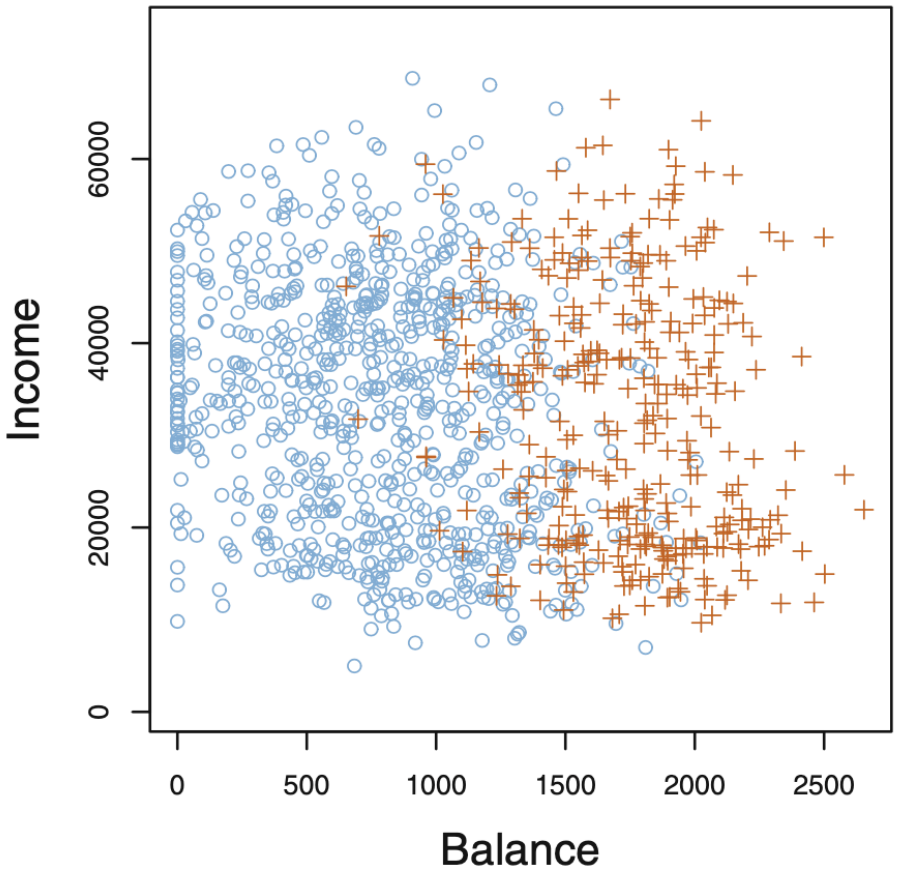
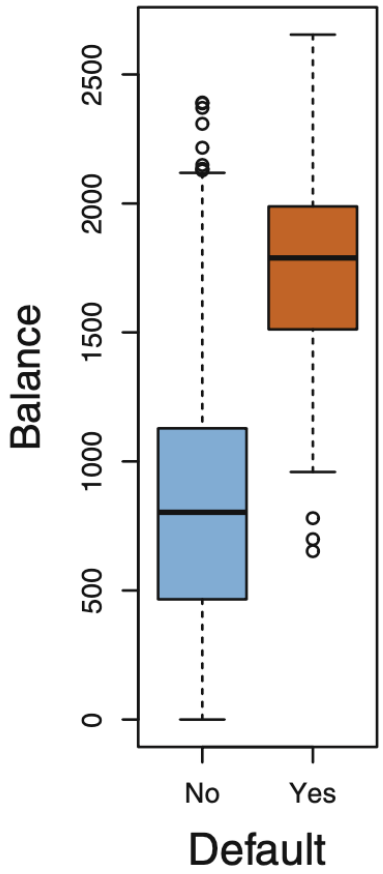
PQHS 471
Lecture 7: Regression methods
GLM, LDA, QDA

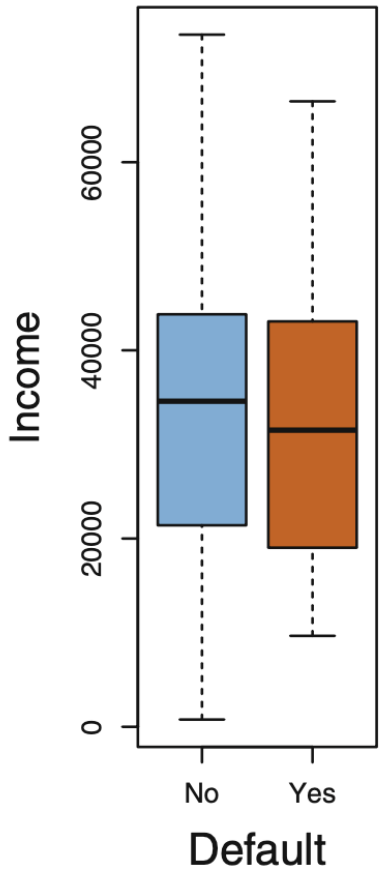
Motivating example

credit card default

Default dataset in R.







Can we use linear regression?

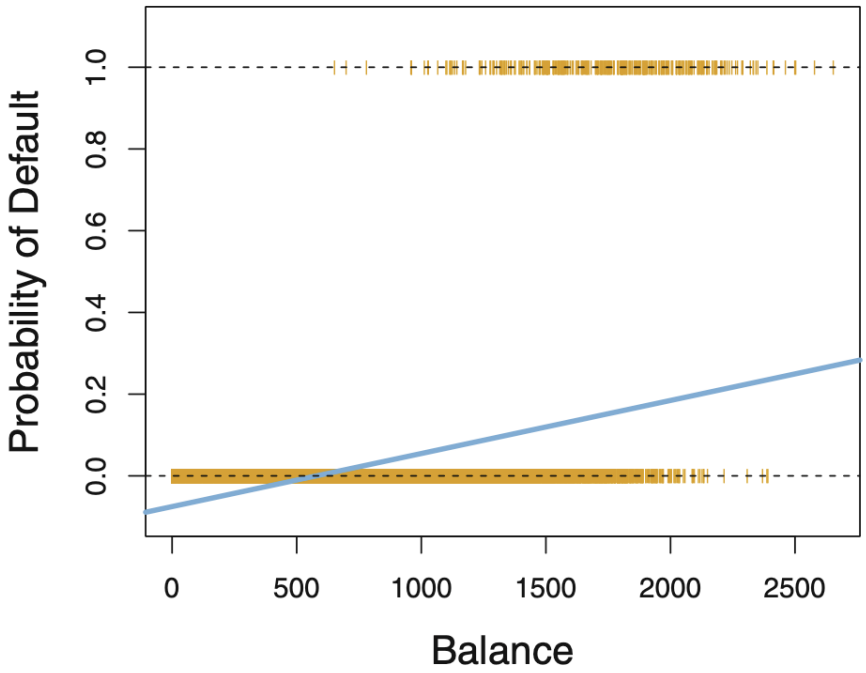
Suppose for the Default classification task that we code

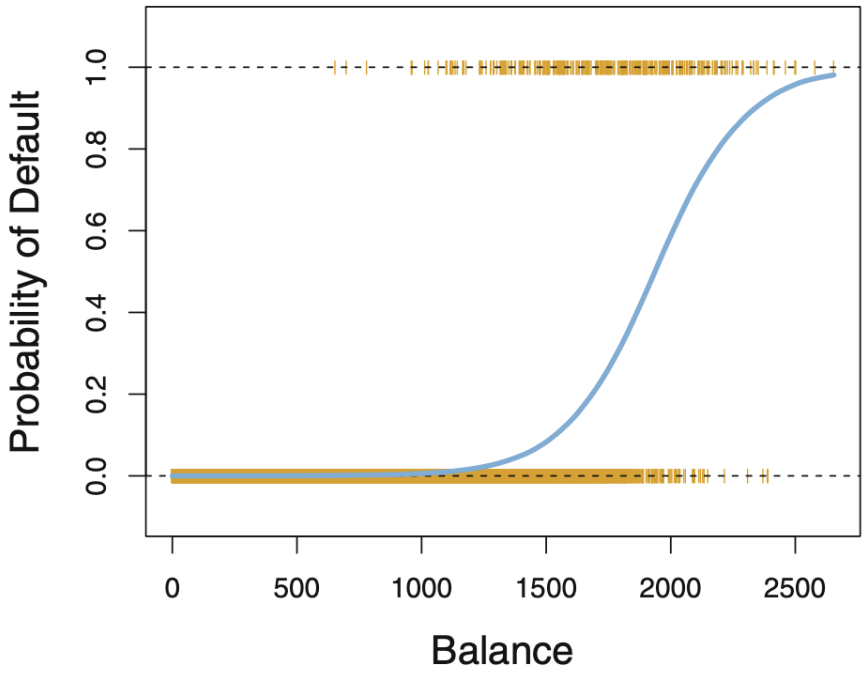
$$Y = \begin{cases} 0 & \text{if No} \\ 1 & \text{if Yes} \end{cases}$$

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

- Since in the population $E(Y|X = x) = Pr(Y = 1|X = x)$, we might think that regression is perfect for this task.
- Linear regression might produce probabilities < 0 or > 1. Logistic regression is more appropriate.

Linear vs. 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? No

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 Regression or Discriminant Analysis are more appropriate.

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

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

- Rather than modeling this response Y directly, logistic regression models the **probability** that Y belongs to a particular category.
- Logistic model is better able to capture the range of probabilities than is the linear regression.

Maximum Likelihood

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

- This likelihood gives the probability of the observed 0's and 1's in the data.
- We pick β_0 and β_1 to maximize the likelihood of the observed data.

Maximum Likelihood

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

- This likelihood gives the probability of the observed 0's and 1's 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

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 $\$2000$?

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

Making Predictions: student

Using student as the only 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

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

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

Multiple Logistic Regression

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

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

	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

Multiple Logistic Regression

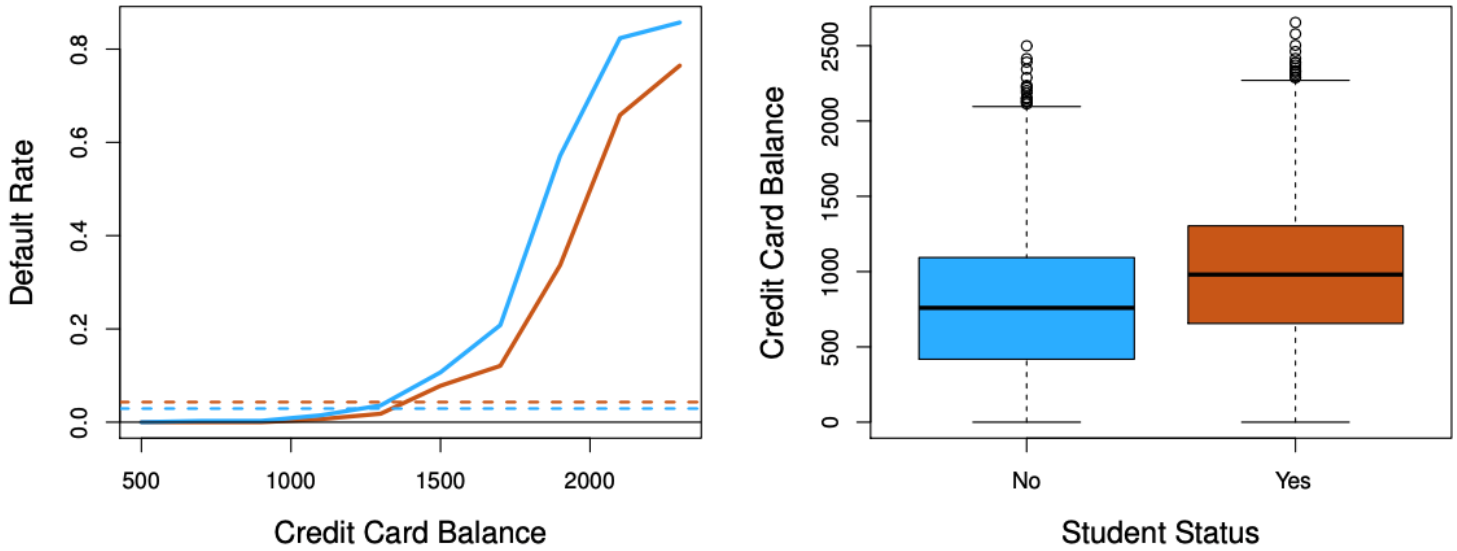
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

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

	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



- 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.
- Different interpretations for univariate- and multiple-Logistic Regression.

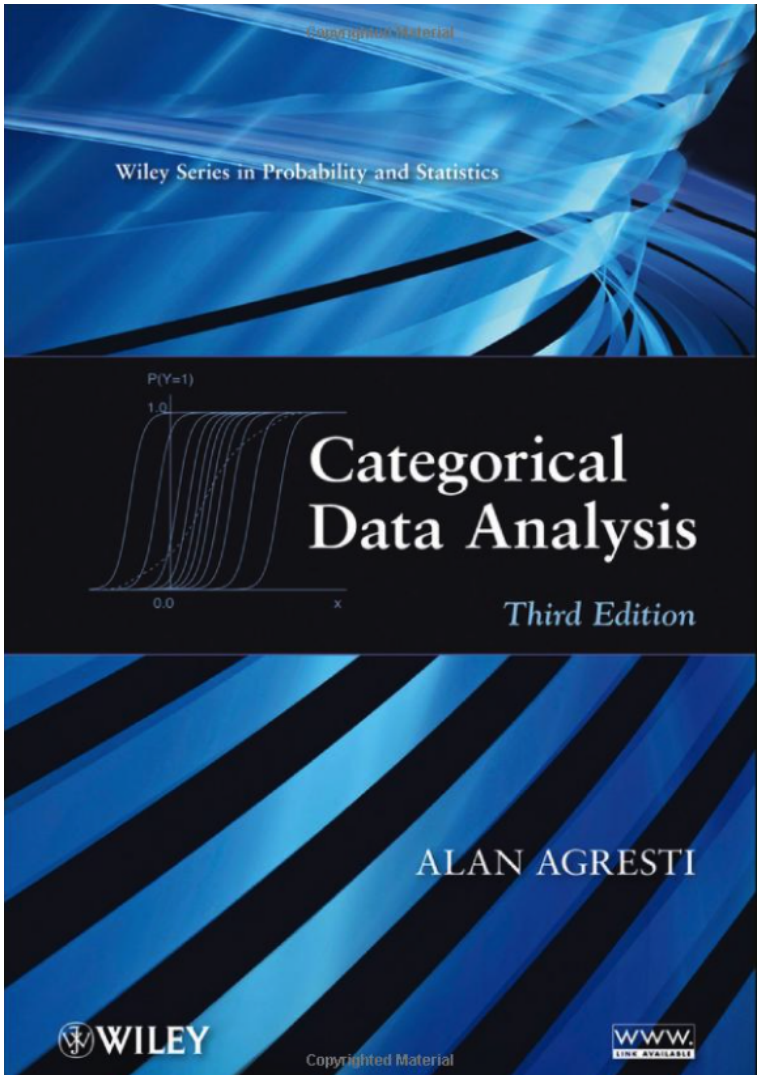
More than two classes

So far we have discussed logistic regression with two classes.
The idea can be generalized to more than two classes, for example:

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

- A linear function for **each** class.
- **multinomial regression** is an example.

Categorical Data Analysis (CDA) book



Highly recommend.

Linear Discriminant Analysis

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/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 revisit

$$P(B) = \sum_{i=1}^{M} P(B|A_i)P(A_i)$$

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

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

One writes this slightly differently for discriminant analysis:

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

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

Discriminant Analysis Overview

Suppose the outcome Y has K classe, and there are p features in X.

- Linear discriminant analysis (**LDA**): For class k, the features $X \sim N_p(\mu_k, \Sigma)$. These K distributions have different means (centers), but they have the same variance–covariance matrix Σ.
- Quadratic discriminant analysis (**QDA**): For class k, the features $X \sim N_p(\mu_k, \Sigma_k)$. These K distributions have different means and possibly different variance–covariance matrices.

Linear Discriminant Analysis for p = 1

Suppose $p = 1$, we have only 1 predictor.
Under the assumption that $f_k(x)$ 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 μ_k and σ_k^2 are the mean and variance parameters for the kth class.
In LDA, we have $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2$.

Linear Discriminant Analysis for p = 1

Suppose $p = 1$, we have only 1 predictor.
Under the assumption that $f_k(x)$ 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 μ_k and σ_k^2 are the mean and variance parameters for the kth class.
In LDA, we have $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2$.

Then, the posterior probability is :

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

The Bayes classifier will then assigning an observation to the class for which $p_k(x)$ is the largest.

Linear Discriminant Analysis for p = 1

discriminant score

Taking the log of $p_k(x)$ and rearrange the terms, we can show that this is equivalent to assigning the observation to the class for which

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

is the largest.

Why the name LDA? Note the $\delta_k(x)$ is a **linear** function of x.

Linear Discriminant Analysis for p = 1

In practice, we need to estimate the parameters μ_k, π_k and σ^2 first:

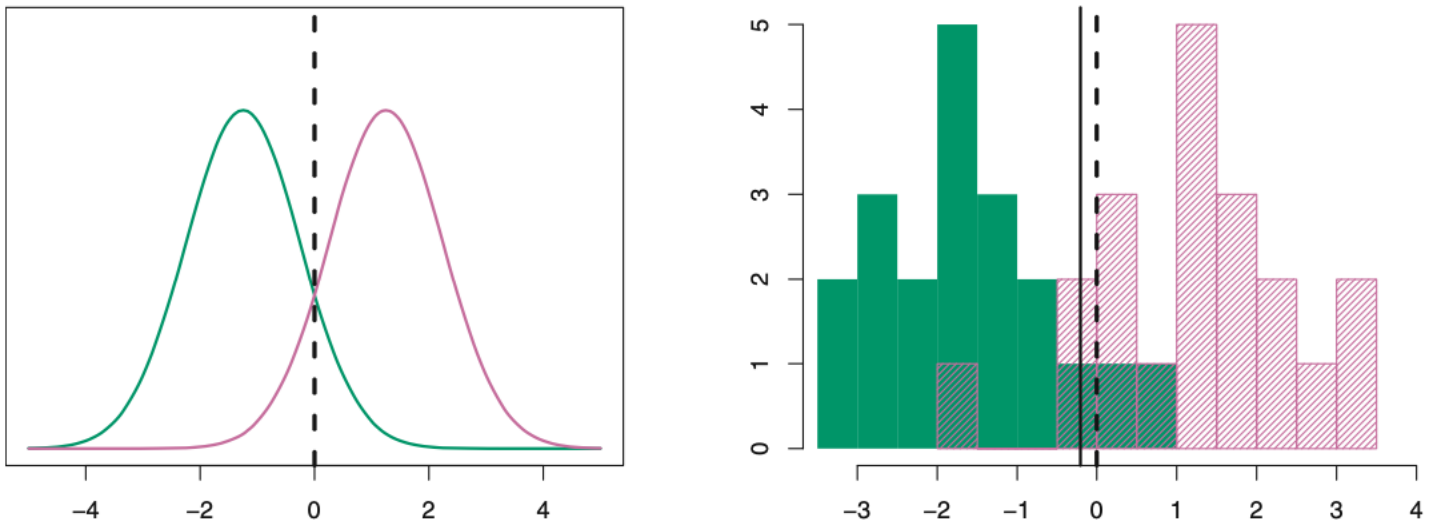
$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

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

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

and then use $\delta_k(x)$ for separation.

Linear Discriminant Analysis for p = 1



Dashed: Bayes decision boundary
Solid: LDA decision boundary
Left panel: Known distribution
Right panel: Simulation of 20 samples each class.

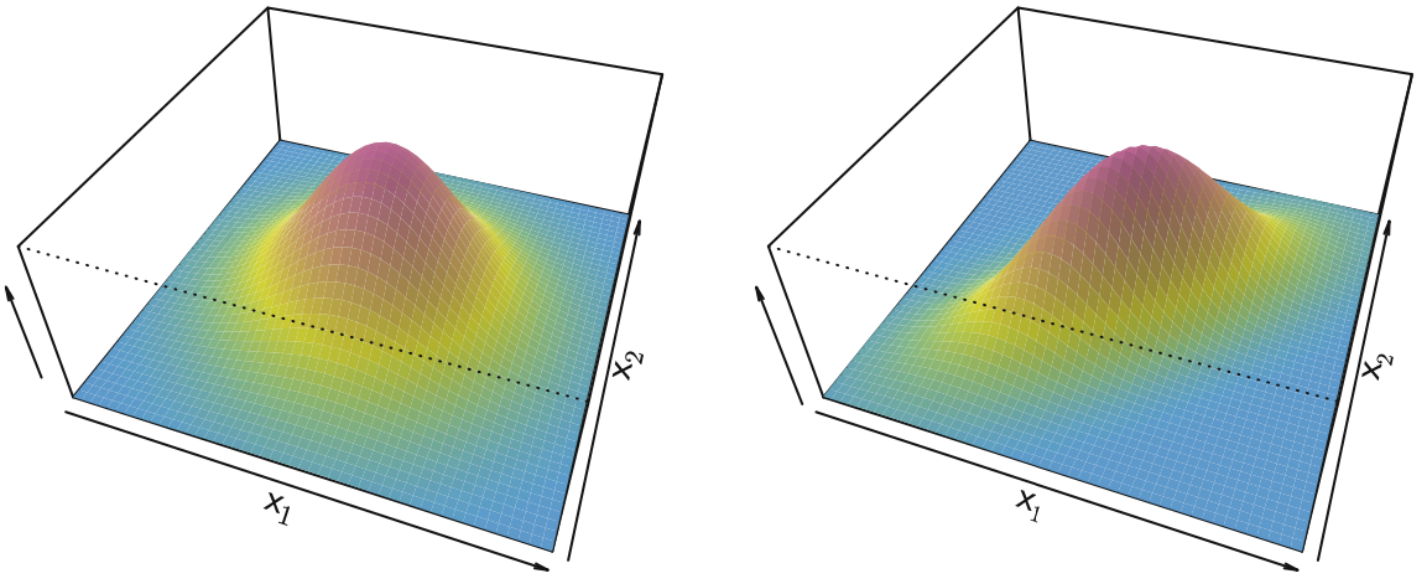
Linear Discriminant Analysis for p = 1

Summary

- Assuming that the observations within each class come from a normal distribution
- Class-specific mean
- Common variance
- Plugging estimates for these parameters into the Bayes classifier

Linear Discriminant Analysis for $p > 1$

Now we have multiple independent variables $\mathbf{X} = (X_1, X_2, ..., X_p)$, drawn from multivariate normal distribution (MVN). We have a class-specific mean vector and a common variance-covariance matrix.



p=2. Left: $x_1 \perp x_2$. Right: cor$(x_1, x_2) = 0.7$

Linear Discriminant Analysis for $p > 1$

In general, MVN $X \sim N(\mu, \Sigma)$:

$$f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)\right)$$

When using LDA for $p > 1$, observations in the kth class are drawn from a MVN $N(\mu_k, \Sigma)$. Here, μ_k is a class-specific mean vector, Σ is a variance-covariance matrix that is common to all K classes.
Similarly, we classify an observation $X = x$ to the class for which

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

is largest.
This is the vector/matrix version of $\delta_k(x)$ from $p = 1$

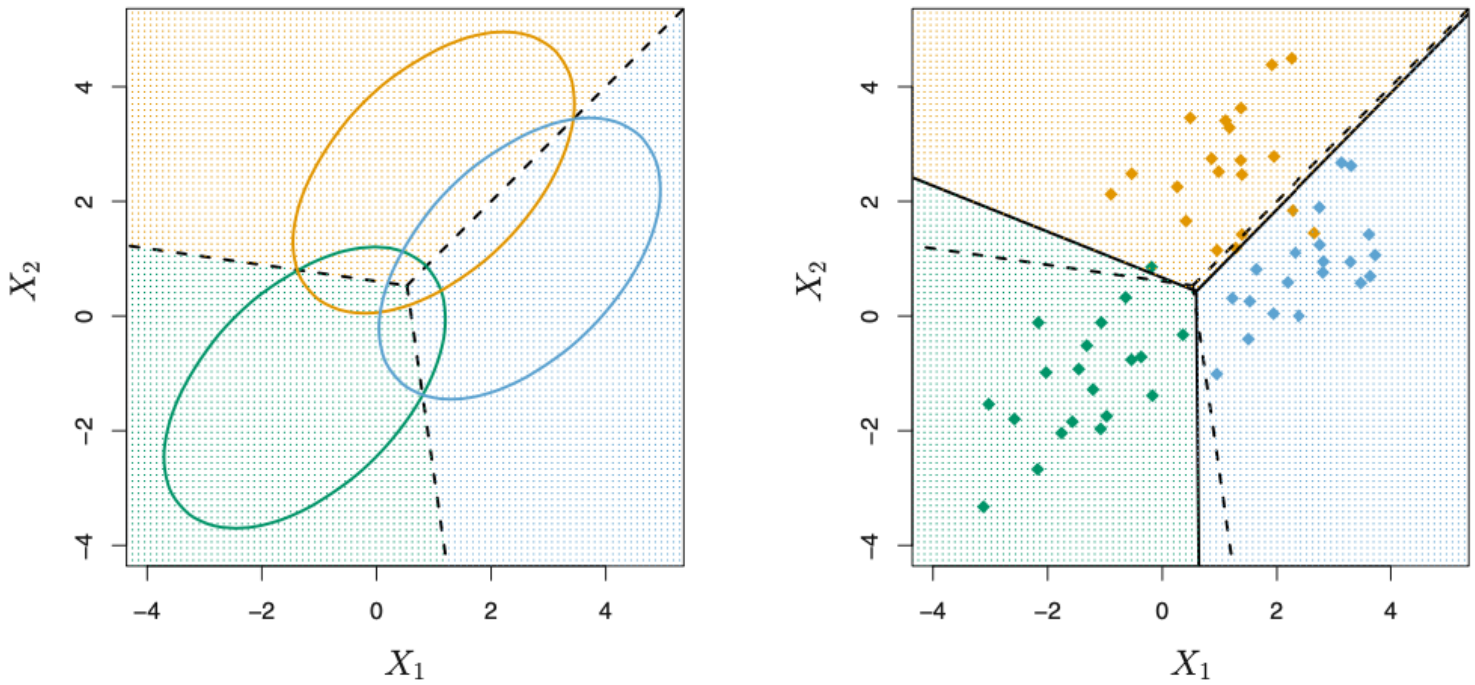
Linear Discriminant Analysis for $p > 1$

The Bayes decision boundary represent the set of values of x s.t. $\delta_k(x) = \delta_l(x)$

$$x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k = x^T \mathbf{\Sigma}^{-1} \mu_l - \frac{1}{2} \mu_l^T \mathbf{\Sigma}^{-1} \mu_l$$

for $k \neq l$. Once again, we need to estimate μ_k, π_k and Σ in practice. Similar procedures are taken as in the univariate situation.

Linear Discriminant Analysis for $p > 1$



Dashed: Bayes decision boundary
Solid: LDA decision boundary
Left panel: Known distribution
Right panel: Simulation of 20 samples each class

Quadratic Discriminant Analysis

Quadratic Discriminant Analysis

In LDA, we used a common variance-covariance matrix Σ for all classes K. In Quadratic Discriminant Analysis(QDA), we dropped that assumption, instead, we assume $X \sim N(\mu_k, \Sigma_k)$, where Σ_k is the variance-covariance matrix for the kth class.

Quadratic Discriminant Analysis

In LDA, we used a common variance-covariance matrix Σ for all classes K. In Quadratic Discriminant Analysis(QDA), we dropped that assumption, instead, we assume $X \sim N(\mu_k, \Sigma_k)$, where Σ_k is the variance-covariance matrix for the kth class.

Under this assumption, the Bayes classifier assigns $X = x$ to the class which

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

is largest.
Note the name **quadratic** came from the fact that $\delta_k(x)$ have quadratic terms of x.

LDA vs QDA

Why would one prefer LDA to QDA, or vice-versa?
Answer: **Bias-variance trade-off.**

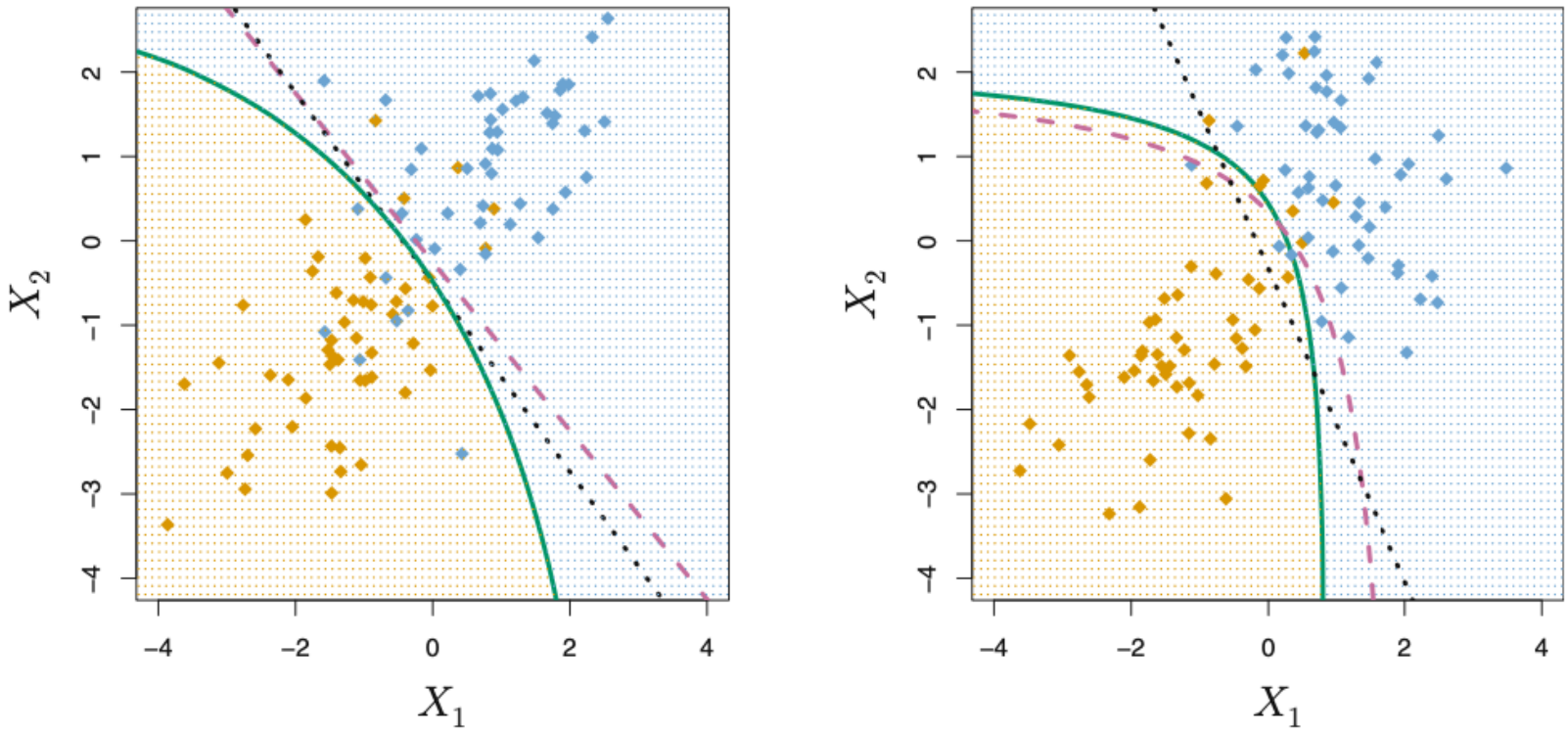
Having p predictors means estimating extra $Kp(p+1)/2$ parameters in Σ_k. This is expensive! \$\$\$
$p = 50$, need some multiple of 1,275!

LDA is much less flexible than QDA, so it has substantially lower variance.
Can lead to improvement in prediction performance.
LDA can suffer from high bias when a common variance matrix is badly off.

LDA vs QDA

- LDA is better if:
  - Relatively few observations (small n), so reducing variance is crucial.
- QDA is better if:
  - Training set is large.
  - Assumption of a common covariance matrix for all K is clearly untenable.

LDA vs QDA



Bayes: purple dashed. LDA: black dotted. QDA: green solid.
Left: $\Sigma_1 = \Sigma_2$. Right: $\Sigma_1 \neq \Sigma_2$

Comparison of Classification Methods

KNN, logistic regression, LDA, QDA

logistic regression and LDA

Different motivation but closely connected.

Suppose $p = 1$. $p_1(x)$ and $p_2(x) = 1 - p_1(x)$ are the prob. that $X = x$ belongs to class 1 and 2, respectively.

In LDA, we can show:

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

where c_0 and c_1 are functions of μ_1, μ_2, and σ^2.

In logistic regression:

$$\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 x$$

Both produce linear decision boundaries.

logistic regression and LDA

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

$$\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 x$$

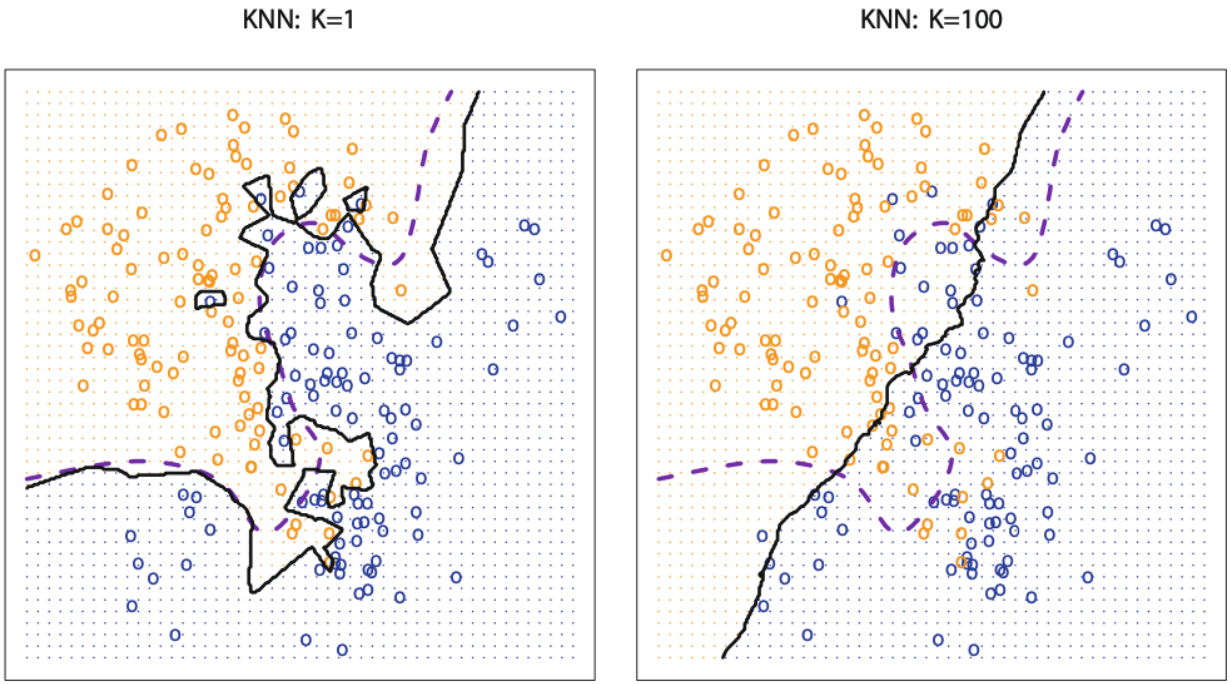
The only difference: β_0 and β_1 are estimated using maximum likelihood, whereas c_0 and c_1 are computed using estimated mean and variance from a normal distribution.
Same connection holds for $p > 1$.

logistic regression and LDA

Since logistic regression and LDA differ only in their fitting procedures, do they produce similar results?
Often, but not always.

LDA outperform logistic regression when Gaussian distribution assumption with a common variance-covariance matrix is reasonable.
Logistic regression outperforms LDA if Gaussian assumptions are not met.

KNN



- KNN is a completely non-parametric approach.
- No assumptions are made about the shape of the decision boundary.
- KNN dominates when the true decision boundary if highly non-linear.
- No info about predictor's importance; no coefficients estimations.

QDA

- Serves as a compromise between the non-parametric KNN method and the linear parametric LDA and logistic regression.

Textbook chapters

- ISLR: chapter 4: 4.1 - 4.5