STATS 415: Classification – Bayes rule, Linear and Quadratic Discriminant Analysis

Prof. Liza Levina

Department of Statistics, University of Michigan

Examples of Classification

- Classifying tumors as benign or malignant.
- Classifying credit card transactions as legitimate or fraudulent.
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil.
- Categorizing news stories as finance, weather, entertainment, sports, etc.

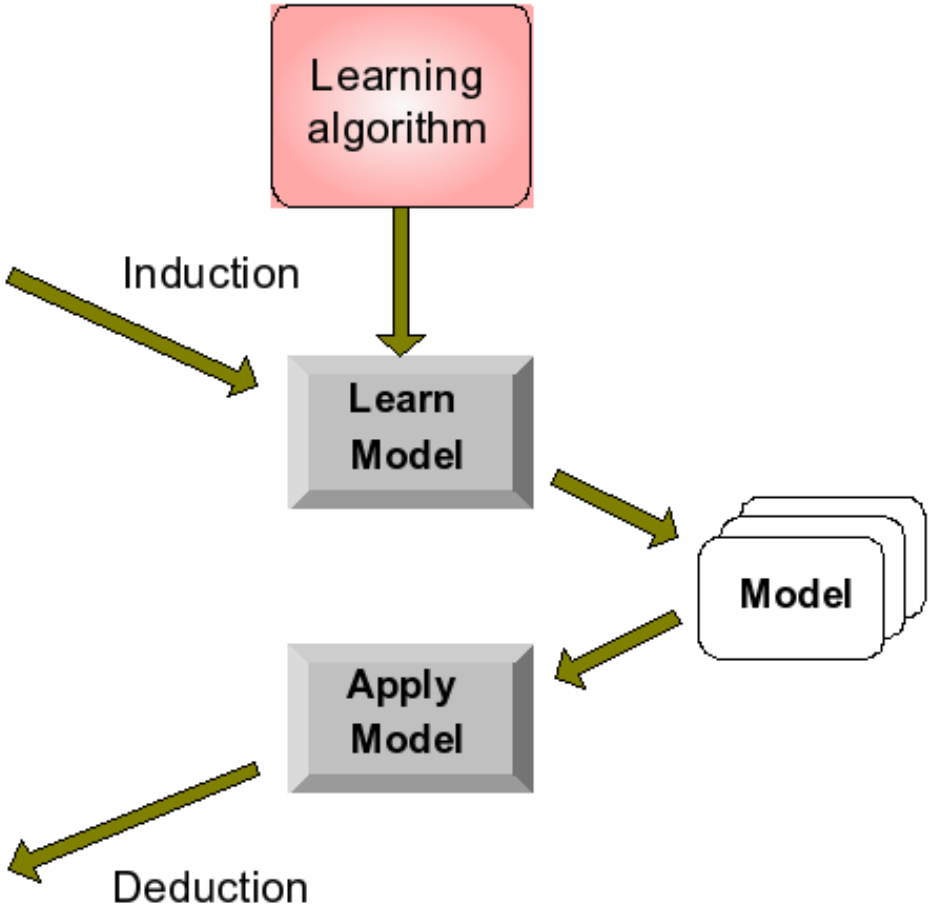
Classification basics

- A set of variables, one of which is designated as response (class), and is categorical
- Given training data, find a model for the class variable as a function of other variables (predictors).
- Just like with regression, the function can be explicit or implicit
- Goal 1 (prediction): *new* data points (test set) should be assigned a class as accurately as possible.
- Goal 2 (inference): understand which of the variables help predict class, and how they are connected

Classification principles

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Mathematical Setup

- Class label y
- Covariates / features: $x = (x_1, x_2, \ldots, x_p)$
- y takes values in a finite, unordered set (categorical)
  - Two-class: $y \in \{c_1, c_2\}$
  - Multi-class: $y \in \{c_1, c_2, \ldots, c_K\}$
- We have training data: $(x_1, y_1), \ldots, (x_n, y_n)$.

Optimal Classifier

- (x, y) have a joint probability distribution (as random variables (X, Y))
- Fundamental assumption: this distribution is the same for the training and test sets
- Choose $\hat{C}$ to have small misclassification error, a.k.a. expected risk on future data points from the same distribution:

$$R(\hat{C}) = P(\hat{C}(X) \neq Y)$$

- Bayes optimal classifier: pick the most likely class

$$\begin{aligned} C^*(x_0) &= \arg\min_C \mathsf{R}(C) \\ &= \arg\max_k P(y = c_k | x = x_0) \end{aligned}$$

The Bayes rule for $K = 2$

- Assume Y takes values 1 or 2, and X is continuous variable with a probability density function which depends on the class Y.
- Prior class probabilities:

$$P(Y = 1) = \pi_1, \; P(Y = 2) = \pi_2, \; \pi_1 + \pi_2 = 1$$

- Class-conditional densities:

$$P(X = x | Y = 1) = p_1(x), \; P(X = x | Y = 2) = p_2(x)$$

- Optimal classifier: compare $P(Y=1|X=x)$ to $P(Y=2|X=x)$ and pick the larger one
- Need to compute the posterior probability: use Bayes rule

$P(Y=1|X=x) =$ proportional to pi(x)πi

depends on p1(x)π1 and p2(x)π2 which is bigger

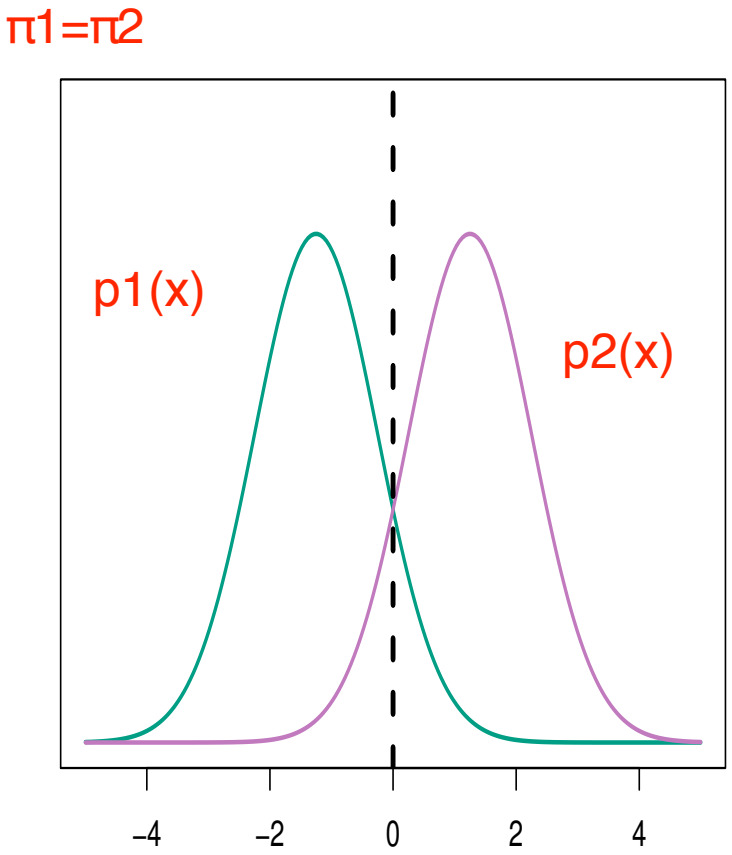
The general case

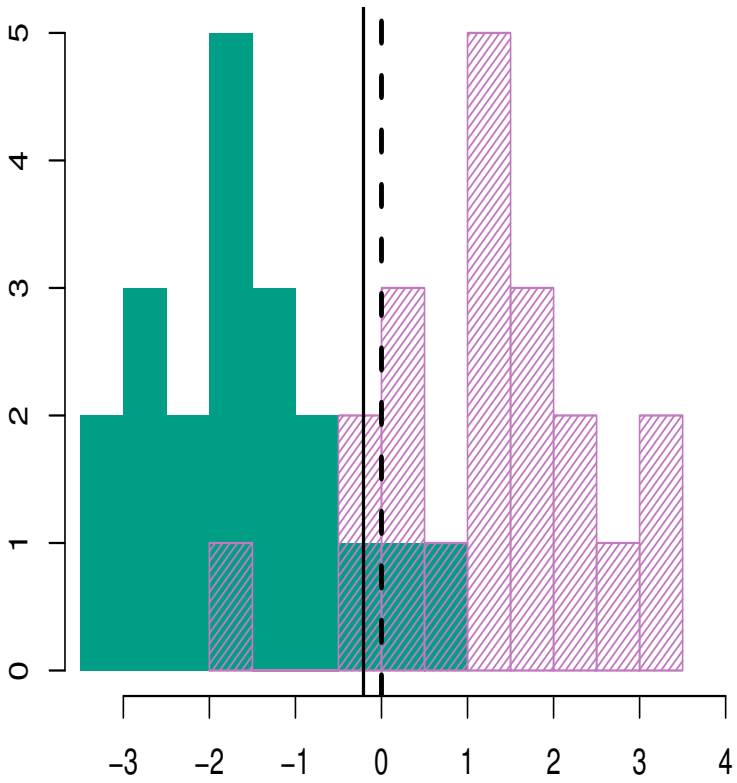
- Let π_k be the prior probability of class k, $P(Y = c_k)$.
- Let $p_k(x)$ be the class-conditional density of X in class k.
- The posterior probability

$$P(Y = c_k | X = x) = \frac{p_k(x)\pi_k}{\sum_{k=1}^{K} p_k(x)\pi_k}$$

- The optimal classifier picks c_k such that $P(Y = c_k | X = x)$ is maximized

Theory vs practice





Generative Classification Methods

- Estimate $p_k(x) = P(X = x|Y = c_k)$ from training data, then plug-in the estimated $\hat{p}_k(x)$ into the Bayes rule:

$$\hat{P}(y = c_k|X = x) \text{ is proportional to } \hat{p}_k(x)\hat{P}(Y = k)$$

- Also need to estimate prior class probabilities P(y=ck)

Common options for estimating priors

$$\hat{\pi}_k = \hat{P}(Y = k)$$

- Assume equal for all classes and drop them
- Estimate from data as

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

  where n_k is the number of observations from class k in the training data, $n = n_1 + n_2 + \cdots + n_K$.
- Use prior knowledge

Discriminative Classification Methods

Estimate $P(Y = k|X = x)$ directly, without going through the Bayes rule

- Logistic regression
- Support vector machines (SVM)
- Classification trees (CART) ??
- Ensemble methods: bagging, boosting, random forests

Parametric generative methods

- Class densities are unknown but we believe/assume that the data come from a parametric model

$$p_k(x) = p_k(x; \theta_k),\ k = 1, \ldots, K$$

- The form of p_k is assumed known, and the distribution is determined by the vector of unknown parameters $\theta_k \in \Theta$.
- Plug-in classifier

$$\hat{p}(k|x) \propto \hat{\pi}_k p_k(x; \hat{\theta}_k)$$

- Assuming normal densities gives linear and quadratic discriminant analysis

Non-parametric generative methods

- Do not assume a parametric model for $p_k(x)$
- Estimate the density $\hat{p}_k(x)$ from training data by non-parametric methods
- K nearest neighbor classifier

Pros and cons of parametric methods

Pros Pros and cons of parametric methods

- The rule is fully determined by a small set of parameters
- Relatively easy to fit even in high dimensions
- Many quantities (e.g. error estimates) can be computed explicitly
- The most popular choice for $\hat{\theta}$ is the maximum likelihood estimator, lots of known good properties

Cons

- Parametric assumptions may not hold
- Even an excellent estimate of $\hat{\theta}$ does not always mean that $p(x;\hat{\theta})$ is a good approximation to $p(x;\theta)$ (requires extra conditions)

Pros and cons of nonparametric methods

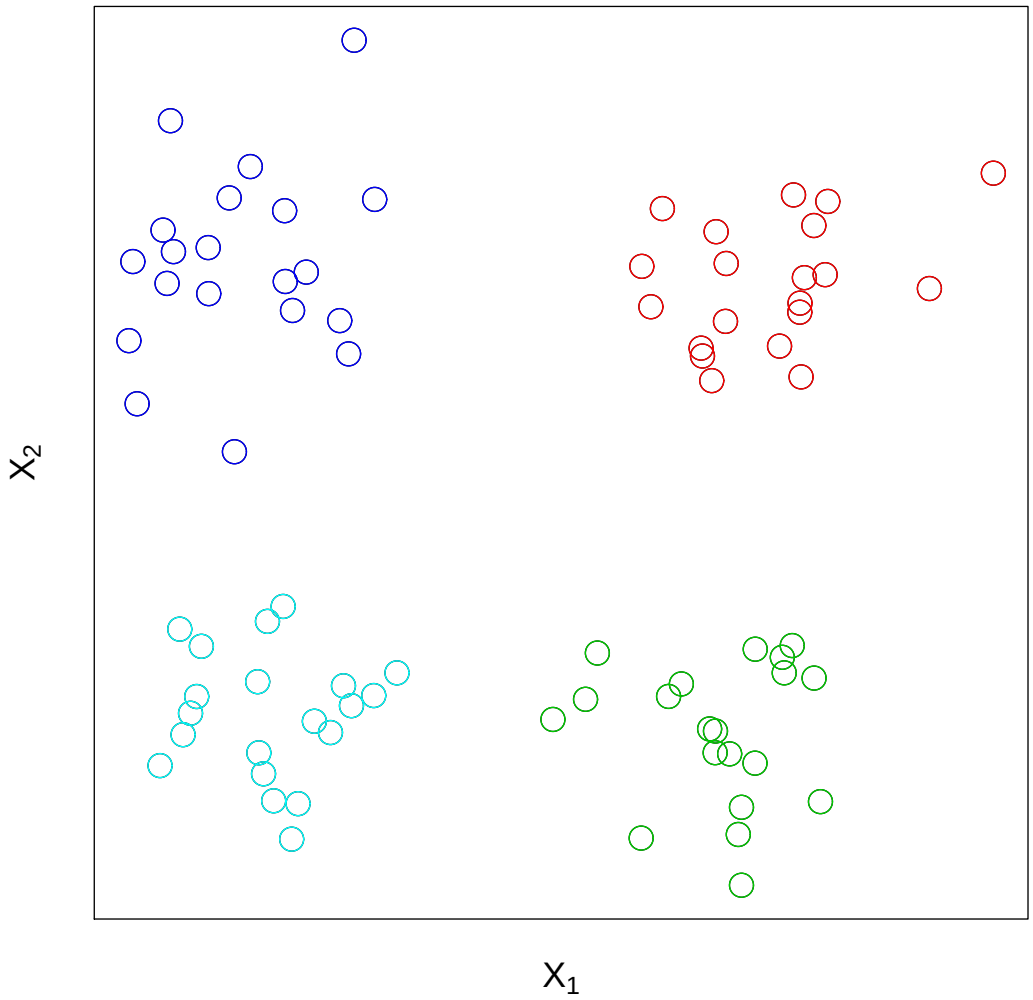
Pros Pros and cons of nonparametric methods

- No need to make assumptions
- Very flexible, in principle can approximate any posterior probabilities

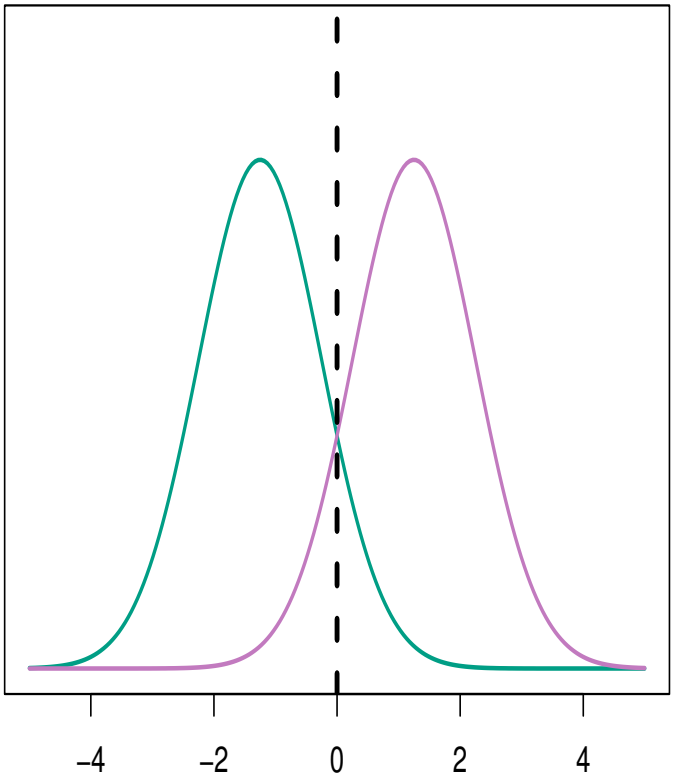
Cons

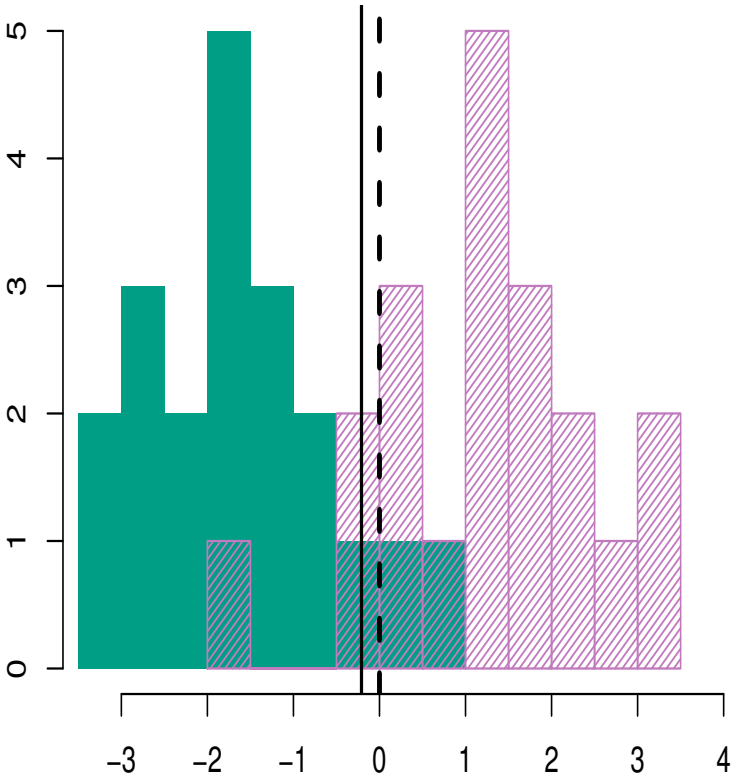
- Usually nothing can be computed explicitly
- Danger of overfitting (unless regularized)
- "Curse of dimensionality" – density estimation fails in high dimensions

??


X_2
X_1

A single Gaussian predictor





The Bayes rule for a Gaussian predictor

- Let X be univariate ($p = 1$), $K = 2$, $p_1(x) \sim N(\mu_1, \sigma_1^2)$, $p_2(x) \sim N(\mu_2, \sigma_2^2)$

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

- Compute the posterior probability: for $k = 1, 2$,

$$P(Y = k|x) = \frac{\pi_k p_k(x)}{\pi_1 p_1(x) + \pi_2 p_2(x)} \propto \pi_k p_k(x)$$

- Assume for simplicity $\pi_1 = \pi_2 = 0.5$. The Bayes rule assigns class 1 if $p_1(x) > p_2(x)$, class 2 otherwise

- Take logs and simplify
- The rule becomes: assign class 1 if

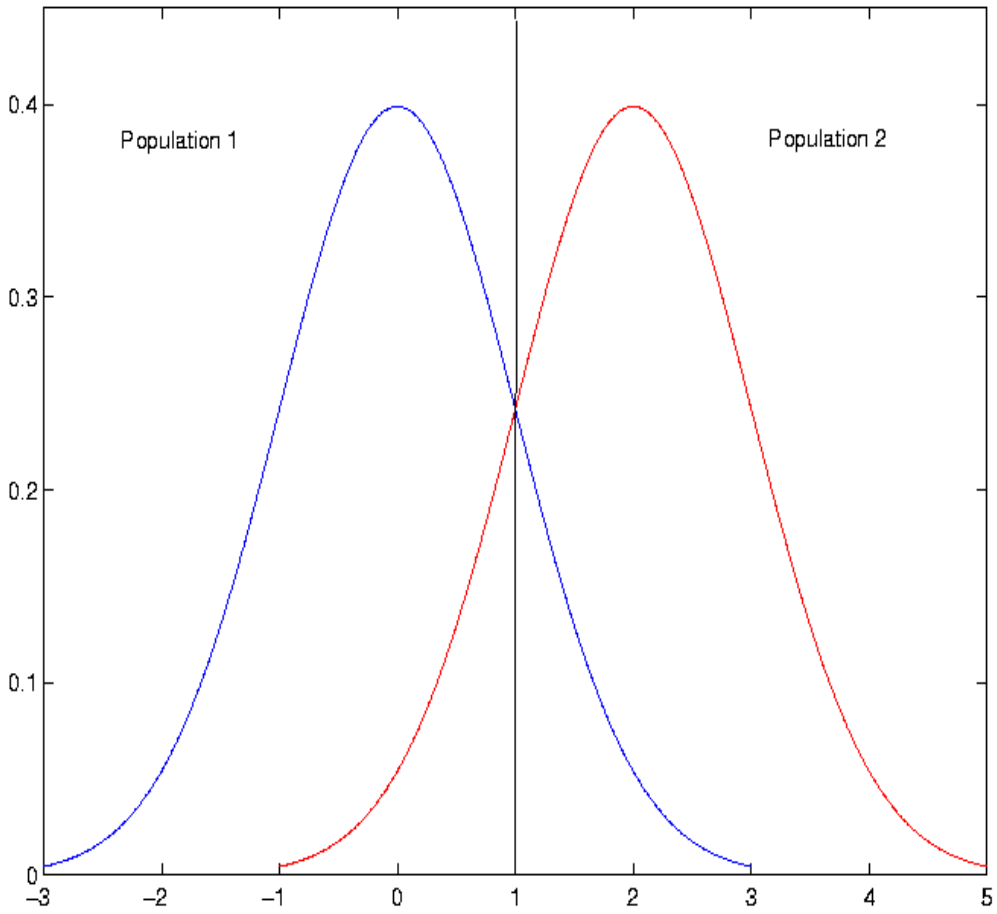
cross out if σ1=σ2

$$x^2 \cancel{\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right)} - 2x\left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}\right) + \left(\frac{\mu_2^2}{\sigma_2^2} - \frac{\mu_1^2}{\sigma_1^2}\right) + 2\log\cancel{\left(\frac{\sigma_2}{\sigma_1}\right)} > 0 \ ,$$

otherwise assign class 2.

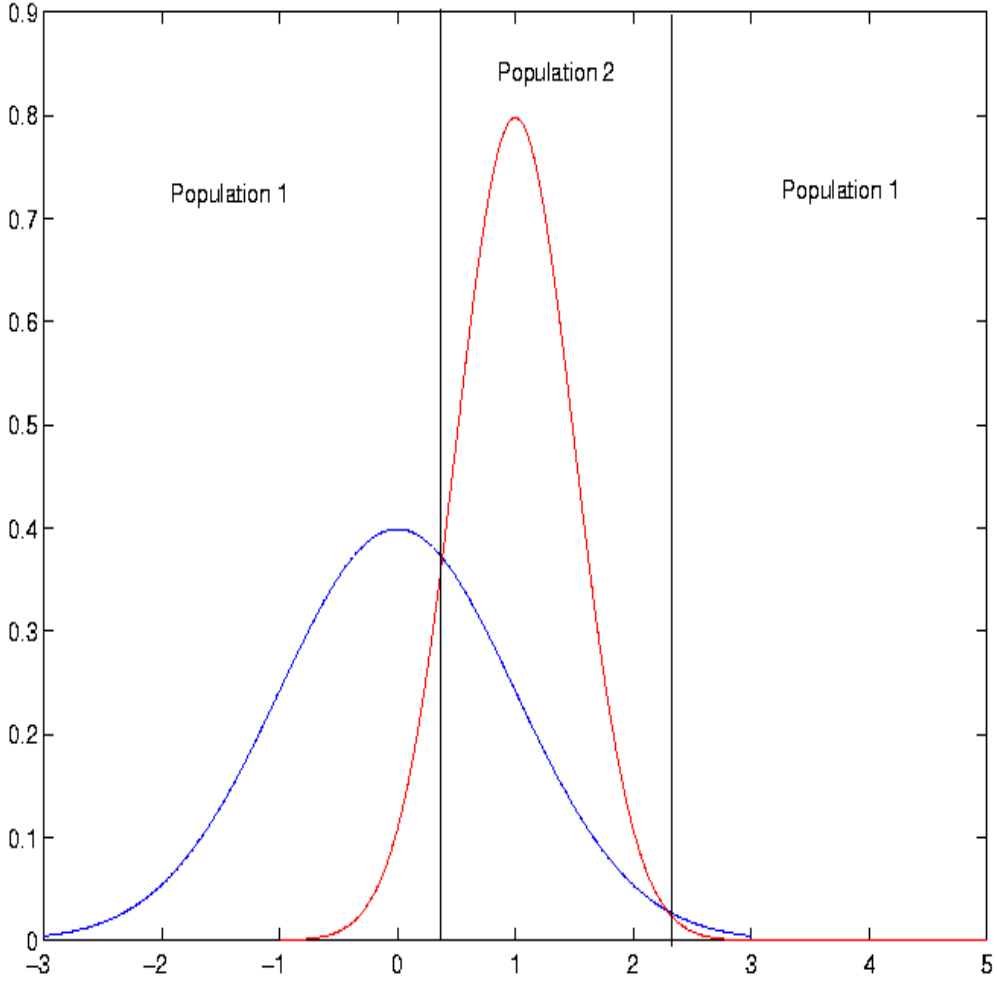
- This is a quadratic function in x ax+b>0
- What if we assume $\sigma_1 = \sigma_2$? see page 22 change class once

Example (equal variances, linear rule)



- $\sigma_1 = \sigma_2 = 1$, $\mu_1 = 0$, $\mu_2 = 2$.
- The Bayes rule: class 1 if $x < 1$, otherwise class 2.

Example (unequal variances, quadratic rule)



- $\sigma_1 = 1$, $\sigma_2 = 1/2$, $\mu_1 = 0$, $\mu_2 = 1$.
- The Bayes rule: class 1 if $x < .381$ or $x > 2.286$, otherwise class 2.

Linear Discriminant Analysis, $p = 1$, $K > 2$

We model X in each class as Gaussian,

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

- LDA: assume $\sigma_k^2 = \sigma^2$ for all k. p=1,K>2
- Compare $P(Y = k|X = x) \propto p_k(x)\pi_k$
- For each k, the discriminant function is

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

- Decision rule

$$\hat{C}(x) = \arg\max_k \delta_k(x)$$

Multivariate Normal Density

Model each class density as multivariate Gaussian $N(\mu_k, \Sigma_k)$:

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k)}.$$

y,x1,x2,x3

mu是一个向量

其中每一个值是x_n的均值

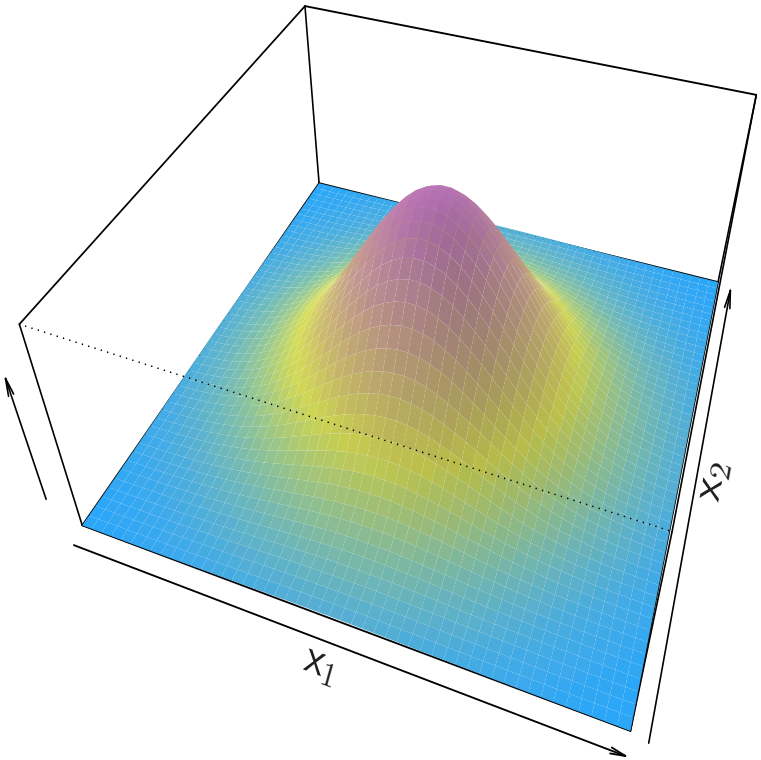
- μ_k: the mean vector for class k matrix_ij=cov(x_i,x_j), 对角线就是var(x_i)
- Σ_k: the covariance matrix for class k; $|\Sigma_k|$ is the determinant of Σ_k
- Each variable x_j is also marginally normal with mean μ_{kj} and variance $\Sigma_{k,jj}$.
- LDA: assume $\Sigma_k = \Sigma$ for all k.
- For each k, the discriminant function is

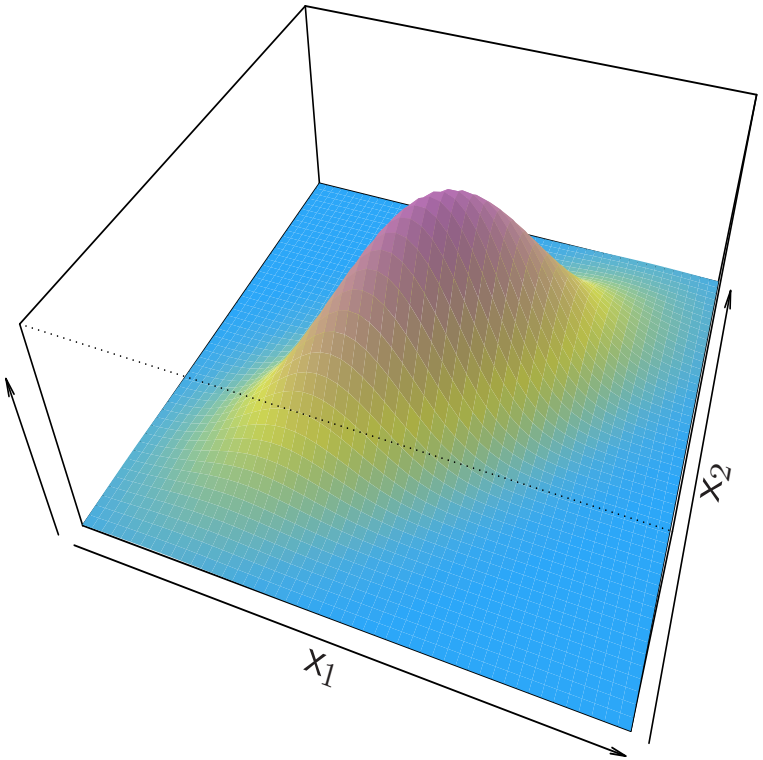
$$\delta_k(x) = x^\top \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^\top \Sigma^{-1}\mu_k + \log \pi_k$$

- Decision rule

$$\hat{C}(x) = \arg\max_k \delta_k(x)$$

Bivariate normal density





Linear Discriminant Analysis, $p > 2$

- Model each class density as multivariate Gaussian $N(\mu_k, \Sigma_k)$
- LDA: assume $\Sigma_k = \Sigma$ for all k.
- For each k, the discriminant function is

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

- Decision rule

$$\hat{C}(x) = \arg\max_k \delta_k(x)$$

- Comparing class k and class k', the decision boundary is given by

$$\{x : x^\top \Sigma^{-1} (\mu_k - \mu_{k'}) + \log \frac{\pi_k}{\pi_{k'}} - \frac{1}{2} (\mu_k + \mu_{k'})^\top \Sigma^{-1} (\mu_k - \mu_{k'}) = 0\}$$

- Linear decision boundary, with the directional vector $\Sigma^{-1}(\mu_k - \mu_{k'})$; generally not in the direction of $(\mu_k - \mu_{k'})$.
- Assuming equal priors, we classify x to the class with the closest centroid to x, using the squared Mahalanobis distance corresponding to Σ:

$$(x - \mu_k)^T \Sigma^{-1} (x - \mu_k)$$

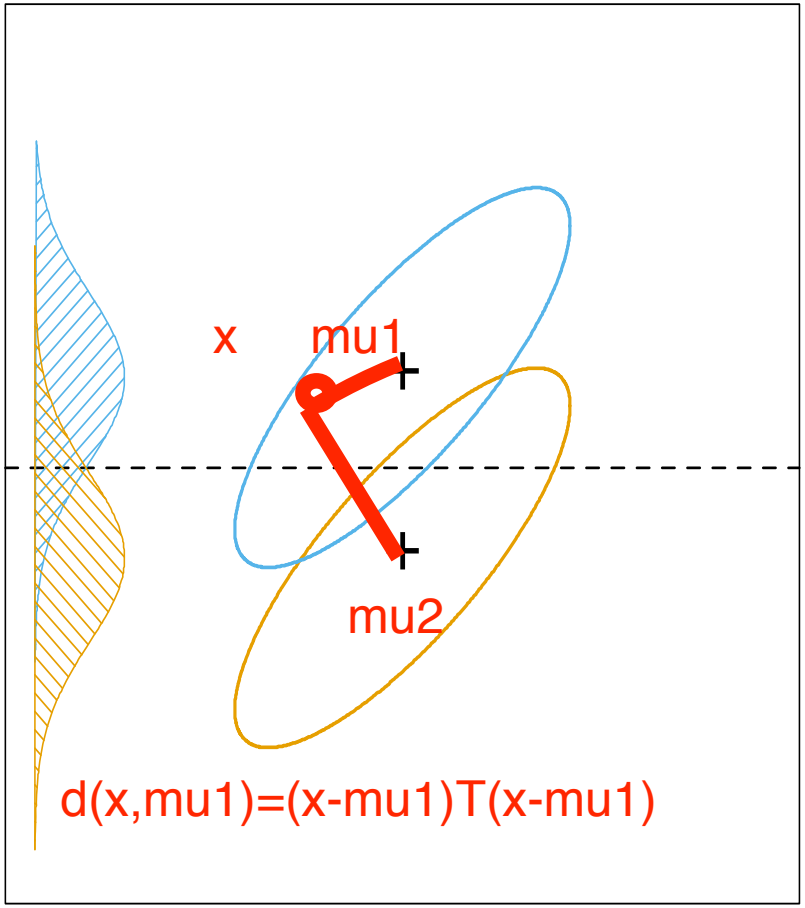
- Special case: $\Sigma = I$; then classify to the class with the closest centroid in Euclidean distance

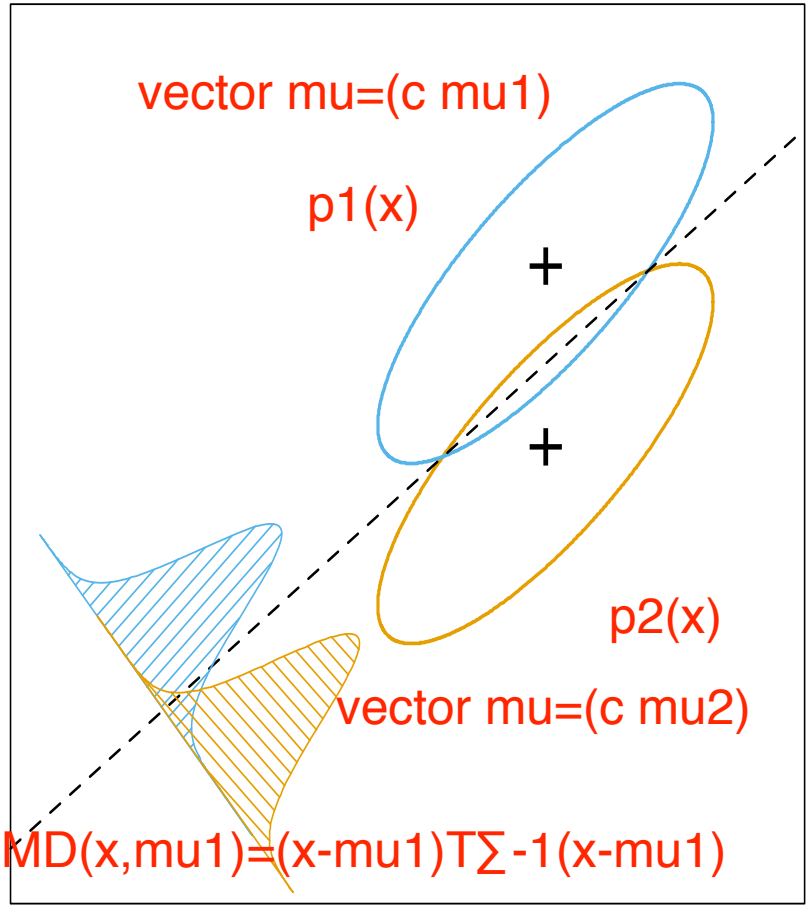
$$(x - \mu_k)^T (x - \mu_k)$$

using x2 alone

The first predictor is useless

LDA



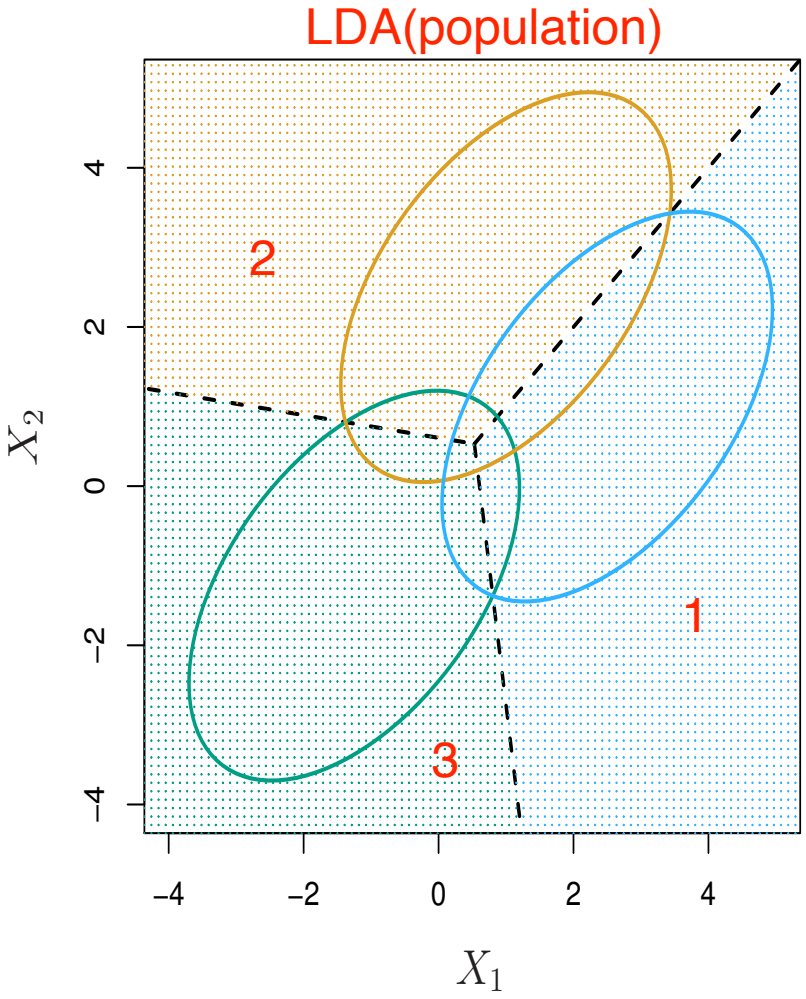


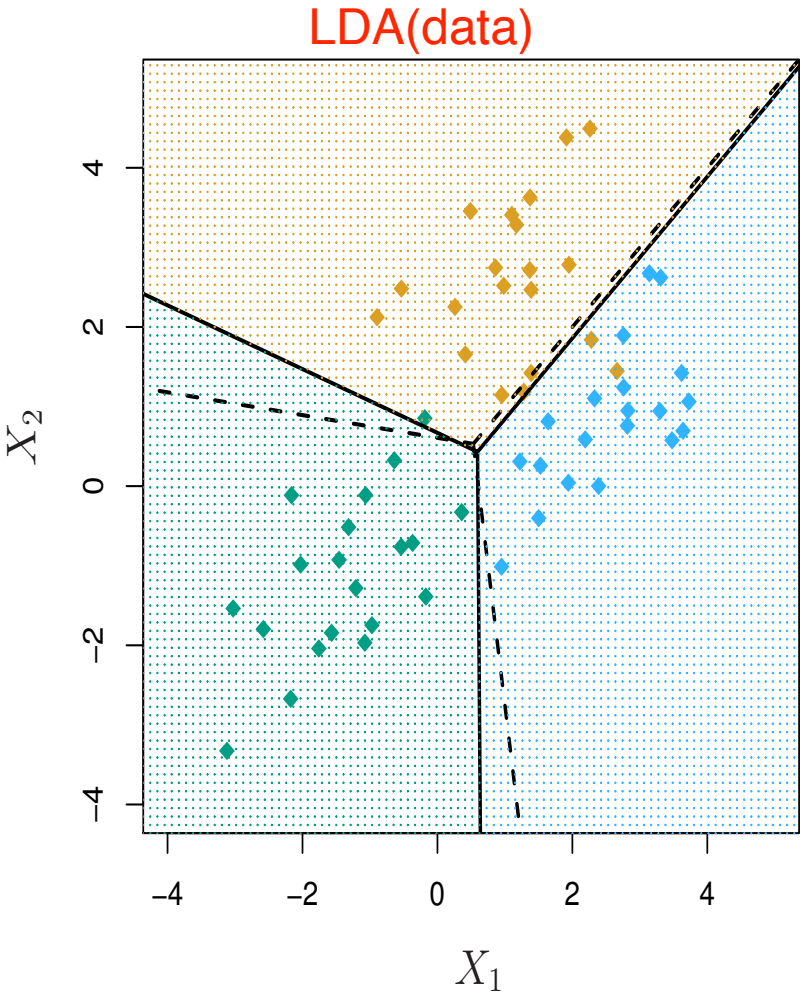
σ=mu1-u2

a+bTx=0

x1

σ=∑-1(mu1-mu2)


LDA(population)
2
1
3
X_1
X_2
−4
−2
0
2
4


LDA(data)
X_1
X_2
−4
−2
0
2
4

Parameter Estimation for LDA

In practice, we estimate parameters from training data.

- Class priors: let n_k be the number of training observations in class k,

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

- Class means:

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

- Class covariances:

$$\hat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{i:y_i=c_k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\mathsf{T}$$

- The pooled covariance

$$\hat{\Sigma} = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{i:y_i=c_k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\mathsf{T}$$

Quadratic Discriminant Analysis

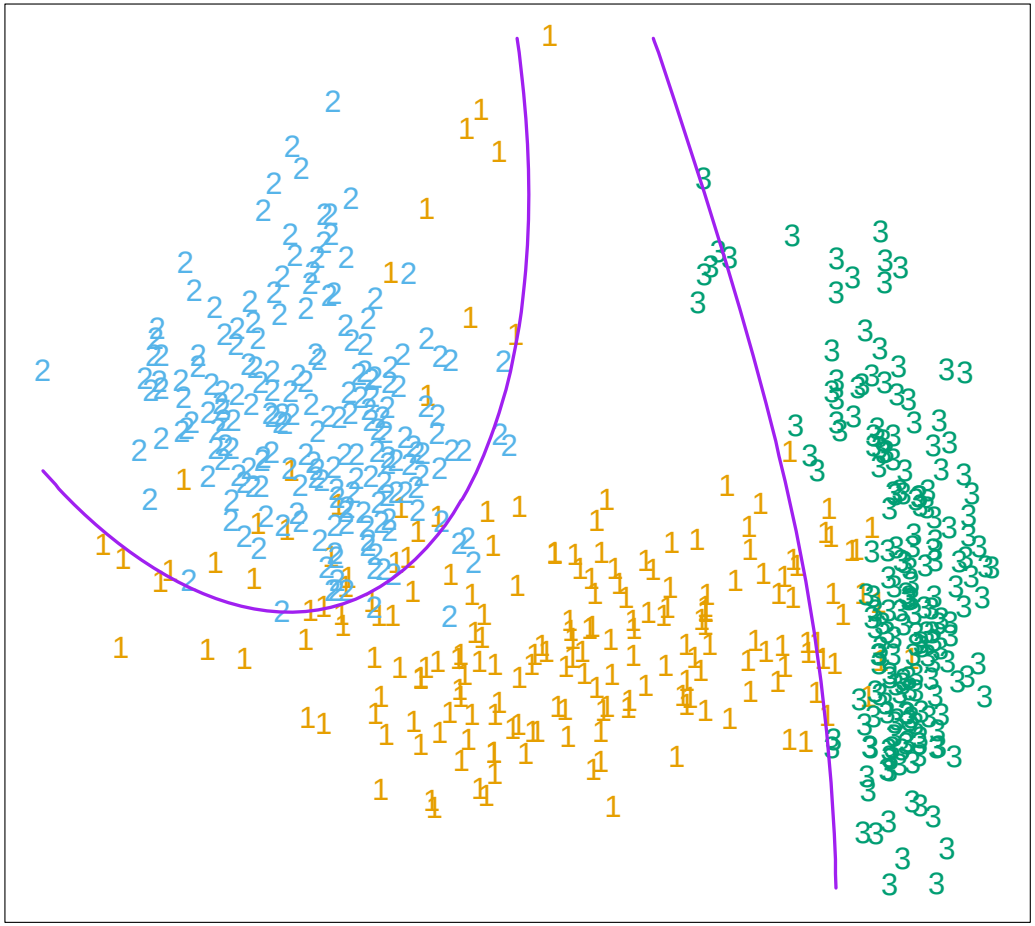
- Σ_k's are allowed to be different. Quadratic discriminant analysis
- Discriminant function

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k) + \log \pi_k \\ &= x^\top W_k x + x^\top w_k + b_k\end{aligned}$$

- The decision boundary between class k and class k' is a quadratic function

$$\{x : x^\top (W_k - W_{k'})x + x^\top (w_k - w_{k'}) + (b_k - b_{k'}) = 0\}$$

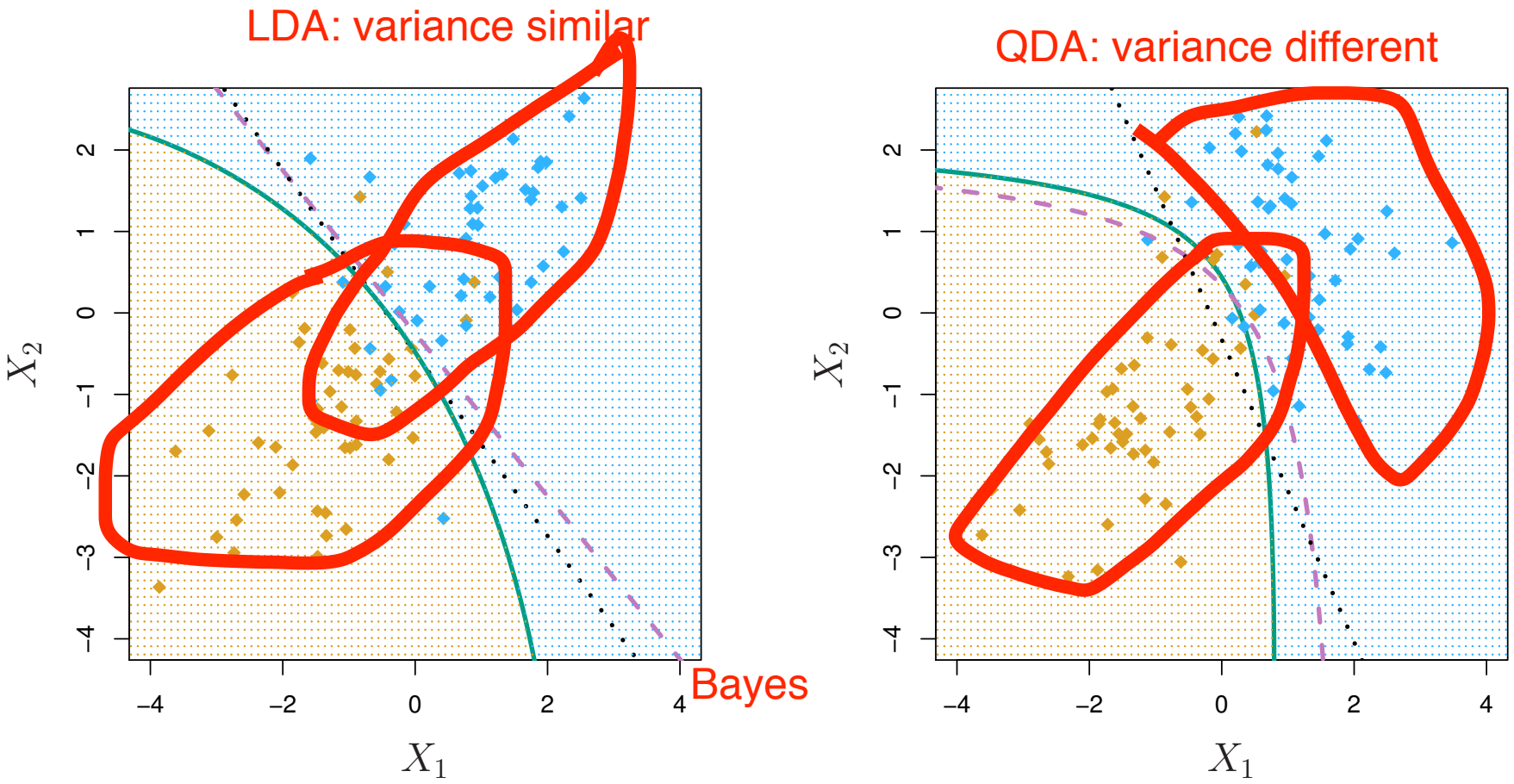
QDA example ($p = 2$, $K = 3$



Which is better, LDA or QDA?

- QDA will work best when the variances are very different between classes and we have enough observations to accurately estimate the variances.
- LDA will work best when the variances are similar among classes or we don't have enough data to accurately estimate the variances.

Comparing LDA to QDA



Naive Bayes Classifier

- Assume independence among input variables when class is given

  Naive Bayes Classifier

$$p_k(x_1, \ldots, x_p) = p_{k,1}(x_1) p_{k,2}(x_2) \ldots p_{k,p}(x_p)$$

- Estimate $p_{k,j}$ for each pair of k and j separately.
- New point is classified to c_k corresponding to the largest $\prod_{j=1}^{p} \hat{p}_{k,j}(x_j) \cdot \pi_k$.
- DQDA (diagonal QDA): assume each Σ_k is diagonal
- DLDA (diagonal LDA): assume the common Σ is diagonal
- Strong assumption, but often classifies well when p is large.

How many parameters?

In all cases, need $K-1$ parameters for the prior probabilities π_k, $K \times p$ parameters for the means μ_k

- LDA: $p(p+1)/2$ covariance parameters; need the total sample size $n > p$ to invert $\hat{\Sigma}$. p+(p-1)+…+1
- QDA: $K \times p(p+1)/2$ covariance parameters; need $n_k > p$ for each class to invert $\hat{\Sigma}_k$
- DLDA: p covariance parameters, only need $n > 1$
- DQDA: Kp covariance parameters, only need $n_k > 1$

LDA

An alternative interpretation of LDA

- Fisher: find linear combinations of variables $a'x$ such that between-class variance (the variance of class centroids) is maximized relative to within-class variance (pooled variance within classes) alternative interpretation of LDA
- This interpretation does not require the normal assumption
- All of classification information for K classes is contained in $K-1$ linear combinations; thus for classification purposes can just compute these $K-1$ linear combinations and use LDA as a dimension reduction technique

Discriminant analysis summary

- LDA and QDA are optimal under the normal assumption
- Work well in practice in many cases, even when the normal assumption is questionable
- Requires estimating a number of parameters proportional to p^2, so needs regularization when $p > n$
- Naive Bayes is a simple and effective regularization when p is much larger than n; there are others.