Chapter 4: Classification

Yonghyun Kwon

Department of Mathematics, Korea Military Academy

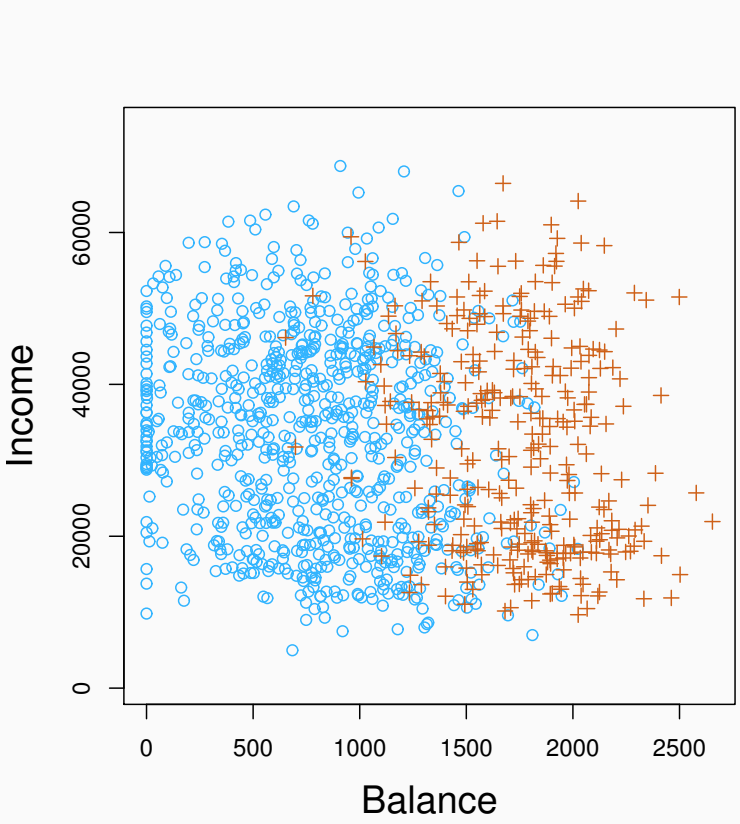
Why Not Linear Regression?

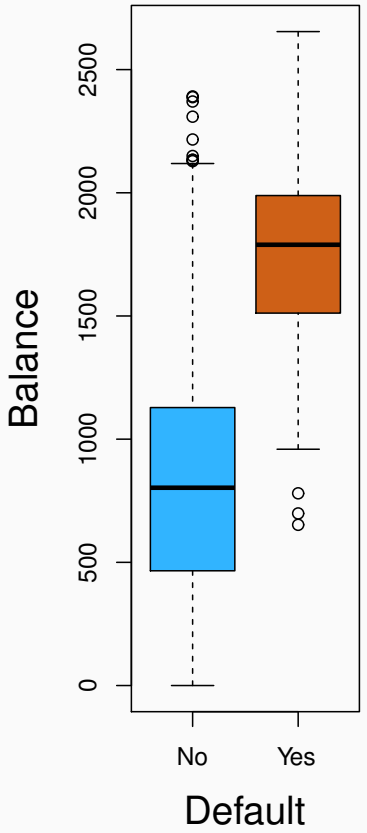
Classification

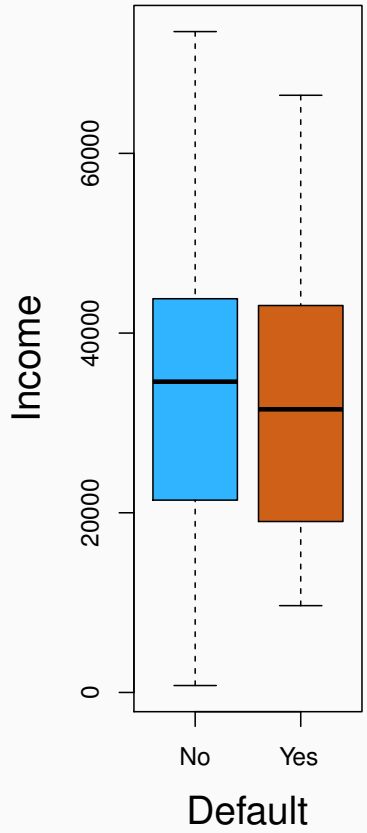
- Qualitative response Y takes values in an unordered set $\mathcal{C}$ (e.g., eye color $\in \{\text{brown, blue, green}\}$; email $\in \{\text{spam, ham}\}$).
- Given features X and qualitative $Y \in \mathcal{C}$, the task is to build a classifier $C(X) \in \mathcal{C}$.
- Often we want the *class probabilities* $\Pr(Y = c \mid X)$, not just the hard label.

When would calibrated probabilities matter more than hard labels?

Example: Credit Card Default







Visualization of `balance`, `income`, and `default`.

Can We Use Linear Regression?

Suppose we code the binary outcome `default` as

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

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

- In the population, $E(Y \mid X = x) = \Pr(Y = 1 \mid X = x)$.
- *Problem:* Linear regression can predict "probabilities" outside $[0, 1]$.
- *Remedy:* Logistic regression models $\Pr(Y = 1 \mid X)$ on $[0, 1]$.

Beyond Two Classes

- Now suppose the response Y can take three categories:

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

- Coding as $1, 2, 3$ imposes a false ordering and equal spacing.
- *Therefore, linear regression is not appropriate.*

Logistic Regression

Logistic Regression

Let $p(X) = \Pr(Y = 1 \mid X)$.

- Using `balance` to predict `default`, logistic regression assumes

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

- For any β_0, β_1, X, we always have $0 < p(X) < 1$.

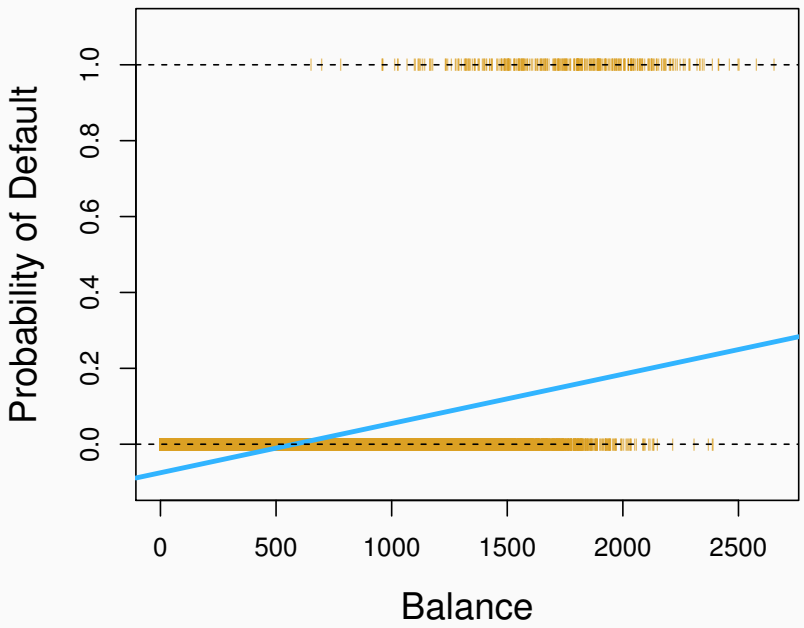
A rearrangement gives the *log-odds (logit)*:

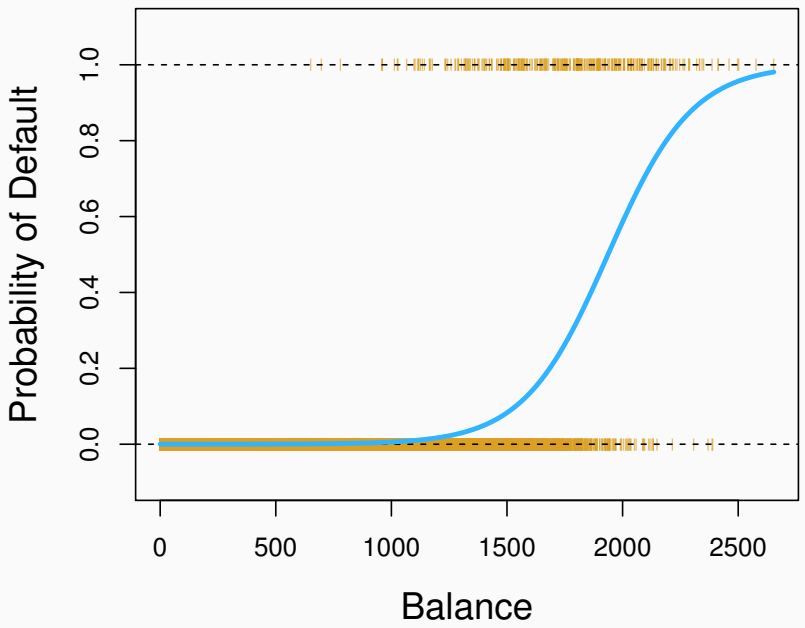
$$\log \left(\underbrace{\frac{p(X)}{1 - p(X)}}_{\text{Odds}} \right) = \beta_0 + \beta_1 X,$$

where $\log(= \ln)$ is the natural logarithm.

Why can't we just fit linear regression for a 0/1 response?

Linear versus Logistic Regression





Orange points mark `default` $= 1$. *Linear regression* does not estimate $\Pr(Y = 1 \mid X)$ well; *logistic regression* is better suited.

Maximum Likelihood Estimation(MLE)

We estimate β_0, β_1 by maximizing the *likelihood*

$$L(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} \{1 - p(x_i)\}.$$

Most statistical packages fit logistic regression via *maximum likelihood estimation(MLE)*; in R, use `glm`(..., family=binomial).

Coefficient	Estimate	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

With $\hat{\beta}_0 = -10.6513$ and $\hat{\beta}_1 = 0.0055$,

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

What is our estimated probability of `default` for someone with

- `balance` $= \$1000$?

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

- `balance` $= \$2000$?

$$\hat{p}(2000) =$$

Categorical Predictor

Using `student` as the predictor,

Coefficient	Estimate	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

$$\widehat{\Pr}(\texttt{default} = \text{Yes} \mid \texttt{student} = \text{Yes}) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1}}$$

$$\widehat{\Pr}(\texttt{default} = \text{Yes} \mid \texttt{student} = \text{No}) = \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}}$$

Odds Ratio(OR) for Categorical Predictor

$$\widehat{\Pr}(Y = 1 \mid X = 1) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1}}$$

$$\Longrightarrow \text{Odds}_{X=1} = \frac{\widehat{\Pr}(Y = 1 \mid X = 1)}{\widehat{\Pr}(Y = 0 \mid X = 1)} = e^{\hat{\beta}_0 + \hat{\beta}_1}$$

$$\widehat{\Pr}(Y = 1 \mid X = 0) = \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}}$$

$$\Longrightarrow \text{Odds}_{X=0} = \frac{\widehat{\Pr}(Y = 1 \mid X = 0)}{\widehat{\Pr}(Y = 0 \mid X = 0)} = e^{\hat{\beta}_0}$$

$$\boxed{\text{Odds Ratio} = \frac{\text{Odds}_{X=1}}{\text{Odds}_{X=0}} = e^{\hat{\beta}_1}}$$

Odds Ratio and contingency table

		Y		
		0	1	Total
X	0	n_{00}	n_{01}	$n_{00} + n_{01}$
	1	n_{10}	n_{11}	$n_{10} + n_{11}$
	Total	$n_{00} + n_{10}$	$n_{01} + n_{11}$	

$$\widehat{\Pr}(Y = 1 \mid X = 1) = \frac{n_{11}}{n_{10} + n_{11}} \Rightarrow \text{Odds}_{X=1} = \frac{\widehat{\Pr}(Y = 1 \mid X = 1)}{\widehat{\Pr}(Y = 0 \mid X = 1)} = \frac{n_{11}}{n_{10}}$$

$$\widehat{\Pr}(Y = 1 \mid X = 0) = \frac{n_{01}}{n_{00} + n_{01}} \Rightarrow \text{Odds}_{X=0} = \frac{\widehat{\Pr}(Y = 1 \mid X = 0)}{\widehat{\Pr}(Y = 0 \mid X = 0)} = \frac{n_{01}}{n_{00}}$$

$$\boxed{\text{OR} = \frac{\text{Odds}_{X=1}}{\text{Odds}_{X=0}} = \frac{n_{11} n_{00}}{n_{01} n_{10}}}, \quad SE(\log(\text{OR})) = \sqrt{\frac{1}{n_{00}} + \frac{1}{n_{01}} + \frac{1}{n_{10}} + \frac{1}{n_{11}}}$$

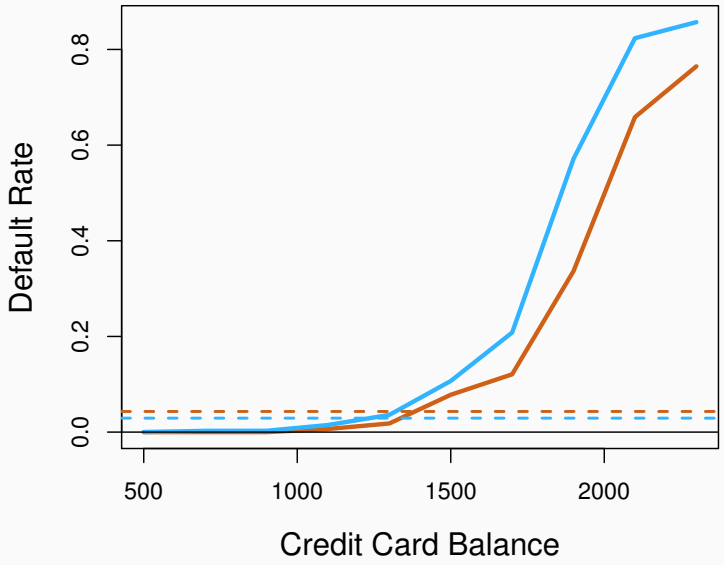
Logistic Regression with Several Variables

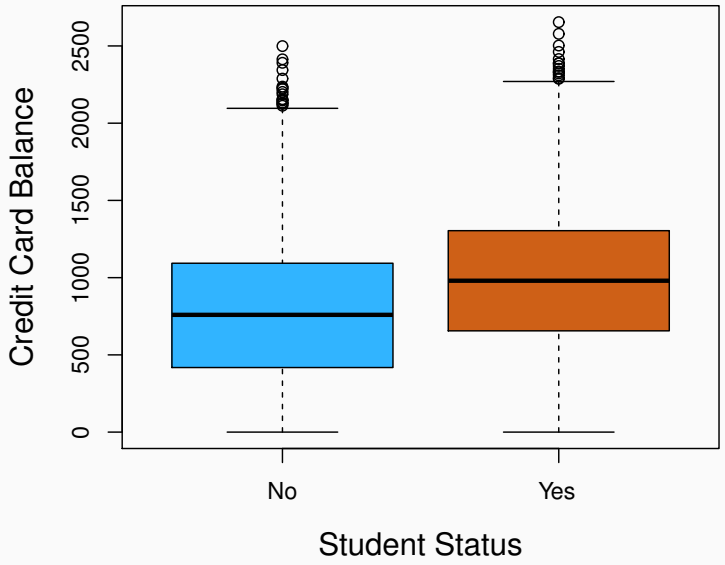
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0+\beta_1 X_1+\cdots+\beta_p X_p, \quad p(X) =$$

Coefficient	Estimate	Std. Err.	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 the coefficient for `student` negative here, whereas it was positive in the simple model?

Confounding





- Students tend to have higher `balance` than non-students.
- Marginally, students have a higher default rate.
- But at each fixed level of `balance`, students default less.
- Multiple logistic regression helps tease out *confounding*.

More than Two Classes

Logistic regression generalizes to K classes (*multinomial regression*):

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

- One linear function per class; only $K-1$ are needed.
- When fitting the model, one may set Kth class as the baseline: $\beta_{0K} = \beta_{1K} = \cdots \beta_{pK} = 0$.
- Estimated by maximizing the *multinomial log-likelihood (cross-entropy)*.
- Implemented in `glmnet` and other packages.

Likelihood-based inference for logistic regression

Note: Suppose $y_i \sim$ Bernoulli(π_i), where $\pi_i = \dfrac{1}{1+\exp\left(-(\beta_0+\beta_1 x_i)\right)}$ for $i = 1, \cdots, n$. The likelihood and the log-likelihood of $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ are

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

$$\ell(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i \log \pi_i + (1-y_i)\log(1-\pi_i)\right).$$

Using $\dfrac{\partial \pi_i}{\partial \boldsymbol{\beta}} = \pi_i(1-\pi_i)\begin{pmatrix}1\\ x_i\end{pmatrix}$, the derivatives of log-likelihood are

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}(y_i-\pi_i)\begin{pmatrix}1\\ x_i\end{pmatrix}, \quad \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} = -\sum_{i=1}^{n}\pi_i(1-\pi_i)\begin{pmatrix}1 & x_i\\ x_i & x_i^2\end{pmatrix}.$$

The MLE $\hat{\boldsymbol{\beta}}$ solves $\frac{\partial \ell(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = 0$. Also, asymptotic properties of the MLE gives

$$\hat{\boldsymbol{\beta}} \dot{\sim} N\left(\boldsymbol{\beta}, I^{-1}\right), \quad \text{where } I = -E\left[\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top}\right] \text{ is the information matrix.}$$

Note: *For inference on $\boldsymbol{\beta}$, writing $\hat{\pi}_i = \pi_i(\hat{\boldsymbol{\beta}})$, we may use*

$$\hat{\boldsymbol{\beta}} \dot{\sim} N\left(\boldsymbol{\beta}, I^{-1}(\hat{\beta})\right), \quad \text{where } I(\hat{\beta}) = \sum_{i=1}^{n} \hat{\pi}_i(1-\hat{\pi}_i) \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}.$$

For instance, to test $H_0 : \beta_1 = 0$, $H_A : \beta_1 \neq 0$, one can consider Wald-type test:

$$Z = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \overset{H_0}{\sim} N(0,1),$$

or Likelihood ratio test(LRT):

$$\begin{aligned} \text{LR statistic} &= 2[\text{(Full model log-lik.) - (Reduced model log-lik.)}] \\ &= 2\left[\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\beta}_0^{(H_0)}, 0)\right] \\ &= 2\sum_{i=1}^{n}\left[y_i \log\frac{\hat{\pi}_i}{\hat{\pi}_0} + (1-y_i)\log\frac{1-\hat{\pi}_i}{1-\hat{\pi}_0}\right] \overset{H_0}{\sim} \chi^2(1). \end{aligned}$$

where $\hat{\pi}_0 = \dfrac{1}{1+\exp\left(-\hat{\beta}_0^{(H_0)}\right)}$ *under the reduced model* $H_0 : \beta_1 = 0$.

Generative Models for Classification

Discriminant Analysis: Modeling $X|Y$

- Strategy: Model the distribution of predictors X within each class Y and use *Bayes' theorem* to obtain $\Pr(Y \mid X)$.
- With class-conditional normals, we obtain *LDA(Linear Discriminant Analysis)* or *QDA(Quadratic Discriminant Analysis)* .
- Other class-conditional distributions are possible, but we focus on normals here.

Bayes Theorem

$$\begin{aligned}\Pr(Y = k \mid X = x) &= \frac{\Pr(X = x \mid Y = k)\ \Pr(Y = k)}{\Pr(X = x)} \\ &= \frac{\Pr(X = x \mid Y = k)\ \Pr(Y = k)}{\sum_{\ell=1}^{K} \Pr(X = x \mid Y = \ell)\ \Pr(Y = \ell)}.\end{aligned}$$

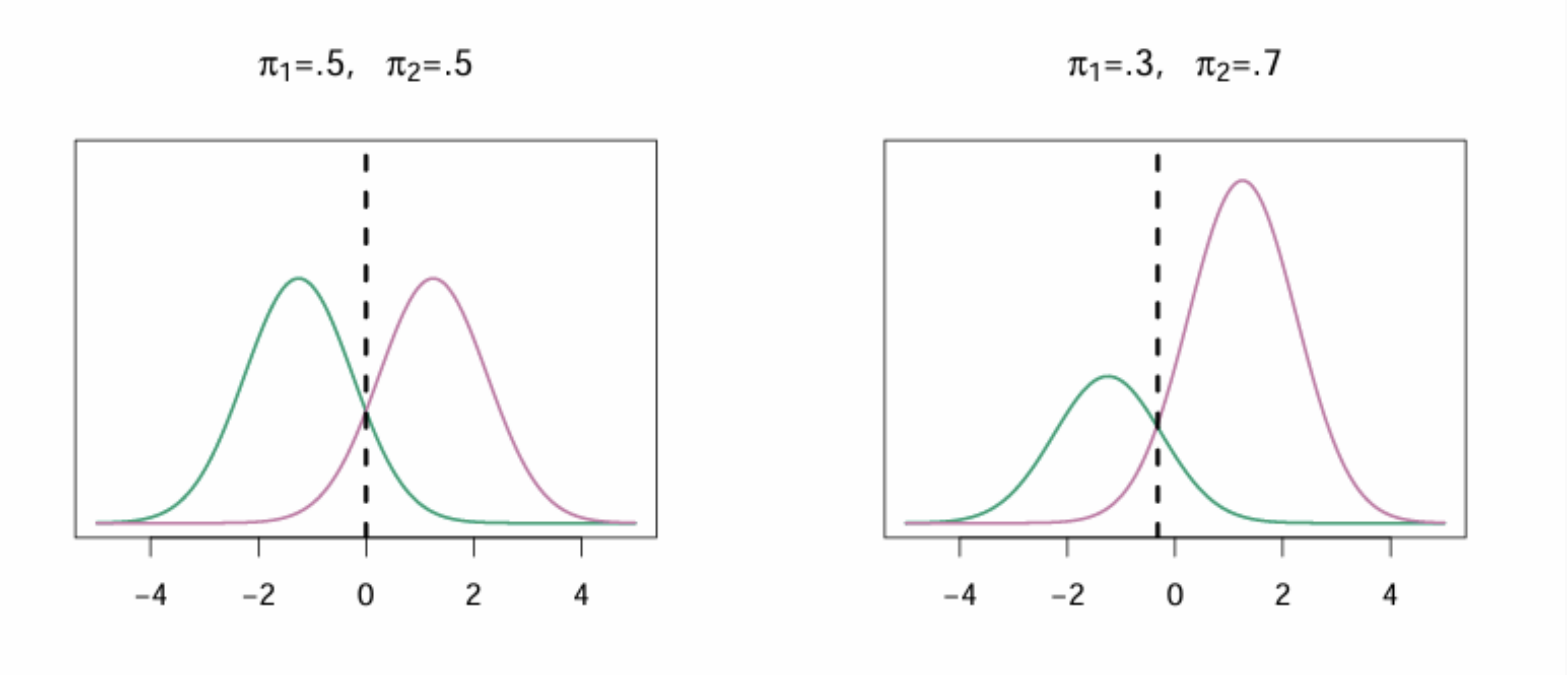
Bayes Theorem for Classification

$$\begin{aligned}\Pr(Y = k \mid X = x) &= \frac{\Pr(X = x \mid Y = k)\,\Pr(Y = k)}{\sum_{\ell=1}^{K} \Pr(X = x \mid Y = \ell)\,\Pr(Y = \ell)} \\ &= \frac{\pi_k\, f_k(x)}{\sum_{\ell=1}^{K} \pi_\ell\, f_\ell(x)}.\end{aligned}$$

- $f_k(x) = \Pr(X = x \mid Y = k)$: density of X in class k.
- $\pi_k = \Pr(Y = k)$: prior (marginal) class probability.

Decision Rule: Highest (Prior-Weighted) Density

Classify x to the class with the largest $\pi_k f_k(x)$.



Why Use LDA?

- When classes are well-separated, logistic regression parameters can be unstable; LDA is more stable.
- For small n and class-conditional normal X, LDA tends to be more stable than logistic regression.
- With $K > 2$ classes, LDA also yields low-dimensional projections for visualization.

Linear Discriminant Analysis when $p = 1$

In class k, the Gaussian density is

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

Here μ_k is the class mean and σ_k^2 the class variance. Assume a common variance, i.e., $\sigma_k = \sigma$ for all k.

Plugging this into Bayes' theorem and denoting $p_k(x) = \Pr(Y = k \mid X = x)$, we obtain

$$p_k(x) = \frac{\pi_k \dfrac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\dfrac{1}{2} \left(\dfrac{x - \mu_k}{\sigma} \right)^2 \right\}}{\displaystyle\sum_{\ell=1}^{K} \pi_\ell \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{1}{2} \left(\frac{x - \mu_\ell}{\sigma} \right)^2 \right\}}.$$

Discriminant functions

To classify an observation with value $X = x$, compare the values $p_k(x)$. Taking logs and discarding terms that do not depend on k, assign x to the class with the largest *discriminant score*:

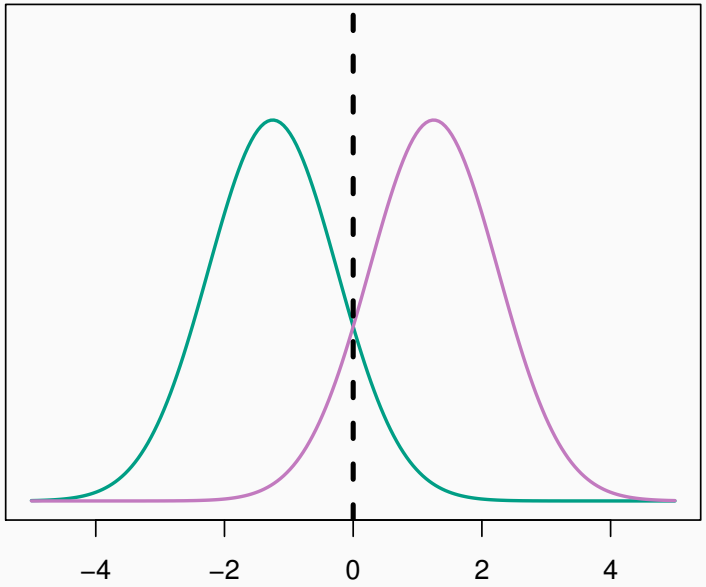
$$\delta_k(x) \;=\; x\,\frac{\mu_k}{\sigma^2} \;-\; \frac{\mu_k^2}{2\sigma^2} \;+\; \log(\pi_k).$$

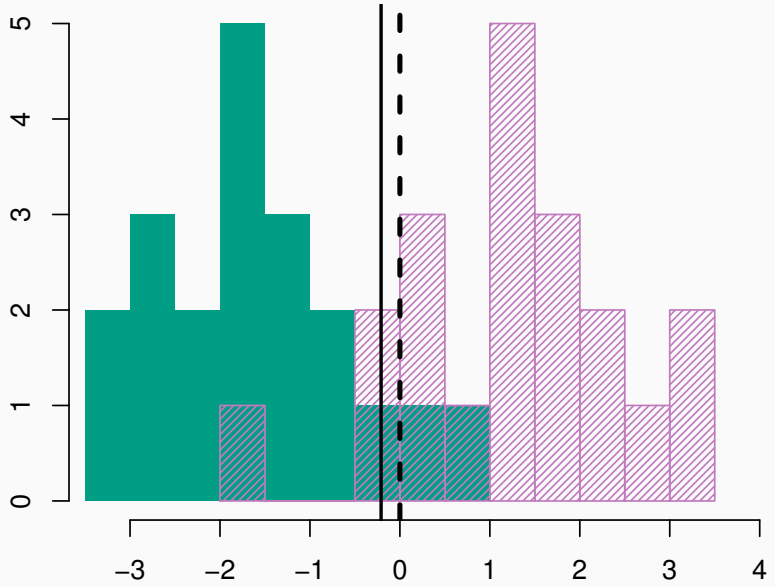
- $\delta_k(x)$ is linear in $x \Rightarrow$ Linear Discriminant Analysis.
- For $K = 2$ with $\pi_1 = \pi_2 = 0.5$, the decision boundary solves $\delta_1(x) = \delta_2(x)$, yielding

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

(Quick exercise to verify.)

Example and Plug-in Classification





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

Typically, we do not know these parameters; we just have the training data. In that case we estimate the parameters and plug them into the rule.

Estimating the Parameters

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

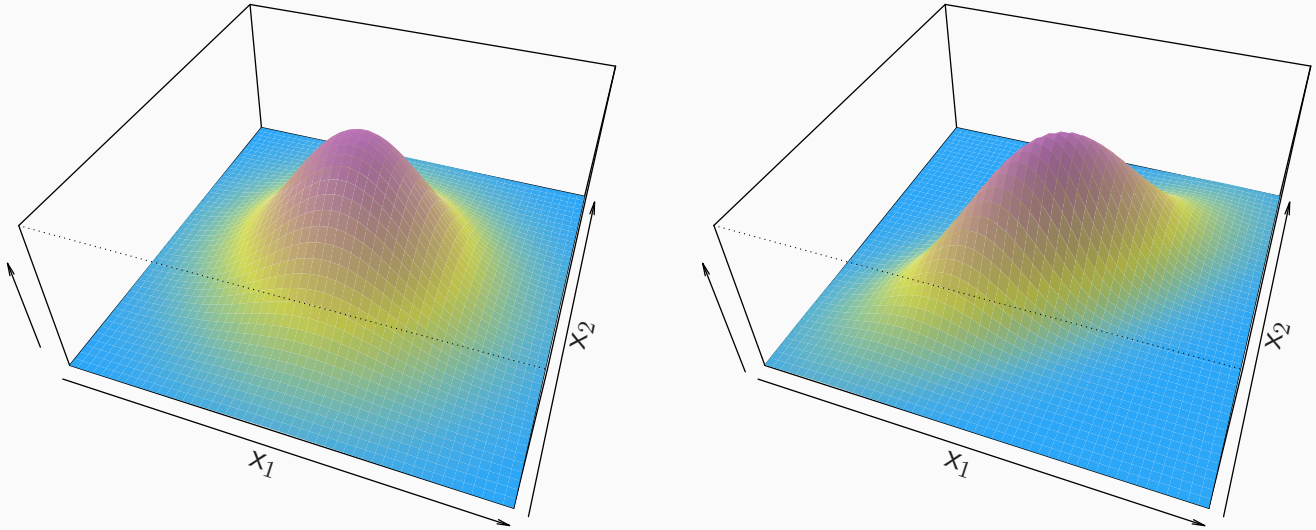
$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i,$$

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

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

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

Linear Discriminant Analysis when $p > 1$



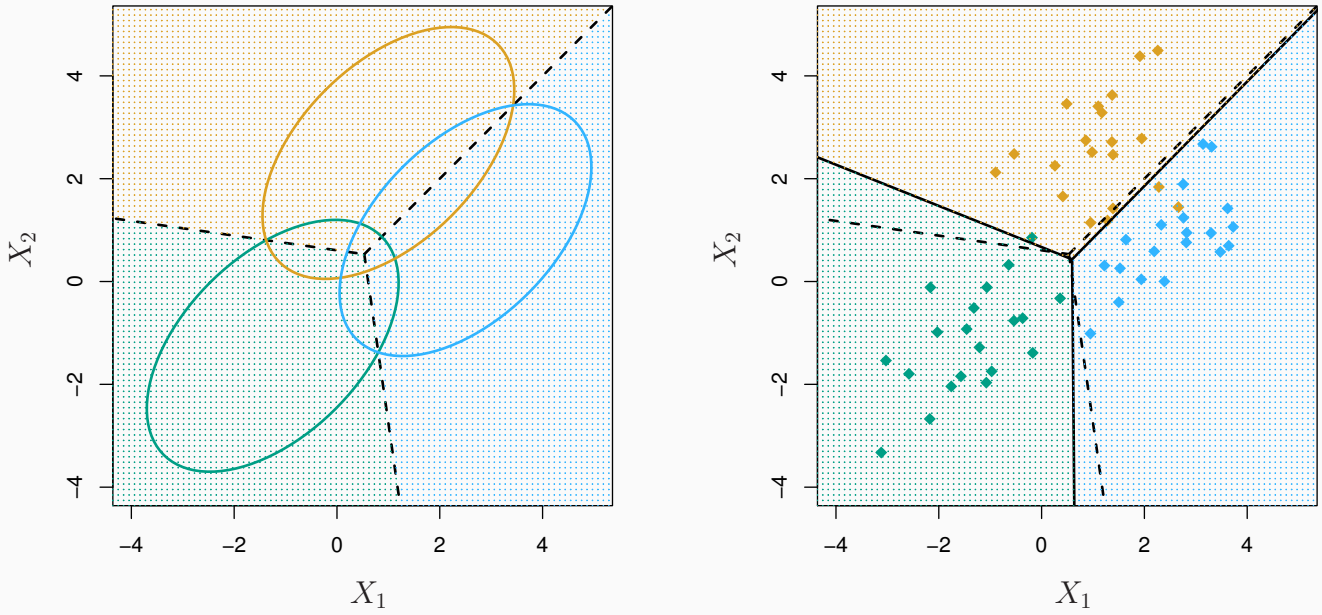
$$\text{Density: } f(x) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} \exp\Big\{ -\tfrac{1}{2}(x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu)\Big\}.$$

$$\text{Discriminant function: } \delta_k(x) = x^{\mathsf{T}}\Sigma^{-1}\mu_k - \tfrac{1}{2}\,\mu_k^{\mathsf{T}}\Sigma^{-1}\mu_k + \log \pi_k.$$

Despite its complex form,

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

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



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

The dashed lines are known as the Bayes decision boundaries. Were they known, they would yield the fewest misclassification errors, among all possible classifiers.

From $\delta_k(x)$ to Probabilities

Once we have estimates $\hat{\delta}_k(x)$, we obtain estimated class probabilities via

$$\Pr(\hat{Y} = k \mid X = x) = \frac{\exp\{\hat{\delta}_k(x)\}}{\sum_{\ell=1}^{K} \exp\{\hat{\delta}_\ell(x)\}}.$$

Thus, classifying to the largest $\hat{\delta}_k(x)$ is the same as classifying to the largest $\Pr(\hat{Y} = k \mid X = x)$.

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

Fisher's LDA

Note: *LDA can be geometrically explained without a distribution assumption. Let* $\boldsymbol{x}_{ik} \in \mathbb{R}^p$ *be the* i*th observation*$(i = 1, \cdots, n_k)$ *from* k*th class* $(k = 1, \cdots, K)$.

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\boldsymbol{x}_{ik} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{ik} - \bar{\boldsymbol{x}})^\top$$

$$= \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\boldsymbol{x}_{ik} - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_{ik} - \bar{\boldsymbol{x}}_k)^\top}_{=W} + \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}})^\top}_{=B},$$

where $\bar{\boldsymbol{x}}_k$ *is the sample mean of class* k, $\bar{\boldsymbol{x}}$ *is the grand sample mean,* W *is the within-group variance, and* B *is the between-group variance. LDA direction* $\boldsymbol{u}$ *(which is orthogonal to separating hyperplane) maximizes the between-group variability while minimizing the within-group variability:*

$$\max_{\boldsymbol{u}\in\mathbb{R}^p} \frac{\boldsymbol{u}^\top B\boldsymbol{u}}{\boldsymbol{u}^\top W\boldsymbol{u}} \tag{1}$$

Fisher's LDA

Note: Substitute $\boldsymbol{x} = W^{1/2}\boldsymbol{u}$. *Then* (1) *becomes* $\dfrac{\boldsymbol{x}^\top W^{-1/2} B W^{-1/2} \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}$. *By Cauchy-Schwarz inequality, we have*

$$\begin{aligned} \boldsymbol{x}^\top W^{-1/2} B W^{-1/2} \boldsymbol{x} &\leq \|\boldsymbol{x}\| \left\| W^{-1/2} B W^{-1/2} \boldsymbol{x} \right\| \\ &\leq \left\| W^{-1/2} B W^{-1/2} \right\| \|\boldsymbol{x}\|^2 , \end{aligned}$$

where the equality holds if $\boldsymbol{x}$ *is the eigenvector of* $W^{-1/2} B W^{-1/2}$ *that corresponds to the maximum eigenvalue of* $W^{-1/2} B W^{-1/2} (= \lambda_{\max})$.

$$W^{-1/2} B W^{-1/2} \boldsymbol{x} = \lambda_{\max} \boldsymbol{x} \iff W^{-1} B \boldsymbol{u} = \lambda_{\max} \boldsymbol{u}$$

Therefore, the solution to (1), $\hat{\boldsymbol{u}}$, *is the eigenvector of* $W^{-1}B$ *that corresponds to the maximum eigenvalue of* $W^{-1}B$. *In a special case where* $K = 2$, *it can be shown that* $\hat{u} \propto W^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$, *which is orthogonal to the separating hyperplane of Bayes LDA.*

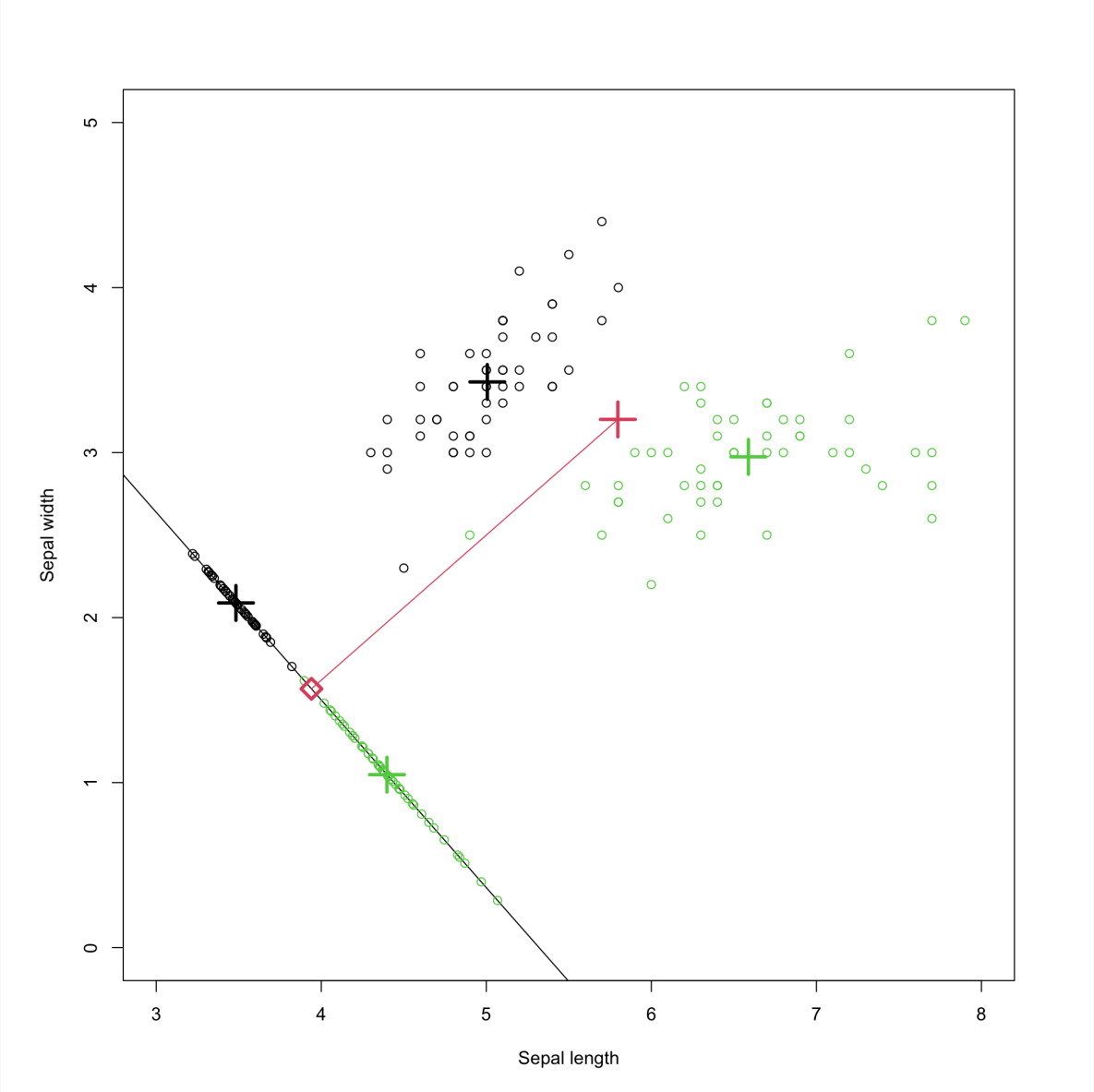


Figure 1: Fisher's LDA on iris data

LDA on `Credit` Data: Confusion Matrix

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

The overall training error is $(23 + 252)/10000 = 2.75\%$.

- Class-specific error rates:
  - No (True) $\rightarrow$ Yes (Predicted) : $23/9667 \approx 0.2\%$.
  - Yes (True) $\rightarrow$ No (Predicted) : $252/333 \approx 75.7\%$.

If we always predict No, overall training error is

Types of Errors & Thresholding

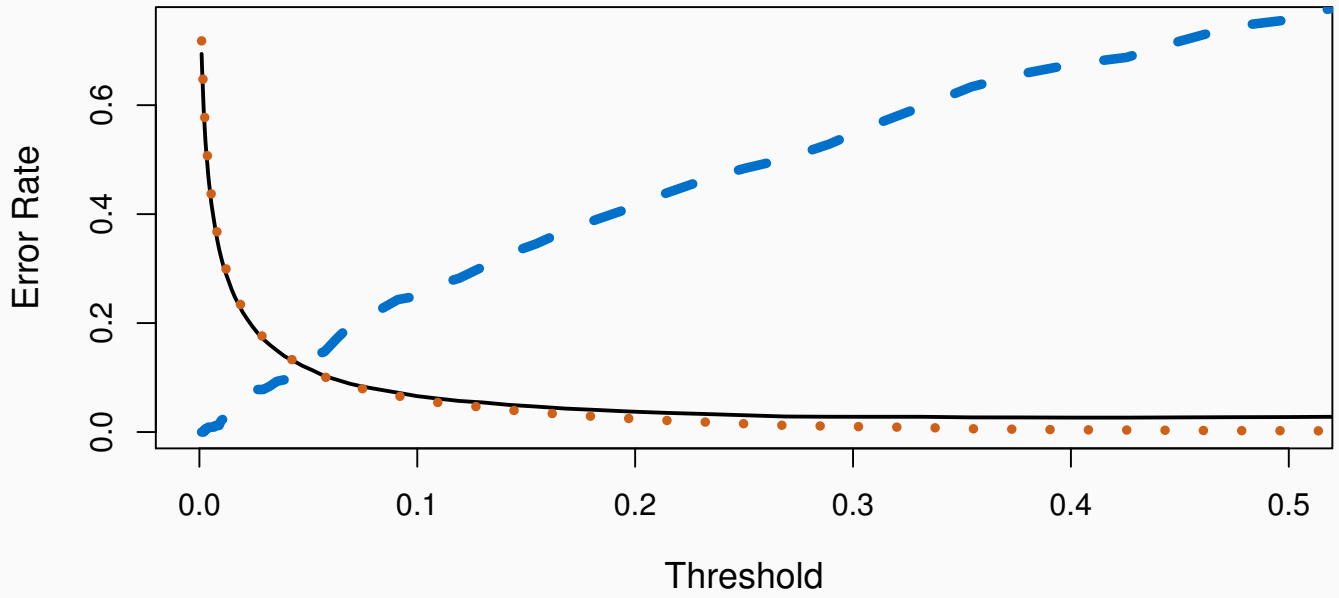
- *False positive rate (FPR)*: fraction of true negatives classified as positive ($\approx 0.2\%$ above).
- *False negative rate (FNR)*: fraction of true positives classified as negative ($\approx 75.7\%$ above).
- *Sensitivity (True Positive Rate)*: $1 - \text{FNR}$.
- *Specificity (True Negative Rate)*: $1 - \text{FPR}$.
- Classification rule:

$$\widehat{\Pr}(\text{Default} = \text{Yes} \mid \text{Balance}, \text{Student}) \geq \textit{threshold}.$$

- Default threshold $= 0.5$; moving the threshold trades off FPR and FNR

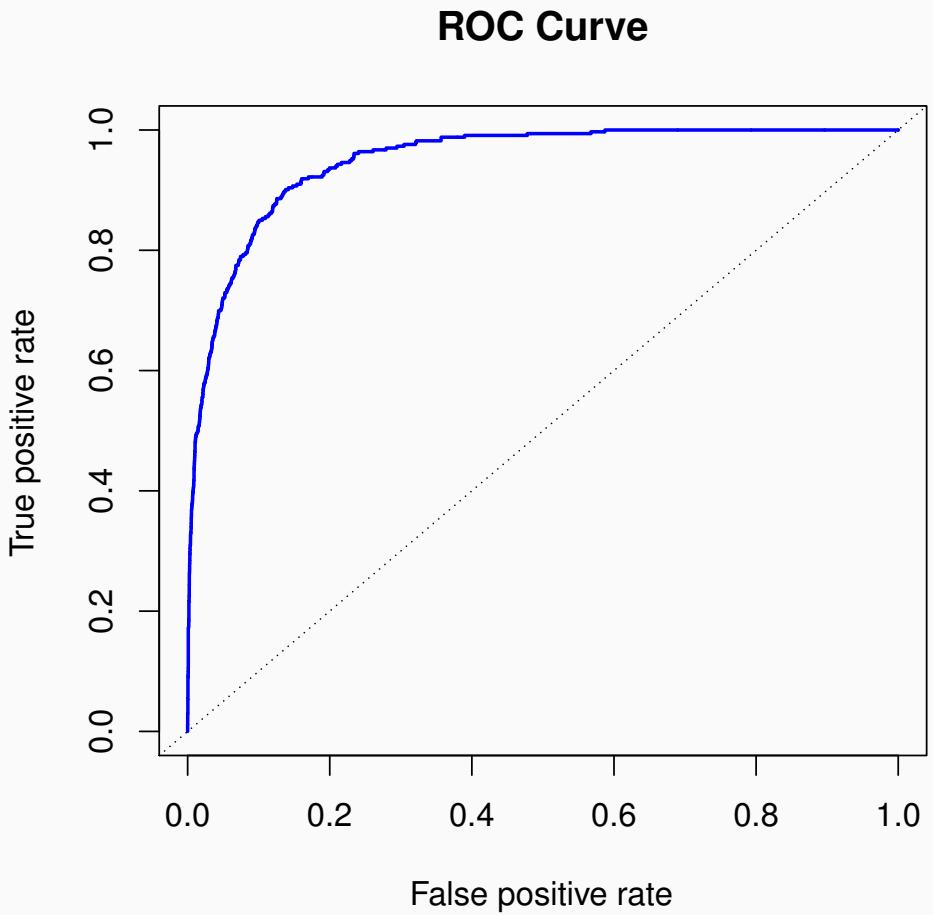
Varying the Threshold

Lowering the threshold generally (increases / decreases) FNR and (increases / decreases) FPR.



Overall error (black solid), FNR (blue dashed), and FPR (orange dotted).

- *ROC curve*: plots *True Positive Rate* vs. *False Positive Rate* over all thresholds.
- *AUC* (Area Under the Curve): scalar summary of performance; higher is better.



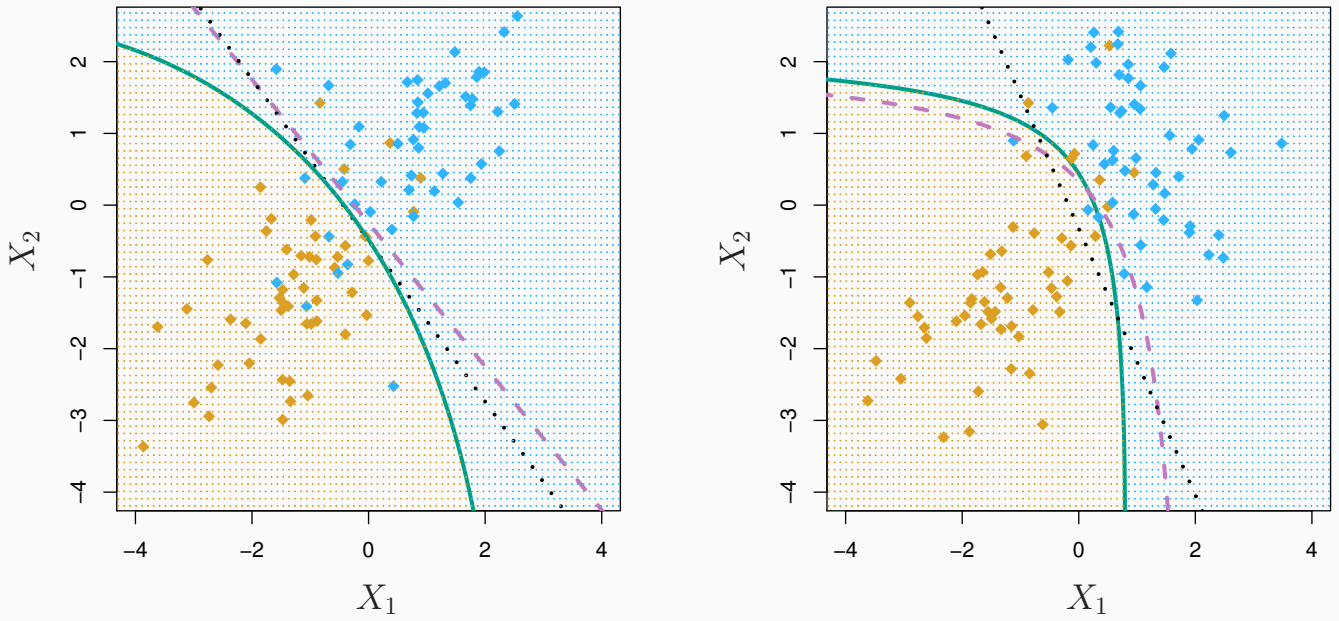
Other Forms of Discriminant Analysis

General form:

$$\Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(x)}.$$

- With Gaussian $f_k(x)$ and a common covariance Σ: *Linear Discriminant Analysis*.
- With Gaussian $f_k(x)$ but class-specific covariances Σ_k: *Quadratic Discriminant Analysis*.
- With conditional independence model $f_k(x) = \prod_{j=1}^{p} f_{kj}(x_j)$: *Naive Bayes*.

Quadratic Discriminant Analysis (QDA)



$$\delta_k(x) = -\tfrac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k) + \log \pi_k - \tfrac{1}{2}\log|\Sigma_k|.$$

Class-specific covariances Σ_k yield quadratic decision boundaries.

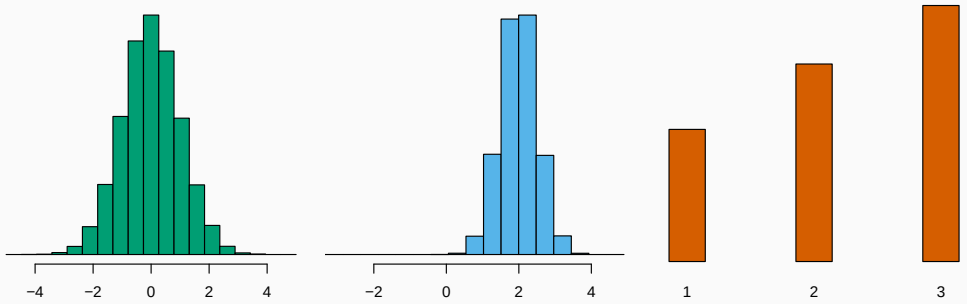
Naive Bayes in More Detail

- Naive Bayes assumes

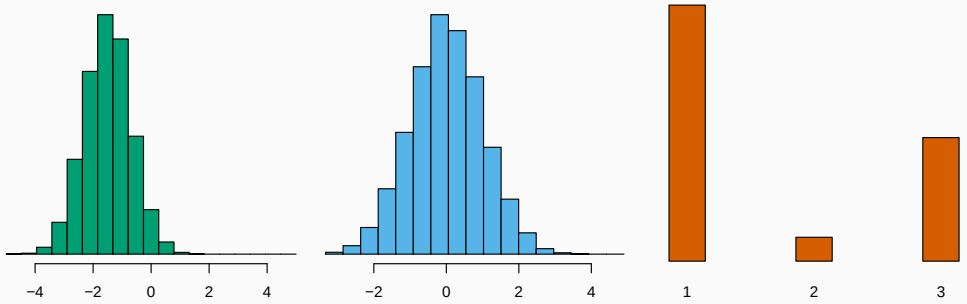
$$f_k(x) = f_{k1}(x_1) \cdot f_{k2}(x_2) \cdots f_{kp}(x_p),$$

  i.e. features are independent given the class.
- Motivation:
  - Modeling high-dimensional densities is hard; easier to work with one-dimensional marginals.
  - Handles *mixed features*: quantitative (e.g. Gaussian or histogram) and qualitative (category frequencies).
- Despite its unrealistic assumption, the reduced variance often yields strong classification performance.

Naive Bayes: Toy Example



Density estimates for class $k = 1$



Density estimates for class $k = 2$

$$x^* = (0.4,\ 1.5,\ 1)^\top,$$

$$\hat{\pi}_1 = \hat{\pi}_2 = 0.5$$

$$\hat{f}_{11}(0.4) = 0.368$$

$$\hat{f}_{12}(1.5) = 0.484$$

$$\hat{f}_{13}(1) = 0.226$$

$$\hat{f}_{21}(0.4) = 0.030$$

$$\hat{f}_{22}(1.5) = 0.130$$

$$\hat{f}_{23}(1) = 0.616$$

$$\Pr(Y = 1 \mid X = x^*) = 0.944, \quad \Pr(Y = 2 \mid X = x^*) = 0.056$$

Naive Bayes and GAMs

$$\log \frac{\Pr(Y=k \mid X=x)}{\Pr(Y=K \mid X=x)} = \log \frac{\pi_k f_k(x)}{\pi_K f_K(x)}$$

$$= \log \frac{\pi_k \prod_{j=1}^{p} f_{kj}(x_j)}{\pi_K \prod_{j=1}^{p} f_{Kj}(x_j)}$$

$$= \log \frac{\pi_k}{\pi_K} + \sum_{j=1}^{p} \log \frac{f_{kj}(x_j)}{f_{Kj}(x_j)}$$

$$= a_k + \sum_{j=1}^{p} g_{kj}(x_j),$$

where $a_k = \log\left(\frac{\pi_k}{\pi_K}\right)$, $g_{kj}(x_j) = \log\left(\frac{f_{kj}(x_j)}{f_{Kj}(x_j)}\right)$.

The Naive Bayes takes the form of a *generalized additive model (GAM)*.

Logistic Regression vs. LDA

- For two classes, LDA implies

$$\log \frac{p_1(x)}{1 - p_1(x)} = c_0 + c_1 x_1 + \cdots + c_p x_p,$$

  same functional form as logistic regression.
- Difference lies in estimation:
  - Logistic regression uses conditional likelihood based on $\Pr(Y \mid X)$.
  - LDA uses full likelihood based on $\Pr(X, Y)$.
- In practice, results are often quite similar.

Summary

- Logistic regression: widely used, especially for $K = 2$.
- LDA: effective when n is small, classes well separated, and Gaussian assumptions reasonable.
- QDA: more flexible boundaries; useful when class covariances differ.
- Naive Bayes: good choice when p is large.

Exercises

1. Prove that

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{2}$$

is equivalent to

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}. \tag{3}$$

Exercises

2. Suppose we collect data for a group of students in a statistics class with variables $X_1 =$ hours studied, $X_2 =$ undergrad GPA, and $Y =$ receive an A. We fit a logistic regression and produce estimated coefficient $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

(a) Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.

(b) How many hours whould the student in part (a) need to study to have a 50% chance of getting an A in the class?

Exercises

3. This problem has to do with odds.

   (a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?

   (b) Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will default?

Exercises

4. (Linear Discriminant Analysis) Show that finding $k \in \{1, \cdots, K\}$ that maximizes

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{\ell=1}^{K} \pi_\ell \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_\ell)^2\right)} \tag{4}$$

is equivalent to finding $k \in \{1, \cdots, K\}$ that maximizes

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k). \tag{5}$$

Also, when $K = 2$, show that this is equivalent to choosing $k = 1$ if

$$\frac{(\mu_2 - \mu_1)}{\sigma^2}\left(x - \frac{\mu_1 + \mu_2}{2}\right) < \log\left(\frac{\pi_1}{\pi_2}\right). \tag{6}$$