LDA, QDA, Naive Bayes

Generative Classification Models

Marek Petrik

2/16/2017

Last Class

- Logistic Regression
- Maximum Likelihood Principle

Logistic Regression

- Predict **probability** of a class: $p(X)$
- Example: $p(\text{balance})$ probability of default for person with balance
- **Linear regression**:

$$p(X) = \beta_0 + \beta_1$$

- **logistic regression**:

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

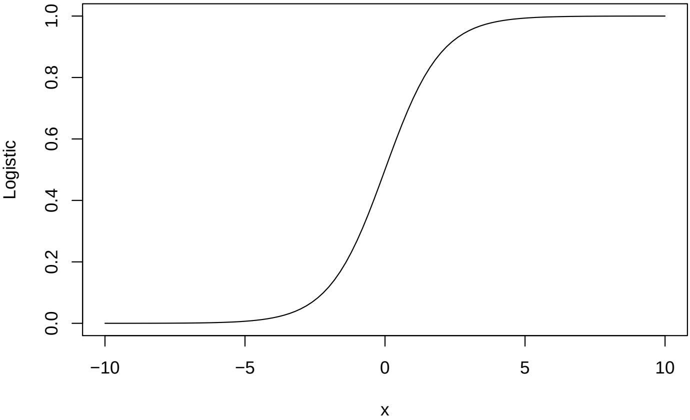
- the same as:

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

- Odds: $p(X) / 1 - p(X)$

Logistic Function

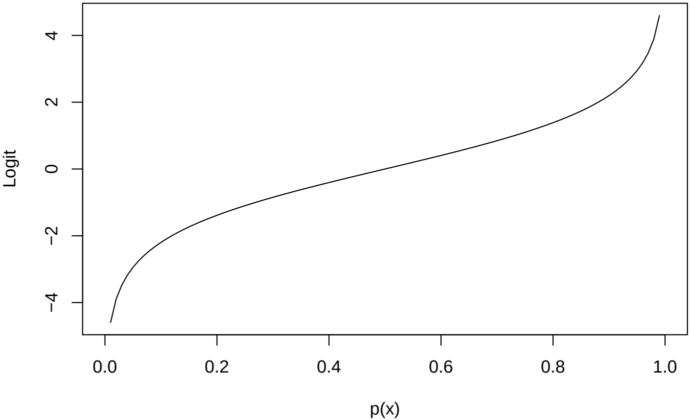
$$y = \frac{e^x}{1 + e^x}$$



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

Logit Function

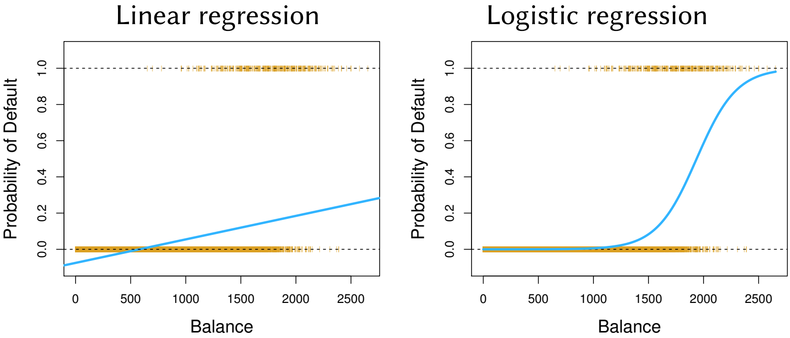
$$\log\left(\frac{p(X)}{1-p(X)}\right)$$



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

Logistic Regression

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



Estimating Coefficients: Maximum Likelihood

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

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

- Find the most likely model:

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

- Likelihood function is difficult to maximize
- Transform it using $\log$ (strictly increasing)

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

- Strictly increasing transformation does not change maximum

Today

1. Classification methods continued
2. Discriminative vs. Generative ML Models
3. Generative classification models:
   - Linear Discriminant Analysis (LDA)
   - Quadratic Discriminant Analysis (QDA)
   - Naive Bayes Classification

Discriminative vs Generative Models

- **Discriminative models**
  - Estimate conditional models $\Pr[Y \mid X]$
  - Linear regression
  - Logistic regression
- **Generative models**
  - Estimate joint probability $\Pr[Y, X] = \Pr[Y \mid X] \Pr[X]$
  - Estimates not only probability of labels but also the features
  - Once model is fit, can be used to generate data
  - LDA, QDA, Naive Bayes

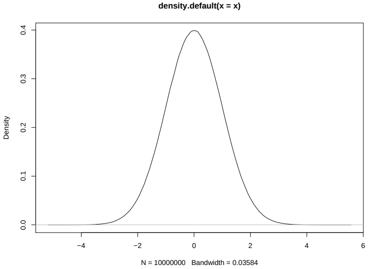
Generative Models

+ Can be used to generate data ($\Pr[X]$)
+ Offers more insights into data
− Often works worse, particularly when assumptions are violated

Normal Distribution

- Density function:

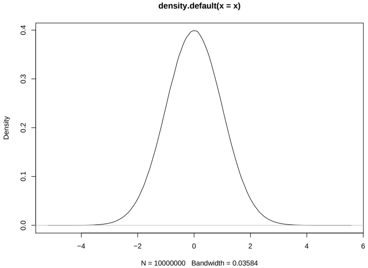
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Normal Distribution

- Density function:

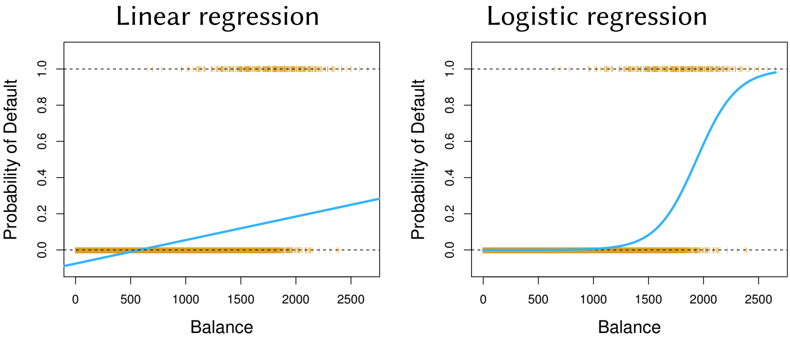
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- **Central limit theorem**: $Z = {}^1\!/_n \sum_{i=1}^{n} X_i$ for i.i.d. X_i is normal with $n \to \infty$

Logistic Regression

$$Y = \begin{cases} 1 & \text{if default} \\ 0 & \text{otherwise} \end{cases}$$

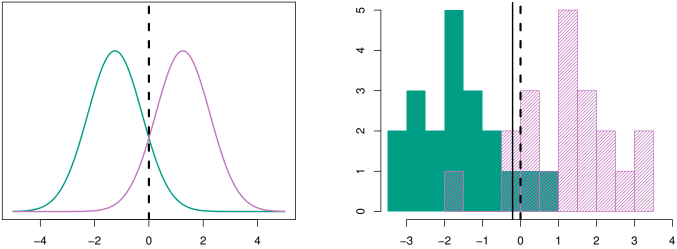


Predict:

$$\Pr[\text{default} = \text{yes} \mid \text{balance}]$$

LDA: Linear Discriminant Analysis

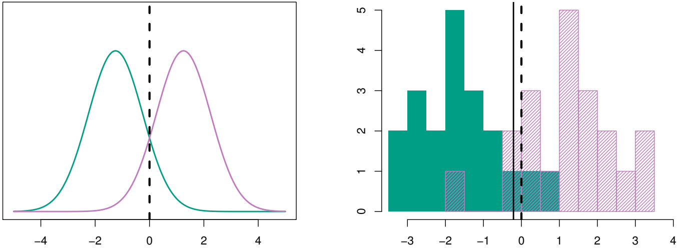
- **Generative model**: capture probability of predictors for each label



- Predict:

LDA: Linear Discriminant Analysis

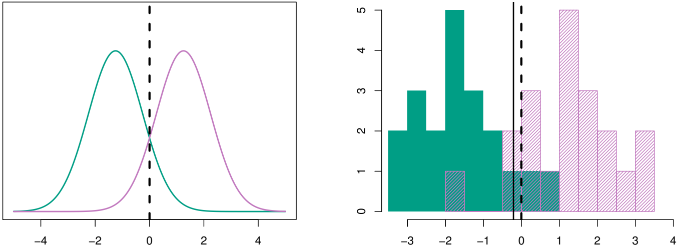
- **Generative model**: capture probability of predictors for each label



- Predict:
  1. $\Pr[\text{balance} \mid \text{default} = \text{yes}]$ and $\Pr[\text{default} = \text{yes}]$

LDA: Linear Discriminant Analysis

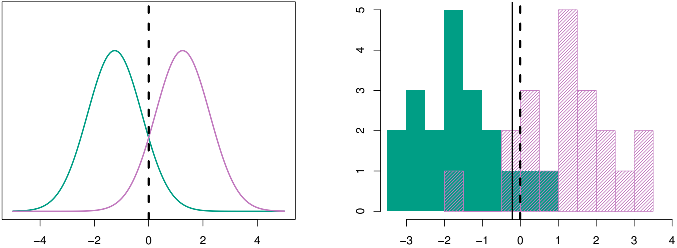
- **Generative model**: capture probability of predictors for each label



- Predict:
  1. $\Pr[\text{balance} \mid \text{default} = \text{yes}]$ and $\Pr[\text{default} = \text{yes}]$
  2. $\Pr[\text{balance} \mid \text{default} = \text{no}]$ and $\Pr[\text{default} = \text{no}]$

LDA: Linear Discriminant Analysis

- **Generative model**: capture probability of predictors for each label



- Predict:
  1. $\Pr[\text{balance} \mid \text{default} = \text{yes}]$ and $\Pr[\text{default} = \text{yes}]$
  2. $\Pr[\text{balance} \mid \text{default} = \text{no}]$ and $\Pr[\text{default} = \text{no}]$
- Classes are normal: $\Pr[\text{balance} \mid \text{default} = \text{yes}]$

LDA vs Logistic Regression

- **Logistic regressions**:

$$\Pr[\text{default} = \text{yes} \mid \text{balance}]$$

- **Linear discriminant analysis**:

$$\Pr[\text{balance} \mid \text{default} = \text{yes}] \text{ and } \Pr[\text{default} = \text{yes}]$$

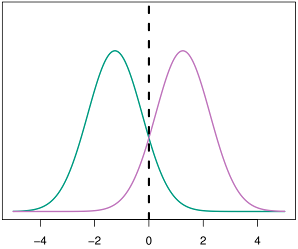
$$\Pr[\text{balance} \mid \text{default} = \text{no}] \text{ and } \Pr[\text{default} = \text{no}]$$

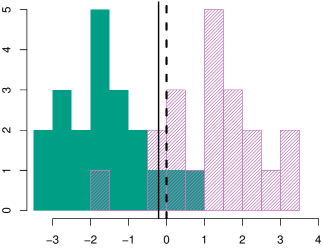
LDA with 1 Feature

- Classes are normal and class probabilities π_k are scalars

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

- **Key Assumption**:Class variances σ_k^2 are the same.





Bayes Theorem

- Classification from label distributions:

$$\Pr[Y = k \mid X = x] = \frac{\Pr[X = x \mid Y = k] \Pr[Y = k]}{\Pr[X = x]}$$

Bayes Theorem

- Classification from label distributions:

$$\Pr[Y = k \mid X = x] = \frac{\Pr[X = x \mid Y = k] \Pr[Y = k]}{\Pr[X = x]}$$

- Example:

$$\Pr[\text{default} = \text{yes} \mid \text{balance} = \$100] =$$
$$\frac{\Pr[\text{balance} = \$100 \mid \text{default} = \text{yes}] \Pr[\text{default} = \text{yes}]}{\Pr[\text{balance} = \$100]}$$

Bayes Theorem

- Classification from label distributions:

$$\Pr[Y = k \mid X = x] = \frac{\Pr[X = x \mid Y = k] \Pr[Y = k]}{\Pr[X = x]}$$

- Example:

$$\Pr[\text{default} = \text{yes} \mid \text{balance} = \$100] = \frac{\Pr[\text{balance} = \$100 \mid \text{default} = \text{yes}] \Pr[\text{default} = \text{yes}]}{\Pr[\text{balance} = \$100]}$$

- Notation:

$$\Pr[Y = k \mid X = x] = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

Classification With LDA

Probability in class k_1 $>$ Probability in class k_2

Classification With LDA

Probability in class k_1 $>$ Probability in class k_2

$$\Pr[Y = k_1 \mid X = x] > \Pr[Y = k_2 \mid X = x]$$

Classification With LDA

$$\text{Probability in class } k_1 > \text{Probability in class } k_2$$

$$\Pr[Y = k_1 \mid X = x] > \Pr[Y = k_2 \mid X = x]$$

$$\frac{\pi_{k_1} f_{k_1}(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} > \frac{\pi_{k_2} f_{k_2}(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

Classification With LDA

$$\text{Probability in class } k_1 > \text{Probability in class } k_2$$
$$\Pr[Y = k_1 \mid X = x] > \Pr[Y = k_2 \mid X = x]$$
$$\frac{\pi_{k_1} f_{k_1}(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} > \frac{\pi_{k_2} f_{k_2}(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$
$$\pi_{k_1} f_{k_1}(x) > \pi_{k_2} f_{k_2}(x)$$

Classification With LDA

$$\text{Probability in class } k_1 > \text{Probability in class } k_2$$

$$\Pr[Y = k_1 \mid X = x] > \Pr[Y = k_2 \mid X = x]$$

$$\frac{\pi_{k_1} f_{k_1}(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} > \frac{\pi_{k_2} f_{k_2}(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

$$\pi_{k_1} f_{k_1}(x) > \pi_{k_2} f_{k_2}(x)$$

$$\log\left(\pi_{k_1} f_{k_1}(x)\right) > \log\left(\pi_{k_2} f_{k_2}(x)\right)$$

Classification With LDA

$$\text{Probability in class } k_1 > \text{Probability in class } k_2$$
$$\Pr[Y = k_1 \mid X = x] > \Pr[Y = k_2 \mid X = x]$$
$$\frac{\pi_{k_1} f_{k_1}(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} > \frac{\pi_{k_2} f_{k_2}(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$
$$\pi_{k_1} f_{k_1}(x) > \pi_{k_2} f_{k_2}(x)$$
$$\log\left(\pi_{k_1} f_{k_1}(x)\right) > \log\left(\pi_{k_2} f_{k_2}(x)\right)$$
$$\hat{\delta}_{k_1}(x) > \hat{\delta}_{k_2}(x)$$

Discriminant function:

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

Derive at home

Estimating LDA Parameters

Estimating LDA Parameters

- Maximum log likelihood!

$$\max_{\mu,\sigma} \log \ell(\mu,\sigma) = \max_{\mu,\sigma} \sum_{i=1}^{N} \log\left(f_{y_i}(x_i)\right) =$$

$$\max_{\mu,\sigma} \sum_{i=1}^{N} \log\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_{y_i})^2\right)\right) =$$

$$\max_{\mu,\sigma} \sum_{i=1}^{N} \left(-\log\sigma - \frac{1}{2\sigma^2}(x_i - \mu_{y_i})^2 + \text{consts}\right)$$

Estimating LDA Parameters

- Maximum log likelihood!

$$\max_{\mu,\sigma} \log \ell(\mu, \sigma) = \max_{\mu,\sigma} \sum_{i=1}^{N} \log\left(f_{y_i}(x_i)\right) =$$

$$\max_{\mu,\sigma} \sum_{i=1}^{N} \log\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_{y_i})^2\right)\right) =$$

$$\max_{\mu,\sigma} \sum_{i=1}^{N} \left(-\log\sigma - \frac{1}{2\sigma^2}(x_i - \mu_{y_i})^2 + \text{consts}\right)$$

- Concave in μ and $1/\sigma^2$, consider a single class with mean μ

$$\frac{\partial}{\partial\mu} \log \ell(\mu, \sigma) = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu) = 0$$

$$\frac{\partial}{\partial\sigma} \log \ell(\mu, \sigma) = \frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{N} (x_i - \mu)^2 = 0$$

Estimating LDA Parameters

- $\log \ell$ is derivatives:

$$\frac{\partial}{\partial \mu} \log \ell(\mu, \sigma) = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu) = 0$$

$$\frac{\partial}{\partial \sigma} \log \ell(\mu, \sigma) = \frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{N} (x_i - \mu)^2 = 0$$

- Therefore:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

Better Parameter Estimates

- Maximum likelihood variance σ^2 is biased:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

- Unbiased estimate:

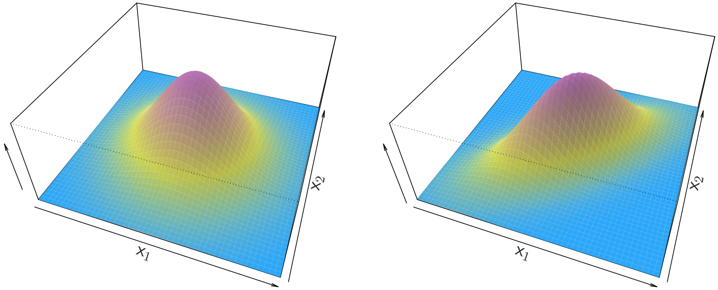
$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \mu)^2$$

- *See ISL for precise formula for more than a single class*

LDA with Multiple Features

- Multivariate Normal Distributions:



- Multivariate normal distribution density (mean vector μ, covariance matrix Σ):

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

Multivariate Maximum Likelihood

- Consider a singe class:

$$\max_{\mu,\sigma} \log \ell(\mu,\Sigma) = \max_{\mu,\Sigma} \sum_{i=1}^{N} \log\left(f_k(x_i)\right) =$$

$$\max_{\mu,\Sigma} \sum_{i=1}^{N} \log\left(\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_i-\mu)^\top \Sigma^{-1}(x_i-\mu)\right)\right) =$$

$$\max_{\mu,\Sigma} -\frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^\top \Sigma^{-1}(x_i-\mu) =$$

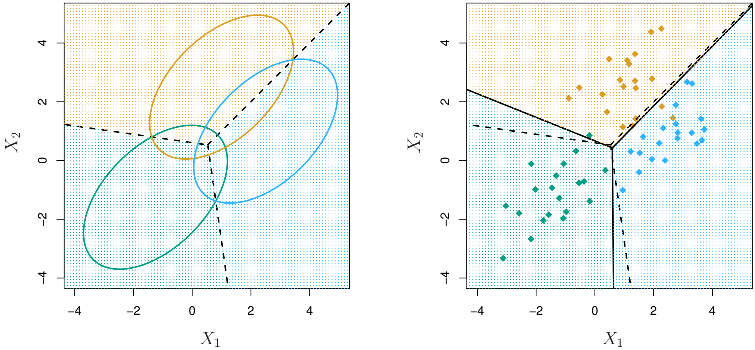
$$\max_{\mu,\Sigma} -\frac{N}{2}\log|\Sigma| - \frac{1}{2}\,\mathrm{Trace}\left(\Sigma^{-1}\sum_{i=1}^{N}(x_i-\mu)^\top(x_i-\mu)\right)$$

- Use $\partial/\partial\Sigma \log|\Sigma| = \Sigma^{-\top}$ and $1/\partial A\, \mathrm{Trace}(AB) = B^\top$

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^\top(x_i-\mu)$$

Multivariate Classification Using LDA

- **Linear**: Decision boundaries are linear



Confusion Matrix: Predict default

		True		
		Yes	No	Total
Predicted	Yes	a	b	$a+b$
	No	c	d	$c+d$
	Total	$a+c$	$b+d$	N

Result of LDA classification: Predict default if
$\Pr[\text{default} = \text{yes} \mid \text{balance}] > 1/2$

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

Confusion Matrix: Predict default

		True		
		Yes	No	Total
Predicted	Yes	a	b	$a+b$
	No	c	d	$c+d$
	Total	$a+c$	$b+d$	N

Result of LDA classification: Predict default if
$\Pr[\text{default} = \text{yes} \mid \text{balance}] > 1/2$

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

Most people who default are predicted as No default

Increasing LDA Sensitivity

Result of LDA classification: Predict default if
$\Pr[\text{default} = \text{yes} \mid \text{balance}] > 1/2$

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

Increasing LDA Sensitivity

Result of LDA classification: Predict default if
$\Pr[\text{default} = \text{yes} \mid \text{balance}] > 1/2$

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

Result of LDA classification: Predict default if
$\Pr[\text{default} = \text{yes} \mid \text{balance}] > 1/2$

		True		
		Yes	No	Total
Predicted	Yes	195	235	403
	No	138	9 432	9 570
	Total	333	9 667	10 000

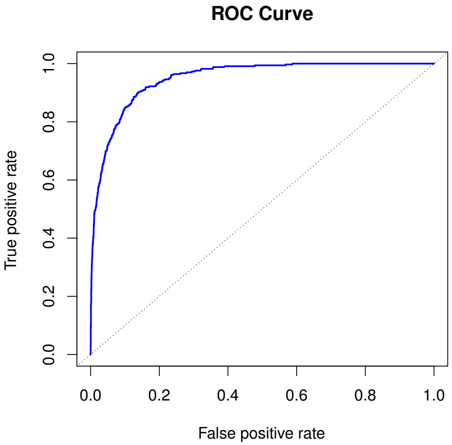
True Positives, etc

		Reality	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

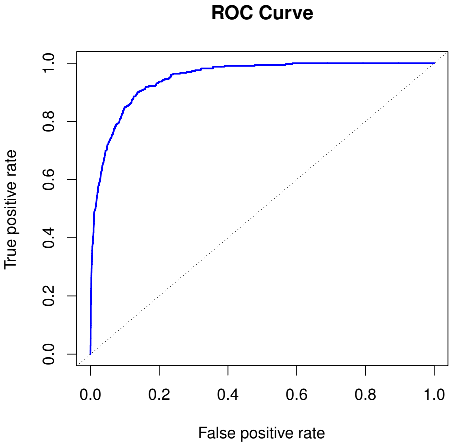
- **Recall/sensitivity** = TP/(TP+FN)
- **Precision** = TP/(TP+FP)
- **Specificity** = TN/(TN+FP)

ROC Curve

		Reality	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative



Area Under ROC Curve



- Larger area is better
- Many other ways to measure classifier performance, like F_1

QDA: Quadratic Discriminant Analysis

- Generalizes LDA

- **LDA**: Class variances $\Sigma_k = \Sigma$ <u>are the same</u>
- **QDA**: Class variances Σ_k <u>can differ</u>

QDA: Quadratic Discriminant Analysis

- Generalizes LDA

- **LDA**: Class variances $\Sigma_k = \Sigma$ are the same
- **QDA**: Class variances Σ_k can differ

- LDA or QDA has smaller training error on the same data?

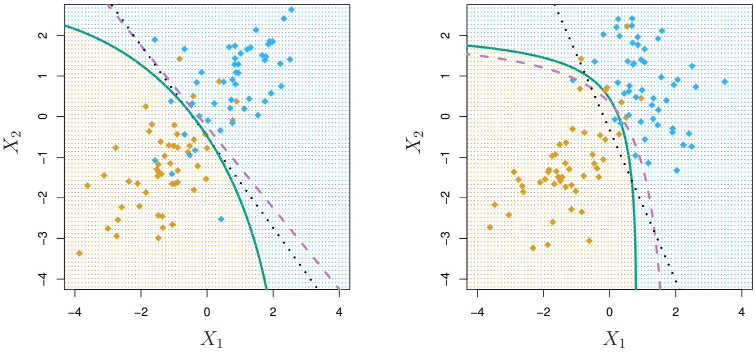
QDA: Quadratic Discriminant Analysis

- Generalizes LDA

- **LDA**: Class variances $\Sigma_k = \Sigma$ are the same
- **QDA**: Class variances Σ_k can differ

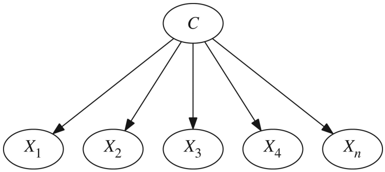
- LDA or QDA has smaller training error on the same data?
- What about the test error?

QDA: Quadratic Discriminant Analysis



Naive Bayes

- Simple Bayes net classification



- With normal distribution over features $X_1, \ldots, X_k$ special case of QDA with diagonal Σ
- Generalizes to non-Normal distributions and discrete variables
- More on it later …