

Ch 4.4.1: Linear Discriminant Analysis

Lecture 8 - CMSE 381

Michigan State University
:::
Dept of Computational Mathematics, Science & Engineering

February 5, 2024

Announcements

Last time:

- Logistic Regression

Announcements:

- Third homework due Friday! Covers:
 - ▶ Mon 2/12 Review Midterm 1
 - ▶ Weds 2/14 No class
 - ▶ Fri 2/16 Midterm 1
- Office hours
 - ▶ Mon: 4-6pm; Tue: 12:30-2:30pm;
Wed: 7-9pm; Thu: 12:30-2:30pm

Covered in this lecture

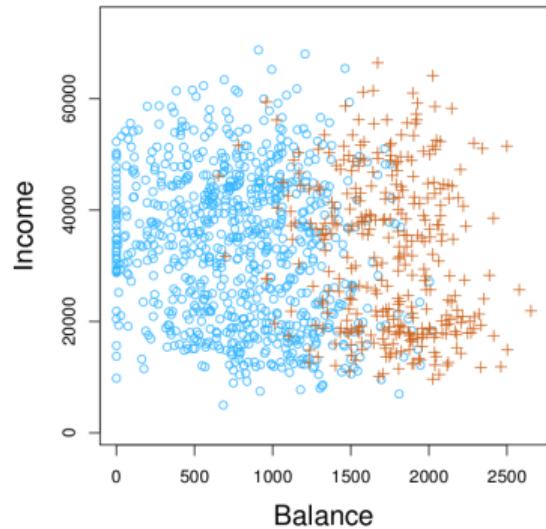
- Bayes theorem
- Linear Discriminant Analysis,
- Quadratic Discriminant Analysis

Section 1

Logistic Regression Review

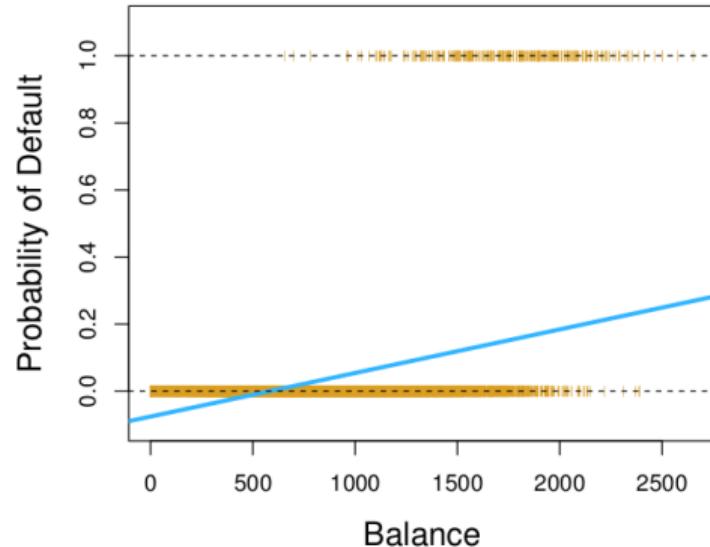
What is classification

- Classification: When the response variable is qualitative
- Goal: Model the probability that Y belongs to a particular category
- Example data:
 $p(\text{balance}) = \Pr(\text{default} = \text{yes} \mid \text{balance})$

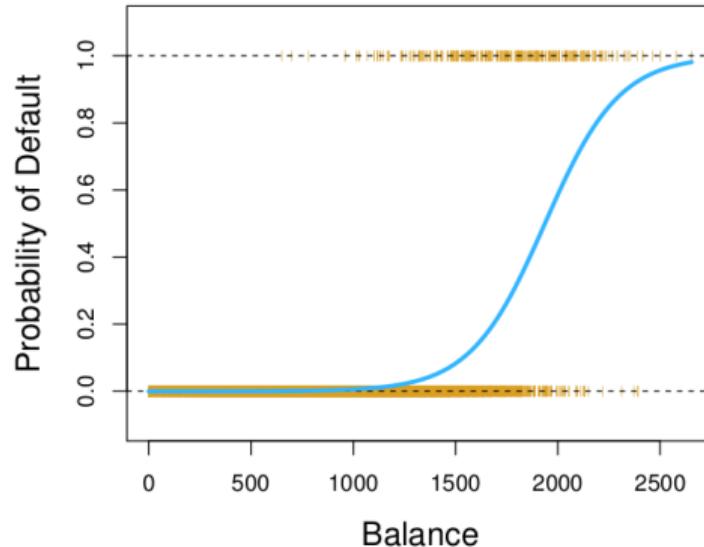


Logistic Regression

$$\Pr(\text{default} = \text{yes} \mid \text{balance}) = \frac{e^{\beta_0 + \beta_1 \text{balance}}}{1 + e^{\beta_0 + \beta_1 \text{balance}}}$$



Linear Regression



Logistic Regression

Odds

$$\frac{p(x)}{1 - p(x)} = \frac{\Pr(Y = 1 | X = x)}{1 - \Pr(Y = 1 | X = x)} = \frac{\Pr(Y = 1 | X = x)}{\Pr(Y = 0 | X = x)}$$

- Logistic function is chosen so that odds are linear
- Can take any value from 0 (low odds) to ∞ (high odds)
- 1 in 4 people with odds of $1/3$ will default since $p(X) = 0.25$ and $0.25/(1 - 0.25) = 1/3$
- 9 in 10 people with odds of 9 will default since $p(X) = 0.9$ and $0.9/(1 - 0.9) = 9$

How to get logistic function

Assume the (natural) log odds (logits) follow a linear model

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

Solve for $p(x)$:

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

Playing with the logistic function: <https://www.desmos.com/calculator/jzsakksqcm>

Interpreting the coefficients

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

β_1 means increasing x by one unit
increases the log odds by 1 unit

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

Estimating Coefficients: Maximum Likelihood Estimation

- **Likelihood:** Probability that data is generated from a model

$$\ell(\text{model}) = \Pr[\text{data} \mid \text{model}]$$

- Find the most likely model

$$\max_{\text{model}} \ell(\text{model})$$

- Hard to maximize likelihood, instead maximize log

$$\max_{\text{model}} \log(\ell(\text{model}))$$

- Strictly increasing log function doesn't change maximum

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

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

$$\ell(\beta_0, \beta_1) = \prod_{i|y_i=1} p(x_i) \prod_{i'|y_{i'}=0} (1 - p(x_{i'}))$$

Multiple Logistic Regression

Multiple features:

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

Equivalent to:

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

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Multinomial Logistic Regression

What if we have a categorical variable with more than two levels (let's say K of them)?

Plan A

Play the dummy variable game:

Make K the baseline:

$$\Pr(Y = k | X = x) = \frac{e^{\beta_{k0} + \beta_{k1}x_1 + \dots + \beta_{kp}x_p}}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0} + \beta_{l1}x_1 + \dots + \beta_{lp}x_p}}$$

$$\Pr(Y = K | X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0} + \beta_{l1}x_1 + \dots + \beta_{lp}x_p}}.$$

Calculated so that log odds between two classes is linear:

$$\log \left(\frac{\Pr(Y = k | X = x)}{\Pr(Y = K | X = x)} \right) = \beta_{k0} + \beta_{k1}x_1 + \dots + \beta_{kp}x_p$$

Example

Predict

$Y \in \{\text{stroke, overdose, seizure}\}$ for hospital visits based on X_p

$$\Pr(Y = \text{stroke} | X = x) = \frac{\exp(\beta_{\text{str},0} + \beta_{\text{str},1}x)}{1 + \exp(\beta_{\text{str},0} + \beta_{\text{str},1}x) + \exp(\beta_{\text{OD},0} + \beta_{\text{OD},1}x)}$$

$$\Pr(Y = \text{overdose} | X = x) = \frac{\exp(\beta_{\text{OD},0} + \beta_{\text{OD},1}x)}{1 + \exp(\beta_{\text{str},0} + \beta_{\text{str},1}x) + \exp(\beta_{\text{OD},0} + \beta_{\text{OD},1}x)}$$

$$\Pr(Y = \text{seizure} | X = x) = \frac{1}{1 + \exp(\beta_{\text{str},0} + \beta_{\text{str},1}x) + \exp(\beta_{\text{OD},0} + \beta_{\text{OD},1}x)}$$

Plan B: Softmax coding

Treat all levels symmetrically

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0} + \beta_{k1}x_1 + \dots + \beta_{kp}x_p}}{\sum_{l=1}^K e^{\beta_{l0} + \beta_{l1}x_1 + \dots + \beta_{lp}x_p}}.$$

Calculated so that log odds between two classes is linear

$$\log \left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = k'|X = x)} \right) = (\beta_{k0} - \beta_{k'0}) + (\beta_{k1} - \beta_{k'1})x_1 + \dots + (\beta_{kp} - \beta_{k'p})x_p.$$

Softmax example

$$\Pr(Y = \text{stroke} | X = x)$$

$$= \frac{\exp(\beta_{\text{str},0} + \beta_{\text{str},1}x)}{\exp(\beta_{\text{str},0} + \beta_{\text{str},1}x) + \exp(\beta_{\text{OD},0} + \beta_{\text{OD},1}x) + \exp(\beta_{\text{seiz},0} + \beta_{\text{seiz},1}x)}$$

$$\Pr(Y = \text{overdose} | X = x)$$

$$= \frac{\exp(\beta_{\text{OD},0} + \beta_{\text{OD},1}x)}{\exp(\beta_{\text{str},0} + \beta_{\text{str},1}x) + \exp(\beta_{\text{OD},0} + \beta_{\text{OD},1}x) + \exp(\beta_{\text{seiz},0} + \beta_{\text{seiz},1}x)}$$

$$\Pr(Y = \text{seizure} | X = x)$$

$$= \frac{\exp(\beta_{\text{seiz},0} + \beta_{\text{seiz},1}x)}{\exp(\beta_{\text{str},0} + \beta_{\text{str},1}x) + \exp(\beta_{\text{OD},0} + \beta_{\text{OD},1}x) + \exp(\beta_{\text{seiz},0} + \beta_{\text{seiz},1}x)}$$

Section 2

Generative Models

Goal:

Another way to approximate
 $Pr(Y = k | X = x)$

How?
BAYES!!!!

Bayes Theorem

$$A: Y=1 \quad B: X=x$$
$$\underline{P(Y=1 | X=x)} = \frac{\underline{P(Y=1)} \underline{P(X=x | Y=1)}}{\underline{P(X=x)}}$$
$$\underline{P(A | B)} = \frac{\underline{P(A)} \cdot \underline{P(B | A)}}{\underline{P(B)}}$$

- $P(A | B)$ (Posterior): probability of A being true given B
- $P(A)$ (Prior): probability of A being true
- $P(B)$ (Marginalization): probability of B being true
- $P(B | A)$ (Likelihood): probability of B true given that A is true

Example: Favorite language by year

Example:

- dangerous fires are rare (1%)
- but smoke is fairly common (10%) due to barbecues,
- and 90% of dangerous fires make smoke

We can then discover the **probability of dangerous Fire when there is Smoke**:

$$\begin{aligned} P(\text{Fire|Smoke}) &= \frac{P(\text{Fire}) P(\text{Smoke|Fire})}{P(\text{Smoke})} \\ &= \frac{1\% \times 90\%}{10\%} \\ &= 9\% \end{aligned}$$

Example: Favorite language by year



	Fresh	Soph	Junior	Senior
Python	9	14	13	17
R	14	15	10	8
	23	29	23	25

$$P(Y = \text{py} | X = \text{jr}) = \frac{P(Y = \text{py}) \cdot P(X = \text{jr} | Y = \text{py})}{P(X = \text{jr})}$$

P(Y = py) P(X = jr)

$\frac{13}{53}$ $\frac{23}{100}$

An equivalent formula

$$P_k(x) = P(Y=k | X=x)$$

$$A: Y=1$$

$$B: X=x$$

$$A^c: Y=0$$

$$\underline{P(Y=1 | X=x)} = \frac{\underline{P(Y=1 | X=x)}}{\underline{P_k(x)}}$$

$$\frac{\pi_1 f_1(x)}{\pi_1 \cdot P(Y=1) + \pi_0 \cdot P(Y=0)} = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_0 f_0(x)}$$

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

$$\Leftrightarrow P(A | B) = \frac{P(A) \cdot P(B | A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}$$

$$P_1(x) = \frac{\pi_1 f_1}{\pi_1 f_1 + \pi_0 f_0}$$

$$P_0(x) = \frac{\pi_0 f_0}{\pi_1 f_1 + \pi_0 f_0}$$

$$P(B) = P(A, B) + P(A^c, B) = P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)$$

Following book notation

- Classify an observation into one of $K \geq 2$ classes
- π_k is the *prior* probability that a randomly chosen observation comes from the k th class, $P(Y = k)$

- $f_k(X) = \Pr(X | Y = k)$ is the density function of X for an observation from the k th class
 - ▶ Large $f_k(x)$ if there is high probability that observation in the k th class has $X \approx x$
 - ▶ Small if unlikely that an observation in the k th class has $X \approx x$

Bayes to the rescue!

Posterior probability that an observation $X = x$ belongs to the k th class:

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

- Second equation is Bayes, $p_k(x)$ is our notation for it
- Plug in estimates for π_k and $f_k(x)$ to get an estimate
- Estimate for π_k is easy with a random sample, just take proportion that are k th class
- Estimating density f_k is hard(er).....

Section 3

Linear Discriminant Analysis for $p = 1$

Assumptions

Assume $f_k(x)$ is normal/Gaussian:

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

- μ_k = mean of k th class
- σ_k^2 = variance of k th class
- Assume $\sigma_1^2 = \dots = \sigma_K^2$

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

Bayes Classifier

Same Bayes person, different Bayes definition

$$-\frac{x^2}{\sum \sigma_k^2}$$

Bayes classifier

Assign the class k for which $p_k(x)$ is largest

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum \pi_k f_k(x)}$$

$$\text{f}_k - \frac{\sum M_k x}{\sum \sigma_k^2}$$

Finding largest k

$$\pi_k \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma^2}(x - \mu_k)^2\right)$$

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

is the same as finding largest k for

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

$$\begin{aligned} p'_k(x) &= \pi_k f_k \\ \rightarrow \delta_k &= \log p'_k = \log \pi_k f_k = \log \pi_k + \log f_k \\ &\approx -\frac{1}{2\sigma^2} (x - \mu_k)^2 = -\frac{1}{2\sigma^2} (x - \mu_k)^2 - 2\mu_k x + \mu_k^2 \end{aligned}$$

Example when $K = 2$, $\pi_1 = \pi_2 = \frac{1}{2}$

$$\delta_1(x) = x \cdot \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \ln(\frac{1}{2})$$

$$\delta_2(x) = x \cdot \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + \ln(\frac{1}{2})$$

Decision boundary:

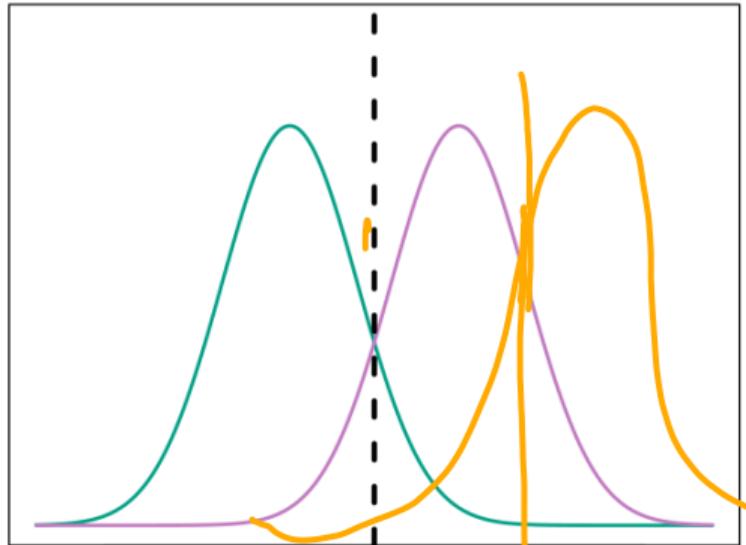
$$x = \frac{\mu_1 + \mu_2}{2}$$

$x:$ $\delta_1(x) = \delta_2(x)$

$$\Leftrightarrow x \cdot \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} = x \cdot \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$

$$\Leftrightarrow x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

$$\pi_1 = \pi_2 = \frac{1}{2}$$



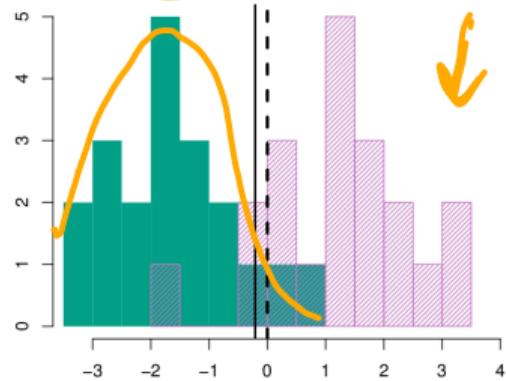
New plan: Linear Discriminant Analysis (LDA)

Estimate

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

$f_k(x)$

$\{(x_i, y_i)\}$



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

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

- $\hat{\pi}_k = n_k / n$

- Black solid line: calculated boundary for assignment
- This example, $n_1 = n_2 = 20$, so $\hat{\pi}_1 = \hat{\pi}_2$, so decision bdry half way between sample means
- Optimal Bayes decision boundary dashed line

Example 1

Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X , last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $\underline{X} = 4$ last year.

$$Y = \begin{cases} 1, & \text{yes} \\ 0, & \text{no} \end{cases}$$

$$M_1 = 10$$

$$M_0 = 0$$

$$\pi_{1|k} = P(Y=k)$$

$$\pi_1 = 0.8 \Rightarrow \pi_0 = 0.2$$

$$\delta_1 = \bar{x} \cdot \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \log \pi_1 = 4 \cdot \frac{10}{36} - \frac{100}{2 \cdot 36} + \log 0.8$$

$$\delta_0 = 4 \cdot \frac{0}{36} - \frac{0^2}{2 \cdot 36} + \log(0.2)$$

Example 2

Assume the probability that a person defaulted on credit card payment is 10%
Based on the given ~~table~~ use LDA to predict the default status of a person with a credit card balance of 1800.

$$\bar{T}_1 = 0.1 \quad \bar{T}_0 = 0.9$$

$$\hat{\mu}_1 = \frac{1000 + 2000 + 2500}{3}$$

$$\hat{\mu}_0 = \frac{0 + 500 + 1500}{3}$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{k=0}^1 \sum_{i|y_i=k} (x_i - \hat{\mu}_k)^2 = \frac{+ \Rightarrow (2500 - \hat{\mu}_1)^2}{(n-2)}$$

Balance	Prediction
---------	------------

$$(0 - \hat{\mu}_0)^2 \text{ No}$$

$$(500 - \hat{\mu}_0)^2 \text{ No}$$

$$(1000 - \hat{\mu}_1)^2 \text{ Yes}$$

$$(1500 - \hat{\mu}_0)^2 \text{ No}$$

$$(2000 - \hat{\mu}_1)^2 \text{ Yes}$$

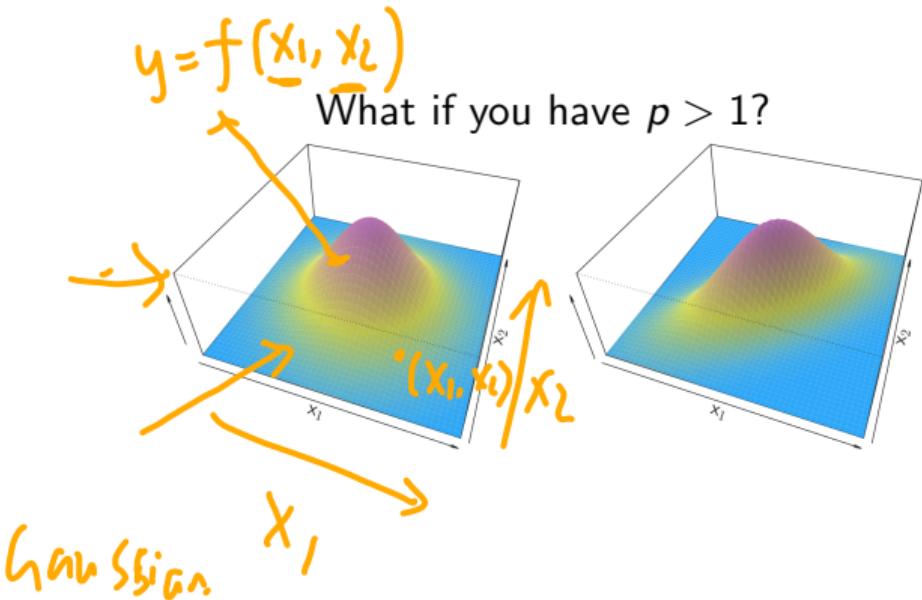
$$(2500 - \hat{\mu}_1)^2 \text{ Yes}$$

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

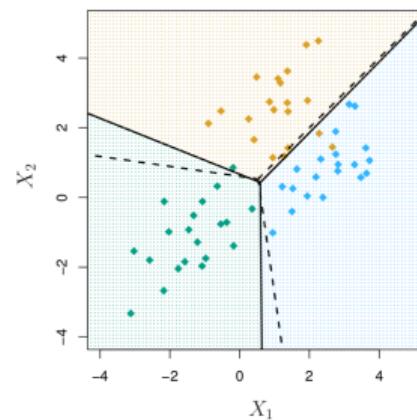
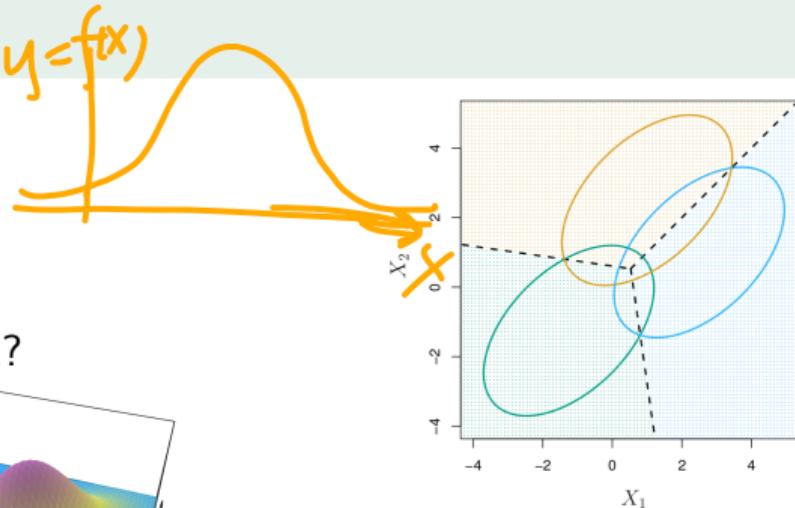
LDA review

- Assume observations in each class come from normal
- Class specific means
- Common variance
- Plug in estimates into Bayes classifier

High dimensional LDA - $p > 1$



$$y = f(x)$$



LDA - $p > 1$

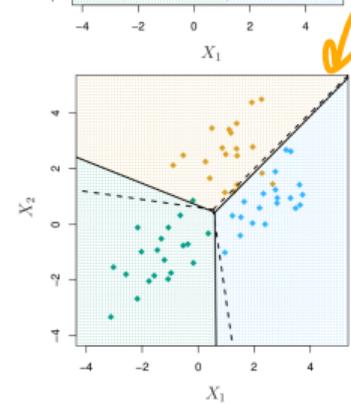
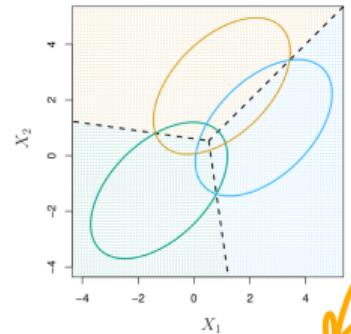
- Assume observations in the k th drawn from multi-variate normal distribution with the same covariance matrix for all k : $N(\mu_k, \Sigma)$
- For new date point x , predict the k for which $p_k(x) = \Pr(Y = k|X = x)$ is largest

- Equivalent to finding k for which $\delta_k(x)$ is largest

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

- Approximate μ_i 's π_i 's and Σ to find boundary where the returned k switches

$$\delta_k = b_k - T_k(x)$$



Quadratic Discriminant Analysis (QDA)

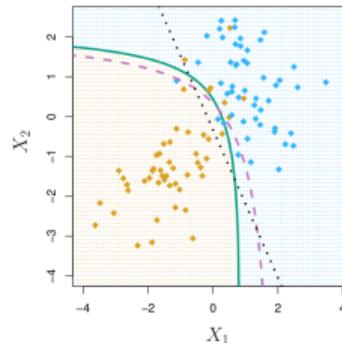
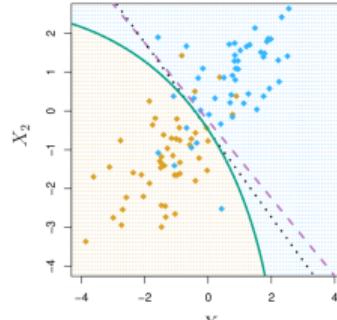
- Same idea as LDA, but don't assume same covariance matrix
- Assume observations in k th class drawn from normal distribution $N(\mu_k, \Sigma_k)$

- Make new predictions based on

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

- This setup means decision boundaries are quadratic

$$P = \frac{\delta_k(x)}{\sum_{k=1}^K \delta_k(x)}$$



Example

Assume the probability that a person defaulted on credit card payment is 10%.
Based on the given use LDA to predict the default status of a person with a credit card balance of 1800.

$$\rightarrow \hat{\mu}_1 = \frac{1}{n} \sum_{i=1, y_i=1} X_i \quad \hat{\mu}_0 = \dots \rightarrow 1000$$

$$\rightarrow \hat{\sigma}_1^2 = \frac{1}{n-1} \sum_{i=1, y_i=1} (X_i - \hat{\mu}_1)^2 \rightarrow 2000$$

$$\rightarrow \hat{\sigma}_0^2 = \dots \rightarrow 2500$$



Balance	Prediction
0	No
500	No
1000	Yes
1500	No
2000	Yes
2500	Yes

$$\pi_1 = 0.1 \\ \pi_0 = 0.9$$

500

No

No

Yes

-

No

Yes

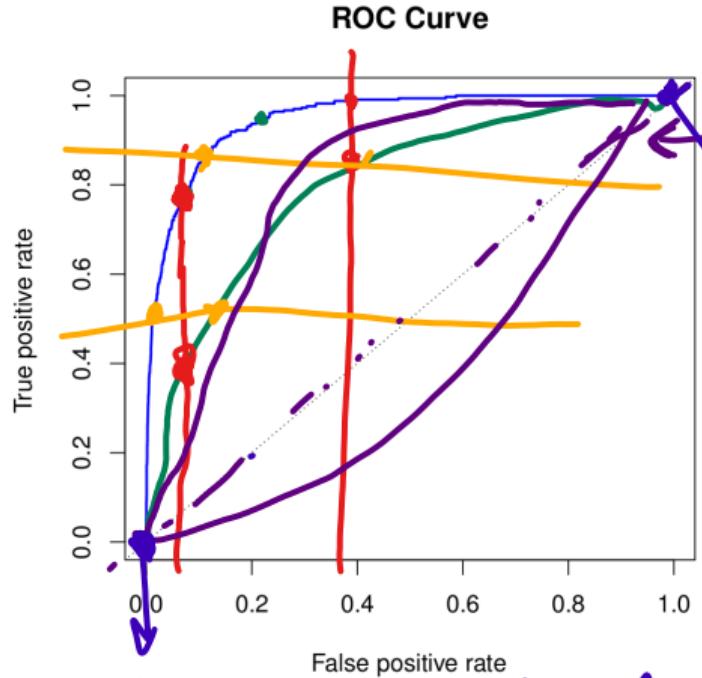
Yes

Confusion matrix and types of Errors

		True default status		
		No	Yes	Total
Predicted default status	No	9432	138	9570
	Yes	235	195	430
Total		9667	333	10000

Predicted class	True class		
	- or Null	+ or Non-null	Total
	True Neg. (TN)	False Neg. (FN)	N*
- or Null	False Pos. (FP)	True Pos. (TP)	P*
Total	N		P
Name	Definition	Synonyms	
False Pos. rate	FP/N	Type I error, 1-Specificity	
True Pos. rate	TP/P	1-Type II error, power, sensitivity, recall	
Pos. Pred. value	TP/P*	Precision, 1-false discovery proportion	
Neg. Pred. value	TN/N*		

ROC curve



AUC → Area under the curve

ROC of a random classifier
classifier always return 1.

classifier always return 0

Multivariate normal distribution

Gaussian (normal) distributions

- $Z \sim N(0, 1)$ means Z follows a standard Gaussian distribution, i.e., has probability density

$$Y.V \quad f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

- If Z_1, Z_2, \dots, Z_d are iid $N(0, 1)$ random variables, then say $\mathbf{Z} = (Z_1, Z_2, \dots, Z_d)$ follows a standard multivariate Gaussian distribution on \mathbb{R}^d , i.e., $\mathbf{Z} \sim N(\mathbf{0}, I_d)$.

(i,j)th entry
rep. correlation between Z_i & Z_j

$d \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}$

variance matrix

- Other Gaussian distributions on \mathbb{R}^d arise by applying (invertible) linear maps and translations to \mathbf{Z} :

$$\mathbf{z} \mapsto \mathbf{Az} \mapsto \mathbf{Az} + \boldsymbol{\mu}.$$

- ▶ $\mathbf{X} := \mathbf{Az} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \mathbf{AA}^T)$ has $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$ and $\text{cov}(\mathbf{X}) = \mathbf{AA}^T$
- ▶ the (i, j) th entry of $\text{cov}(\mathbf{X})$ represents the correlation between X_i and X_j .

Appendix: Review of Multivariate normal distribution

- $\mathbf{X} \sim N(\mu, \Sigma)$ has the probability density function

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)}$$

$$-\frac{(\mathbf{x}-\mu)^T}{2\alpha\Sigma}$$

- Estimate μ and Σ from data: maximum likelihood estimators

- ▶ $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$
- ▶ $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})^T$
- ▶ $\hat{\mu}$ is unbiased and $\hat{\Sigma}$ is slightly biased