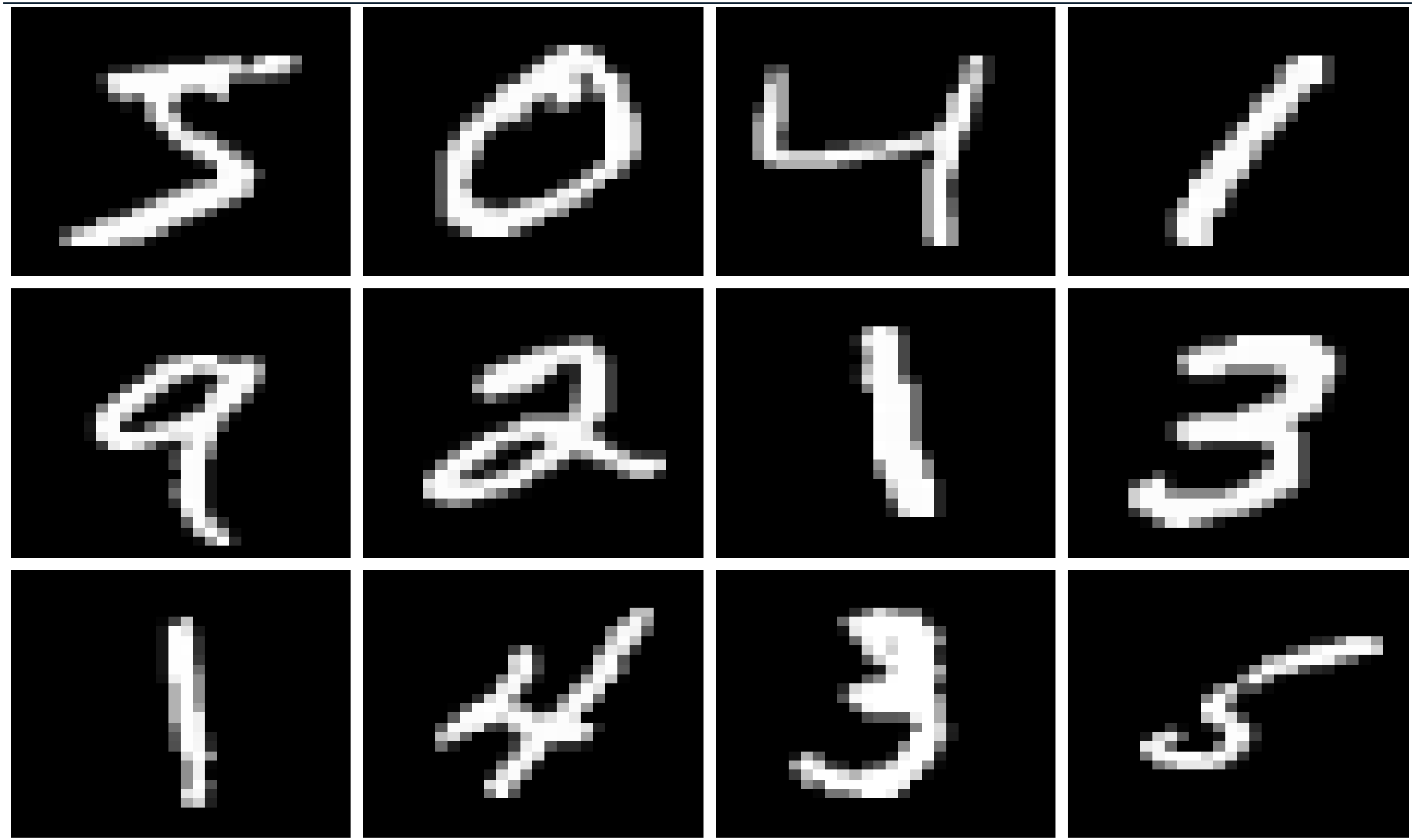
Math 514 - Probability 1

(updated: 2019-01-31)

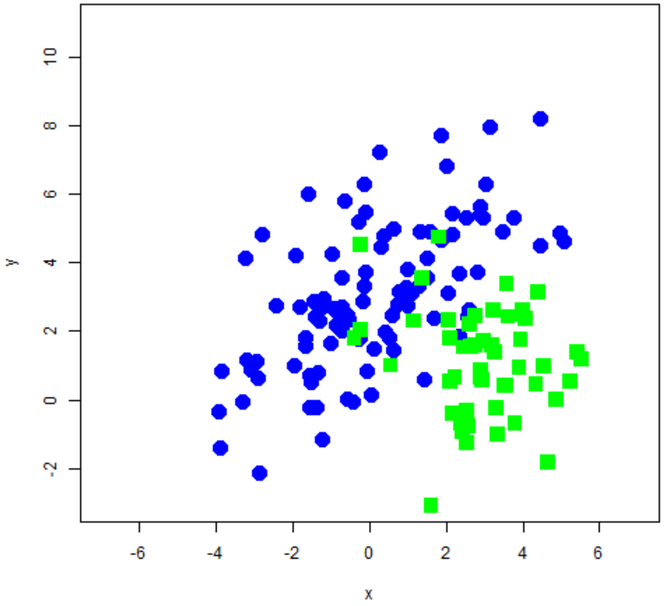
Probability for Machine Learning

- Machine Learning is fundamentally probabilistic
  - If outcome is sure, then memorization, not learning
- When outcomes are probabilistic, *can't* ask what will be the next value
  - Can ask what is the most likely next value (**expected value**)
  - Can ask about variation in values (**variance**)
- Thinking probabilistically requires intuitive understanding of distributions
  - Knowledge of relevant distributions gives optimal solution to classification problems
  - What is the distribution describing cat images? - relevant distributions are seldom know apriori

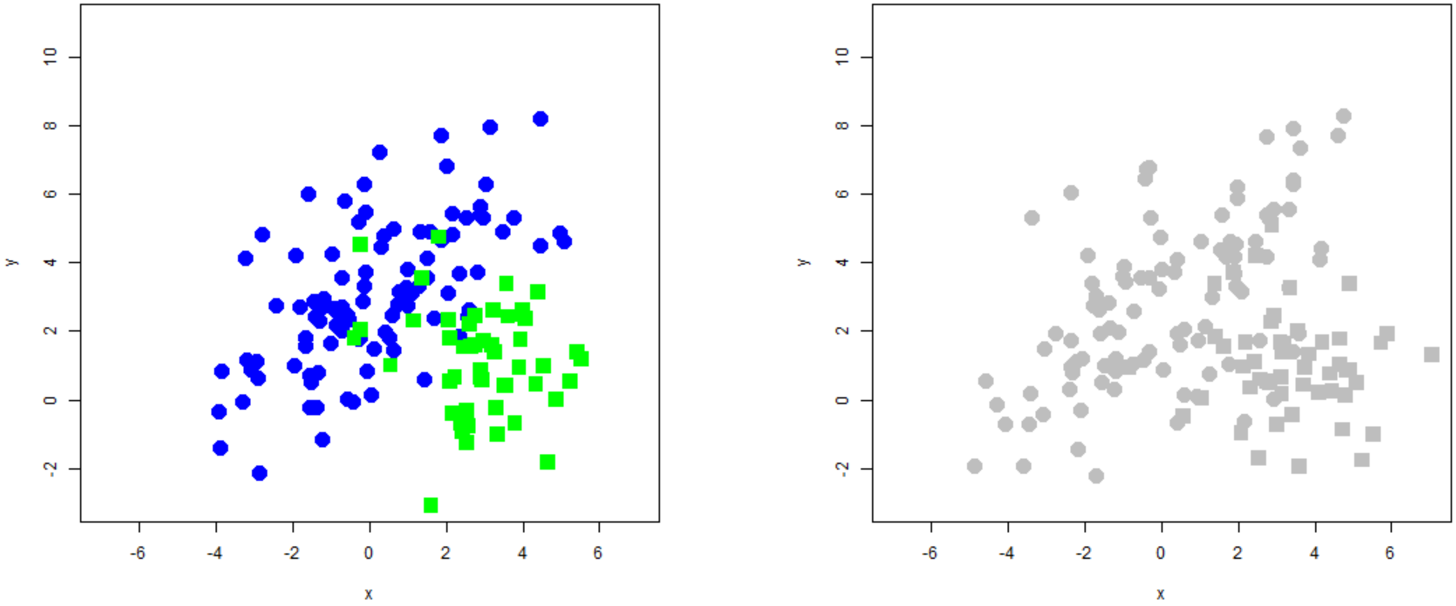
Classification - MNIST Digits



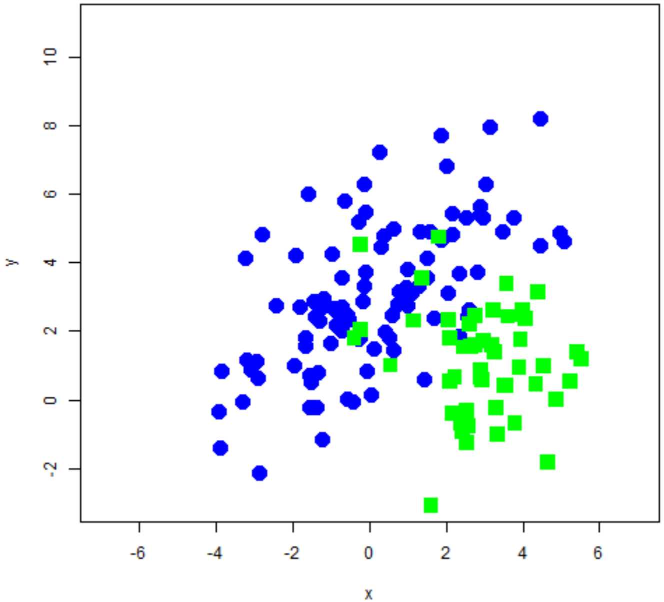
Classification and Bayes' Error

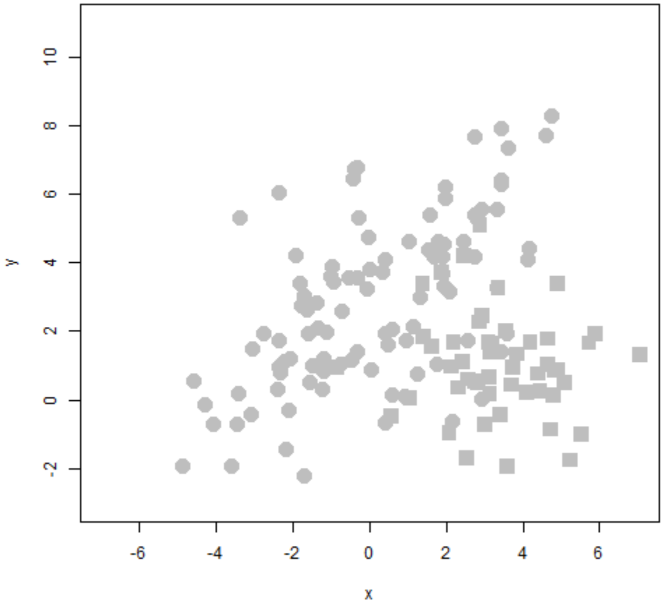


Classification and Bayes' Error



Classification and Bayes' Error





Optimal Classifier:

$$P(blue|\mathbf{x}) > P(green|\mathbf{x})$$

Scatter Plot Code

```
require(mvtnorm,quietly = TRUE)
set.seed(1)
x1=rmvnorm(100,c(0,3),matrix(c(5,3,3,5),nrow=2))
plot(x1,xlim=c(-7,7),ylim=c(-3,11),pch=16,cex=2,col="blue",
     xlab="x",ylab="y")
x2=rmvnorm(50,c(3,1),matrix(c(2,0,0,2),nrow=2))
points(x2,pch=15,cex=2,col="green")
```

Lecture Goals

The lecture covers enough probability theory to understand the first homework assignment. The assignment is on Baye's Error and the development of a 1-dimensional classifier.

Assume there are two coins C_1 and C_2 with known but different probabilities to be heads p_1 and p_2. Also assume that when a coin is chosen at random that C_1 is chosen with probability $P(C_1)$ and C_2 is chosen with probability $P(C_2) = \; 1 - P(C_1)$.

A coin is chosen as described and flipped N times. The problem is to decide which coin is most likely given the data.

- Typically you would not know the probability to be heads or the frequency with which coins are chosen and they would have to be estimated from the data.

Lecture Goals

This simple setup requires a surprising amount of probability. Intuitively, more heads will tend to favor the coin with the higher p_i. The question is how to choose a decision boundary.

- Construction of the classifier uses Bayes' Rule
  - Probability, Conditional Probability, Bayes's Rule etc.
- The experimental data has a binomial distribution
  - Distribution, Bernoulli distribution, binomial distribution
- To compute the decision boundary, will approximate the binomial with a normal distribution
  - Central Limit Theorem, binomial is sum of Bernoullis distributions, normal distribution
- Bayes' error is irreducible error

Probability and Intuition

> "On voit, par cet Essai, que la thÃ©orie des probabilitÃ©s n'est, au fond, que le bon sens rÃ©duit au calcul;"[1]
>
> --- Pierre-Simon Laplace

translation

> "One sees, from this Essay, that the theory of probabilities is basically just common sense reduced to calculus;"

[1]Essai philosophique sur les ProbabilitÃ©s (1814). Ouvres complÃ¨tes de Laplace, tome VII, p. cliii, Paris: Gauthier-Villars, 1878-1912

Probability and Intuition

From 1982 Tversky and Kahneman[1] study wiki

- The study describes Linda as:

  > "31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations".

- Study participants are then asked which is more probable:
  - Linda is a bank teller
  - Linda is a bank teller and is active in the feminist movement

Probability and Intuition

From 1982 Tversky and Kahneman[1] study wiki

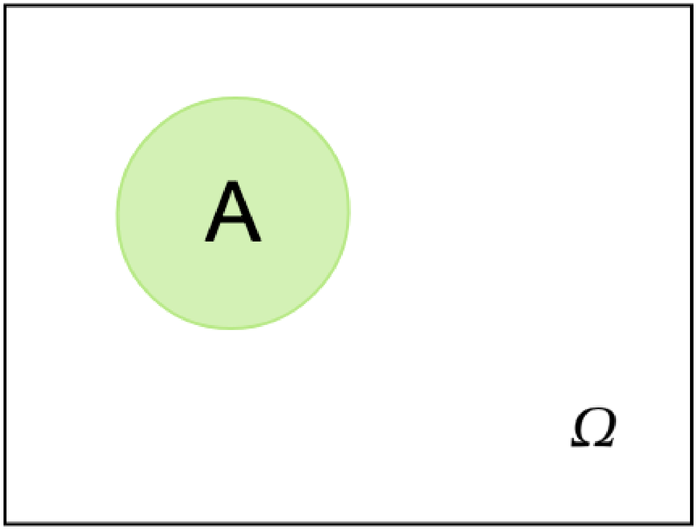
- The study describes Linda as:

  > "31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations".

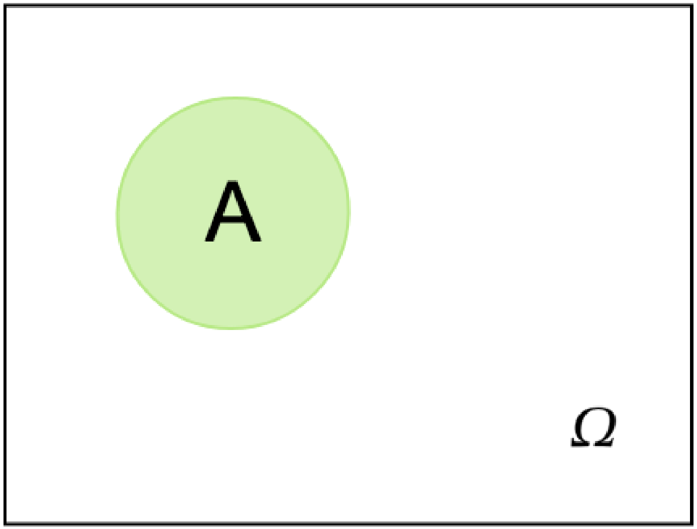
- Study participants are then asked which is more probable:
  - Linda is a bank teller
  - Linda is a bank teller and is active in the feminist movement
- $A \cap B \subset A \Rightarrow P(A \cap B) < P(A)$
- 85% of study participants chose the second option

[1] Kahneman was awarded the 2002 Nobel Prize in Economics

Probability



Probability



Outcomes

$$\Omega = \{\omega_i\}\,, \quad \omega_i \cap \omega_j = \emptyset \quad i \neq j$$

Events

$$A \subset 2^{|\Omega|}$$

Probability Model

$$0 \leq P(A) \leq 1$$

$$P(\omega_i \cup \omega_j) = P(\omega_i) + P(\omega_j)$$

Example Toss a coin 3 times

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

Events: Exactly 1 head, 2 or more heads, ...

Model 1: $P(\omega_i) = 1/8$

Model 2: $P(\omega_i) = p^{n_h}(1-p)^{(3-n_h)}$

Probability

- There are many possible probabilities
  - prior, posterior, learned, empirical, estimated, etc.
- **Axioms of probability** (discrete)
  - $P(A) \in [0,1] \; \forall A \; \subset \Omega$
  - $P(\Omega) = 1$
  - $P(\cup A_i) = \Sigma_i A_i$ if $A_i \cap A_j = \emptyset$
- **Example** If the only possible weather forecasts are rainy (R) or sunny (S) then ...
  - $P(R \cap S) = 0$
  - $P(R \cup S) = 1$
  - $P(R) \in \{0, 1\}$

Probability Distribution

- A **discrete probability distribution** is called a Probability Mass Function (PMF). Typically identifed by upper case P.
  - The values of a PMF are probabilities, so values sum to 1.
  - An impossible event e.g. $P(A \cap A^c)$ has probabiltiy 0. A sure event has probability 1.
  - The PMF assigns a probability for every possible outcome.
  - **Examples** Bernoulli, binomial

Probability Distribution

- A **discrete probability distribution** is called a Probability Mass Function (PMF). Typically identifed by upper case P.
  - The values of a PMF are probabilities, so values sum to 1.
  - An impossible event e.g. $P(A \cap A^c)$ has probabiltiy 0. A sure event has probability 1.
  - The PMF assigns a probability for every possible outcome.
  - **Examples** Bernoulli, binomial
- A **continuous probability distribution** is called a probability density function (PDF). Typically identified with lower case p.
  - A PDF is a probability density, NOT a probability. Values of a pdf can exceed 1.
  - Probabilities are computed from pdf's by integration: $P(A) = \int_A p(x)dx$
  - A pdf satisfies: $\int_{\mathbb{R}} p(x)dx = 1$
  - **Examples** uniform, normal

Discrete Probability Distribution

$$\begin{array}{c|c} & \begin{matrix}1 & 2 & 3 & 4 & 5 & 6\end{matrix} \\ \begin{matrix}1\\2\\3\\4\\5\\6\end{matrix} & \begin{bmatrix}2 & 3 & 4 & 5 & 6 & 7\\3 & 4 & 5 & 6 & 7 & 8\\4 & 5 & 6 & 7 & 8 & 9\\5 & 6 & 7 & 8 & 9 & 10\\6 & 7 & 8 & 9 & 10 & 11\\7 & 8 & 9 & 10 & 11 & 12\end{bmatrix}\end{array}$$

Toss a pair of dice

- Observe both dice
  - All outcomes equally likely, $P(i,j) = \frac{1}{36}$

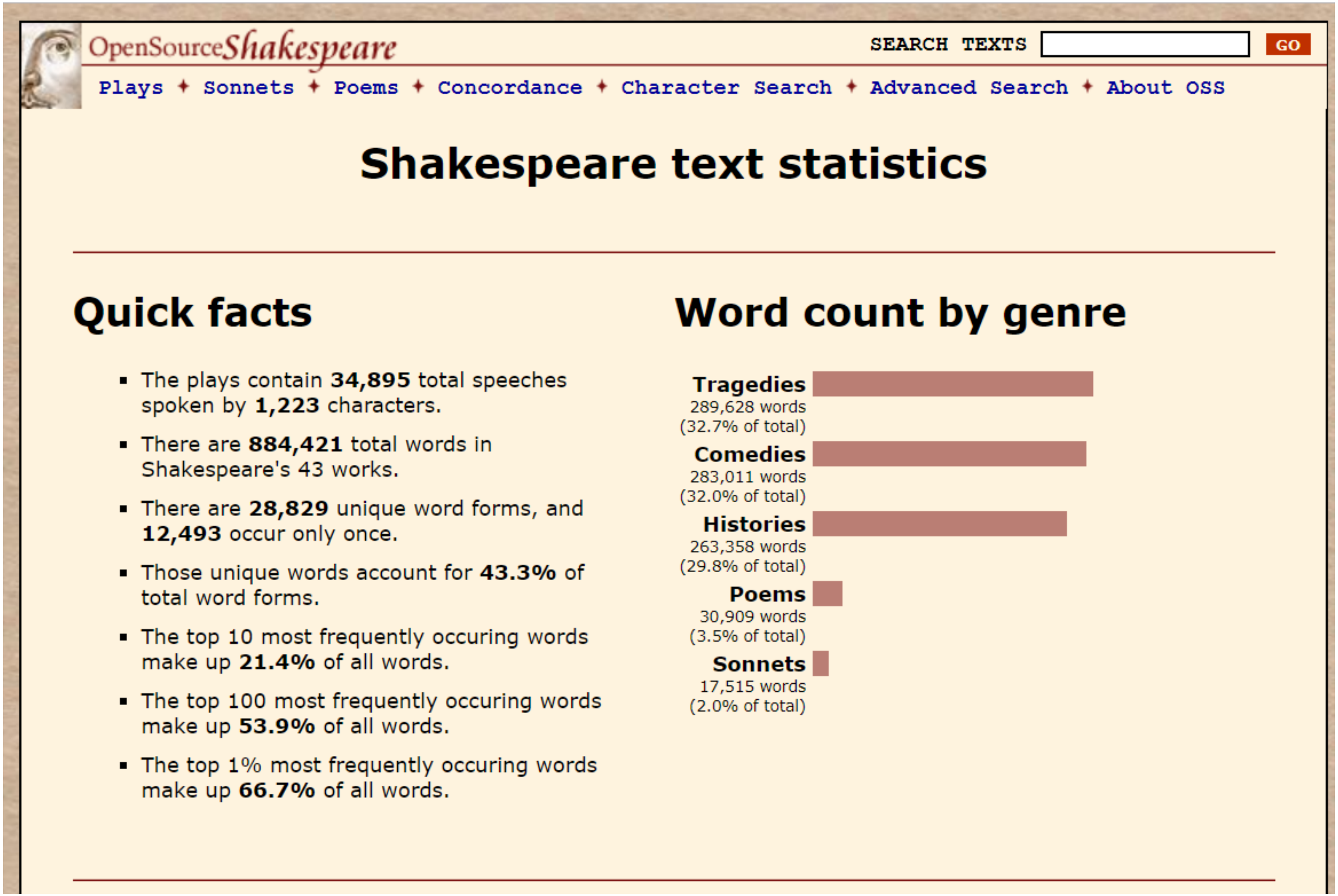
Toss a pair of dice

- Observe sum of dice
  - Probabilities given by counting the anti-diagonals

$$P(2) = P(12) = \frac{1}{36}$$
$$P(3) = P(11) = \frac{2}{36}$$
$$P(4) = P(10) = \frac{3}{36}$$
$$P(5) = P(9) = \frac{4}{36}$$
$$P(6) = P(8) = \frac{5}{36}$$
$$P(7) = \frac{6}{36}$$

Check: $2 \cdot (1+2+3+4+5) + 6 = 36$

Open Source Shakespeare



OpenSourceShakespeare

SEARCH TEXTS GO

Plays ✦ Sonnets ✦ Poems ✦ Concordance ✦ Character Search ✦ Advanced Search ✦ About OSS

Shakespeare text statistics

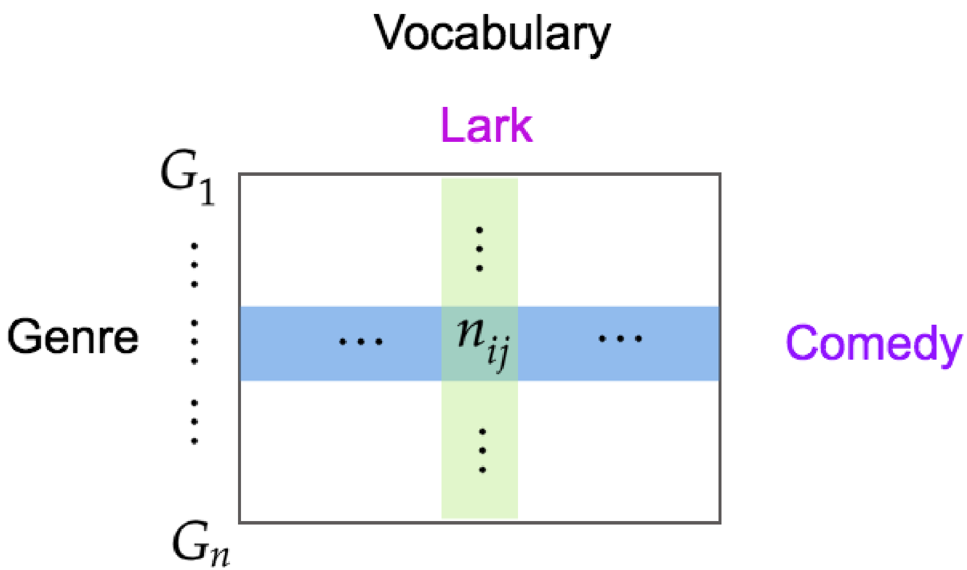
Quick facts

- The plays contain **34,895** total speeches spoken by **1,223** characters.
- There are **884,421** total words in Shakespeare's 43 works.
- There are **28,829** unique word forms, and **12,493** occur only once.
- Those unique words account for **43.3%** of total word forms.
- The top 10 most frequently occuring words make up **21.4%** of all words.
- The top 100 most frequently occuring words make up **53.9%** of all words.
- The top 1% most frequently occuring words make up **66.7%** of all words.

Word count by genre

Distribution from Data (Word Frequencies)

- Corpus $\mathcal{C}$ containing $n_c = |\mathcal{C}|$ words. $|\mathcal{C}|$=884,421
- Corpus $\mathcal{C}$ is divided into sub-corpora by genre
- Corpus $\mathcal{C}$ uses vocabulary $\mathcal{V}$ with $n_v = |\mathcal{V}|$ words. $|\mathcal{V}|$=28,829
- $P('lark') = (\sum_{i \in genres} n_{i,1})/n_c$ (assumes 'lark' is at position 1 in $\mathcal{V}$)
- $P('lark'|'comedy') = n_{1,1}/(\sum_{j \in \mathcal{V}} n_{i,j})$



Random Variables

A **random Variable** (RV) is a real valued function of an experimental outcome.

- For example, a random variable X could give the number of times heads appears in a sequence of n coin tosses. Or number of tosses until first head
- The probability space/probability model specifies outcomes and probabilities associated with outcomes
- The values of a random variable then "inherit" probabilities from the assocated probability space

$$X(\omega_i) = x_j \in \mathbb{R}$$

$$X^{-1}(x_j) = \{\omega_i \mid X(\omega_i) = x_j\}$$

Let the random variable X take on the values $\{x_k\}$, then the collection of probabilities $P(X = x_k)$ is the distribution of the random variable.

Can essentially forget about the probability space

Discrete Random Variable

- Random variables are mappings from outcomes to $\mathbb{R}$. $X \sim Ber(p)$
- Discrete random variable X has a PMF $P(x) = \mathbb{P}(X = x)$
  - The **expected (mean) value** of X is: $E[X] = \Sigma_i x_i \, P(x_i)$ provided the sum is absolutely convergent
  - Given a sample $x^{(i)}$ $i \in 1, \ldots, M, \quad \mathbb{E}[X]$ can by approximated by $(1/M)\Sigma_i x^{(i)}$.
  - Expected value is linear: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$
  - The **variance** of X is the expected value of the squared deviation from the mean:

$$\mathbb{V}[X] = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mu^2$$

- The standard deviation $\sigma_x = \sqrt{\mathbb{V}[X]}$
- Note: if $X \sim Ber(p)$ then $\mathbb{E}[X] = 0 \cdot P(X = 0) \; + \; 1 \cdot P(X = 1) = P(X = 1)$

Bernoulli Distribution

X is a **Bernoulli** random variable with parameter $p \in [0, 1]$ if $P(X = 1) = p$ and $P(X = 0) = 1 - p$. Denoted $X \sim Bernoulli(p)$.

$$\Sigma_{x_i} Ber(x_i|p) = p + (1 - p) = 1 \text{ as required}$$

- For $x \in \{0, 1\}$

$$Ber(x|\, p) = p^x (1 - p)^{1-x}$$

- Expected value

$$\sum_{x_i} Ber(x_i|\, p)\, x_i = 0 \cdot (1 - p) + 1 \cdot p = p$$

- Variance

$$\mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - p^2 = 0^2 \cdot (1 - p) + 1^2 \cdot p - p^2 = p \cdot (1 - p) = pq$$

Binomial Distribution

Toss a coin n times, what is the probability of getting $0, \cdots, n$ heads.

The possible outcomes are n-long sequences of H/T or 0/1 -let p be the probability of H -Tosses are independent so probabilities are products

$$P(HTH) = p(H)p(T)p(H) = p^2(1-p)$$

Each n-long sequence has probability $p^{n_h}(1-p)^{n-n_h)}$

Where n_h is the number of heads.

There is only 1 way to toss all heads, so $P(H, H, \cdots, H) = p^n$

There are n-ways to toss 1 head:

$$P(\text{one head}) = P(H, T, \cdots, T) + P(T, H, T, \cdots) + \cdots = np^1(1-p)^{n-1}$$

In general, the probability of k heads is

$$B(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial Distribution

The binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Comes from the expansion

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

Setting $a = p, b = 1 - p$ shows that binomial distribution meets the requirement

$$\sum_k B(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$$

Binomial Distribution

X is binomial random variable with parameters n, p if it gives the probability for the number of successes in n binary trials with p the probability of success on each trial.

- For $k \in 0, 1, \ldots, n$

$$B(k|n,p) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

- Expected Value (the hard way). If $X \sim B(k|n,p)$

$$E[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{(n-k)}$$

$$= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{(n-k)}$$

$$= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{(k-1)} (1-p)^{(n-k)}$$

Shift sum by 1

$$= np \sum_{k=0}^{n-1} \frac{(n-1)!}{k!(n-k-1)!} p^k (1-p)^{(n-k-1)}$$

Set $m = n-1$

$$= np \sum_{k=0}^{m} \frac{m!}{k!(m-k)!} p^k (1-p)^{(m-k)}$$

$$= np$$

Properties of Binomial Distribution

X is binomial random variable with parameters n, p if it gives the probability for the number of successes in n binary trials with p the probability of success on each trial.

Let Y be a Bernoulli RV with paramter p so $E[Y] = p$ and $Var[Y] = p(1-p)$. In terms of Y,

$$X = Y_1 + Y_2 \ldots + Y_n$$

so

$$E[X] = E[\Sigma_i Y_i] = \sum_i (E[Y_i]) = n\,p$$

The Y_i are independent (covered later) so

$$Var[X] = Var[\Sigma_i Y_i] = \sum_i (Var[Y_i] = n\,p\,(1-p)$$

If n is large enough, and p is not too small, the **Central Limit Theorem** says that X can be approximanted by a Normal distribution.

Continuous Random Variable

- A continuous random variable X has a pdf $p(x)$ satisfying $\mathbb{P}(a < x < b) = \int_a^b p(x)dx$
- The expected value of X is: $\mathbb{E}[X] = \int_{\mathbb{R}} x\, p(x)dx$ provided the integral is absolutely convergent
- If $X \sim U[0,1]$ then $p(x) = 1$ on $[0,1]$ and 0 otherwise, and

$$\mathbb{E}[X] = \int_0^1 x\, p(x)dx = \frac{x^2}{2}|_0^1 = 1/2$$

- The variance $\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[(X - \mu)^2$

$$X \sim U[0,1],\ \ Var[X] = \int_0^1 x^2\, p(x)dx\ -\ (\frac{1}{2})^2 = \frac{x^3}{3}|_0^1 - (\frac{1}{2})^2 = 1/12$$

- The standard deviation $\sigma_x = \sqrt{\mathbb{V}[X]}$

Continuous Distribution

Uniform Distribution

For $a < b$, the uniform pdf is

$$U(x|a,b) = p_U(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

If $X \sim U(x|0,1)$ then

$$F_x(x) = \int_0^x p_u(x)dx = \int_0^x dx = x$$

$$E[X] = \int_0^1 x\, p_u(x)dx = \int_0^1 x\, dx = \frac{x^2}{2}|_0^1 = 1/2$$

$$X \sim U[0,1],\ Var[X] = \int_0^1 x^2\, p(x)dx \ - \ (\frac{1}{2})^2 = \frac{x^3}{3}|_0^1 - (\frac{1}{2})^2 = 1/12$$

Continuous Distribution

Normal (Gaussian) Distribution

$$N(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$\phi(x) = N(x|0,1) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$$N(x|\mu,\sigma^2) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)$$

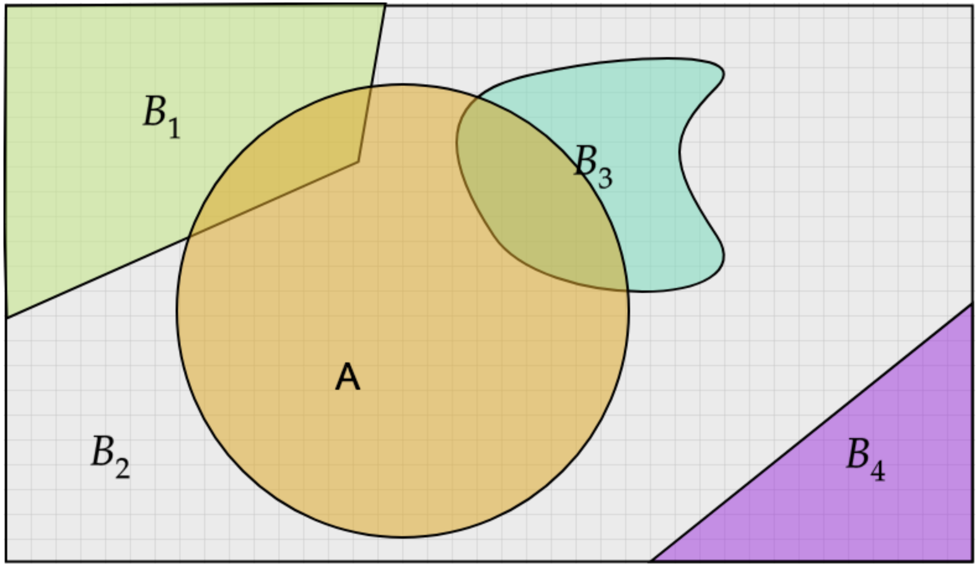
$$\phi^{'}(x) = -x\,\phi(x)$$
$$\phi^{''} = (x^2-1)\,\phi(x)$$

You should be able to show that if $X \sim N(\mu,\sigma^2)$ then $(X-\mu)/\sigma \sim N(0,1)$

Functions of Random Variables

- Functions of random variables are random variables
- Discrete: $\mathbb{E}[f(X)] = \Sigma_i f(x_i)\, P(x_i)$
- Continuous: $\mathbb{E}[f(x)] = \int_{\mathbb{R}} f(x)\, p(x) dx$

Total Probability



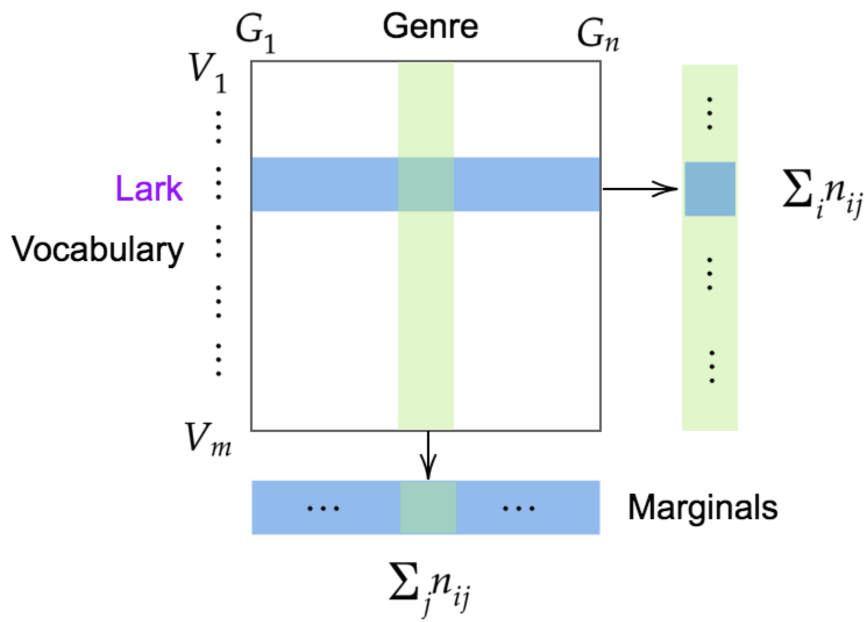
$$P(B_i \cap B_j) = 0 \quad \text{for} \quad i \neq j$$

$$P(\cup_i B_i) = 1$$

$$P(A) = \sum_i P(A \cap B_i)$$

Joint Distributions

- A 2d table is an example of a joint distribution
- Joint distributions lead to conditional distributions and conditional probabilites



Joint Distributions

Most often data is multidimensional. PMFs and PDFs can be extended to describe probabilities involving more that one random variable. The relations shown can be extended to more than two variables

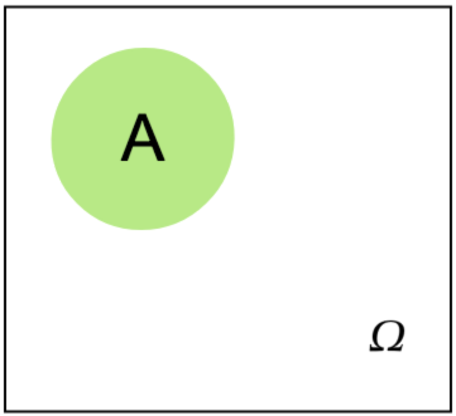
- If P is the PMF describing the joint behavior of random variables X and Y then
- $P(x, y) = \mathbb{P}(X = x,\, Y = y)$
- $\Sigma_x\, \Sigma_y P(x, y) = 1$, likewise,for continuous RV $\int_x \int_y p(x, y)\, dx\, dy = 1$
- If $Z = f(x, y)$ then $\mathbb{E}[Z] = \Sigma_x \Sigma_y f(x, y) P(x, y)$
- The sum rule (marginal probability). The collection of events $Y = y_i$ are disjoint and decompose the probability space so $\{X = x\} = \cup_i(\{X = x\} \cap \{Y = y_i\})$. This gives

$$P(x) = \Sigma_y\, P(x, y)$$

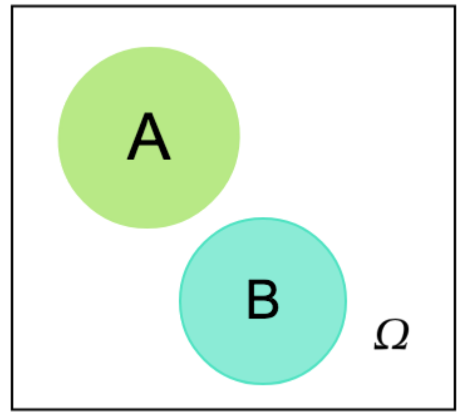
- For example, if $Z = a + bX + cY$ then $\mathbb{E}[Z] = a + b\Sigma_{x,y}\, x\, p(x, y) + c\Sigma_{x,y}\, y\, p(x, y)$. Using the sum rule resuls in

$$\mathbb{E}(a + b\, X + c\, Y) = a + b\mathbb{E}[X] + c\mathbb{E}[Y]$$

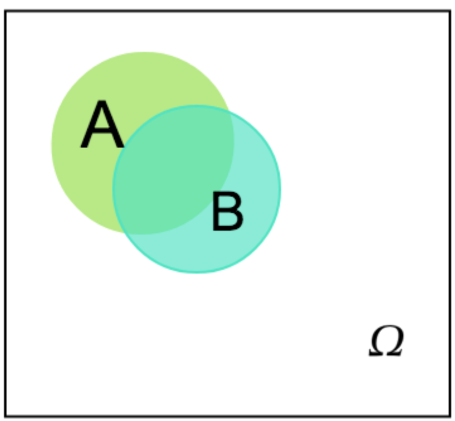
Conditional Probability



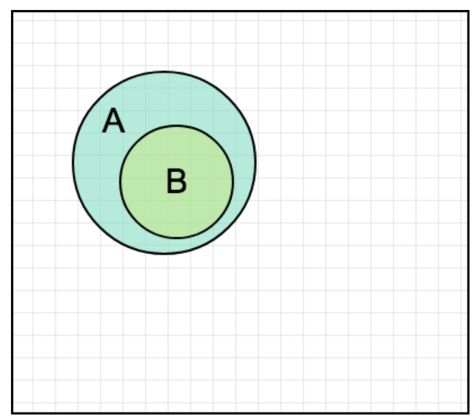
$P(A)$



$P(A|B) = 0$



$P(A|B) = .5$



$P(A|B) = 1$

Conditional Probability

The (unconditional) probability of an event $P(A)$ is evaluated within the full event space Ω

The conditional probabiltiy $P(A|B)$ is the probability of A assuming focus is restricted to B

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

Example Toss a coin 3 times and let X_i be the number of heads after i tosses. Assume a fair coin.

$$P(X_3 = 3) = P(H, H, H) = \frac{1}{8}$$

$$P(X_3 = 3|X_1 = 1) = \frac{P(X_1 = 1, X_3 = 3)}{P(X_1 = 1)} = \frac{1/8}{4/8} = \frac{1}{4}$$

$$P(X_3 = 3|X_2 = 1) = \frac{1/8}{4/8} = \frac{1}{4}$$

Discrete Conditional Probability

$$
\begin{array}{c c}
 & \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} &
\begin{bmatrix}
2 & 3 & 4 & 5 & 6 & 7 \\
3 & 4 & 5 & 6 & 7 & 8 \\
4 & 5 & 6 & 7 & 8 & 9 \\
5 & 6 & 7 & 8 & 9 & 10 \\
6 & 7 & 8 & 9 & 10 & 11 \\
7 & 8 & 9 & 10 & 11 & 12
\end{bmatrix}
\end{array}
$$

Toss a pair of dice

- Let A be the event that the sum S is odd, and let B be the event that the sum is prime
  - Compute $P(A), P(B), P(B|A)$

$$P(A) = P(S \in 3, 5, 7, 9, 11) = \frac{18}{36}$$

$$P(B) = P(S \in 2, 3, 5, 7, 11) = \frac{15}{36}$$

$$P(B|A) = \frac{P(A, B)}{P(A)} = \frac{14/36}{18/36} = \frac{7}{9}$$

Bayes' Rule

Multiplication Rule

$$P(\bigcap_{i=1}^{n} A_i) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \ldots P(A_n|\bigcap_{i=1}^{n-1} A_i)$$

$$= P(A_1)\frac{P(A_1 \cap A_2)}{P(A_1)}\frac{P(A_1 \cap A_2 \cap A_3)}{p(A_1 \cap A_2)} \cdots \frac{P(\bigcap_{i=1}^{n} A_i)}{P(\bigcap_{i=1}^{n-1} A_i)}$$

Bayes' Rule

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

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

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

Total Probability Law

Total Probability

If A_i are disjoint events, $A_i \cap A_j = \emptyset, i \neq j$, such that

$$\bigcup_{i=1}^{n} A_i = \Omega$$

Then

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n) \\ &= P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n) \end{aligned}$$

Marginal Distribution - Discrete case

Let $P(x, y)$ be a joint PMF. The set of events $Y = y_j$ are disjoint and their union spans Ω. Total probability gives

$$\begin{aligned} P(X = x) &= P(X = x, Y = y_1) + \cdots + P(X = x, Y = y_n) \\ &= \sum_{y_j} P(X = x, Y = y_j) \\ &= \sum_{y} P(x, y) \end{aligned}$$

Marginal distribution Again

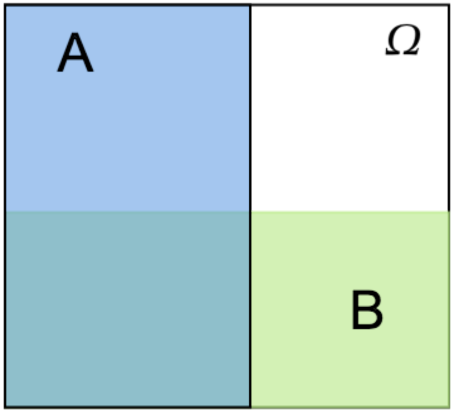
- Conditional probability determines marginal distribution

$$
\begin{aligned}
p_X(x) &= \sum_y p_{X,Y}(X = x, Y = y) \\
&= \sum_y p(x|y)p(y)
\end{aligned}
$$

Continuous Distribution

$$
\begin{aligned}
p(x) &= \int_y p(x, y)dy \\
&= \int_y p(x|y)p(y)dy
\end{aligned}
$$

Independent Events

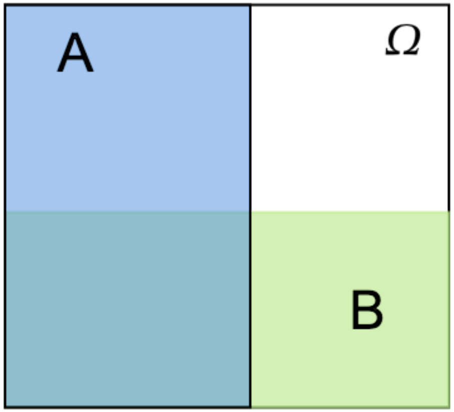


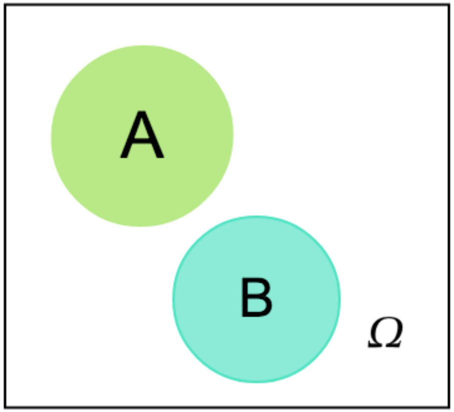
$$P(A) = \frac{1}{2}$$

$$P(B) = \frac{1}{2}$$

$$P(A \cap B) = \frac{1}{4} = P(A)P(B)$$

Independent Events





$$P(A) = \frac{1}{2}$$

$$P(B) = \frac{1}{2}$$

$$P(A \cap B) = \frac{1}{4} = P(A)P(B)$$

$$P(A) = p_a \neq 0$$

$$P(B) = p_b \neq 0$$

$$P(A \cap B) = 0 \neq P(A)P(B)$$

Independence

The intuitive idea that unrelated events do not affect each other's probabilities is called indepence. For example, when tossing a fair coin twice, the probability that the second toss results is H does not depend on the outcome of the first toss.

- If X is outcome of first toss and Y is outcome of the second toss, then for a fair coin all outcomes are equally likely.
- $P(X=i, Y=j) = 1/4 = P(X=i)\,P(Y=j)$ for $i, j \in 0, 1$
- The mathematical condition for independence is

$$P(X=x, Y=y) = P(X=x)\,P(Y=y)$$

- For continuous random variables:

$$p(x, y) = p(x)\,p(y)$$

- This implies that $\mathbb{E}[X\,Y] = \int\int p(x,y)dxdy = \int\int p(x)dx\,p(y)dx = \mathbb{E}[X]\mathbb{E}[Y]$

Independent Events

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array}
\begin{array}{c}
\begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\
\begin{bmatrix}
2 & 3 & 4 & 5 & 6 & 7 \\
3 & 4 & 5 & 6 & 7 & 8 \\
4 & 5 & 6 & 7 & 8 & 9 \\
5 & 6 & 7 & 8 & 9 & 10 \\
6 & 7 & 8 & 9 & 10 & 11 \\
7 & 8 & 9 & 10 & 11 & 12
\end{bmatrix}
\end{array}$$

Toss a pair of dice

- Let A be event that the sum is 6. Let B be the event that the first toss was 4. Are A,B independent?

$$P(A) = \frac{5}{36}, \quad P(B) = \frac{6}{36}$$
$$P(A \cap B) = \frac{1}{36} \neq \frac{5 \cdot 6}{36 \cdot 36}$$

- Let A be event that the sum is 7. Let B be the event that the first toss was 4. Are A,B independent?

$$P(A) = \frac{6}{36}, \quad P(B) = \frac{6}{36}$$
$$P(A \cap B) = \frac{1}{36} = \frac{6 \cdot 6}{36 \cdot 36}$$

Empirical Mean

Given data points $x^{(i)}$ sampled from a distribution, is it obvious that the following is true?

$$\frac{1}{m}\sum_{i=1}^{m} x^{(i)} \underset{m\to\infty}{\longrightarrow} E[X]$$

By definition (discrete)

$$E[X] = \sum_{j=1}^{k} p(x_j)x_j$$

Each sample $x^{(i)} = x_j$ for some j. Grouping the empirical average by values

$$\frac{1}{m}\sum_{i=1}^{m} x^{(i)} = \frac{1}{m}\Big[\sum_{x^{(i)}=x_1} x^{(i)} + \cdots \sum_{x^{(i)}=x_k} x^{(i)}\Big]$$

$$= \frac{n_1}{m}x_1 + \cdots + \frac{n_k}{m}x_k$$

where $n_1 + \cdots n_k = m$

As $m \to \infty$, $\frac{n_1}{n} \to p(x_1)$ etc

Notation for Empirical Expected Value

Assume that C is a cost function. Given a set of data $\mathcal{D}$, the expected value is given by

$$\begin{aligned} E_{(\mathbf{x},y)\sim\mathcal{D}}[C(y,f(\mathbf{x}))] &= \sum_{(x,y)\in\mathcal{D}} p(\mathbf{x},y)\, C(y,f(\mathbf{x})) \quad \text{discrete} \\ &= E[C] \end{aligned}$$

If $\mathbf{x}^{(i)}, y^{(i)}$ drawn randomly from $\mathcal{D}$ then

$$\frac{1}{n}\sum_i C(y^{(i)}, f(\mathbf{x})^{(i)}) \to E[C]$$

Empirical Covariance

$Cov[X, Y] = E[(X - E[X])(Y - E[Y])] = E[XY]$ when $E[X] = E[Y] = 0$

Given data vectors $\mathbf{x}^{(i)}$, $i = 1, \cdots, m$ want to compute the empirical covariance of the $\mathbf{x}$ components

Let $\tilde{\mathbf{x}}^{(i)}$ be the centered (zero mean) version of the data $\mathbf{x}^{(i)}$

$$\tilde{\mathbf{x}}^{(i)} = \mathbf{x}^{(i)} - \frac{1}{m} \sum \mathbf{x}^{(i)}$$

Now the (j, k) element of the symmetric covariance matrix will be the mean of $\tilde{x}_j$ times $\tilde{x}_k$

$$C_{j\,k} = \frac{1}{m} \sum_{i=1}^{m} \tilde{x}_j \tilde{x}_k$$

Note: This empirical equation is biased, should divide by $m - 1$ to get the unbiased estimate. The difference is seldom important.

Empirical Covariance

The matrix outer product gives cross-products of a vector

$$\begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_d^{(i)} \end{bmatrix} [x_1^{(i)}, \cdots, x_d^{(i)}] = \begin{bmatrix} x_1^{(i)} x_1^{(i)} & \cdots & x_1^{(i)} x_d^{(i)} \\ \vdots & & \vdots \\ x_d^{(i)} x_1^{(i)} & \cdots & x_d^{(i)} x_d^{(i)} \end{bmatrix}$$

Will show that one way to interpret matrix multiplication AB is

$AB = \sum_i col_i(A) \bigotimes row_i(B)$

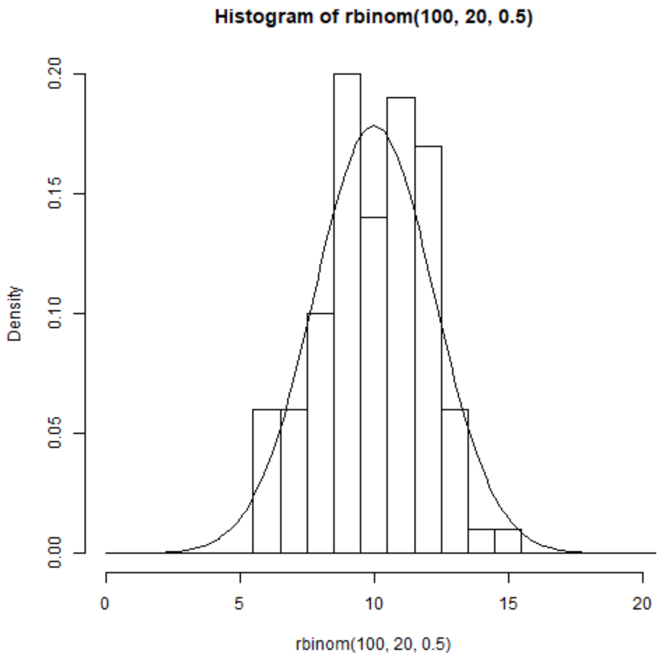
It follows that the biased empiricial covariance matrix is $(col_i(X) = \mathbf{x}^{(i)})$

$$\begin{aligned} C &= \frac{1}{m} \sum_i col_i(x)\, row_i(x^T) \\ &= \frac{1}{m} X X^T \end{aligned}$$

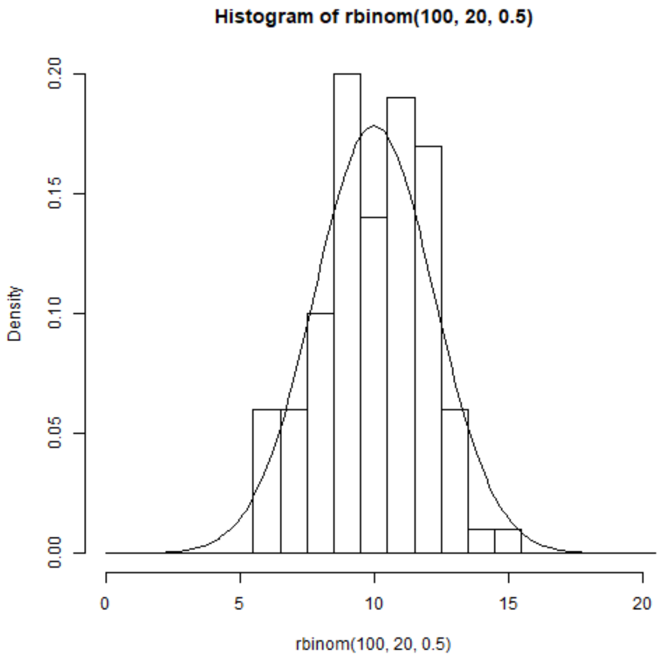
note 1: It is the transpose of this if data is arranged by rows.

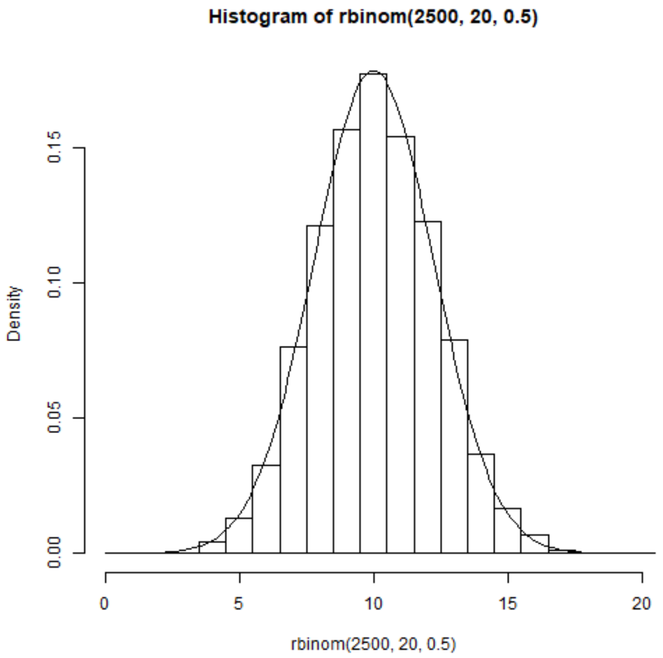
note 2: The dimension of the covariance matrix should be $d \times d$ if the data vectors have d elements

Normal Approximation to Binomial Distribution



Normal Approximation to Binomial Distribution





Code for Normal Approximation Plot

```
# 100 samples, number of successes in 20 trials with p=.5
hist(rbinom(100,20,0.5),xlim=c(0,20),probability=T,
     breaks=seq(0.5,20.5,1))
xgrid=seq(0,20,.1)
lines(xgrid,dnorm(xgrid),10,sqrt(5)))
```

Homework 1

Discriminant

$$d(x) = P(C_1|x) - P(C_2|x) = \begin{cases} \geq 0 & \text{then } C_1 \\ < 0 & \text{then } C_2 \end{cases}$$

We don't know $P(C_i|x)$ but we do know $P(x|C_i)$ (class-conditional distribution) and $P(C_i)$ so apply Bayes' Rule

$$P(C_i|x) = \frac{P(C_i)P(x|C_i)}{p(x)}$$

Substituting into $d(x)$ gives

$$d(x) = P(C_1)P(x|C_1) - P(C_2)P(x|C_2)$$

The discriminant function $d(x) = 0$ at the optimal cut point

If the class-conditional probabilities are approximated by normals then $d(x)$ is continuous and a root finder can be used to find $\mathbf{x}^*$ such that $d(\mathbf{x}^*) = 0$.

Homework 1

Bayes' Error

If we know the joint distribution of the data and the label then a Bayes' classifier can be constructed

$$b_{opt}(\mathbf{x}) = arg\max_{C_i} p(\mathbf{x}, C_i)$$

The probability of misclassification is $1 - b_{opt}(\mathbf{x})$. For each point $\mathbf{x}$, no other classifier can have a smaller probability of error.

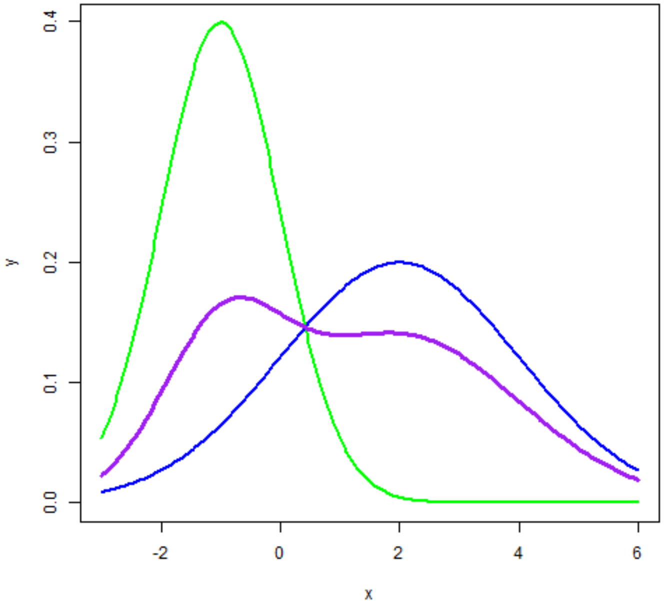
The Bayes' error is the probability of error associated with b_{opt}.

Assume the data is a mixture of Gaussians

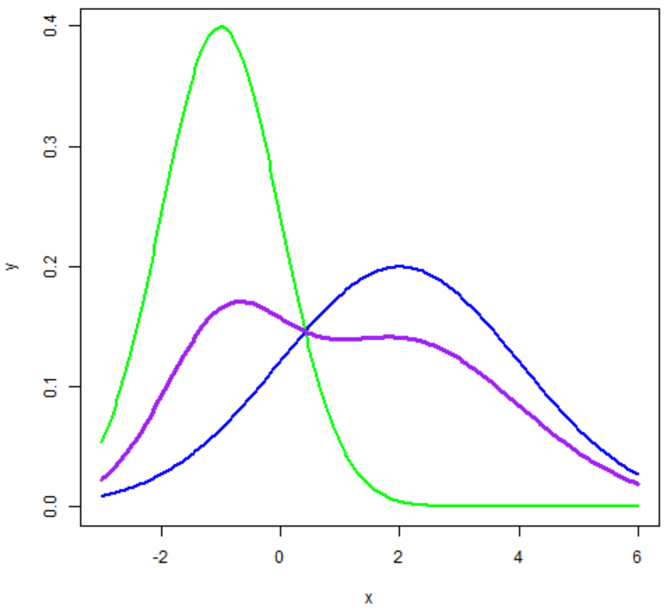
$$p(\mathbf{x}, c) = p(C_1)N(\mathbf{x}; \mu_1, \sigma_1) + p(C_2)N(\mathbf{x}; \mu_2, \sigma_2)$$

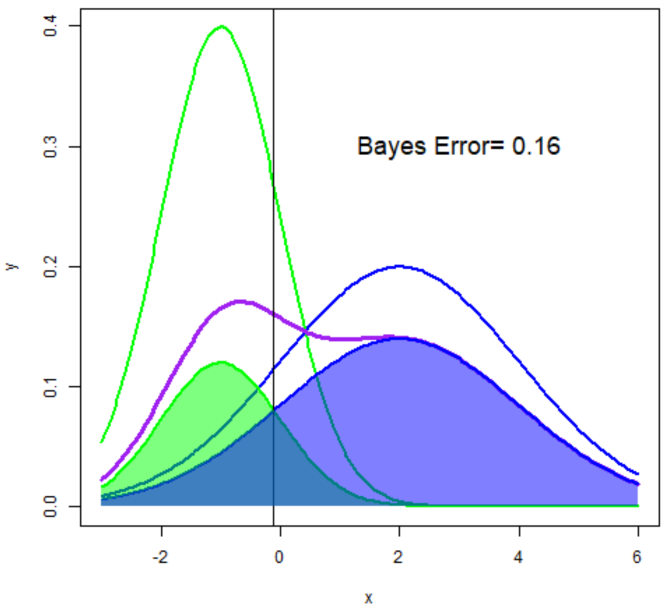
where $P(C_2) = 1 - P(C_1)$

Bayes' Error



Bayes' Error





Code for Bayes Error Plot 1

```
require(zeallot,quietly = TRUE)

# define two normal distributions
# data set 1: 30 points, mean -1, sigma 1
n1=30; m1=-1; s1=1; n2=70; m2=2; s2=2

shade.under.curve=function(xgrid,fcn,rgb.color){
  for(i in 1:(length(xgrid)-1)){
    dx=c(xgrid[i],xgrid[i+1])
    dx=c(dx,rev(dx))
    dy=c(fcn(xgrid[i]),fcn(xgrid[i+1]),0,0)
    polygon(dx,dy,col=rgb.color,border=NA)
  }
}

mixture=function(x,n1,m1,s1,n2,m2,s2){
  (n1*dnorm(x,m1,s1)+n2*dnorm(x,m2,s2))/(n1+n2)
}
discriminant.cl = function(n1,m1,s1,n2,m2,s2){
  function(x){dnorm(x,m2,s2)*n2-(dnorm(x,m1,s1)*n1)}
}
discriminant=discriminant.cl(n1,m1,s1,n2,m2,s2)
```

Code for Bayes Error Plot 1 Continued

```
bayes.error=function(cut,n1,m1,s1,n2,m2,s2){
  w1=n1/(n1+n2)
  w2=1-w1
  if(m1 < m2){
    error=w2*pnorm(cut,m2,s2)+w1*(1-pnorm(cut,m1,s1))
  }else{
    error=w1*pnorm(cut,m1,s1)+w2*(1-pnorm(cut,m2,s2))
  }
  error
}

grid=seq(m1-2*s1,m2+2*s2,length.out = 200)
plot(grid,dnorm(grid,m1,s1),col="green",type="l",lwd=2,
     xlab="x",ylab="y")
lines(grid,dnorm(grid,m2,s2),col="blue",type="l",lwd=2)
lines(grid,mixture(grid,n1,m1,s1,n2,m2,s2),col="purple",
      type="l",lwd=3)
```

Code for Bayes Error Plot 2

```
grid=seq(m1-2*s1,m2+2*s2,length.out = 200)
plot(grid,dnorm(grid,m1,s1),col="green",type="l",lwd=2,
     xlab = "x",ylab="y")
lines(grid,dnorm(grid,m2,s2),col="blue",type="l",lwd=2)
lines(grid,mixture(grid,n1,m1,s1,n2,m2,s2),col="purple",
      type="l",lwd=3)
f12.cl=function(w1,m1,s1){function(x){w1*dnorm(x,m1,s1)}}
f1=f12.cl(n1/(n1+n2),m1,s1)
f2=f12.cl(n2/(n1+n2),m2,s2)
lines(grid,n1*dnorm(grid,m1,s1)/(n1+n2),col="green",
      type="l",lwd=2)
lines(grid,n2*dnorm(grid,m2,s2)/(n1+n2),col="blue",
      type="l",lwd=2)
bayes.cut=uniroot(discriminant,c(-2,2))$root
abline(v=bayes.cut)
shade.under.curve(seq(-3,2,length.out = 30),f1,rgb(0,1,0,.5))
shade.under.curve(seq(-3,6,length.out = 50),f2,rgb(0,0,1,.5))
error=bayes.error(bayes.cut,n1,m1,s1,n2,m2,s2)
text(3,.3,paste("Bayes Error=",round(error,2)),cex=1.5)
```

Bayes Multidimensional Discriminant Function

The Baye's decision rule classifies data point x as $\mathcal{C}_1$ if $P(\mathcal{C}_1|x) \geq P(\mathcal{C}_2|x)$ and as $\mathcal{C}_2$ otherwise.

Let

$$y_k(x) = P(\mathcal{C}_k \mid x) = P(x|\, \mathcal{C}_k) \cdot P(\mathcal{C}_k)$$

Using $y(x) = y_1(x) - y_2(x)$ the decision rule becomes $x \in \mathcal{C}_1$ if $y(x) \geq 0$ and $\mathcal{C}_2$ otherwise. An alternative is to take the ln of y_k and let $y(x) = \ln y_1(x) - \ln y_2(x)$. This gives

$$\begin{aligned} y(x) &= \ln y_1(x) - \ln y_2(x) \\ &= \ln \frac{p(x|\mathcal{C}_1)}{p(x|\mathcal{C}_2)} + \ln \frac{P(\mathcal{C}_1)}{P(\mathcal{C}_2)} \end{aligned}$$

This form will be convenient when the class conditional probabilities are Gausian.

Multidimensional Gaussian

Let X be a random vector with d components $\mathrm{X}_1, \ldots, \mathrm{X}_d$. Let Σ be the $d \times d$ covariance matrix $\Sigma_{i,j} = \mathbb{E}[(\mathrm{X_i} - \mu_i)(\mathrm{X}_j - \mu_j)]$ where μ is the vector mean of X. In d dimensions the general multivariate Gaussian probability density function is:

$$\mu = \mathbb{E}[\mathrm{X}]$$

$$\Sigma = \mathbb{E}[(\mathrm{X} - \mu)(\mathrm{X} - \mu)^T]$$

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

$$\mathrm{D}^2 = (x - \mu)^T \Sigma^{-1} (x - \mu)$$

Note that Σ^{-1} is symmetric because Σ is symmetric.

Bayes Decision Rule for Gaussian Data

If each of the class conditional densities $p(x|\mathcal{C}_k)$ are independed and Gaussian then the discriminant functions become:

$$\ln y_k(x) = -\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k) - \frac{1}{2}\ln|\Sigma_k| + \ln P(\mathcal{C}_k)$$

This can be simplified if the class conditional covariance matrices Σ_k are equal. With that simplification, after dropping terms that don't depend on k, have

$$\ln y_k(x) = \mathbf{w}_k^T\mathbf{x} + b_k$$

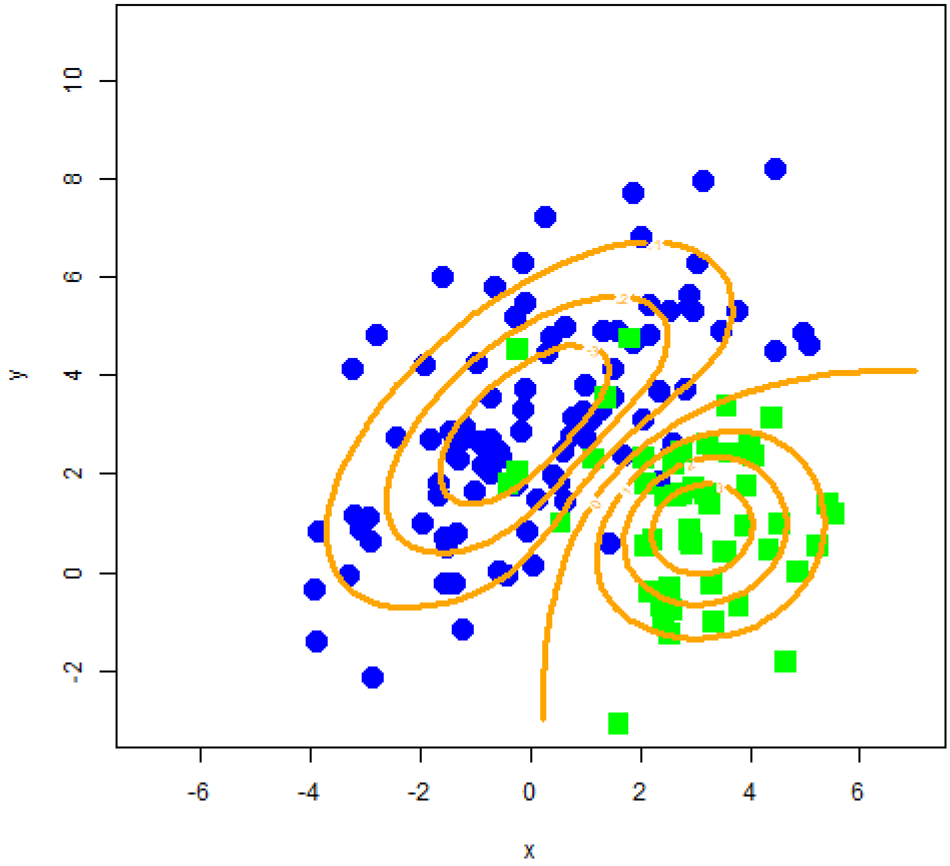
$$\mathbf{w}_k^T = \mu_k^T\Sigma^{-1}$$

$$b_k = -\frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \ln P(\mathcal{C}_k)$$

Since each term in the discriminant is linear in $\mathbf{x}$, the decision boundaries will by hyperplanes

See Bishop section 4.2.1

Bayes Classifier, 2d



Code for Bayes Classifier 2d

```
require(mvtnorm,quietly = TRUE)
set.seed(1)
n1=100; n2=50
m1=c(0,3)
S1=matrix(c(5,3,3,5),nrow=2)
m2=c(3,1)
S2=matrix(c(2,0,0,2),nrow=2)
x1=rmvnorm(n1,m1,S1)
plot(x1,xlim=c(-7,7),ylim=c(-3,11),pch=16,cex=2,col="blue",
     xlab="x",ylab="y")
x2=rmvnorm(n2,m2,S2)
points(x2,pch=15,cex=2,col="green")
```

Code for Bayes Classifier 2d Continued

```
discriminant.cl=function(n1,m1,S1,n2,m2,S2){
  function(x){
    dmvnorm(x,m2,S2)*n2-dmvnorm(x,m1,S1)*n1
  }
}
discriminant=discriminant.cl(n1,m1,S1,n2,m2,S2)
nxg=50
xg=seq(-7,7,length.out = nxg)
yg=seq(-3,11,length.out = 50)
x=as.matrix(expand.grid(xg,yg))
z=matrix(discriminant(x),nrow=nxg)
contour(xg,yg,z,add=TRUE,lwd=3,col="orange")
```