Linear Classification

CS 534: Machine Learning

Slides adapted from Lee Cooper, Joydeep Ghosh, and Sham Kakade

Review: Linear Regression

Regression

- Given an input vector $\mathbf{x}^\top = (x_1, x_2, \ldots, x_p)$, we want to predict the quantitative response Y

- Linear regression form:

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{p} x_i \beta_i$$

- Least squares problem:

- $$\min_{\beta} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \implies \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Feature Selection

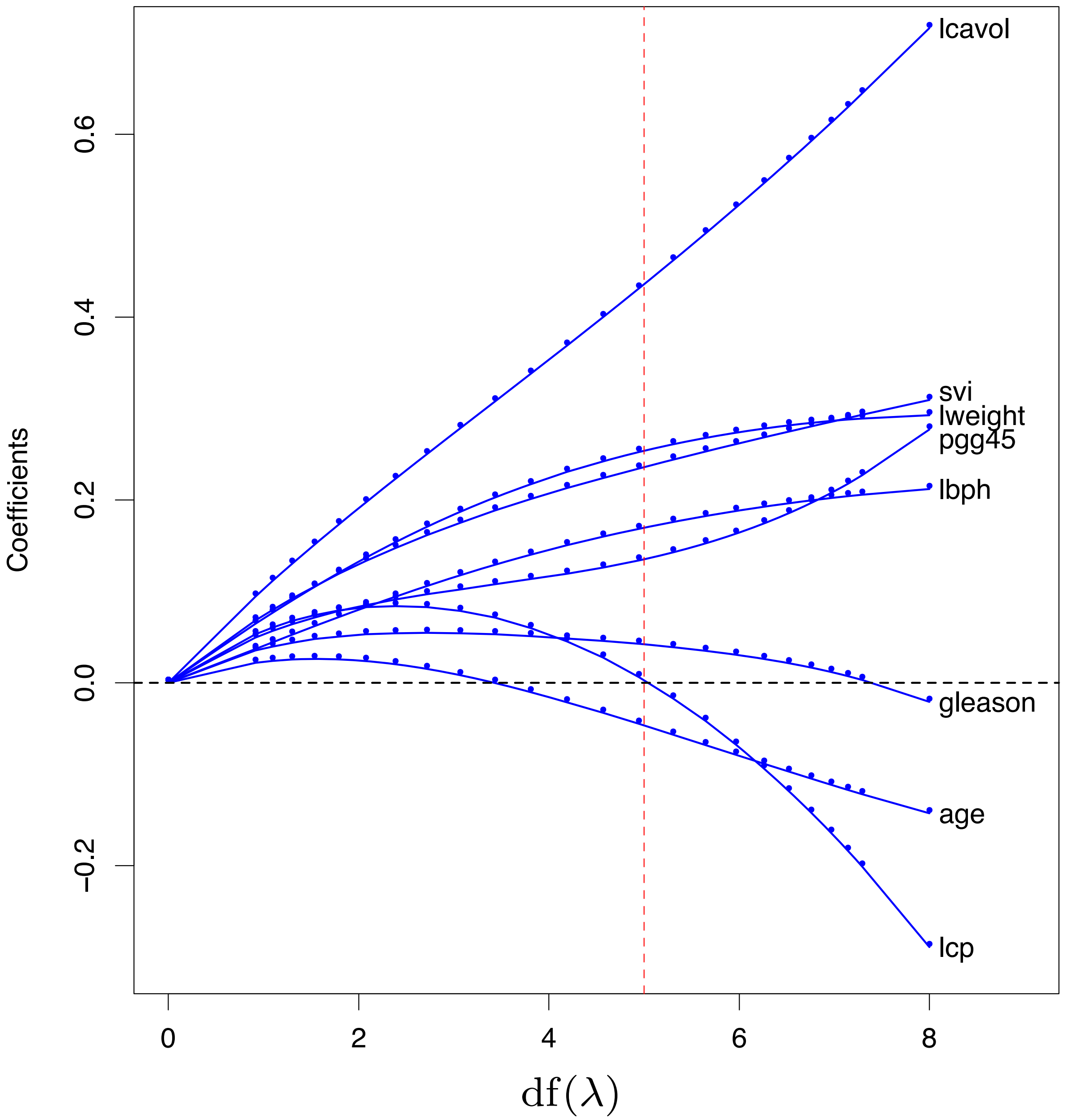
- Brute force is infeasible for large number of features
- Algorithms
  - Best subset selection — beyond 40 features is impractical
  - Stepwise selection (forward and backward)

Regularization

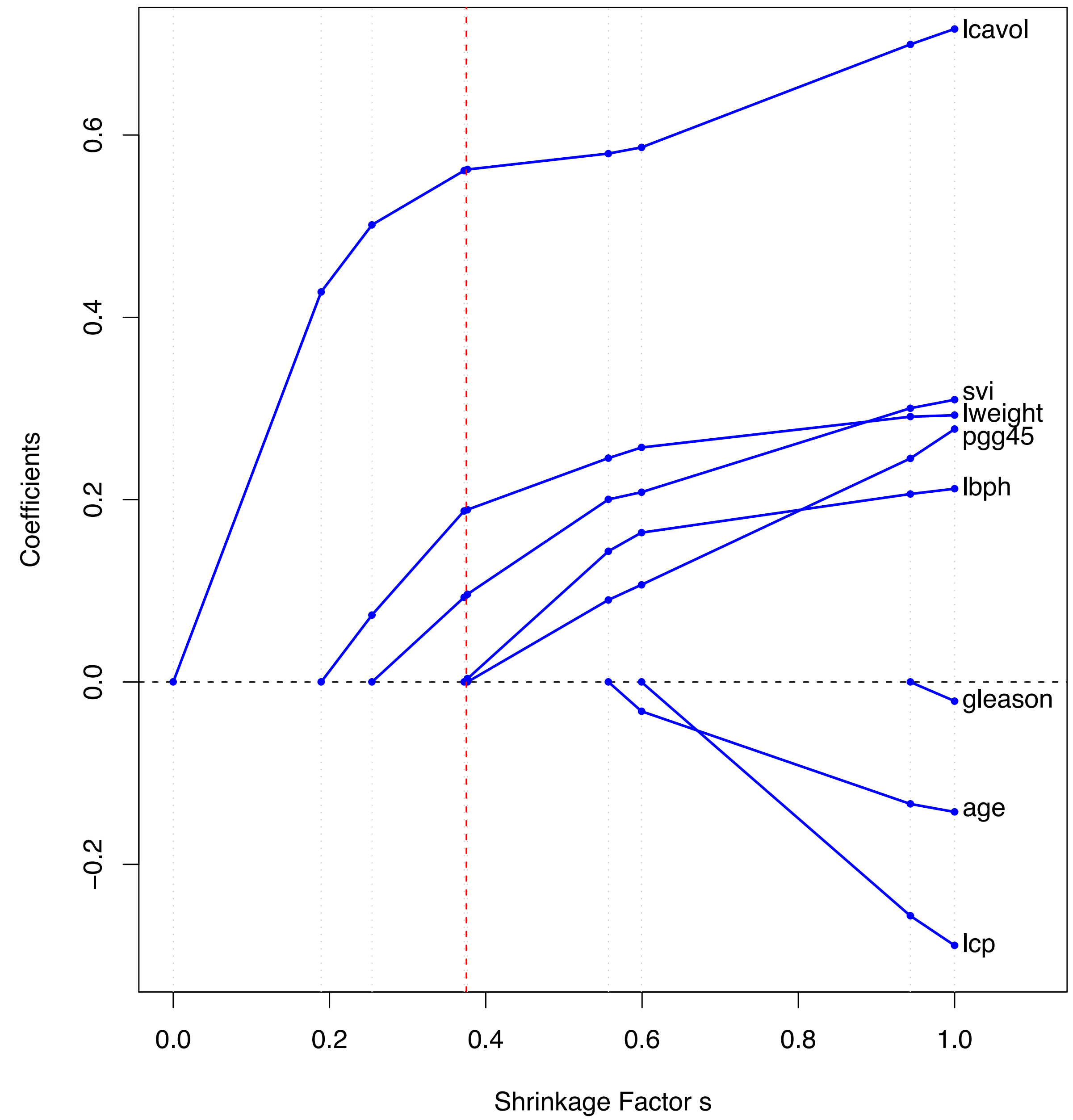
- Add penalty term on model parameters to achieve a more simple model or reduce sensitivity to training data
- Less prone to overfitting

$$\min_{\beta} L(\mathbf{X}\boldsymbol{\beta}, \mathbf{y}) + \lambda \mathrm{penalty}(\boldsymbol{\beta})$$

Ridge & Lasso Regularization



Ridge



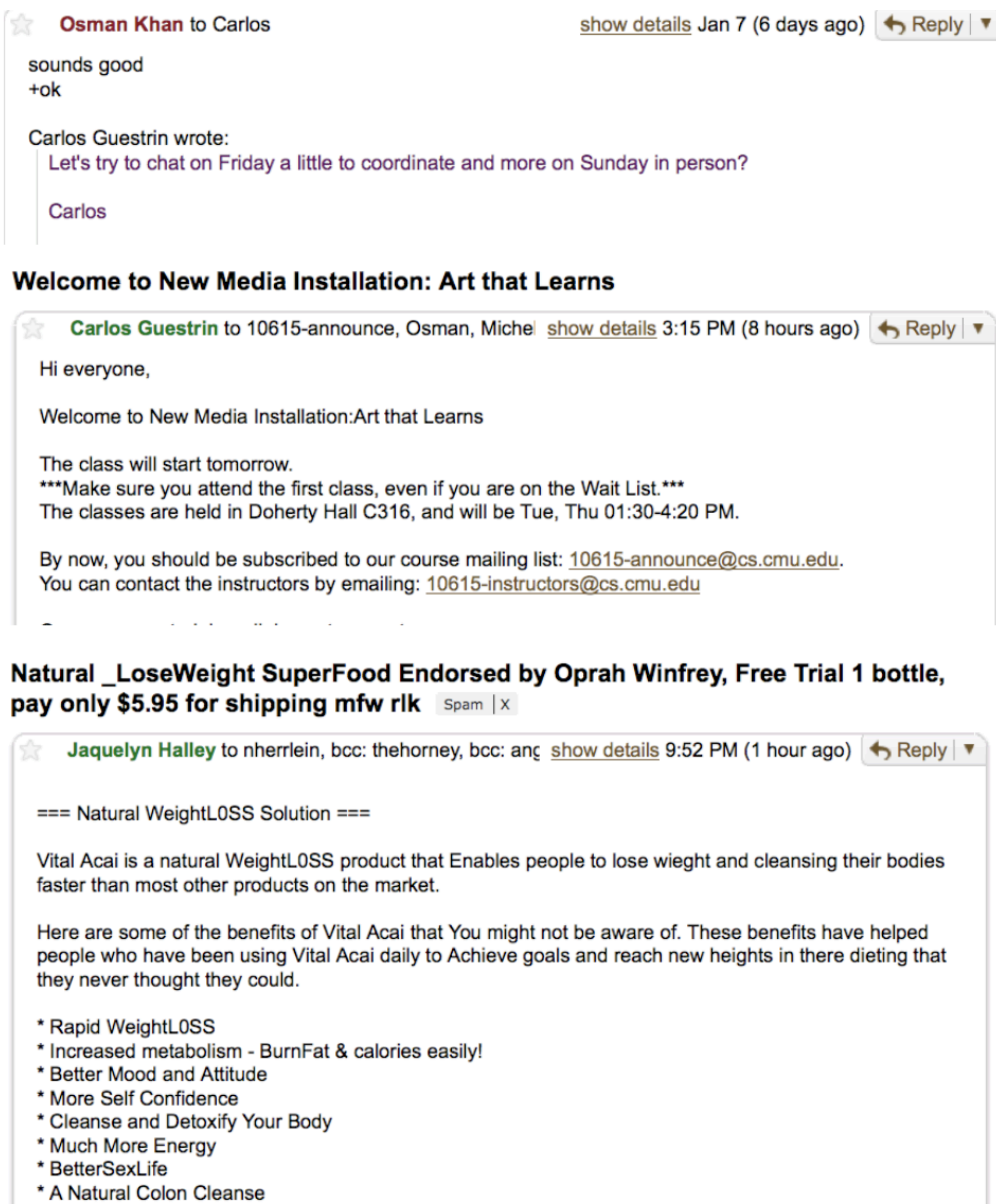
Lasso

Figures 3.8 & 3.10 (Hastie et al.)

Thus far, regression: Predict a continuous value given some inputs or features

Linear Classification

Linear Classifiers: Spam Filtering



Osman Khan to Carlos — show details Jan 7 (6 days ago) — Reply

sounds good
+ok

Carlos Guestrin wrote:
Let's try to chat on Friday a little to coordinate and more on Sunday in person?

Carlos

Welcome to New Media Installation: Art that Learns

Carlos Guestrin to 10615-announce, Osman, Michel — show details 3:15 PM (8 hours ago) — Reply

Hi everyone,

Welcome to New Media Installation:Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu

Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rlk Spam | X

Jaquelyn Halley to nherrlein, bcc: thehorney, bcc: ang — show details 9:52 PM (1 hour ago) — Reply

=== Natural WeightL0SS Solution ===

Vital Acai is a natural WeightL0SS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* Rapid WeightL0SS
* Increased metabolism - BurnFat & calories easily!
* Better Mood and Attitude
* More Self Confidence
* Cleanse and Detoxify Your Body
* Much More Energy
* BetterSexLife
* A Natural Colon Cleanse

spam
vs
not spam

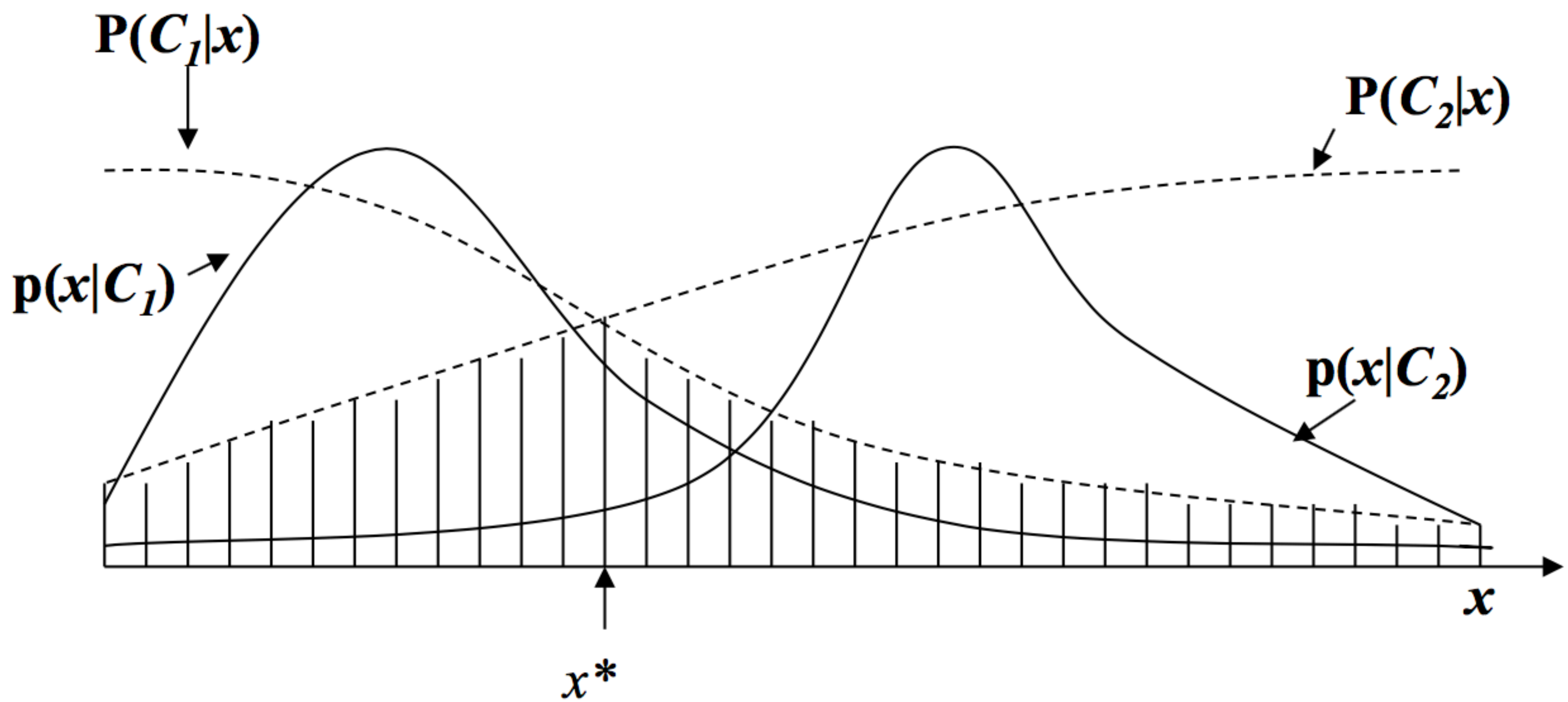
Linear Classifiers: Weather Prediction

Hour	Weather	Temp.	Precip.	Wind
10pm	Mostly Clear	41°F	0 in	NW - 5 mph
12am	Mostly Clear	39°F	0 in	NW - 3 mph
02am	Mostly Clear	39°F	0 in	NW - 3 mph
04am	Mostly Clear	37°F	0 in	NW - 3 mph
06am	Mostly Clear	36°F	0 in	NW - 3 mph
08am	Mostly Sunny	43°F	0 in	WNW - 3 mph
10am	Mostly Sunny	50°F	0 in	W - 2 mph
12pm	Mostly Sunny	55°F	0 in	SW - 2 mph
02pm	Mostly Sunny	57°F	0 in	S - 3 mph
04pm	Mostly Sunny	57°F	0 in	S - 3 mph
06pm	Mostly Clear	54°F	0 in	SSE - 3 mph
08pm	Mostly Clear	50°F	0 in	S - 3 mph
10pm	Partly Cloudy	46°F	0 in	S - 4 mph

Notation

- Number of classes: K
- A specific class: k
- Set of classes: G
- Prior probability of class k: $\pi_k = \Pr(G = k)$
- $$\sum_{j=1}^{K} \pi_j = 1$$

Bayes Decision Theory



$P(C_i|x)$ – *a posteriori* probability

$p(x|C_i)$ – (class conditional) likelihood function

$P(C_i)$ – class priors

Statistical Decision Theory Revisited

- Natural rule of classification:

$$f(\mathbf{x}) = \text{argmax}_{j=1,\ldots,K} \Pr(G = k | \mathbf{X} = \mathbf{x})$$

- Application of Bayes' rule:

$$\Pr(G = k | \mathbf{X} = \mathbf{x}) = \frac{\Pr(\mathbf{X} = \mathbf{x} | G = k) \Pr(G = k)}{\Pr(\mathbf{X} = \mathbf{x})}$$

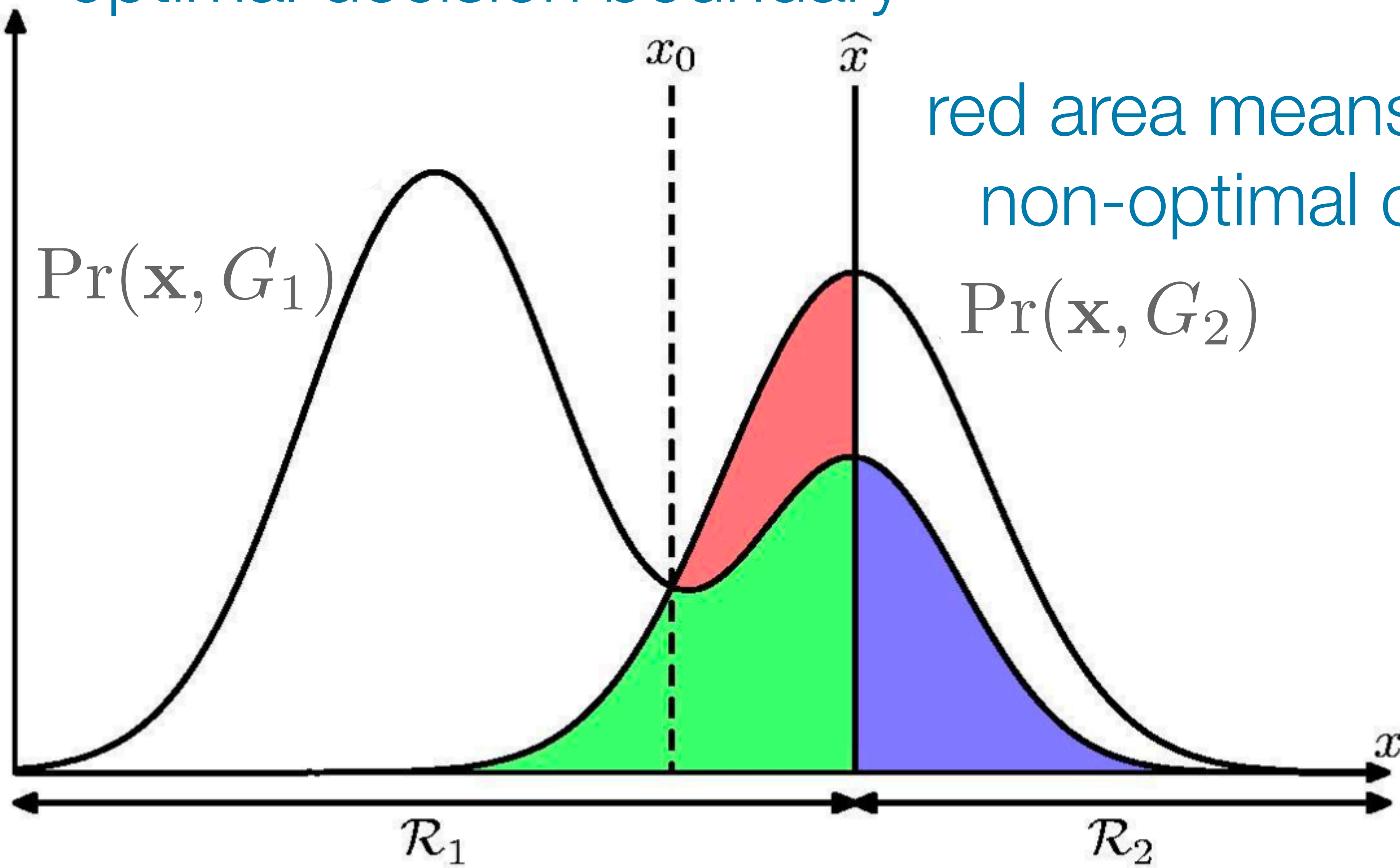
- Since denominator same across all classes

$$f(\mathbf{x}) = \text{argmax}_{j=1,\ldots,K} \Pr(\mathbf{X} = \mathbf{x} | G = k) \pi_k$$

Classification Evaluation

Misclassification Rate

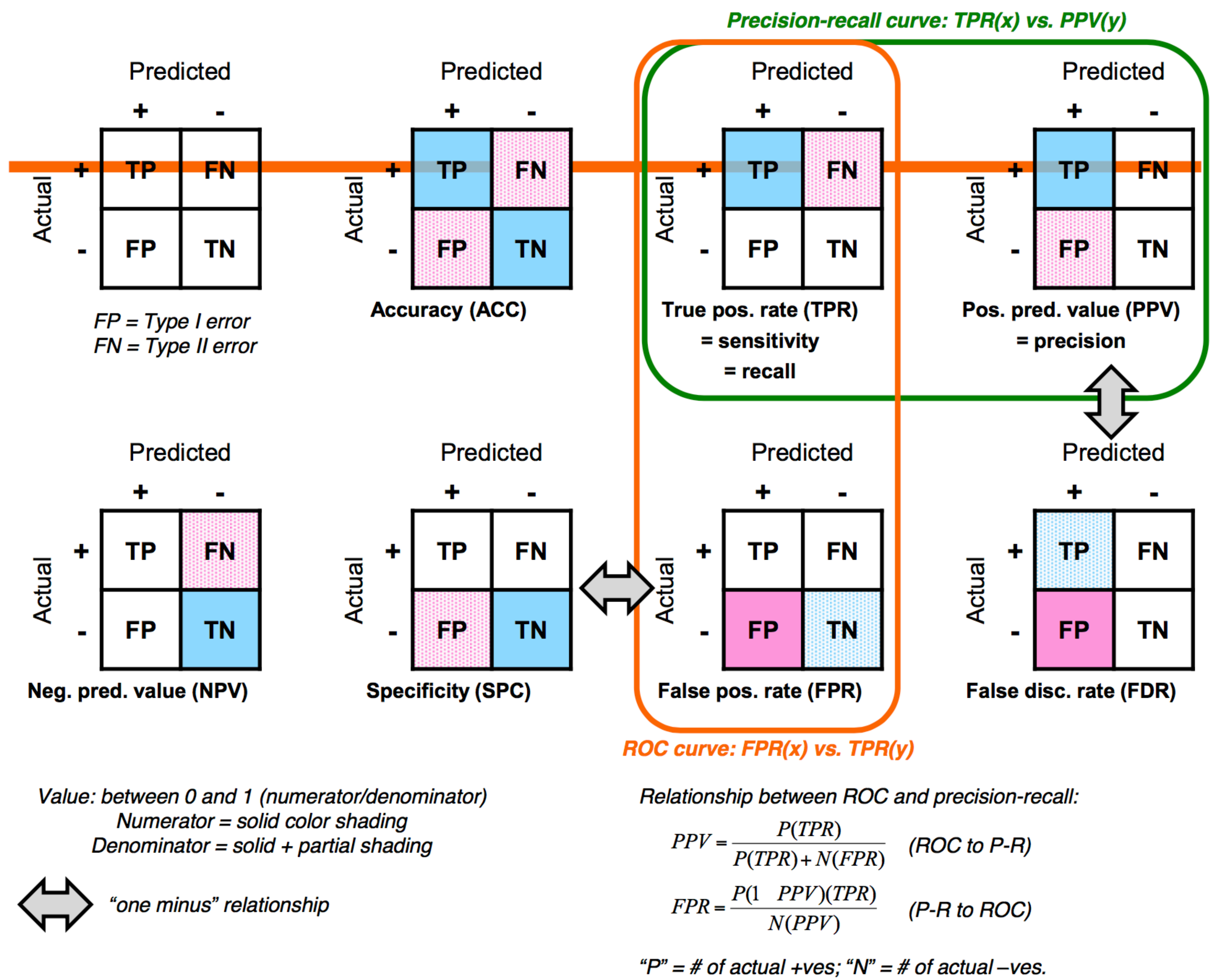
optimal decision boundary



red area means it is an non-optimal design

$$\Pr(\text{mistake}) = \int_{\mathcal{R}_1} \Pr(\mathbf{x}, G_2) d\mathbf{x} + \int_{\mathcal{R}_2} \Pr(\mathbf{x}, G_1) d\mathbf{x}$$

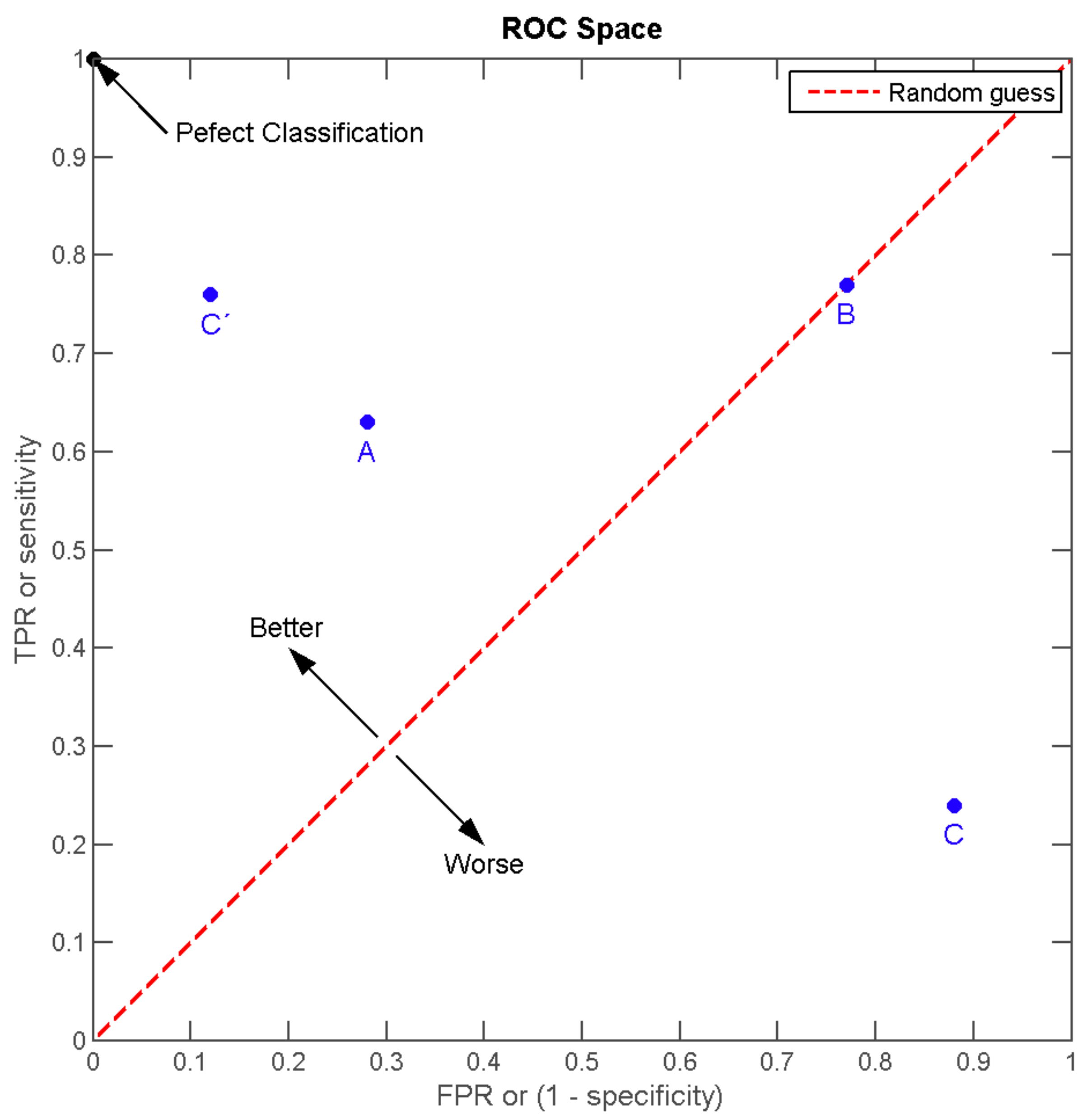
Confusion Matrix & Metrics



Problems with Accuracy

- Assumes equal cost for both types of error
  - FN = FP
- Is 99% accuracy?
  - Depends on the problem and the domain
  - Compare to the base rate (i.e., predicting predominant class)

Receiver Operating Characteristic Curve



AUC = area under ROC curve

https://en.wikipedia.org/wiki/Receiver_operating_characteristic

ROC Curves

- Slope is always increasing
- Each point represents different tradeoff (cost ratio) between FP and FN
- Two non-intersecting curves means one method dominates the other
- Two intersecting curves means one method is better for some cost ratios, and other method is better for other cost ratios

Area Under ROC Curve (AUC)

- > 0.9: excellent prediction — something potentially fishy, should check for information leakage
- 0.8: good prediction
- 0.5: random prediction
- <0.5: something wrong!

AUC is more robust to class imbalanced situation

Discriminant Analysis

Bayes Classifier

- MAP classifier (maximum a posterior)
- Outcome: partitioning of the input space
- Classifier is optimal: statistically minimizes the error rate

Why not use Bayes classifier all the time?

Discriminant Functions

- Each class has a discriminant function: $\delta_k(\mathbf{x})$
- Classify according to the best discriminant:

$$\hat{G}(\mathbf{x}) = \text{argmax}_{j=1,\ldots,K}\delta_k(\mathbf{x})$$

- Can be formulated in terms of probabilities

$$\hat{G}(\mathbf{x}) = \text{argmax}_{j=1,\ldots,K}\Pr(G = k|\mathbf{X} = \mathbf{x})$$

Discriminant Analysis

- Bayes' rule:

$$\Pr(G|X)\Pr(X) = \Pr(X|G)\Pr(G)$$

- Application of Bayes theorem:

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

- Use log-ratio for a two class problem:

$$\log \frac{\Pr(G = k|X = x)}{\Pr(G = \ell|X = x)} = \log \frac{f_k(x)}{f_\ell(x)} + \log \frac{\pi_k}{\pi_\ell}$$

Linear Regression Classifier

- Each response category coded as indicator variable
- Fit linear regression model to each column of response indicator matrix simultaneously
- Compute the fitted output and classify according to largest component
- Serious problems occurs when number of classes greater than or equal to 3!

Linear Discriminant Analysis (LDA)

- Assume each class density is from a multivariate Gaussian

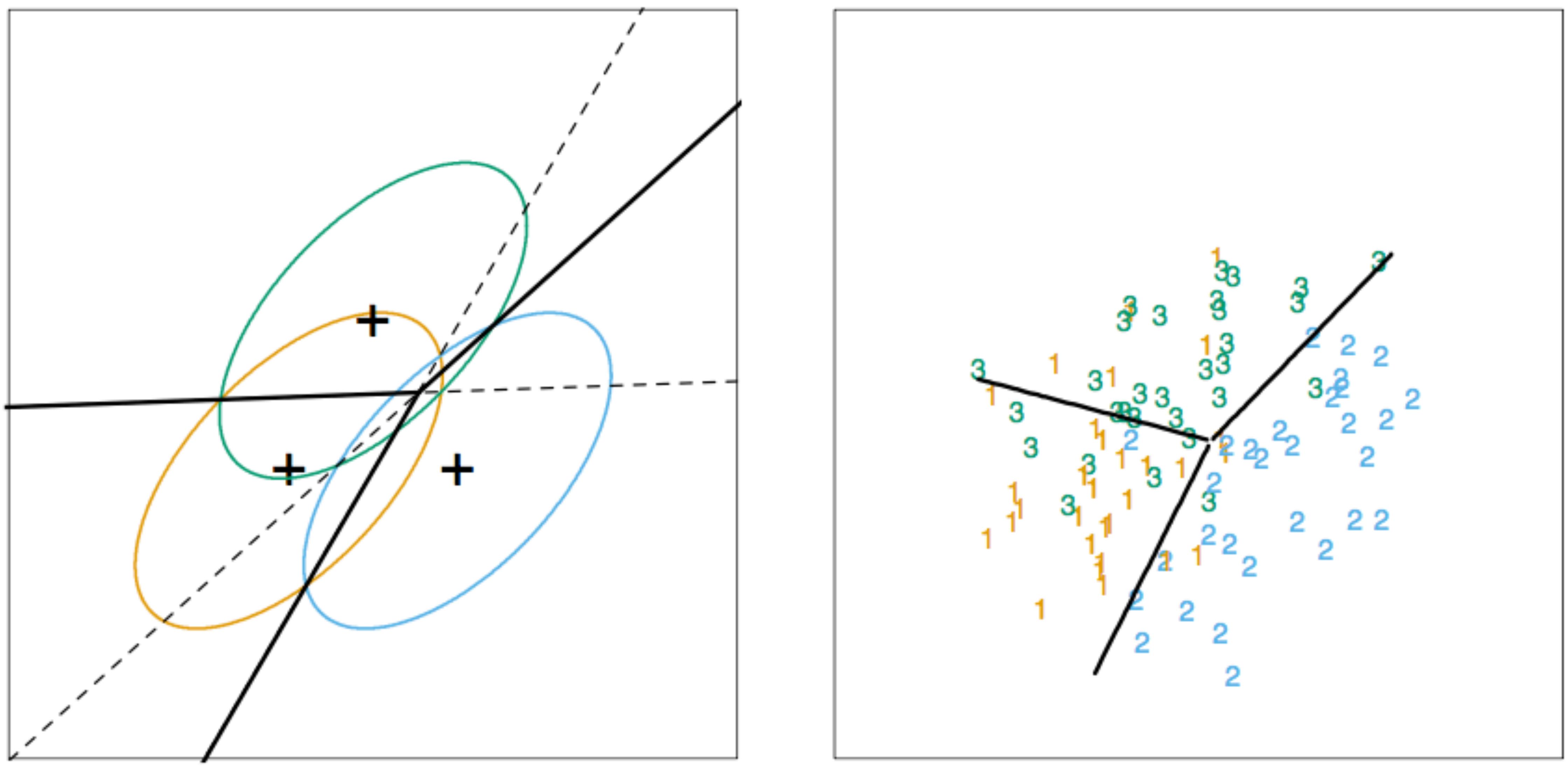
$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right)$$

- LDA assumes class have common covariance matrix

- Discriminant function:

$$\delta_k(\mathbf{x}) = \mathbf{x}^\top \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^\top \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_k + \log \pi_k$$

LDA Decision Boundaries



True distributions with same covariance and different means

Estimated boundaries

Figure 4.5 (Hastie et al.)

LDA vs Linear Regression

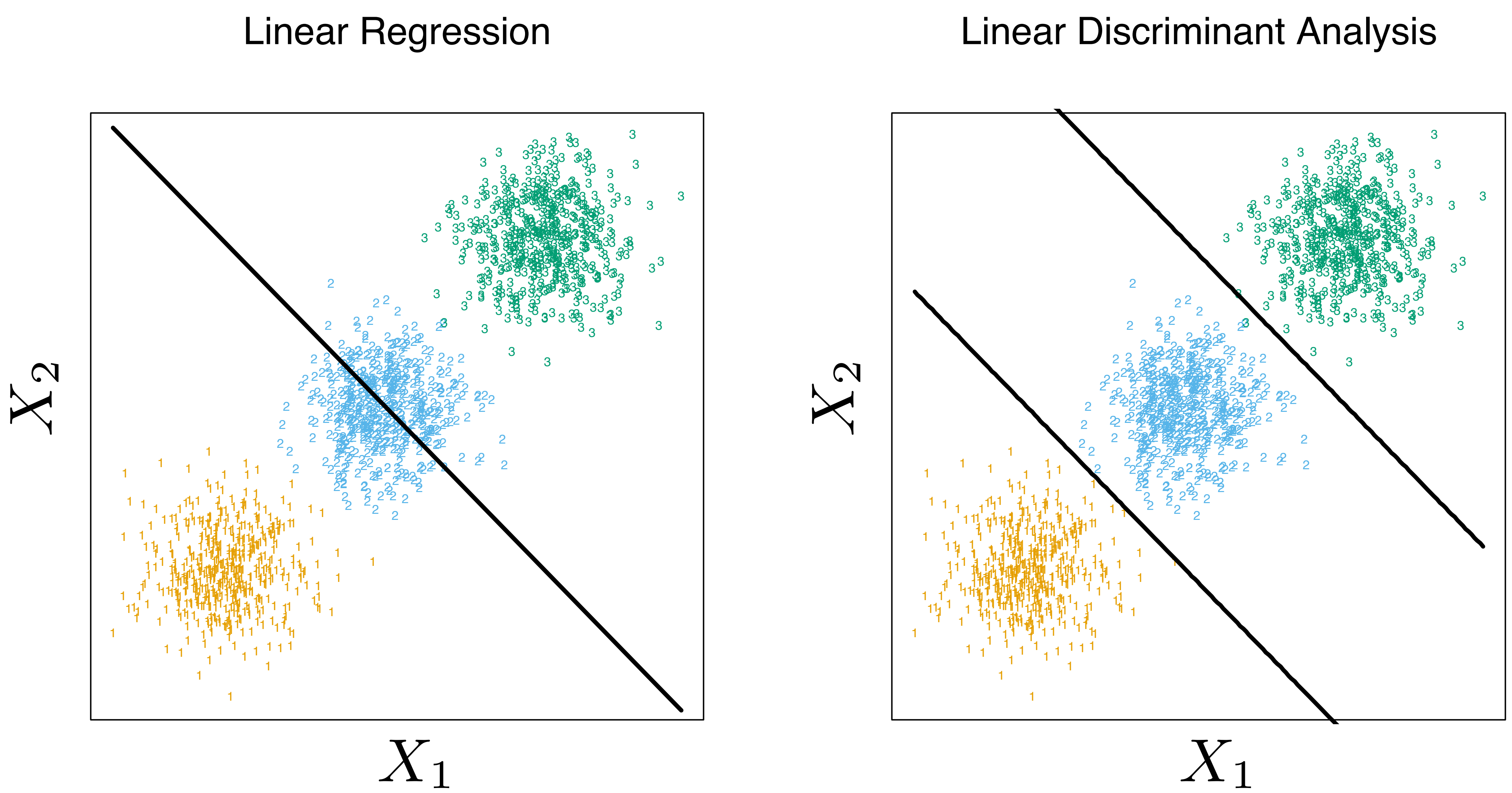


Figure 4.2 (Hastie et al.)

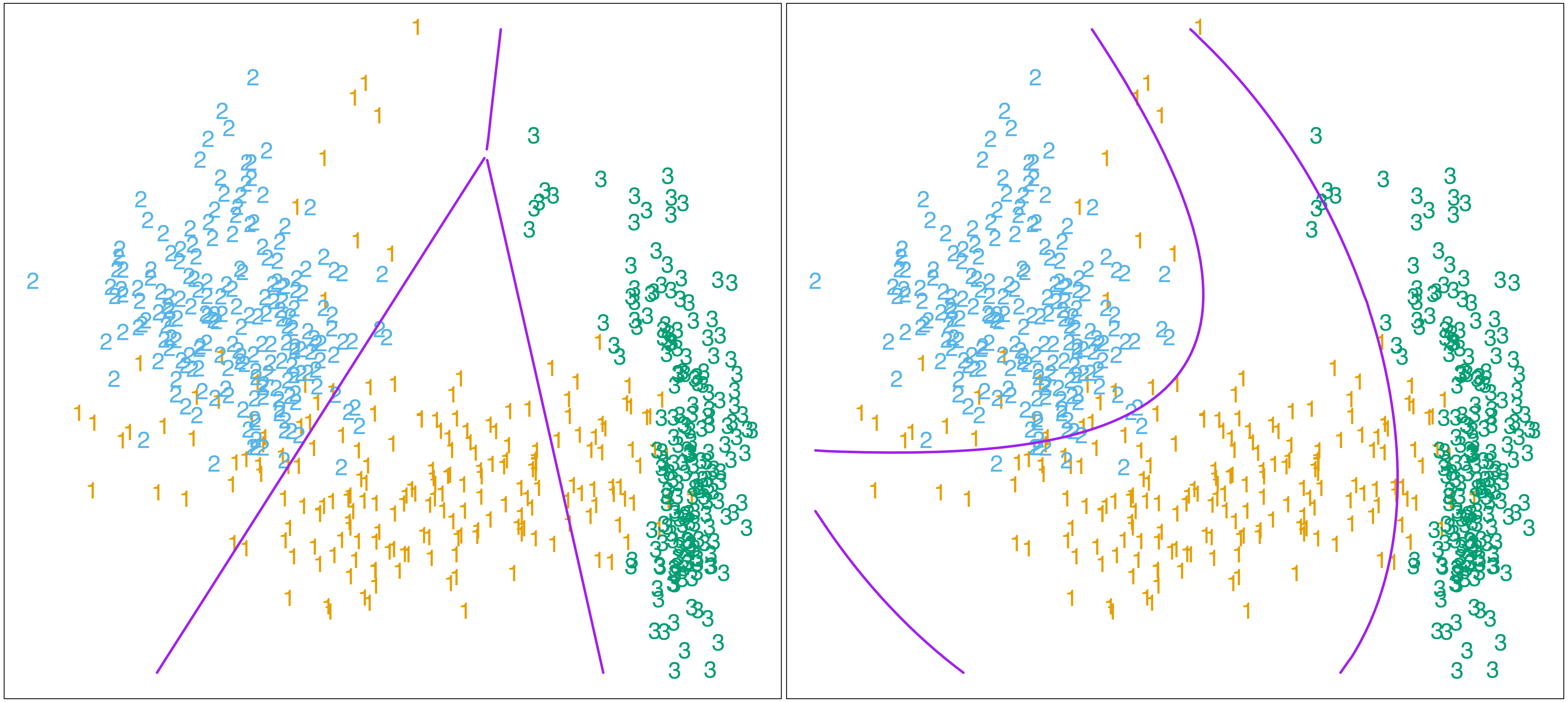
Quadratic Discriminant Analysis (QDA)

- What if the covariances are not equal?
- Quadratic discriminant functions:

$$\delta_k(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{\Sigma}_k| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \log \pi_k$$

- Quadratic decision boundary
- Covariance matrix must be estimated for each class

LDA vs. QDA Decision Boundaries



LDA QDA

Figure 4.1 (Hastie et al.)

Gaussian Parameter Values

- In practice, the parameters of multivariate normal distribution are unknown
- Estimate using the training data
  - Prior distribution $\hat{\pi}_k = N_k / N$
  - Mean $\hat{\boldsymbol{\mu}}_k = \sum_{g_i = k} \mathbf{x}_i / N_k$
  - Variance $\boldsymbol{\Sigma} = \sum_{k=1}^{K} \sum_{g_i = k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^\top / (N - K)$

Regularized Discriminant Analysis

- Compromise between LDA and QDA
- Shrink separate covariances of QDA towards common covariance like LDA
- Similar to ridge regression

$$\hat{\mathbf{\Sigma}}_k(\alpha) = \alpha\hat{\mathbf{\Sigma}}_k + (1-\alpha)\hat{\mathbf{\Sigma}}$$

Example: Vowel Data

- Experiment recorded 528 instances of spoken words
- Words fall into 11 classes (“vowels”)
- 10 features for each instance

Regularized Discriminant Analysis

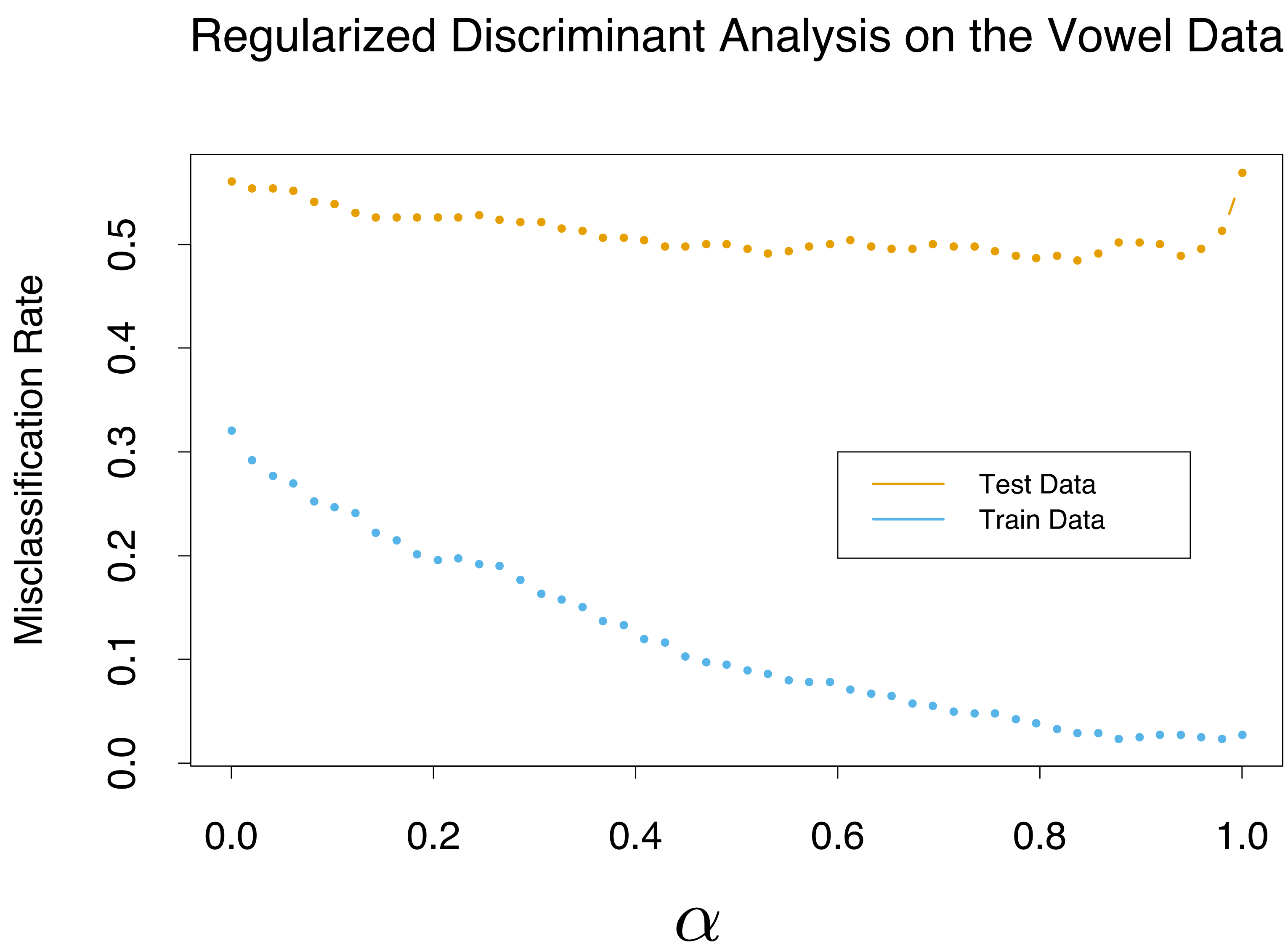


Figure 4.7 (Hastie et al.)

Optimum for test occurs close to QDA

Reduced-rank LDA

- What if we want to further reduce the dimension to L where $L < K - 1$?
- Why?
  - Visualization
  - Regularization — some dimensions may not provide good separation between classes but just noise

Fisher's Linear Discriminant

- Find projection that maximizes ratio of between class variance to within class variance

$$\frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{(\mathbf{a}^\top(\mu_1 - \mu_2))^2}{\mathbf{a}^\top(\Sigma_1 + \Sigma_2)\mathbf{a}}$$

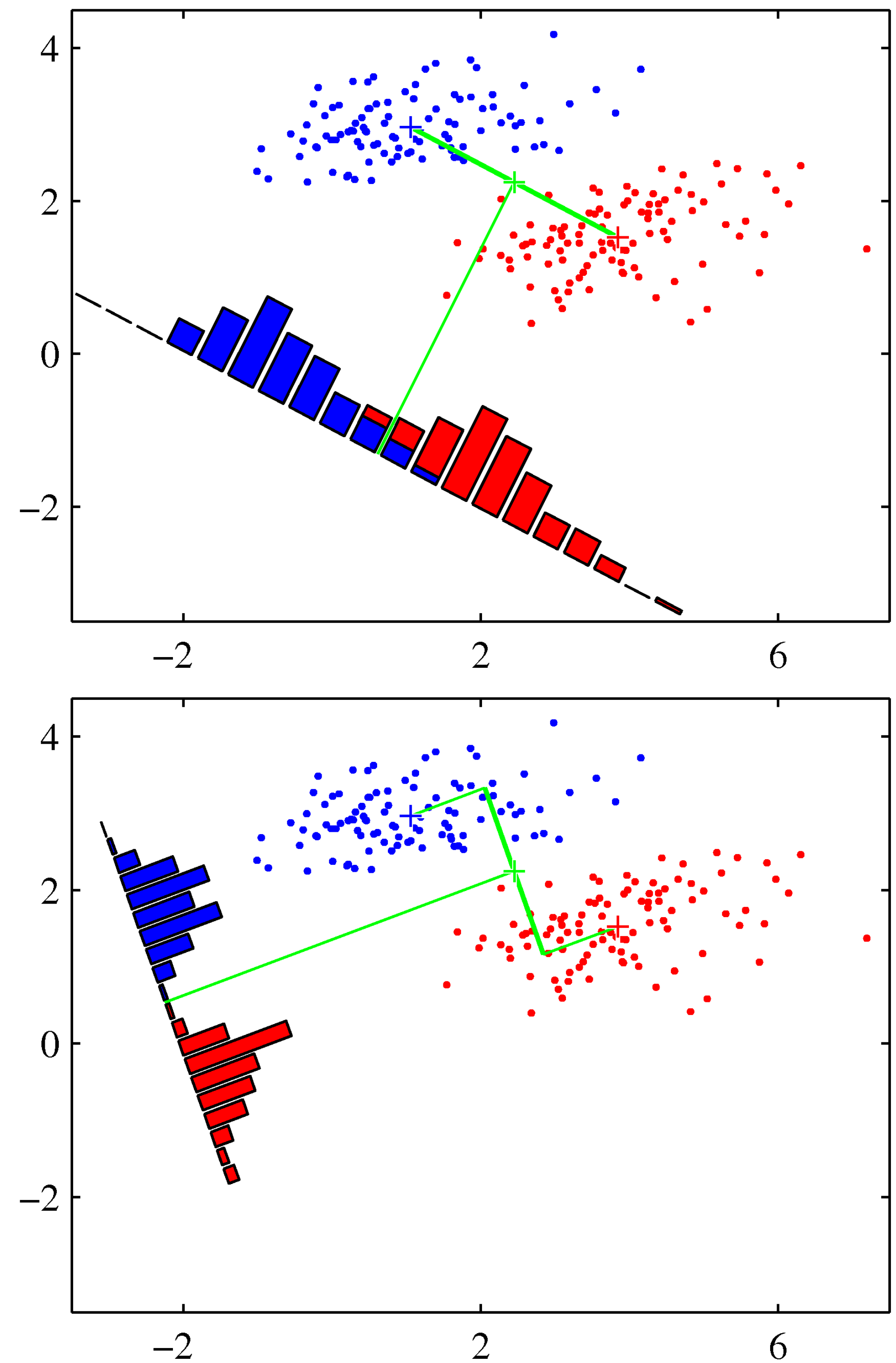
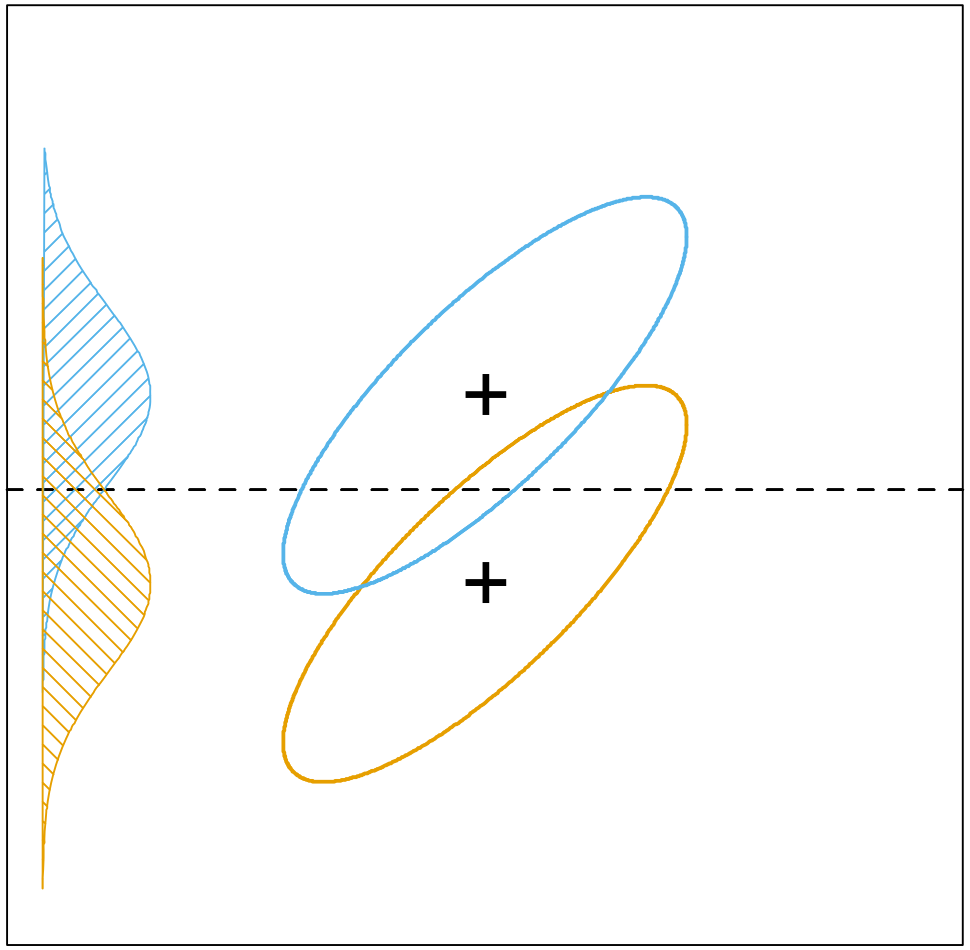


Figure 4.6 (Bishop)

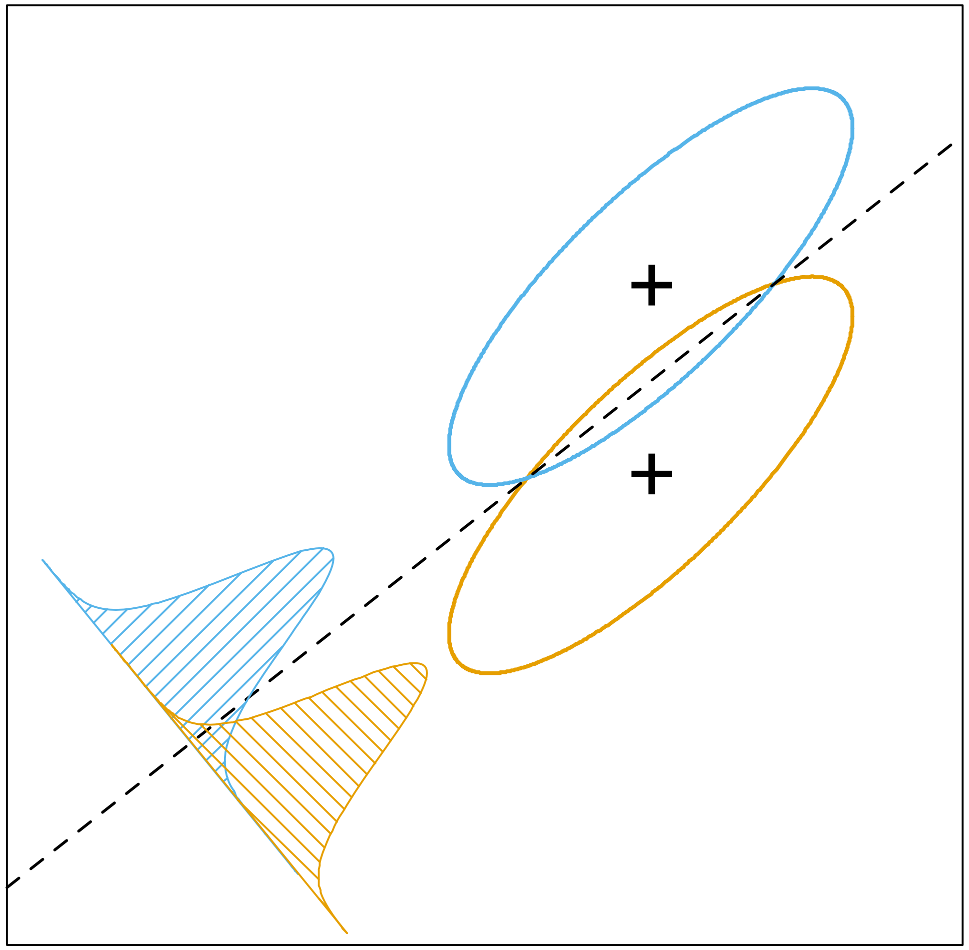
Why Fisher Makes Sense

- Following information is taken into account
  - Spread of class centroids — direction joining centroids separates the mean
  - “Shape” of data defined by covariance — minimum overlap can be found

Why Fisher Makes Sense: Graphically



Projected data maximizing between class only



Discriminant direction

Figure 4.9 (Hastie et al.)

Vowel Data: 2-D Subspace

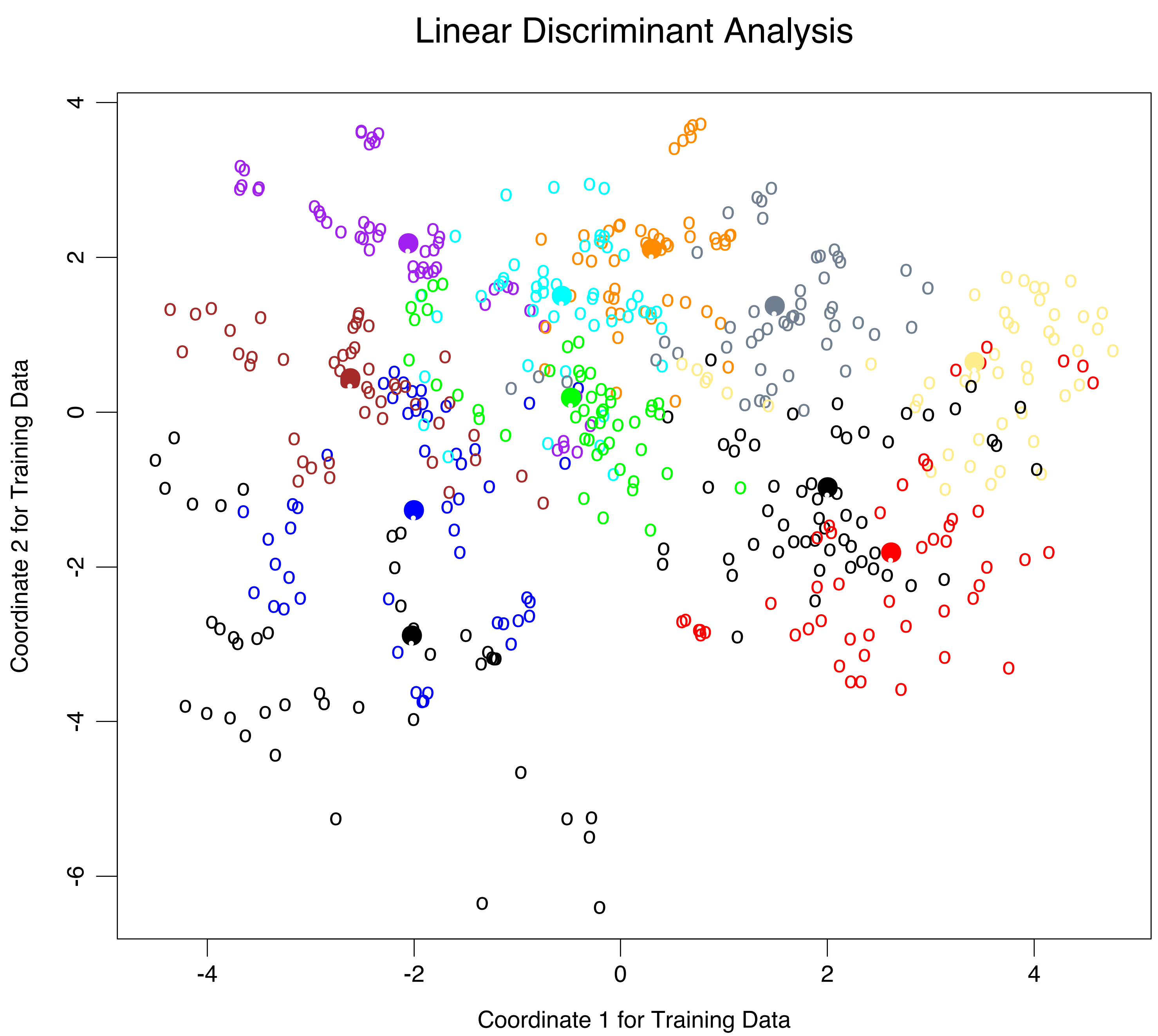


Figure 4.4 (Hastie et al.)

Vowel Data: Reduced-rank LDA

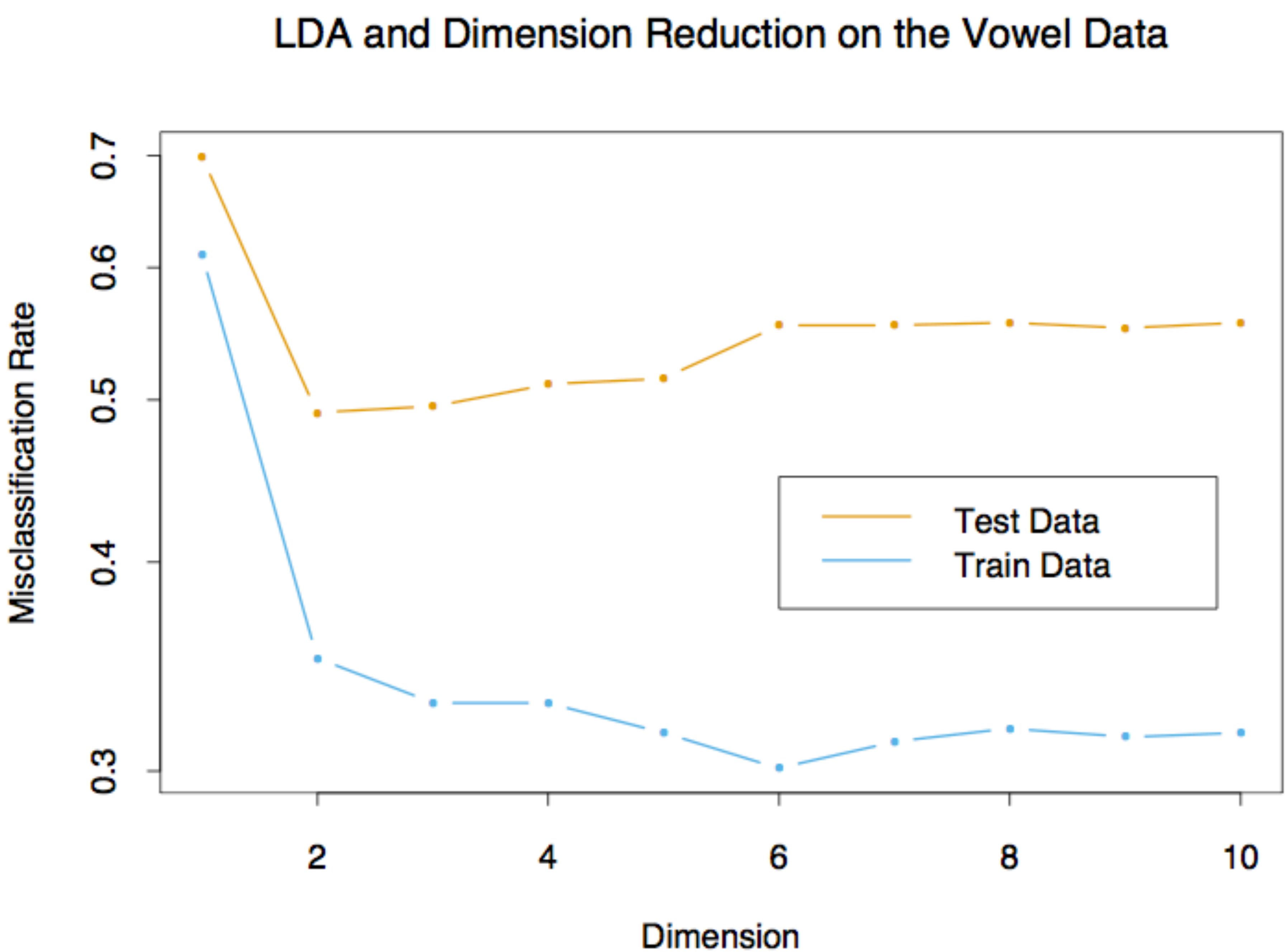


Figure 4.10 (Hastie et al.)

Vowel Data: Reduced-rank LDA (2)

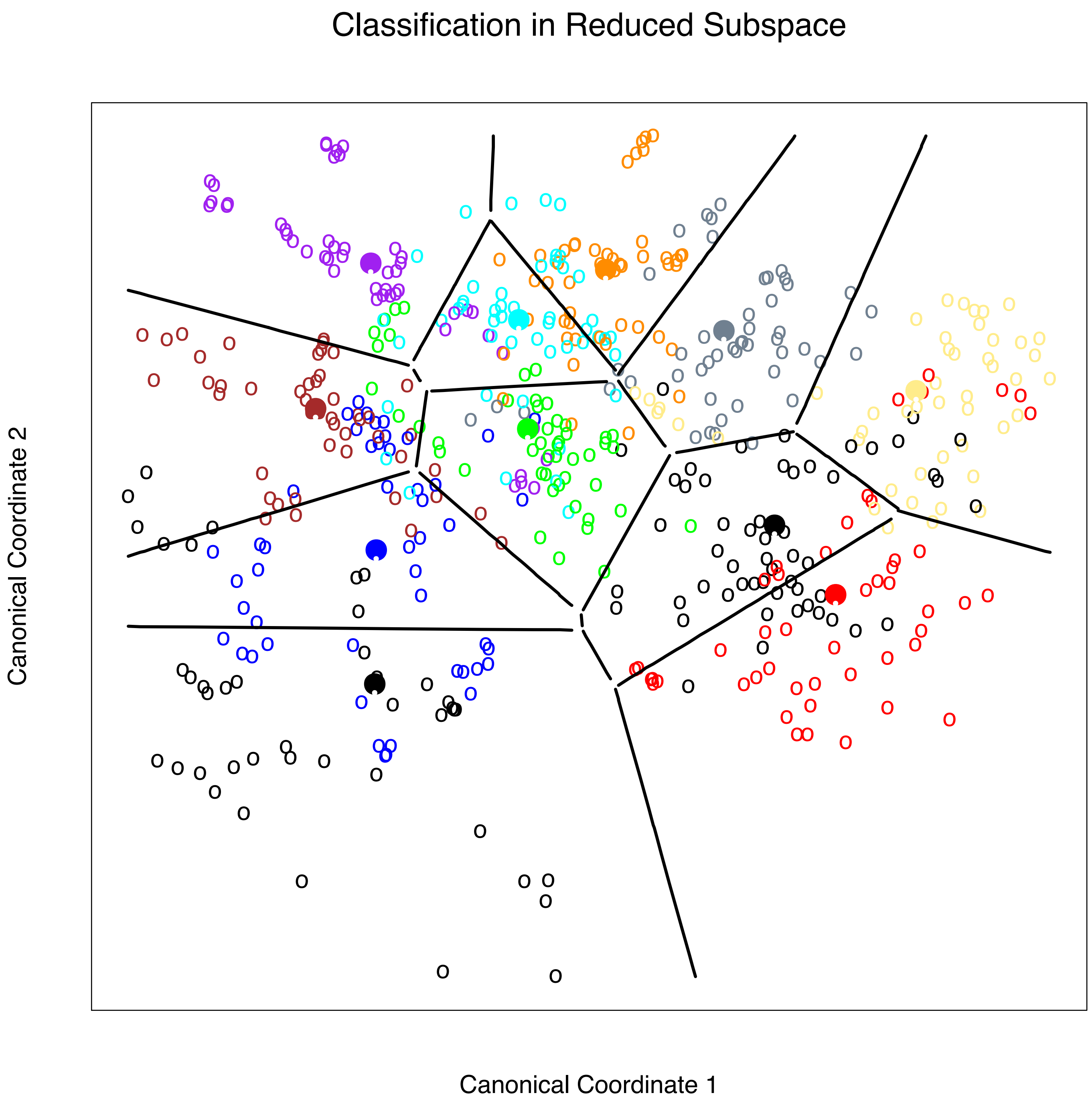


Figure 4.11 (Hastie et al.)

Logistic Regression

Revisiting LDA for Binary Classes

- LDA assumes predictors are normally distributed

$$
\begin{aligned}
\log \frac{\Pr(G = k | X = x)}{\Pr(G = \ell | X = x)} &= \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\boldsymbol{\mu}_k + \boldsymbol{\mu}_\ell)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k + \boldsymbol{\mu}_\ell) \\
&\quad + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) \\
&= \alpha_0 + \boldsymbol{\alpha}^\top \mathbf{x}
\end{aligned}
$$

- Log odds of class 1 vs 2 is a linear function
  - Why not estimate coefficients directly?

Link Functions

- How to combine regression and probability?
  - Use regression to model the posterior
- Link function
  - Map from real values to [0,1]
  - Need probabilities to sum to 1

Logistic Regression

- Logistic function (or sigmoid)

$$f(z) = \frac{1}{1 + \exp(-z)}$$

- Apply sigmoid to linear function of the input features

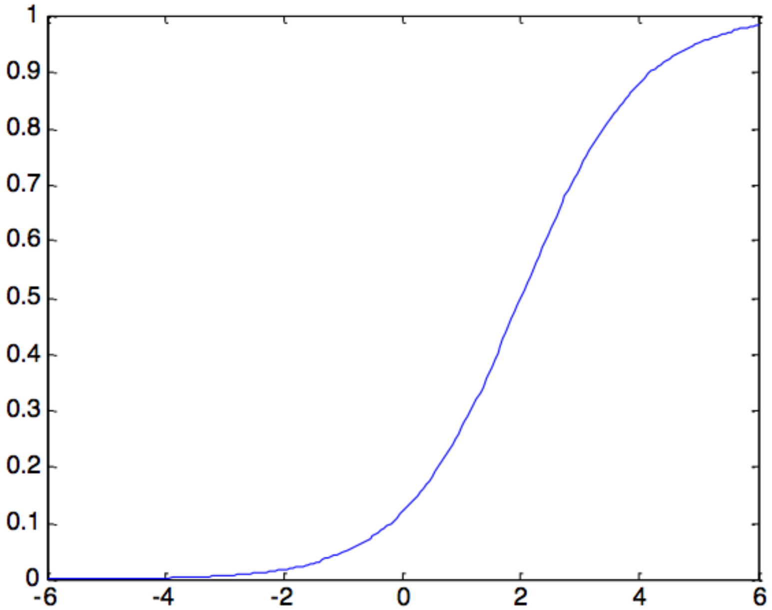
$$\Pr(G = 0 | \mathbf{X}, \boldsymbol{\beta}) = \frac{1}{1 + \exp(\mathbf{X}\boldsymbol{\beta}^\top)}$$

$$\Pr(G = 1 | \mathbf{X}, \boldsymbol{\beta}) = \frac{\exp(\mathbf{X}\boldsymbol{\beta}^\top)}{1 + \exp(\mathbf{X}\boldsymbol{\beta}^\top)}$$

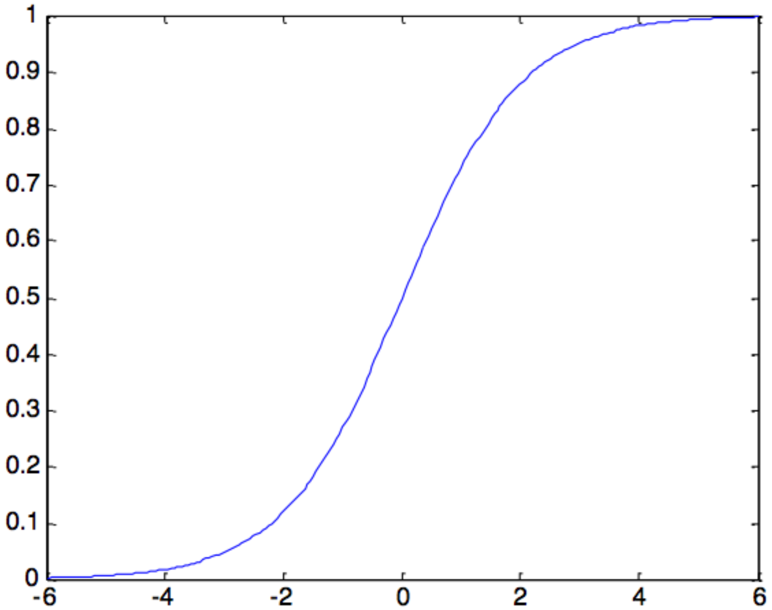
Sigmoid Function

$$f(x) = \frac{1}{1 + \exp^{(w_0 + w_1 x)}}$$

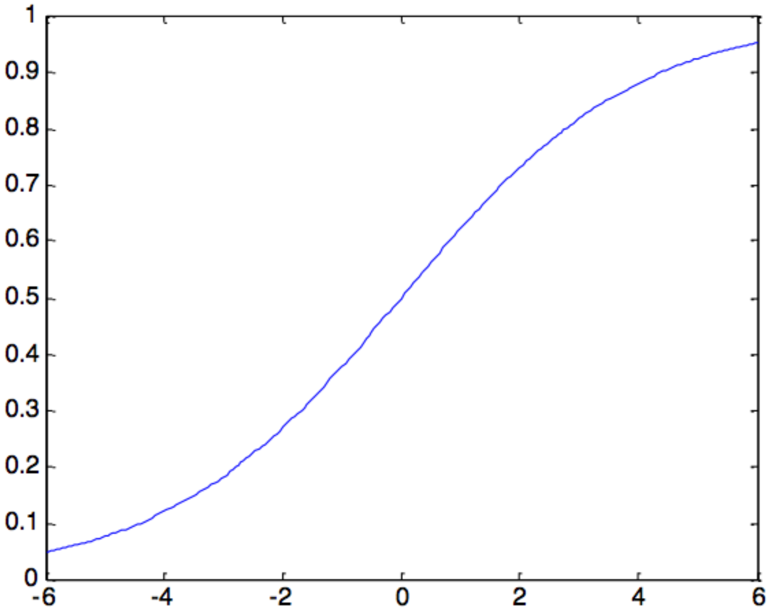
w_0=-2, w_1=-1



w_0=0, w_1=-1



w_0=0, w_1=-0.5



Fitting Logistic Regression Models

- No longer straightforward (not simple least squares)
- See book for discussion of two-class case
- Use optimization methods (Newton-Raphson)
- In practice use a software library

Optimization: Log Likelihood

- Maximize likelihood of your training data by assuming class labels are conditionally independent

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \Pr(G = k | \mathbf{X} = \mathbf{x}_i), \theta = \{\beta_0, \boldsymbol{\beta}\}$$

- Log likelihood

$$\ell(\theta) = \sum_{i=1}^{n} \Pr(G = k | \mathbf{X} = \mathbf{x}_i)$$
$$= p_k(\mathbf{x}; \theta)$$

Optimization: Logistic Regression

- Log likelihood for logistic regression

$$\ell(\theta) = \sum_{i=1}^{n} (y_i \boldsymbol{\beta}^\top \mathbf{x}_i - \log(1 + \exp^{(\boldsymbol{\beta}^\top \mathbf{x}_i)}))$$

- Simple gradient descent using derivatives

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i (y_i - p(\mathbf{x}; \boldsymbol{\beta}))$$

- Book illustrates Newton-Raphson which uses 2nd order information for better convergence

Logistic Regression Coefficients

- How to interpret coefficients?
  - Similar to interpretation for linear regression
- Increasing the ith predictor x_i by 1 unit and keeping all other predictors fixed increases:
  - Estimated log odds (class 1) by an additive factor β_i
  - Estimated odds (class 1) by a multiplicative factor $\exp \beta_i$

Example: South African Heart Disease

- Predict myocardial infarction – “heart attack”
- Variables:
  - sbp – Systolic blood pressure
  - tobacco – Tobacco use
  - ldl – Cholesterol measure
  - famhist – Family history of myocardial infarction
  - obesity, alcohol, age

Example: South African Heart Disease

	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

Table 4.2 (Hastie et al.)

Example: South African Heart Disease

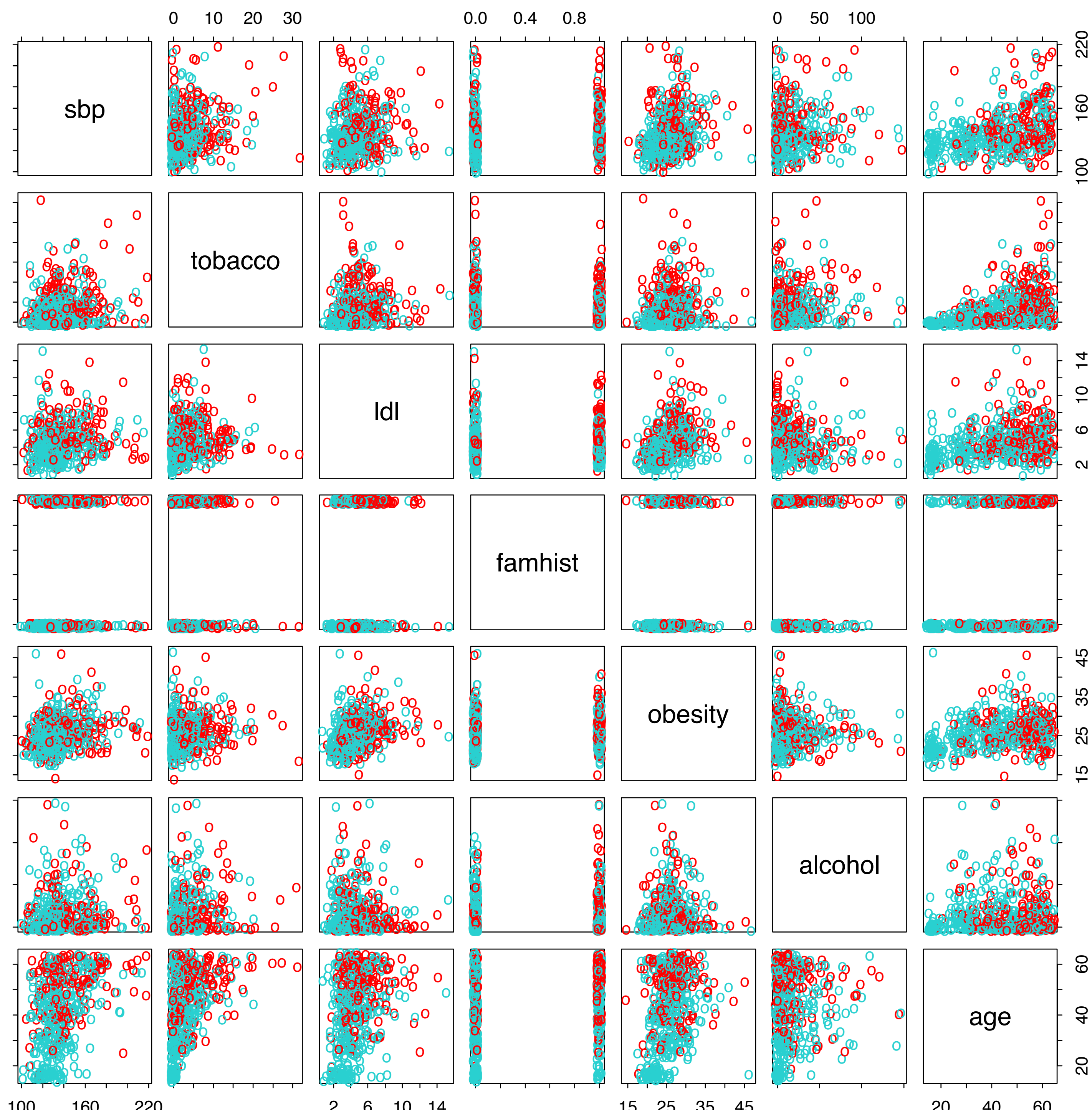


Figure 4.12 (Hastie et al.)

Example: South African Heart Disease

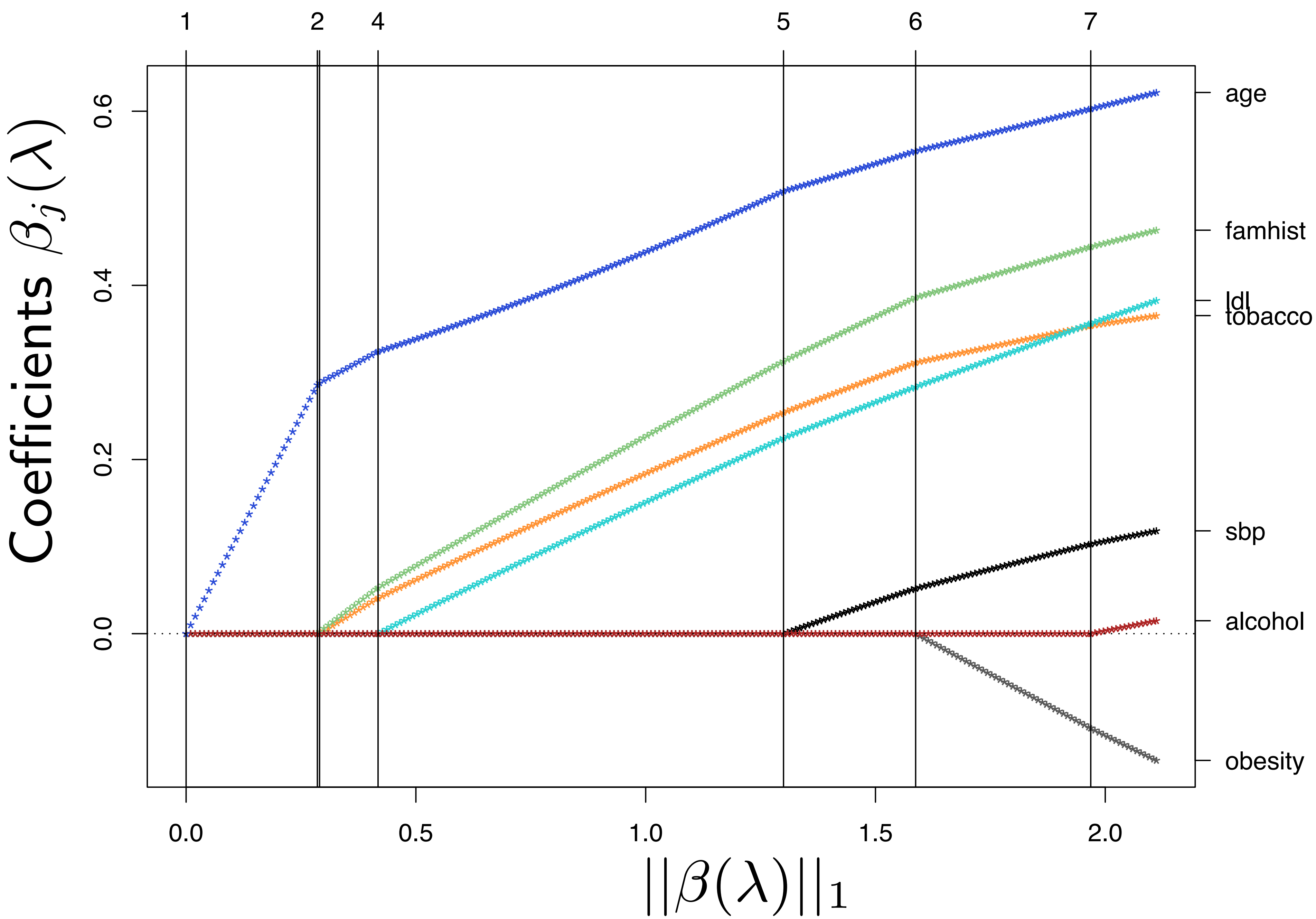
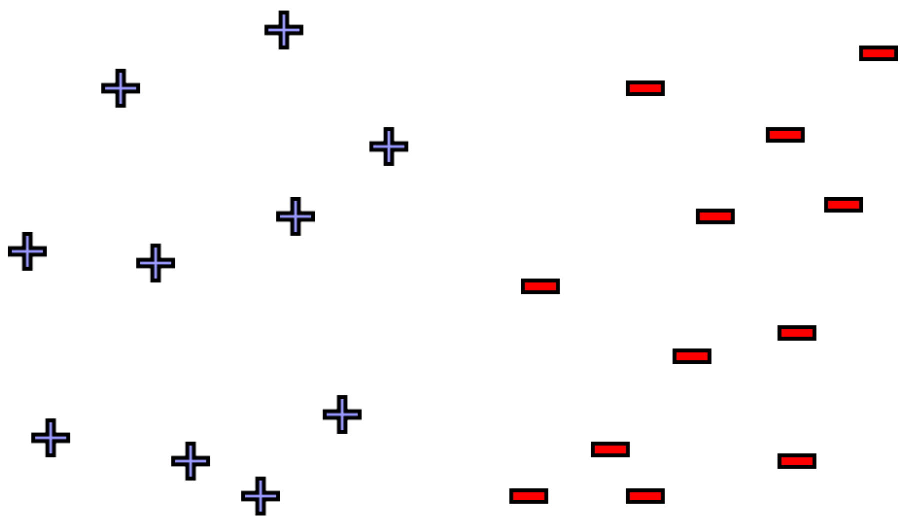


Figure 4.13 (Hastie et al.)

Linear Separability & Logistic Regression

- What happens in the case when my data is completely separable?



  - Weights go to infinity
  - Infinite number of MLEs

Use some form of regularization to avoid this scenario

LDA vs Logistic Regression

- LDA estimates the Gaussian parameters and prior (easy!)
- Logistic regression estimates coefficients directly based on maximum likelihood (harder!)
- Both have linear decision boundaries that are different — why?
  - LDA assumes normal distribution within class
  - Logistic regression is more flexible and robust to situations with outliers and not normal class conditional densities

Multiclass Logistic Regression

- Extension to K classes: use K - 1 models

$$\log \frac{\Pr(G = j|X = x)}{\Pr(G = K|X = x)} = \beta_{0j} + \boldsymbol{\beta}_j^\top \mathbf{x}$$

  - Model the log odds of each class to a base class

  - Fit coefficients jointly by maximum likelihood

- Put them together to get posteriors

$$\Pr(G = i|\mathbf{x}) = \frac{\exp(\beta_{0i} + \boldsymbol{\beta}_i^\top \mathbf{x})}{1 + \sum_j \exp(\beta_{0j} + \boldsymbol{\beta}_j^\top \mathbf{x})}, i \neq j$$

Logistic Regression Properties

- Advantages
  - Parameters have useful interpretations — the effect of unit change in a feature is to increase the odds of a response multiplicatively by the factor $\exp \beta_i$
  - Quite robust, well developed
- Disadvantages
  - Parametric, but works for entire exponential family of distributions
  - Solution not closed form, but still reasonably fast

Logistic Regression Additional Comments

- Example of a generalized linear model with canonical link function = logit, corresponding to Bernoulli
  - For more information, see short course by Heather Turner (http://statmath.wu.ac.at/courses/heather_turner/glmCourse_001.pdf)
- Old technique but still very widely used
- Output layer for neural networks

Comparison on Vowel Recognition

Technique	Error Rates	
	Training	Test
Linear regression	0.48	0.67
Linear discriminant analysis	0.32	0.56
Quadratic discriminant analysis	0.01	0.53
Logistic regression	0.22	0.51

Table 4.1 (Hastie et al.)

Generative vs Discriminative

- Generative: separately model class-conditional densities and priors
  - Example: LDA, QDA
- Discriminative: try to obtain class boundaries directly through heuristic or estimating posterior probabilities
  - Example: Decision trees, logistic regression

Generative vs Discriminative Analogy

- Task is to determine the language someone is speaking
  - Generative: Learn each language and determine which language the speech belongs to
  - Discriminative: Determine the linguistic differences without learning any language