Lecture 7: Classification

Statistical Learning (BST 263)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

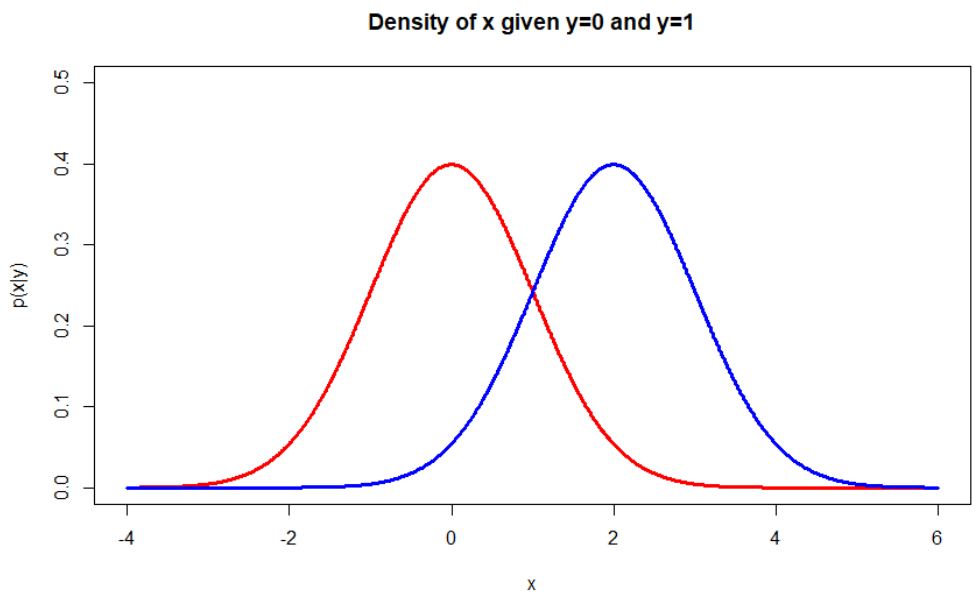
Outline

Example: ELISA test

- The enzyme-linked immunosorbent assay (ELISA) is a chemical test for the presence of an antigen of interest.

- Widely used for the detection of diseases such as malaria, HIV, West Nile virus, and celiac disease.

- We will use this to illustrate some concepts in this lecture.
- See R code file: *classification.r* under Files/Code.



Outline

Loss functions

- If our goal is to get the correct answer as often as possible, then we want to construct $\hat{f}$ to minimize the test error rate.

- However, what if certain types of errors are more costly than others?
- For instance, if we are detecting cancer, then a false positive may result in unnecessary additional tests, whereas a false negative may result in loss of life due to lack of treatment.

- To handle such situations, we use a *loss function* $L(\hat{y}, y)$ to quantify the cost of predicting $\hat{y}$ when the actual class is y.

- Example: The *0-1 loss* is $L(\hat{y}, y) = \mathrm{I}(\hat{y} \neq y)$.

Decision theory

- The *expected loss* is $\mathrm{E}(L(\hat{Y}_0, Y_0))$, where (X_0, Y_0) is a random data point distributed according to the true data generating process, and $\hat{Y}_0 = \hat{f}(X_0)$ is the predicted outcome value.

- Example: The expected 0-1 loss is the test error rate.

- The *decision theory* approach to choosing $\hat{f}$ is to try to minimize expected loss.
- This approach applies to both regression and classification.

- Regression example:
  - The *square loss* is $L(\hat{y}, y) = (\hat{y} - y)^2$.
  - The expected square loss is the test MSE.

Outline

Confusion matrix

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

TABLE 4.4. *A confusion matrix compares the LDA predictions to the true default statuses for the* 10,000 *training observations in the* `Default` *data set. Elements on the diagonal of the matrix represent individuals whose default statuses were correctly predicted, while off-diagonal elements represent individuals that were misclassified. LDA made incorrect predictions for* 23 *individuals who did not default and for* 252 *individuals who did default.*

Confusion matrix

		Actual	
		0 (negative)	1 (positive)
Predicted	0 (negative)	true neg. (TN)	false neg. (FN)
	1 (positive)	false pos. (FP)	true pos. (TP)

- In the TP, FP, TN, FN terminology:
  - "True" / "False" = prediction is correct/incorrect,
  - "Positive" / "Negative" = predicted class is positive/negative.

- *False positive rate* (FPR) = FP / (FP + TN)
  - Fraction of actual negatives that were predicted to be positive.
  - *Specificity* = 1− FPR

- *True positive rate* (TPR) = TP / (TP + FN)
  - Fraction of actual positives that were predicted to be positive.
  - *Sensitivity* = TPR

Confusion matrix example: ELISA test

(See R code file: *classification.r*.)

Expected loss in binary classification setting

Loss matrix:

		Actual	
		0 (negative)	1 (positive)
Predicted	0 (negative)	$L(0,0)$	$L(0,1)$
	1 (positive)	$L(1,0)$	$L(1,1)$

Expected loss:

$$E(L(\hat{Y}_0, Y_0)) = L(0,0)P_{\mathsf{TN}} + L(0,1)P_{\mathsf{FN}} + L(1,0)P_{\mathsf{FP}} + L(1,1)P_{\mathsf{TP}}$$

where $P_{\mathsf{TN}} = \mathbb{P}(\hat{Y}_0 = 0,\, Y_0 = 0)$, etc.

Outline

ROC curve

- Binary classification (e.g., $y \in \{0, 1\}$).
- Most binary classification methods have a detection threshold (i.e., a cutoff) that can be adjusted.
- ELISA example: $f(x) = \mathrm{I}(x > \text{cutoff})$.
- KNN: $\hat{f}(x) = \mathrm{I}\big(\hat{p}_1(x) > \text{cutoff}\big)$.
- Can choose cutoff to try to maximize performance (i.e., to minimize expected loss).

ROC curve example

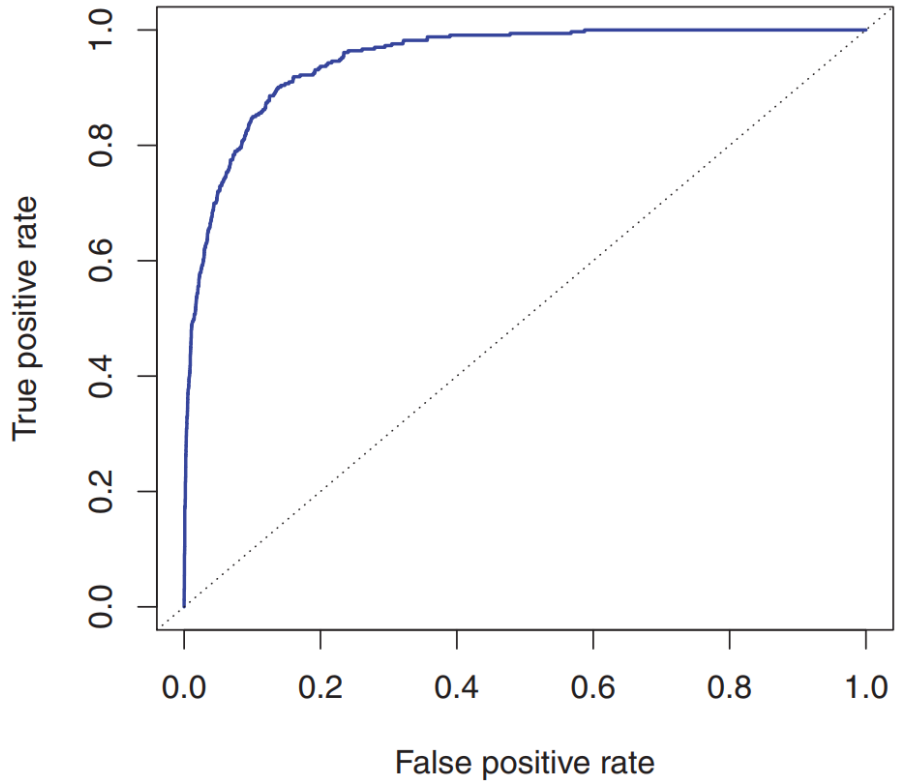


FIGURE 4.8. *A ROC curve for the LDA classifier on the* `Default` *data. It traces out two types of error as we vary the threshold value for the posterior probability of default. The actual thresholds are not shown. The true positive rate is the sensitivity: the fraction of defaulters that are correctly identified, using a given threshold value. The false positive rate is 1-specificity: the fraction of non-defaulters that we classify incorrectly as defaulters, using that same threshold value. The ideal ROC curve hugs the top left corner, indicating a high true positive rate and a low false positive rate. The dotted line represents the "no information" classifier; this is what we would expect if student status and credit card balance are not associated with probability of default.*

ROC curve example: ELISA test

(See R code file: *classification.r*.)

ROC curve

- ROC curves are often used to compare classification methods.
- The area under the ROC curve ("AUC" or "AUROC") summarizes the ROC curve in a single number.
- If the loss function is known, though, we should compare expected loss rather than ROC curves or AUC.

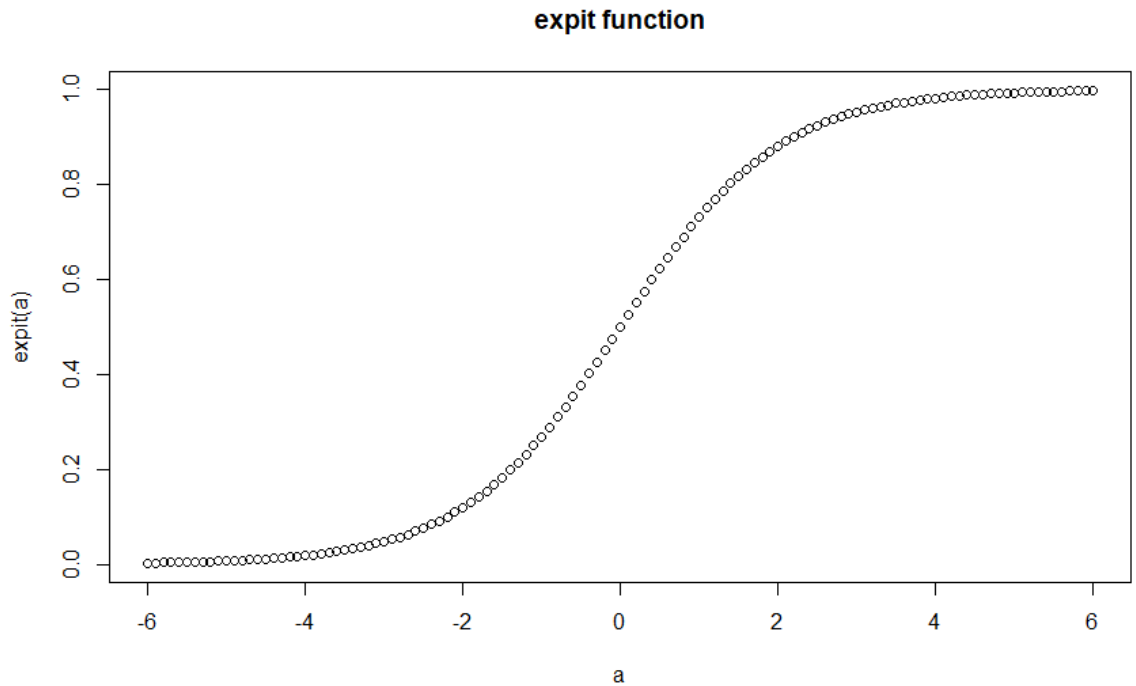
Outline

Logistic regression

- Logistic regression is a binary classification method.

- For $a \in \mathbb{R}$, the *expit function* (a.k.a. *logistic function*) is

$$\text{expit}(a) = \frac{e^a}{e^a + 1} = \frac{1}{1 + e^{-a}}.$$

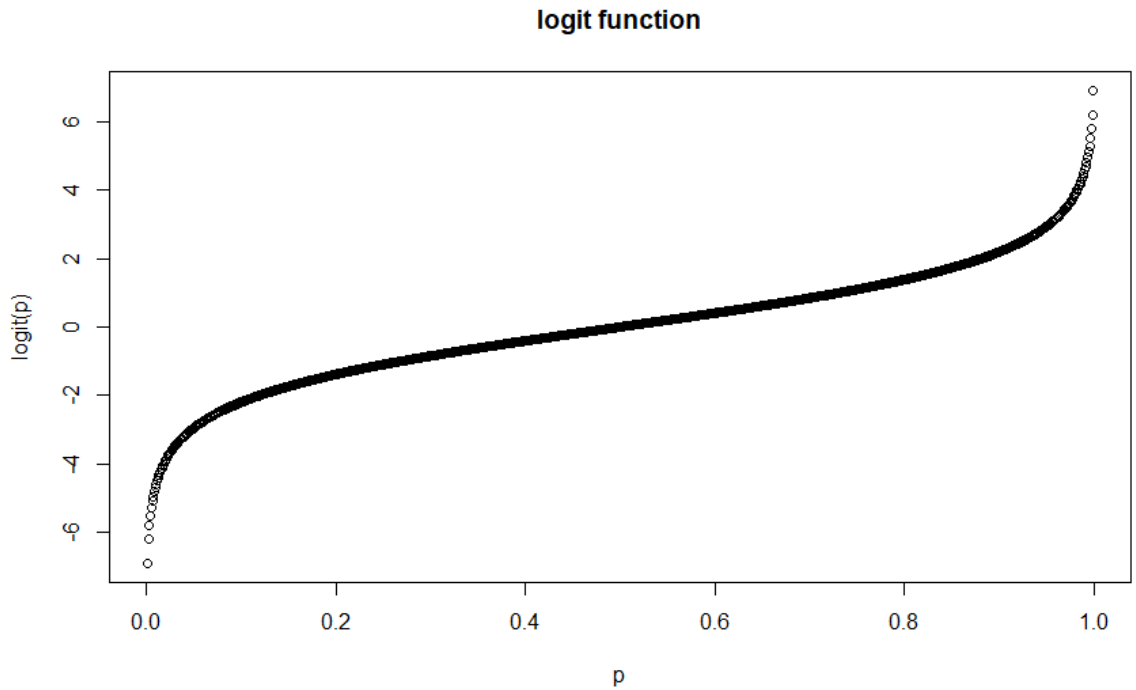


(See R code file: *classification.r*.)

Logistic regression: expit and logit

- For $p \in (0, 1)$, the *logit function* is

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right).$$



- logit and expit are inverses, i.e.,

$$\text{logit}(\text{expit}(a)) = a \quad \text{and} \quad \text{expit}(\text{logit}(p)) = p.$$

Logistic regression model

- Training data $(x_1, y_1), \ldots, (x_n, y_n)$.
- Inputs x_i are mapped to feature vectors $\varphi(x_i) \in \mathbb{R}^p$.
- Outcomes y_i are binary, $y_i \in \{0, 1\}$.

- The outcomes y_i are modeled as random variables

$$Y_i \sim \text{Bernoulli}\big(\pi_\beta(x_i)\big)$$

where $\pi_\beta(x_i) = \text{expit}\big(\varphi(x_i)^\intercal \beta\big)$. Parameters: $\beta \in \mathbb{R}^p$.

- In other words,

$$\mathbb{P}(Y_i = 1 \mid \beta, x_i) = \pi_\beta(x_i) = \frac{1}{1 + \exp(-\varphi(x_i)^\intercal \beta)}.$$

- Equivalently, $\text{logit}\big(\mathbb{P}(Y_i = 1 \mid \beta, x_i)\big) = \varphi(x_i)^\intercal \beta$.

Estimation for logistic regression

- Likelihood function:

$$p(y_{1:n} \mid \beta, x_{1:n}) = \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i \mid \beta, x_i)$$
$$= \prod_{i=1}^{n} \pi_\beta(x_i)^{y_i} \big(1 - \pi_\beta(x_i)\big)^{1-y_i}.$$

- Notation: $x_{1:n} = (x_1, \ldots, x_n)$.

- Can't analytically maximize likelihood with respect to β.

- Iterative Reweighted Least Squares (IRLS) algorithm is used to find the MLE for β.
  - Each step is similar to computing the MLE for linear regression.
  - See ESL 4.4 for details.

Prediction with logistic regression

- Can use estimate $\hat{\beta}$ to predict y_0 for a future x_0.
- Estimated probability of y_0 being class 1 given x_0 is

$$\mathbb{P}(Y_0 = 1 \mid \hat{\beta}, x_0) = \pi_{\hat{\beta}}(x_0).$$

- Can threshold probability at a cutoff to make predictions:

$$\hat{f}(x_0) = \mathrm{I}\big(\pi_{\hat{\beta}}(x_0) > \text{cutoff}\big).$$

Pros/cons of logistic regression

Pros

- Interpretable
- Tends to have lower variance (but this depends on φ)
- Relatively simple and easy to use

Cons

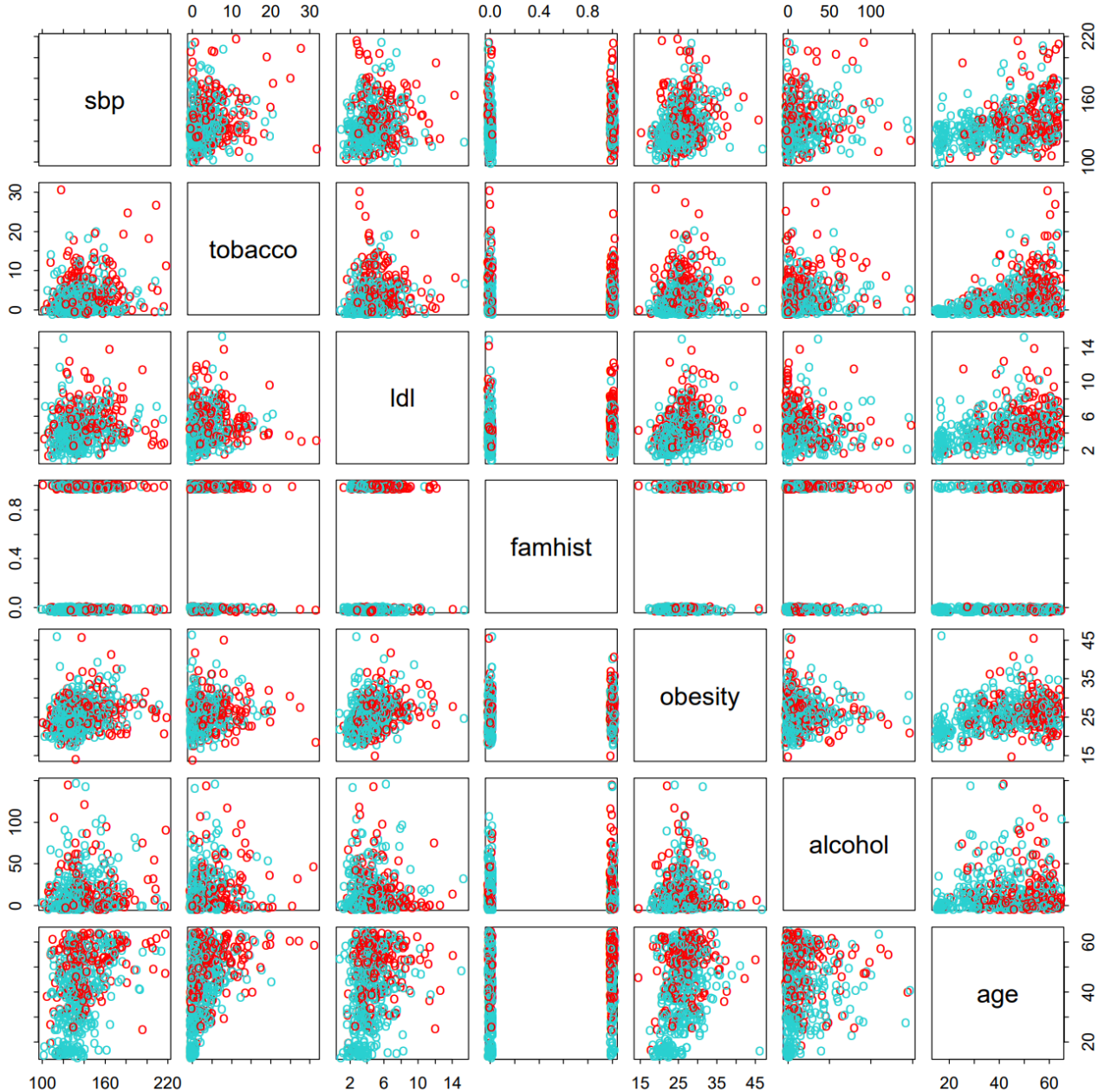
- Tends to have higher bias (but this depends on φ)
- MLE $\hat{\beta}$ can be unstable
  - e.g., if classes are well separated or predictors are collinear.
  - Can fix this with regularization.

Logistic regression example: South African Heart Disease

- Western Cape, South Africa
- Coronary Risk Factor Study (CORIS)
- High incidence of myocardial infarction (MI) in region: 5.1%

- 160 cases, 302 controls. Ages 15-64.
- Outcome y_i is presence/absence of MI at time of survey.

Logistic regression example: South African Heart Disease



Logistic regression example: South African Heart Disease

TABLE 4.2. *Results from a logistic regression fit to the South African heart disease data.*

	Coefficient	Std. Error	Z Score
(Intercept)	−4.130	0.964	−4.285
sbp	0.006	0.006	1.023
tobacco	0.080	0.026	3.034
ldl	0.185	0.057	3.219
famhist	0.939	0.225	4.178
obesity	-0.035	0.029	−1.187
alcohol	0.001	0.004	0.136
age	0.043	0.010	4.184

See ESL Section 4.4.2 for more details.

Outline

Linear/Quadratic Discriminant Analysis (LDA/QDA)

- Linear Discriminant Analysis (LDA) is one of the oldest classification methods.
- Basic idea: Model the classes using multivariate normal distributions, and use the Bayes optimal classifier.
- In ELISA example, we implicitly used LDA!

- LDA and QDA are generative models for classification.

- *Generative*: Model $p(y)$ and $p(x|y)$, and derive $p(y|x)$.
  How do you derive $p(y|x)$ from $p(y)$ and $p(x|y)$?
- *Discriminative*: Model $p(y|x)$ directly.
  - Examples: KNN, logistic regression.

Linear Discriminant Analysis (LDA)

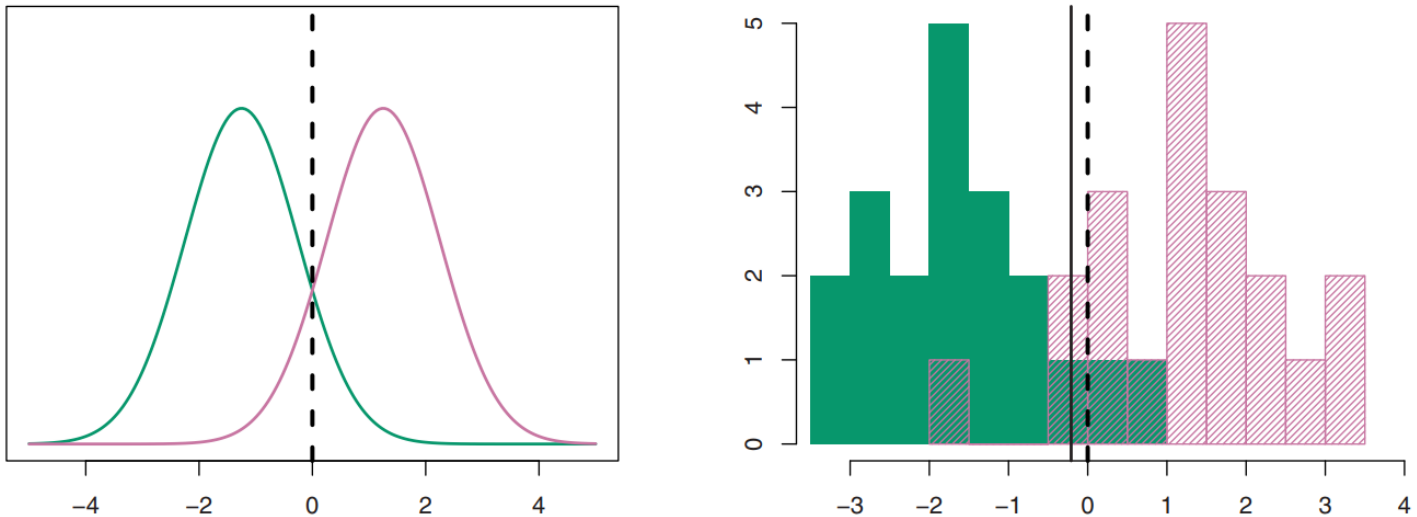


FIGURE 4.4. Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.* Right: *20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.*

Linear Discriminant Analysis (LDA)

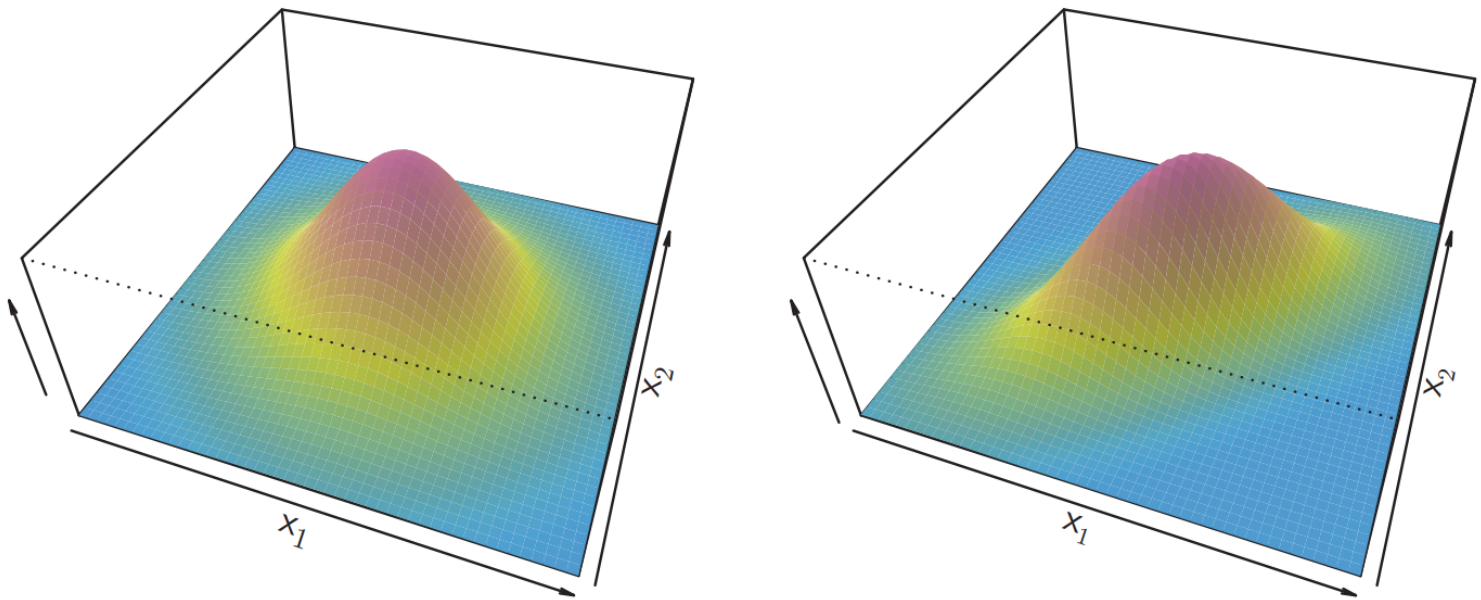
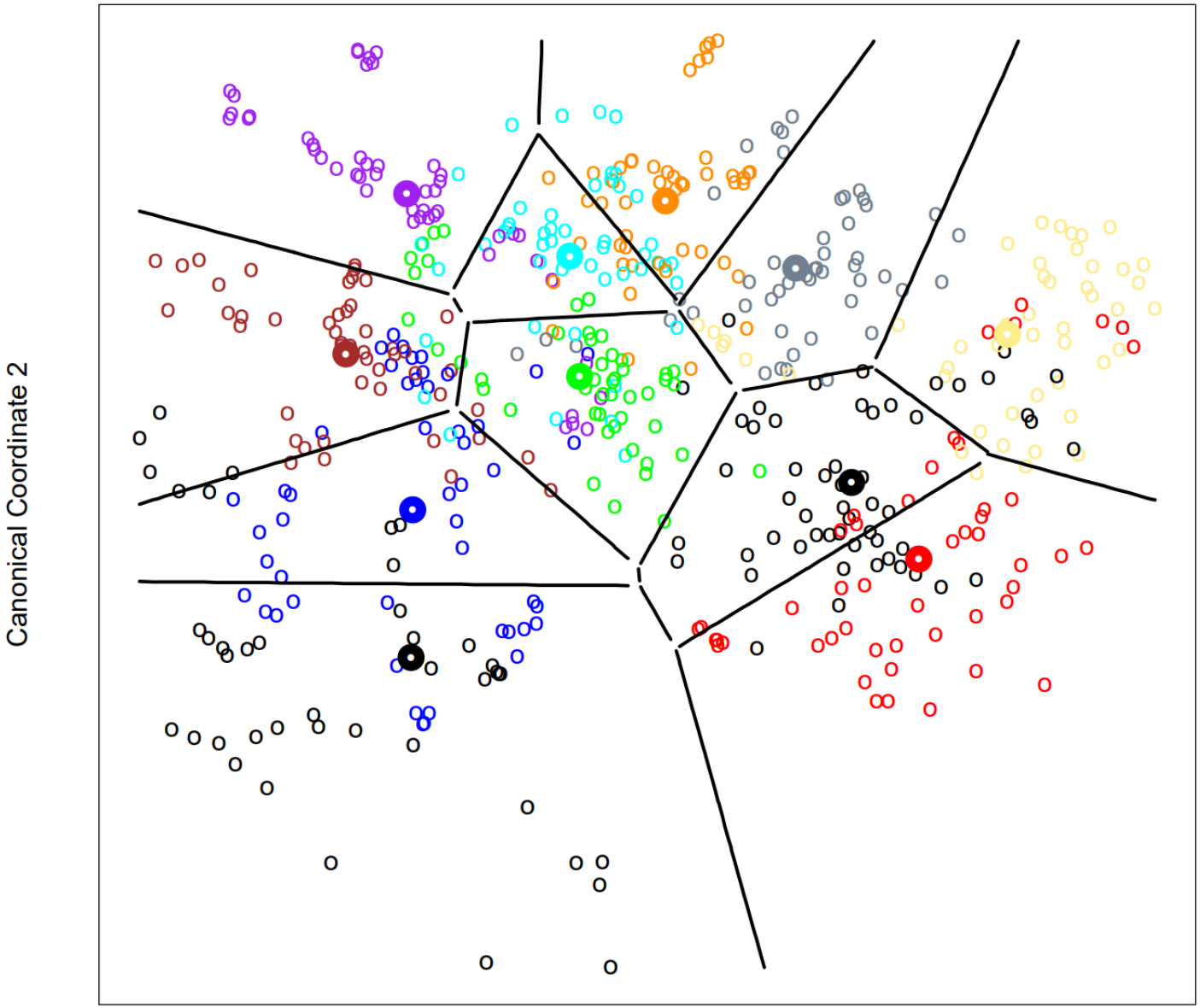


FIGURE 4.5. *Two multivariate Gaussian density functions are shown, with* $p = 2$. Left: *The two predictors are uncorrelated.* Right: *The two variables have a correlation of* 0.7.

Linear Discriminant Analysis (LDA)



LDA versus QDA

- LDA uses a linear decision rule. (less flexible)
- QDA uses a quadratic decision rule. (more flexible)

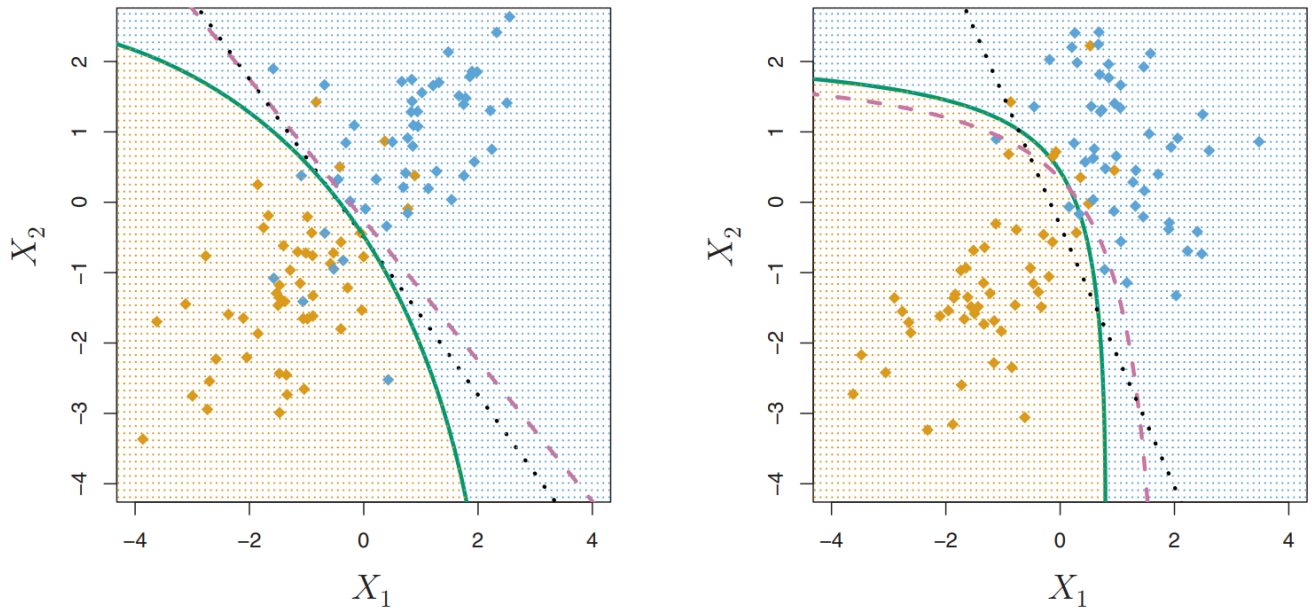


FIGURE 4.9. Left: *The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with* $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. *The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA.* Right: *Details are as given in the left-hand panel, except that* $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. *Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.*

Probabilistic model for LDA/QDA

- Training data $(x_1, y_1), \ldots, (x_n, y_n)$.
- Inputs $x_i \in \mathbb{R}^p$ are real-valued vectors.
- Outputs y_i are categorical, $y_i \in \{1, \ldots, K\}$.

- $\pi_k = \mathbb{P}(Y_i = k) =$ prior probability of class k.

- $p(x_i \mid Y_i = k) = \mathcal{N}(x_i \mid \mu_k, C_k)$ (multivariate normal)
  "Class conditional distribution" for class k.

- LDA constrains $C_1 = \cdots = C_K$. Let's denote $C = C_k$.
  I.e., LDA uses same covariance matrix for each class.

- QDA allows each class k to have a different C_k.

Prediction using the LDA/QDA model

- How can we use this model to predict the y_0 for a future x_0?

- Suppose we know the parameters π_k, μ_k, and C_k for each k.
- Then LDA just uses the Bayes optimal classifier:

$$\text{Choose } k \text{ to maximize } \mathbb{P}(Y_0 = k \mid x_0).$$

- How do we compute $\mathbb{P}(Y_0 = k \mid x_0)$? Note that

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_{y'} p(x|y')p(y')}.$$

- This is called *Bayes' theorem*, but it is just basic probability. *Can you derive this from our probability basics?*

- Thus, in the LDA/QDA model,

$$\mathbb{P}(Y_0 = k \mid x_0) = \frac{\mathcal{N}(x_0 \mid \mu_k, C_k)\,\pi_k}{\sum_{j=1}^{K} \mathcal{N}(x_0 \mid \mu_j, C_j)\,\pi_j}.$$

Estimation for LDA and QDA

- How can we estimate π_k, μ_k, and C_k for each k?
- Define $n_k = \#\{i : y_i = k\}$ (# training points in class k)
- Estimates: $\hat{\pi}_k = n_k/n$ (fraction of training points in class k)
- $\hat{\mu}_k = \frac{1}{n_k} \sum_{i:\, y_i = k} x_i$ (sample mean for class k)
- QDA: $\hat{C}_k = \frac{1}{n_k} \sum_{i:\, y_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\intercal$

  (sample covariance matrix for class k)
- LDA: $\hat{C} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i:\, y_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\intercal$

LDA versus QDA

LDA is less flexible than QDA...

fewer parameters to estimate, lower variance, higher bias.

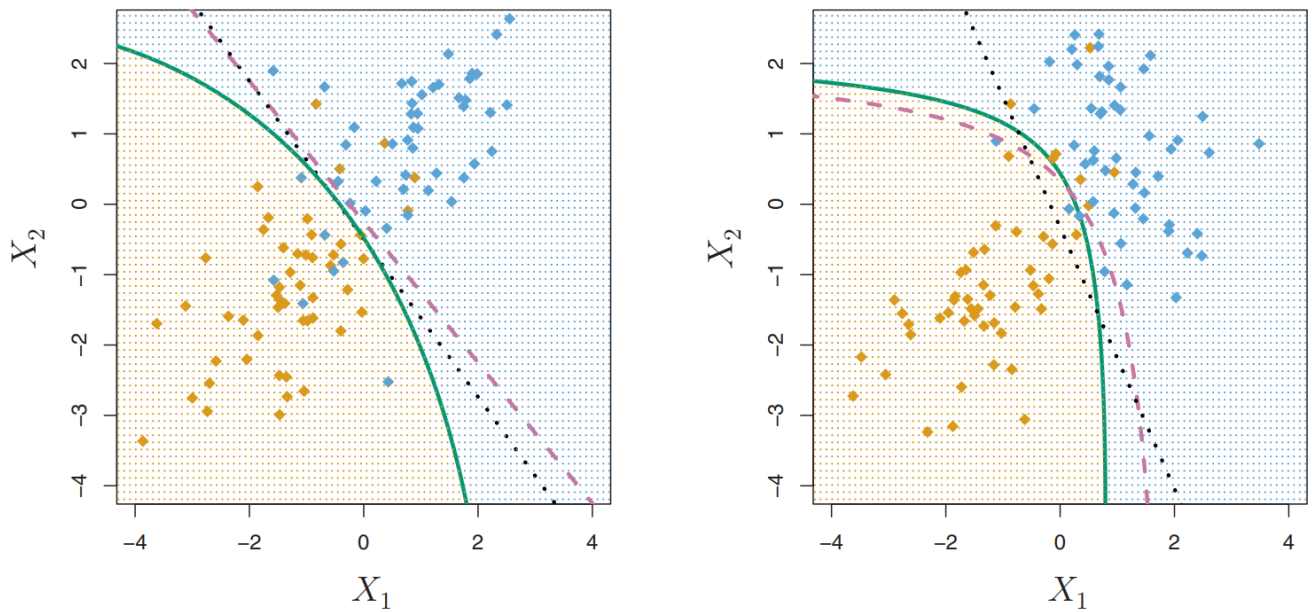


FIGURE 4.9. Left: *The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with* $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. *The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA.* Right: *Details are as given in the left-hand panel, except that* $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. *Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.*

Comparison of KNN, Logistic regression, LDA, and QDA

Typically, we would expect:

- KNN – good when complex boundaries and n is sufficiently large.

- Logistic regression and LDA – good when linear boundaries or p is big relative to n.
  - LDA extends better to multi-class problems
  - LDA is more stable during estimation
  - Logistic regression is more robust to outliers

- QDA – good when quadratic (or moderately complex) boundaries and n is moderately big.