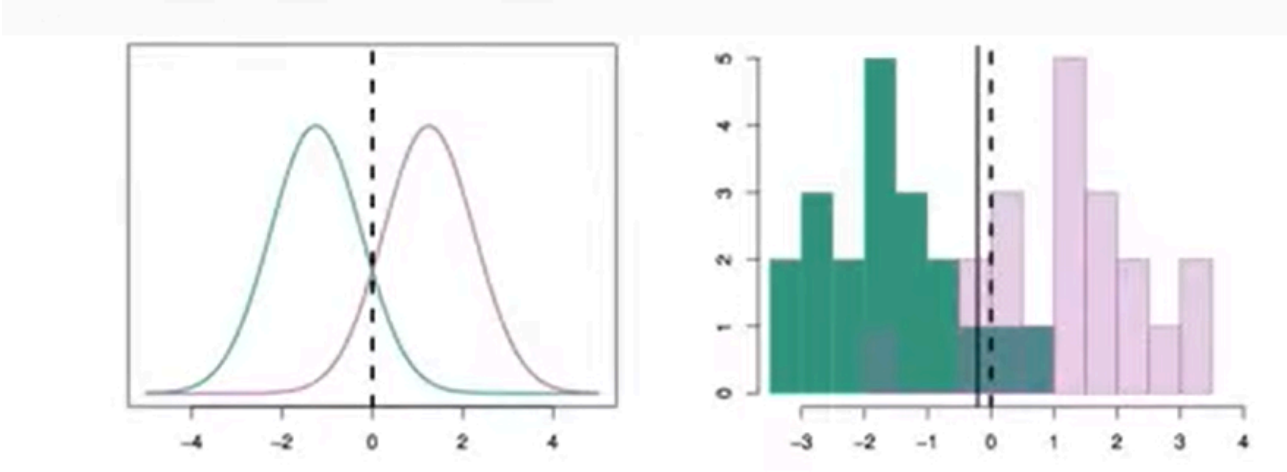
Discriminant Functions

- To classify at the value $X = x$, we need to see which of the $p_k(x)$ is the largest.
- Taking logs, and discarding terms that don't depend on k , we see that this is equivalent to assigning x to the class with 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 above 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

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



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

- Look at the above example where 20 observations are drawn. Typically you don't have these parameters(true mean, true densities et al); we just have the training data. In that case we simply estimate the parameters and plug them into the rule.

ESTIMATING PARAMETERS

$$\hat{\pi_k} = \frac{n_k}{n}$$

$$\hat{\mu_k} = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu_k})^2 = \sum_{k=1}^{K} \frac{n_k - 1}{n-K} \cdot \hat{\sigma_k^2}$$

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

**You will plug these values back into the formula and notice that instead of the being $x = 0$ the estimated LDA lies slightly left of it as shown in the above fig.

Linear Discriminant Analysis when $p > 1$

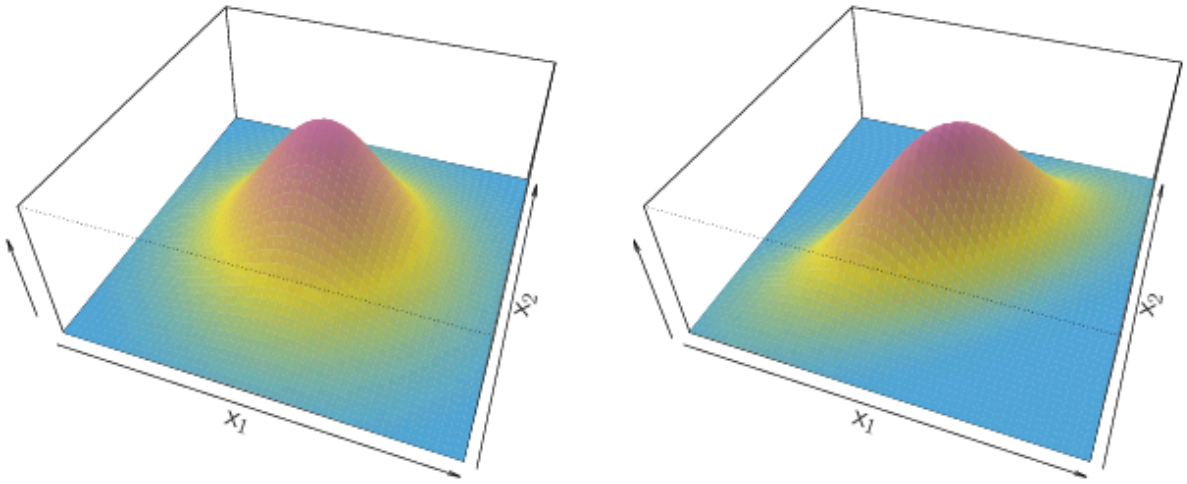


FIGURE 4.5. *Two multivariate Gaussian density functions are shown, with* $p = 2$. Left: *The two predictors are uncorrelated.* Right: *The two variables have a correlation of* 0.7.

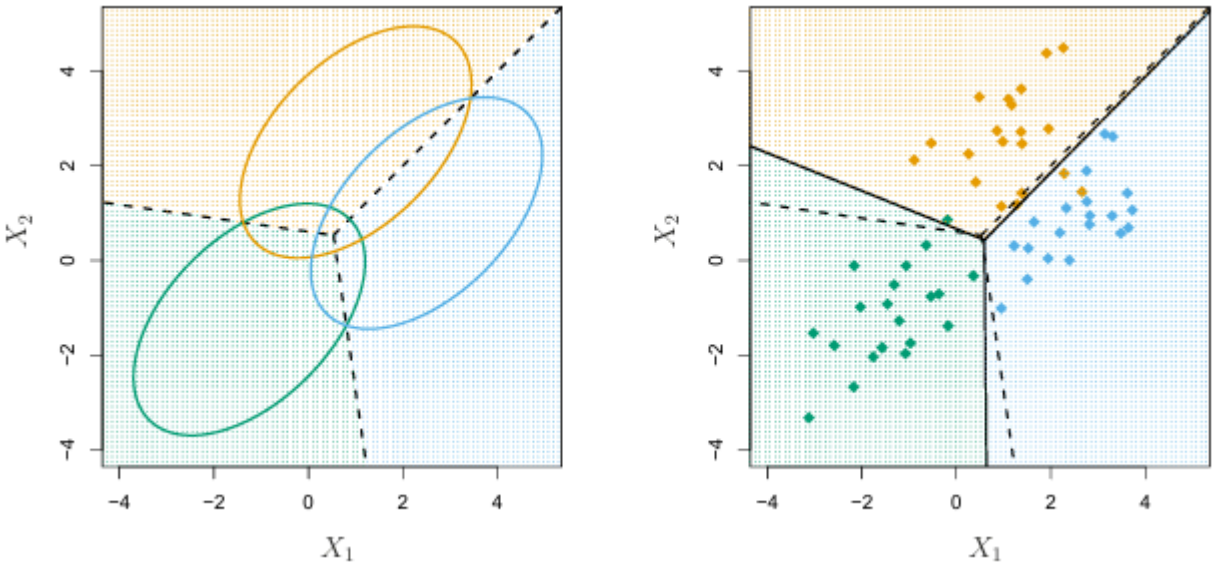
Gaussian Density : $f(x) = \frac{1}{(2\pi)^{p/2}|\sum|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)}$, where $\sum$ is called the **co-variance matrix** .

After some simplifications, **discriminant function** comes out to be

$$\delta_k(x) = x^T(\sum)^{-1}\mu_k - \frac{1}{2}\mu_k^T(\sum)^{-1}\mu_k + \log \pi_k$$

Important to note that it's again linear in x .

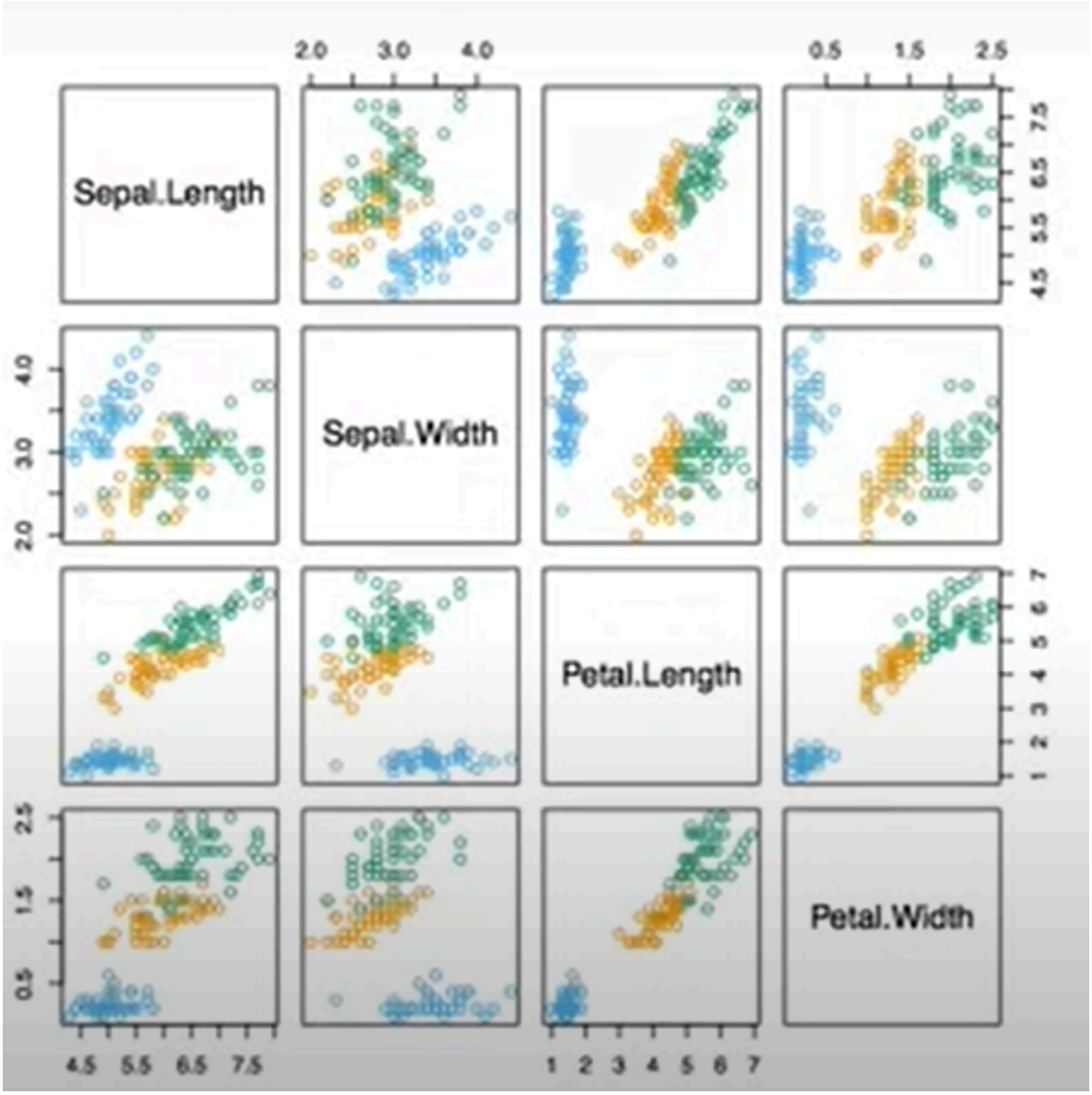
Illustration: $p = 2$ and $K = 3$ classes



Above is an example with three classes. The observations from each class are drawn from a multivariate Gaussian distribution with $p = 2$, with a class-specific mean vector and a common covariance matrix.

- Left: Ellipses that contain 95% of the probability for each of the three classes are shown. The dashed lines are the Bayes decision boundaries.
- Right: 20 observations were generated from each class, and the corresponding LDA decision boundaries are indicated using solid black lines. The Bayes decision boundaries are once again shown as dashed lines
  Here $\pi_1 = \pi_2 = \pi_3 = 1/3$.
  Were the Bayes Decision Boundaries known, they would yield the fewest misclassification errors, among all classifiers. (these are the true decision boundaries)
  After using the formulas derived earlier, you get these solid lines : Linear Discriminant Analysis decision boundaries.

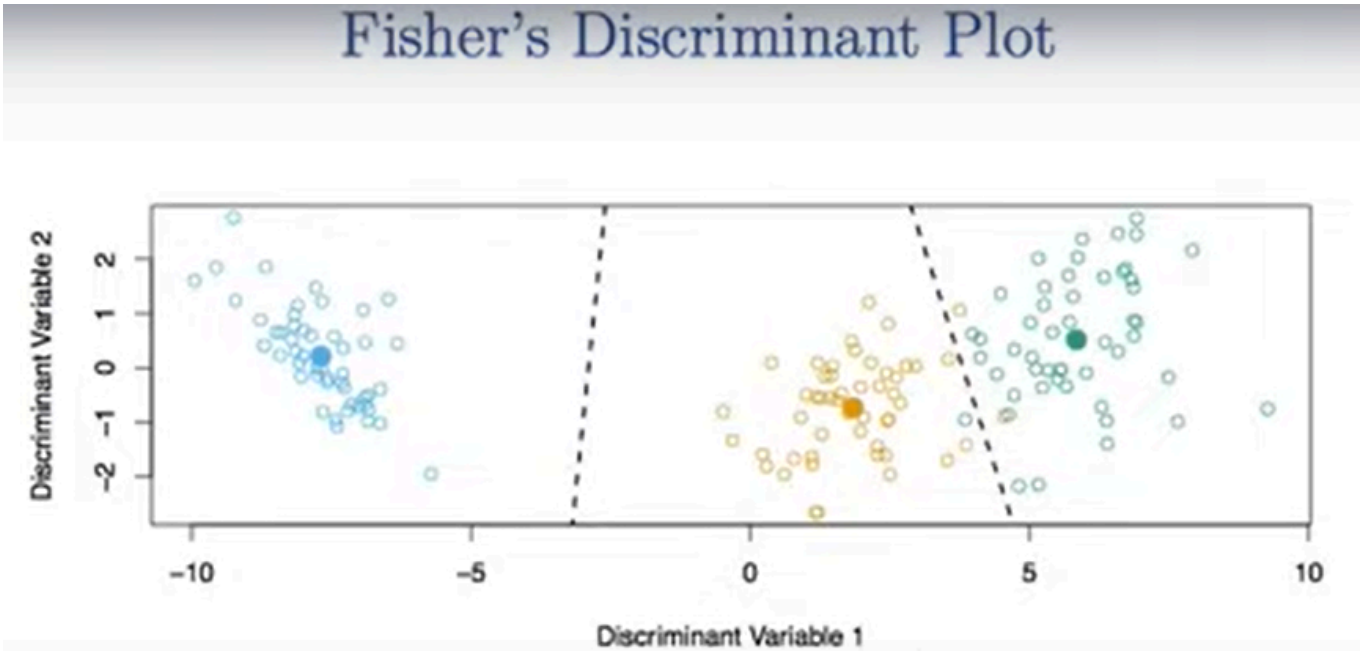
Fisher's Iris Data



4 variables; 3 species; 50 samples per class

- Setosa
- Versicolor
- Virginicia

LDA classifies all but 3 of the 150 training samples correctly. We have something called as:



A plot that captures the classification info for all classes.
When there are K classes, LDA can be viewed exactly in a $K-1$ dimensional plot.
Why?
Basically what **Fisher's Discriminant Analysis** or **Gaussian LDA** is doing is that it's measuring which centroid is the closest. But it's measuring it in a distance where it takes into account the co-variance of the variables.
The centroids span a $K-1$ dimensional plane, leading to these low-dimensional plots.
Even when $K > 3$, we can find the "best" 2-dimensional plane for visualizing the discriminant rule.

From $\delta_k(x)$ to probabilities

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

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

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

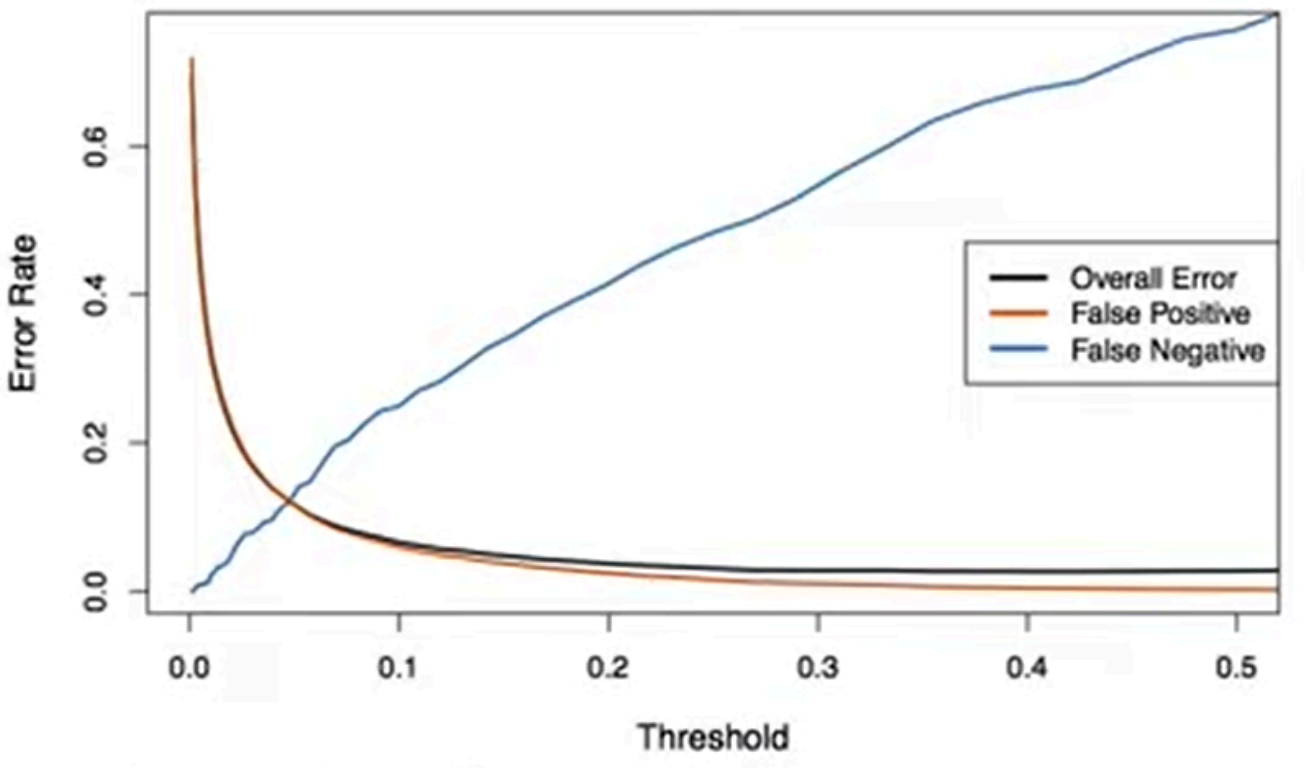
Types of errors

LDA on Credit Data

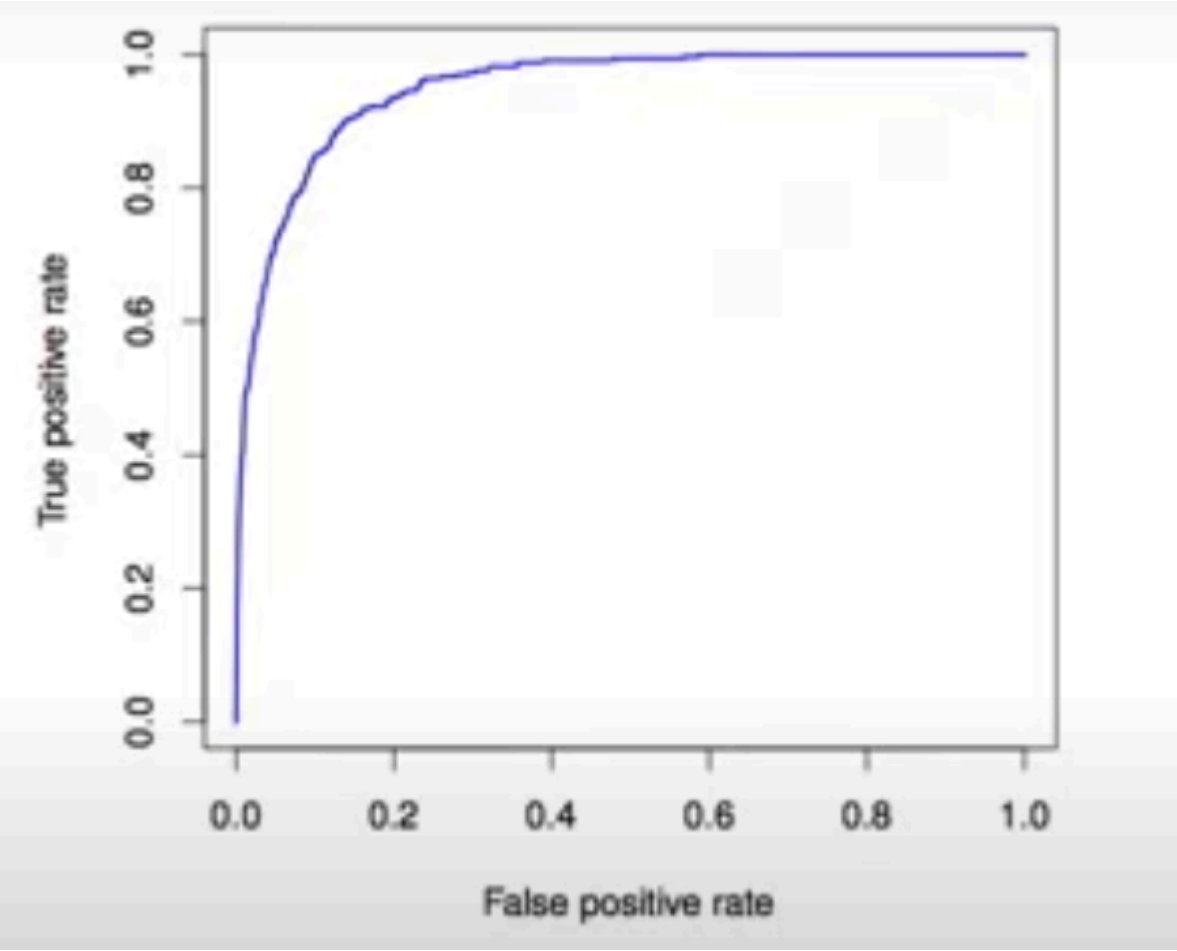
		True Default Status		
		No	Yes	Total
Predicted Default Status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

- **False Positive Rate:** The fraction of negative examples that are classified as positive - 0.2% in example.
- **False Negative Rate:** The fraction of positive examples that are classified as negative - 75.7% in example.
  Remember this table was produced by taking 0.5 as threshold which can always be changed to some other value in [0,1]. On varying the threshold we see that :

Varying the *threshold*



In order to reduce the false negative rate, we may want to reduce the threshold to 0.1 or less. We would like the false positive rate to be low and true positive rate to be high. There's another plot that displays both simultaneously: **ROC Curve**



NOTE:

- In the best cases the curve is deep into the top-left corner.
- This single curve captures the behavior of the classification rule for all possible thresholds.
- And you can compare different classifiers by comparing their ROC curves.
- Sometimes we use the **area under the curve** or **AUC** to summarise the overall performance . Higher **AUC** is good.

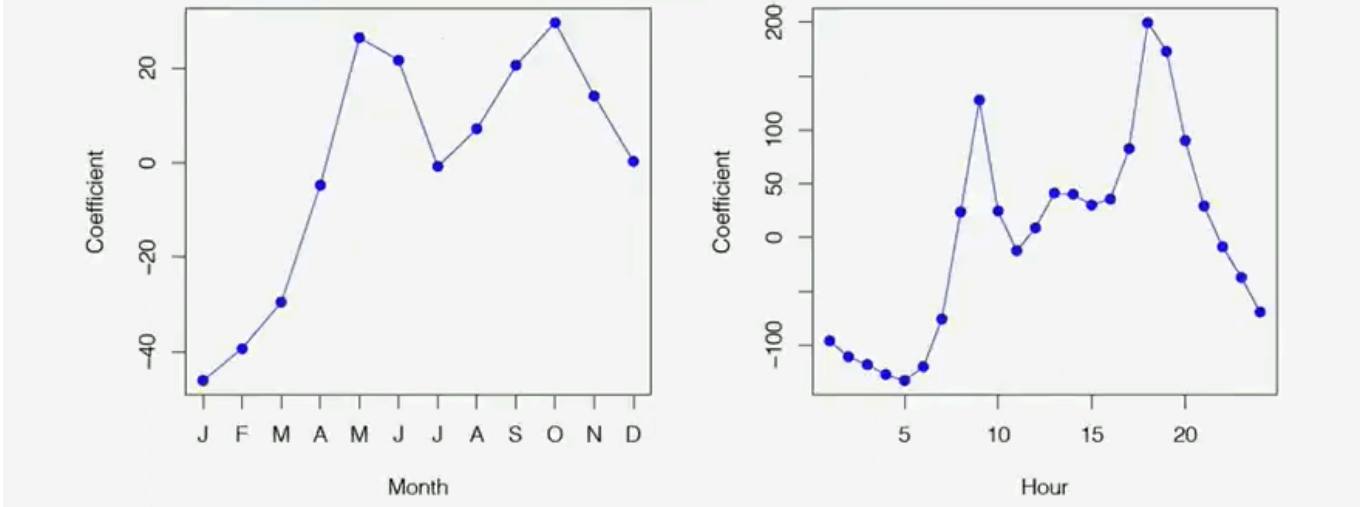
Generalised Linear Models

- Linear regression is used for quantitative responses.
- Linear Logistic Regression is the counterpart for a binary response, and models the logit of the probability as a linear model.
- Other response types exist, such as non-negative responses, skewed distributions, and more.

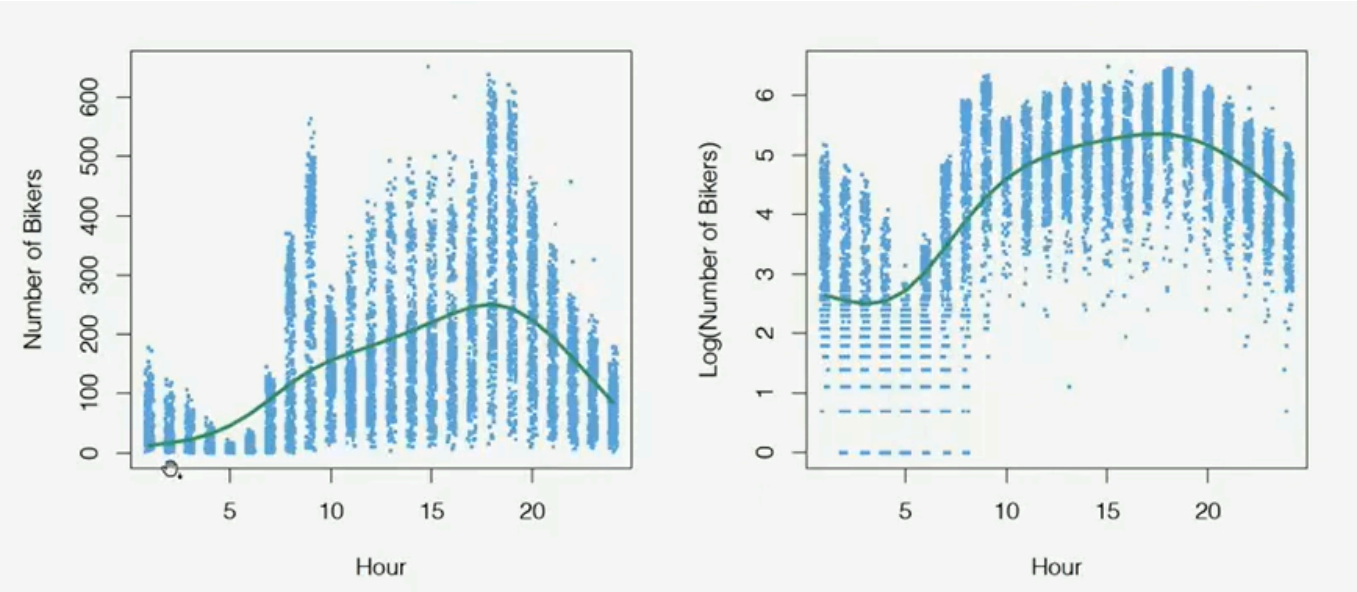
- **Generalised linear models** provide a unified framework for dealing with many different response types

We'll be using the Bikeshare Data for explanation of an important member of this family called **Poisson Regression**. Linear regression with response **bikers** : number of hourly users in bikeshare program in Washington, DC. The x-axis of the plots denote other variables in the dataset like months and hour of use. y

	Coefficient	Std. error	z-statistic	p-value
Intercept	73.60	5.13	14.34	0.00
workingday	1.27	1.78	0.71	0.48
temp	157.21	10.26	15.32	0.00
weathersit[cloudy/misty]	-12.89	1.96	-6.56	0.00
weathersit[light rain/snow]	-66.49	2.97	-22.43	0.00
weathersit[heavy rain/snow]	-109.75	76.67	-1.43	0.15



Mean/Variance Relationship



- In the left plot, a smooth spline has been fit with the data. When mean is low, the spread is also small but spread increases with an increase in the mean i.e. variance mostly increases with the mean.
- It's not shown here but 10% of the linear model predictions were negative(there's no constraint on the linear model that the predictions should be +ve).
- Taking log(**bikers**) alleviates this but has its own problems: predictions on the wrong scale, and some counts are zero.
  So this is where the Poisson model comes in, just like we use binomial model for 0/1 data, the Poisson regression or distribution is good for modelling counts.

Poisson Regression Model

- Poisson distribution is useful for modeling counts. Here's is the **pmf** :

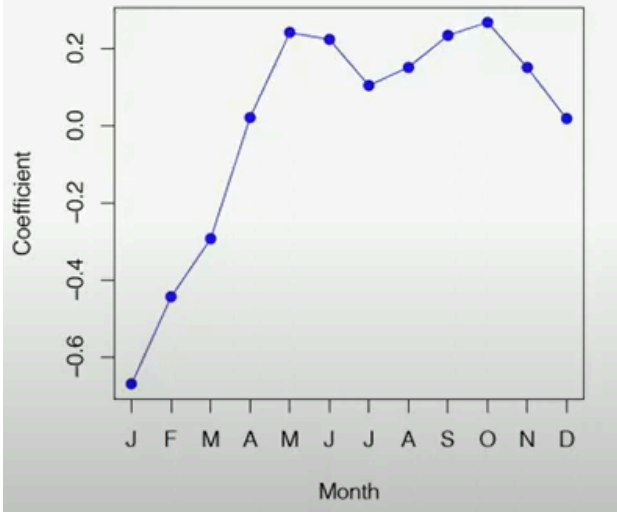
$$Pr(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!},$$

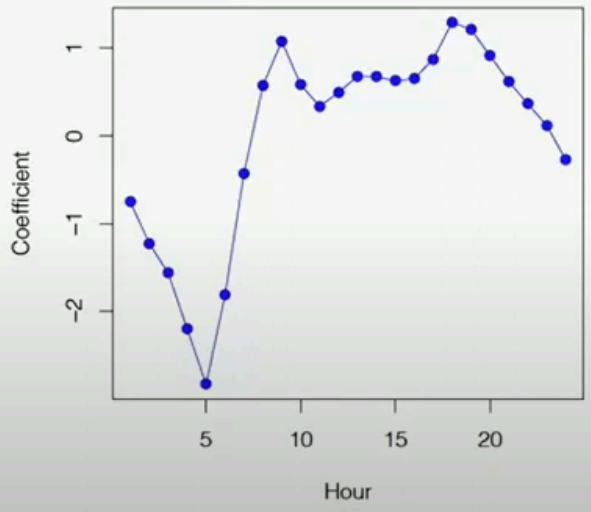
  for $k = 0, 1, 2, \ldots$
- $\lambda = E(Y) = Var(Y)$ i.e. there is a mean/variance dependence.
- BTW for binomial distribution too, the mean and variance are dependent . If the mean is p then the variance is $p(1-p)$.
- With co-variates

$$\log\left(\lambda(X_1, X_2, \cdots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Poisson Regression on Bikeshare data

	Coefficient	Std. error	z-statistic	p-value
Intercept	4.12	0.01	683.96	0.00
workingday	0.01	0.00	7.5	0.00
temp	0.79	0.01	68.43	0.00
weathersit[cloudy/misty]	-0.08	0.00	-34.53	0.00
weathersit[light rain/snow]	-0.58	0.00	-141.91	0.00
weathersit[heavy rain/snow]	-0.93	0.17	-5.55	0.00





-> In this case the variance is somewhat larger than the mean - a situation known as *overdispersion* - so the p-values are misleadingly small.

Note: SO WHEN YOU FIT THE DATA WITH POISSON MODEL, IT TAKES INTO ACCOUNT THE CHANGE IN VARAINCE.
REMEMBER THAT THE WHILE DEALING WITH THE **LINEAR DISCRIMINANT ANALYSIS**, WE ASSUMED THAT THE VARIANCE ARE SAME FOR EAXCH CLASS.

- So the generalized linear models covered up to now are Gaussian , Binomial and Poisson.
- They each have a characteristic *link* function . This is the transformation of the mean that is represented by a linear model:

$$\eta(E(Y|X_1, X_2, \ldots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  The link functions for linear ,logistic and Poisson regression are $\eta(\mu) = \mu, \eta(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$, and $\eta(\mu) = \log(\mu)$ respectively.
- They also each have a characteristic *variance* functions.
- Other GLMs include **Gamma, Negative-binomial, Inverse-Gaussian** and more.

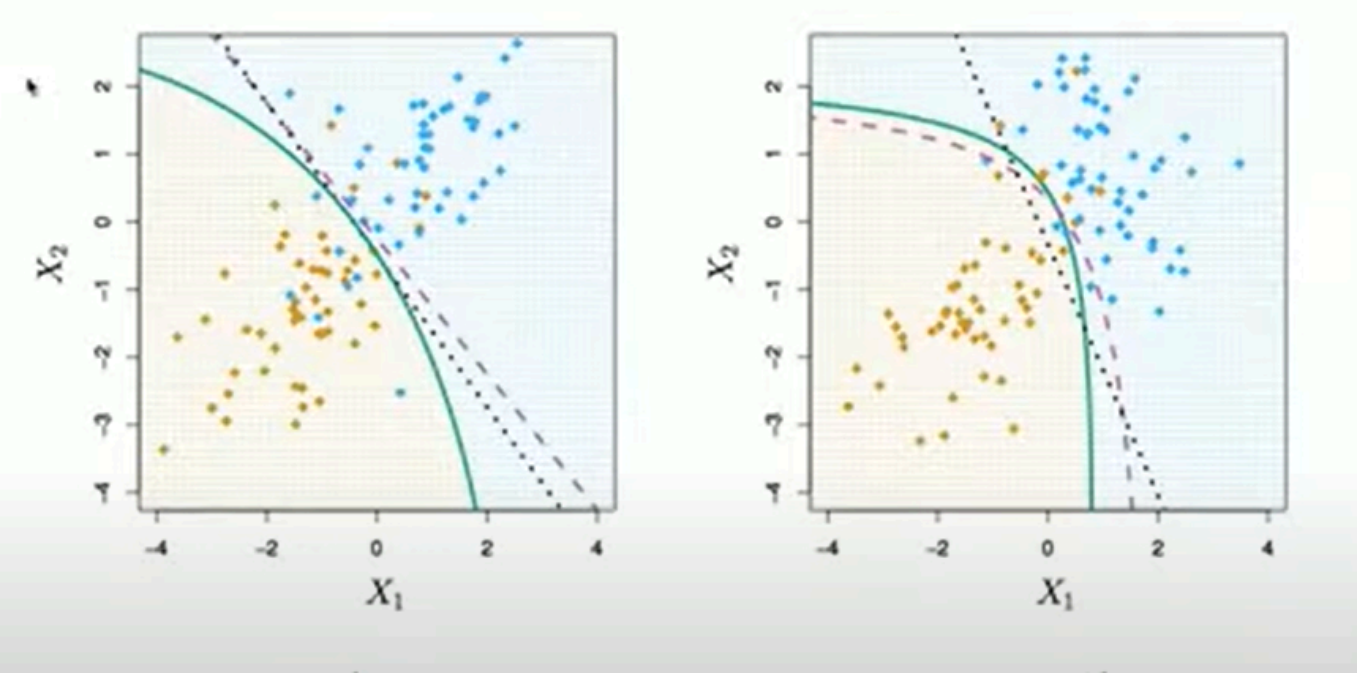
Quadratic Discriminant Analysis and Naive Bayes

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

When $f_k(x)$ are Gaussian Densities, with the same covariance matrix $\sum$ in each class, we get Linear Discriminant Analysis. By altering the forms for $f_k(x)$, we get different classifiers:

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

Quadratic Discriminant Analysis



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

Because the $(\sum)_k$ are different , the quadratic terms matter. You can see in the right image how the quadratic discriminant analysis decision boundary almost traces the Bayes Decision boundary.

Naive Bayes

Assume 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 $(\sum)_k$ is diagonal:

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

- can use for **mixed** feature vectors(qualitative & 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.

Logistic Regression vs LDA

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

$$log\frac{p_1(x)}{1-p_1(x)} = log\frac{p_1(x)}{p_2(x)} = c_0 + c_1x_1 + \cdots + c_px_p$$

So it has the same form as logistic regression. The difference is in how 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.

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.