CS 445/545
Machine Learning
Winter, 2009

- Course overview:
  - Instructor
    - Melanie Mitchell
  - Readings
    - No textbook
    - Readings will be downloadable from the Web

  - Homework
    - One assignment for each topic
    - Written and computer exercises
    - Do on your own; no teamwork, but you can discuss general ideas with classmates
  - Presentation
    - You'll be assigned one of the readings to present in class
  - Midterm exam
    - In class, closed book exam
  - Final exam
    - Take-home open-book exam.

- Grading
  - Homeworks: 50%
  - Presentation: 10%
  - Midterm: 20%
  - Final: 20%

- Mailing list: MachineLearning2009@cs.pdx.edu
  (please give me your e-mail address)

- Academic Integrity
  - You can discuss concepts with other students, but all work you hand in must be your own.
  - You may not copy any text from other students, papers, the internet, or any other source, without clearly referencing it.
  - See syllabus for URL with PSU policy.

What is machine learning?

- Textbook definitions of "machine learning":
  - Detecting patterns and regularities with a good and useful approximation ("model" or "hypothesis")
  - Execution of a computer program to optimize the parameters of the model using the training data or past experience.

A B C D E

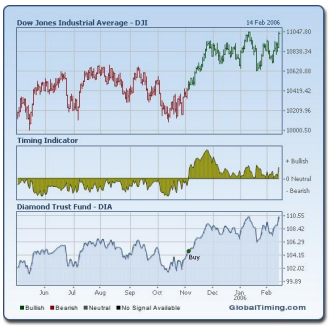
A B C D E

A B C D E

A B C D E

A B C D E

- Example:
  - Task: Letter recognition
  - Performance measure: Fraction of letters recognized correctly
  - Training experience: Different examples of letters in different typefaces



• Example:

– Task: Stock price prediction

– Performance measure: Amount of profit earned

– Training experience: Past time series of stock prices

Types of machine learning tasks

- Classification
  - Output is one of a number of classes (e.g., 'A')
- Regression
  - Output is a real value (e.g., '$35/share")

Difference between “artificial intelligence” and “machine learning”?

Key Ingredients for Any Machine Learning Method

- Underlying representation for “hypothesis”, “model”, or “target function”:
  - mathematical expression, bit string, neural network, decision tree, logical description, if-then rules
- Features (or “attributes”)
  - Elements of underlying representation that describe which aspects of problem instances (or training examples) should be used in learning.

Key Ingredients for Any Machine Learning Method

- Space to search:
  - coefficient values, bits, weights, topologies of networks, topologies of trees, possible logic sentences
- Search method:
  - gradient descent, genetic algorithm, greedy algorithm, etc.
- Data: Divide into three parts.
  - Training data
    - Used to train the model
  - Validation data
    - Used to select model complexity, to determine when to stop training, or to alter training method
  - Test data
    - Used to evaluate trained model

Types of Machine Learning Methods

- Supervised
  - provide explicit training examples with correct answers
    - e.g. neural networks with back-propagation
- Unsupervised
  - no feedback information is provided
    - e.g., unsupervised clustering based on similarity

Types of Machine Learning Methods, continued

- "Semi-supervised"
  - feedback information is provided, but is not detailed
    - examples:
      - genetic algorithm: calculates single-valued "fitness" of individual in population
      - reinforcement learning: reinforcement single is single-valued assessment of current state

Assumption of all ML methods:

Inductive learning hypothesis:

Any hypothesis that approximates target concept well over sufficiently large set of training examples will also approximate the concept well over other examples outside of the training set.

Difference between "induction" and "deduction"?

What questions does ML ask?

- Which is the best method for a given learning problem?
- What can be proved about convergence, performance?
- How much training data or feedback is needed for good generalization performance?

- How to select/design training examples or feedback?
- How can prior knowledge be used to guide learning?
- How can meta-learning be used? (I.e., learner automatically modifies its representation, selection of training examples, etc.)

Syllabus

Review of Probability Theory

Optional reading: P. Sebastiani, A Tutorial on Probability Theory

(link from class website)

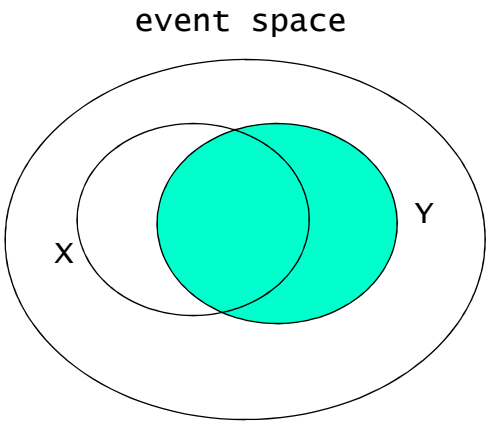
Conditional Probability

- Probability of an event given the occurrence of some other event.

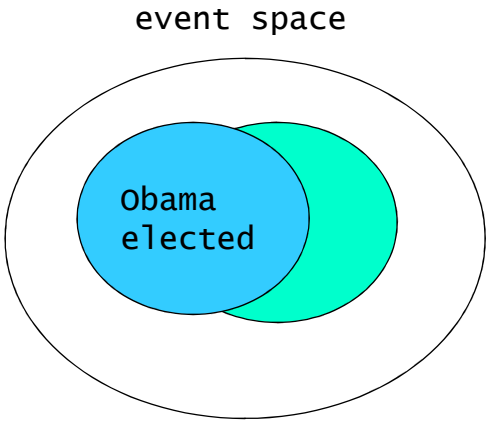
E.g.,

What is the probability that a liberal Supreme Court Justice will be appointed in the next four years, given that Barack Obama was elected President?

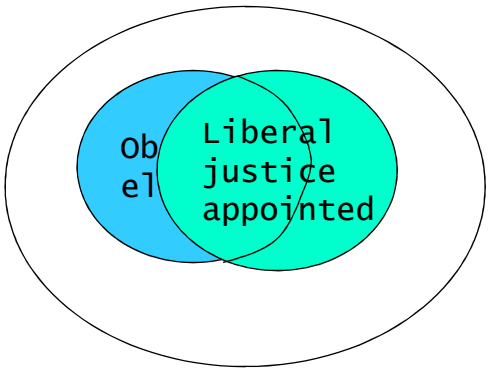
$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}$$



$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}$$

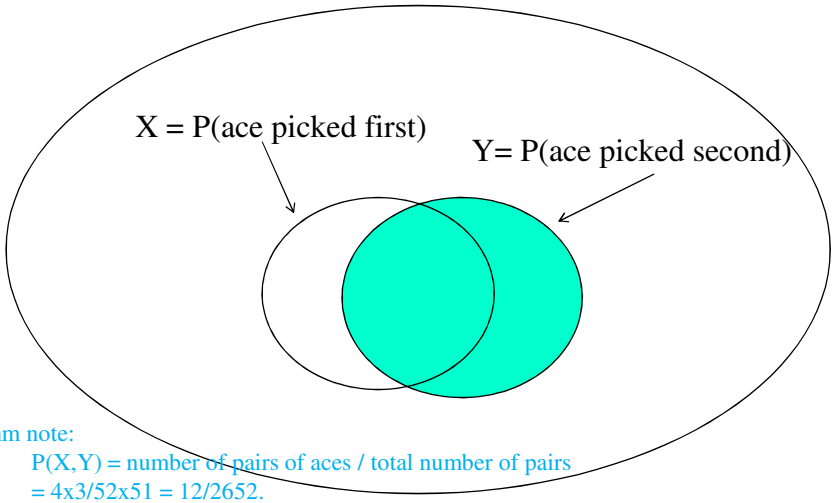


$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{P(X, Y)}{P(Y)}$$



- Consider choosing a card from a well-shuffled standard deck of 52 playing cards. Given that the first card chosen is an ace, what is the probability that the second card chosen will be an ace?

event space: all possible pairs of cards

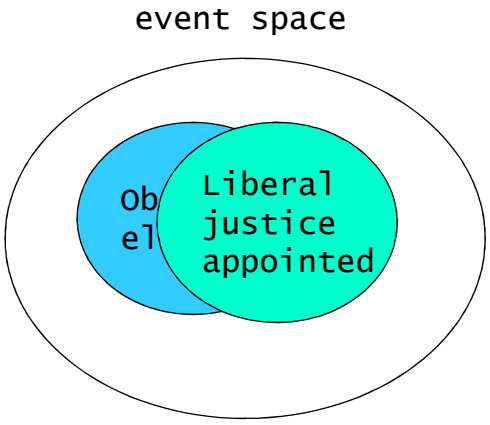


mm note:

P(X,Y) = number of pairs of aces / total number of pairs
= 4x3/52x51 = 12/2652.

P(Y) = 4 / 52
P(X | Y) = (12/2652) / (4 / 52) = 3/51.

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}$$



Relationships among joint, conditional, posterior, and marginal probabilities

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{P(X, Y)}{P(Y)}$$

$$P(X, Y) = P(X \mid Y)P(Y) = P(Y \mid X)P(X)$$

Bayes rule :

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

Application to Machine Learning

- In machine learning we have a space H of hypotheses:
  $h_1, h_2, \ldots, h_n$
- We also have a set D of data
- We want to calculate P(h|D)

Terminology

- ***Prior probability* of *h*:**
  - *P(h)*: Probability that hypothesis *h* is true given our prior knowledge
  - If no prior knowledge, all $h \in H$ are equally probable
- ***Posterior probability* of *h*:**
  - *P(h|D)*: Probability that hypothesis *h* is true, given the data *D*.
- ***Likelihood* of *D*:**
  - *P(D|h)*: Probability that we will see data *D*, given hypothesis *h* is true.

Bayes rule says: $$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

The Monty Hall Problem

You are a contestant on a game show.

There are 3 doors, A, B, and C. There is a new car behind one of them and goats behind the other two.

Monty Hall, the host, asks you to pick a door, any door. You pick door A.

Monty tells you he will open a door , different from A, that has a goat behind it. He opens door B: behind it there is a goat.

Monty now gives you a choice: Stick with your original choice A or switch to C.

Should you switch?

http://math.ucsd.edu/~crypto/Monty/monty.html

Bayesian probability formulation

Hypothesis space H:

h_1 = Car is behind door A

h_2 = Car is behind door B

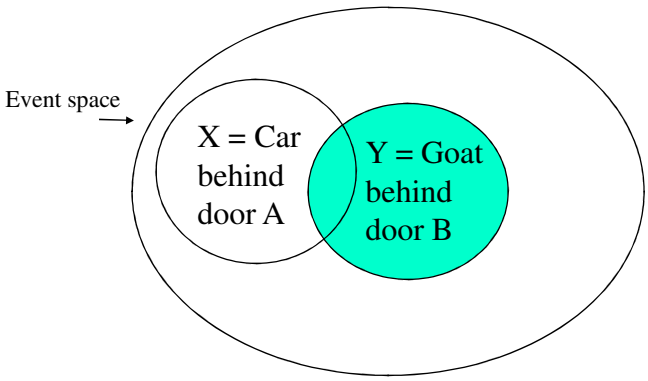
h_3 = Car is behind door C

Data D = Monty opened B

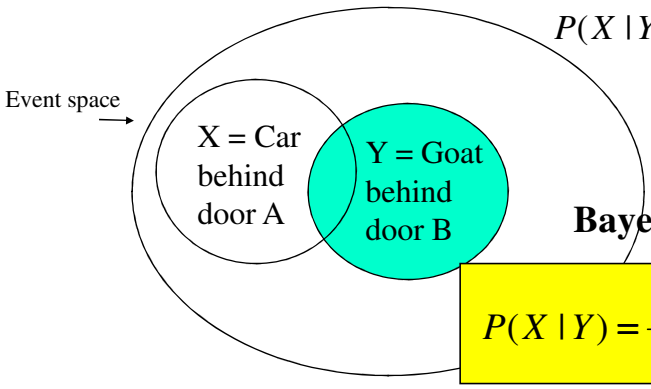
What is $P(h_1 \mid D)$?

What is $P(h_2 \mid D)$?

What is $P(h_3 \mid D)$?



Event space = All possible configurations of cars and goats behind doors A, B, C



$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}$$

Bayes Rule:

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

Another example

It is known that about 0.0038 of the U.S. population is infected with HIV.

The current test for HIV has a 0.002 probability of returning a false positive, and a 0.004 probability of returning a false negative.

Joe takes the test; the result is positive.

Let h_1 be the hypothesis that Joe is actually HIV positive and h_2 be the hypothesis that Joe is actually HIV negative. Which of these is the maximum *a posteriori* hypothesis h_{MAP} ,where

$$h_{MAP} = \arg\max_{h \in H} P(data \mid h)P(h)$$

Given the information above, what is the probability that Joe is HIV positive?

Multiplication (or "chain") rule

$$P(X_1, X_2, ..., X_n) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1, X_2)...P(X_3 \mid X_1, X_2, ..., X_{n-1})$$

(note error in eq. 5 in reading)

Independence and Conditional Independence

- Two random variables, X and Y, are independent if

$$P(X,Y) = P(X)P(Y)$$

- Two random variables, X and Y, are independent *given* Z if

$$P(X,Y \mid C) = P(X \mid C)P(Y \mid C)$$

- Examples?

Classification

- General description of task:
  - Given a *feature vector*, $\boldsymbol{x} = \langle x_1, x_2, \ldots, x_n \rangle$, representing a possible instance of a class, classify x as a positive (1) or negative (0) example.
- General description of learning problem:
  - Given a set of *training examples* $(\boldsymbol{x}, c(\boldsymbol{x}))$, where $c(\boldsymbol{x})$ is the correct classification of $\boldsymbol{x}$, construct a hypothesis that will correctly classify the training examples.

Example: Detecting spam

```
From: Alibris <books@alibris.m0.net>
Reply-to: books@alibris.m0.net
To: mm@cse.ogi.edu
Subject: Melanie, reminding you to save $10 at Alibris

HOLIDAY SPECIAL: SAVE UP TO $10 ON YOUR PURCHASES
(order now and receive by Christmas)


With the holiday season rapidly approaching, we want to remind
you of our most generous sale of the year. As a valued
customer, we invite you to save up to $10 off your Alibris
purchases with three ways to save:

$2 off your order of $20 or more: GIFT2
$5 off your order of $50 or more: GIFT5
$10 off your order of $100 or more: GIFT10

Simply enter the coupon codes above* at checkout. But hurry,
this limited time offer expires on December 16, 2003. Visit
Alibris now and save!

Save money on shipping too! Now through December 9, 2003,
every item listed on our site should be delivered to
continental U.S. and Canadian addresses by December 24th via
standard shipping (our lowest cost option) or get FREE
shipping when you order $49 of In Stock books.
Don't delay, start your holiday shopping now.
http://alibris.m0.net/m/S.asp?HB10950943733X2869462X274232X
```

```
From: "Basil Lutz" <0eynsozueb@a-city.de>
Reply-To: "Basil Lutz" <0eynsozueb@a-city.de>
To: <mm@santafe.edu>, <bonabeau@santafe.edu>
Subject: **SPAM 10.70** This tool will make your website more productive hukm

<html>
<head>
<title>hd36 8 ekj 0O9 92O 2                        </title>
<meta http-equiv=3D"Content-Type" content=3D"text/htm; charset=3Diso-8859=
-1">
</head>

<body>
<p><font face=3D"Arial, Helvetica, sans-serif">Can your website answer que=
stions
  in real time 24 hours a day, 7 days a week? Our clients websites do and =
we're
  not talking about some stale FAQ sheet either. Add <a href=3D"http://www=
dreamscaper.co.mn@click.net-click.net.ph/click.php?id=3Ddrcomnm">live
  operator support</a> to your website today and dramatically increase you=
r revenues.</font></p>
<p> </p>
<p> </p>
<p> </p>
<p> </p>
<p> </p>
<p> </p>
<p> </p>
<p> </p>
<p> </p>
<p><a href=3D"http://www.dreamscaper.co.mn@click.net-click.net.ph/click.ph=
p?id=3Ddrcomnx">stop</a>
  sending me emails</p>
</body>
</html>
```

```
From: =?iso-8859-1?q?james=20ken?= <ja_ken2004@yahoo.fr>
Subject: URGENT ASSISTANCE
To: ja_ken2004@yahoo.fr


 FROM:JAMES KEN.

ATTN:

Dear Respectful one,


I know this proposal letter may come to you as a
surprise considering the fact that we have not had any
formal acquaintance before .but all the same I would
want you for the sake of God to give this an immediate
attention in view of the fact that the security of our
live and possession is at stake .

I am Mr JAMES KEN 28 years old from war ravaged SIERRA
LEONE but presently domiciled in Abidjan  Ivory coast
with my sister JANET who is 18 years old .My father Mr
KEN who before his untimely assassination by the
rebels was the Director of SIERRA LEONE Diamond
corporation (SLDC) .He was killed in our government
residential house along side two of my other brothers
,two house maids and one government attached security
guard fortunately for I, younger sister and mother ,we
were on a week end visit to our home town As we got
the news of the tragedy .We immediately managed to ran
into neighbouring Ivory coast for refuge .But
unfortunately .As Fate would have it ,we lost our dear
mother (may soulrest in peace) as a result of what the
Doctor called cardiac arrest .
```

As we were coming into this country ,we had some documents of a deposit of $ 11 700 000 USD (eleven million seven hundred thousand USD) made by my late father in a security and trust company .According to my father, he intended to use this fund for his international business transaction after his tenure in office but was unfortunately murdered .We had located the security company where the money is deposited with the help of an attorney and established ownership .please right now ,with the bitter experiences we had in our country and the war still going on especially in diamond area which incidentally is where we hail from .coupled with the incessant political upheavals and hostilities in this country Ivory coast ,we desire seriously to leave here and live the rest of our life into a more peaceful and politically stable country like yours Hence this proposal and request .We therefore wish you can help us in the following regards :

1)To provide us with a good bank account to transfer the money into.
2)To help us invest the money into a lucrative business .
3)To assist my sister Janet get a college admission to further her education.

Please I know that , this letter may sound strange and incredible to you but the CNN and the BBC African bulletin normally have it as their major news features .Therefore for the sake of God and humanity give an immediate positive consideration and reply to me via our e-mail address. I will willingly agree to any suitable percentage of the money you will propose as your compensation for your assistance with regards to the above .please in view of our sensitive refugee status and as we are still conscious of our father 's enemies .I would like you to give this a highly confidential approach .

Best Regards .
JAMES KEN.

Features/attributes to use for Spam classification?

Spamassassin results

```
X-Spam-Report: ---- Start SpamAssassin results
          6.70 points, 4 required;
          *  0.4 -- BODY: Offers a limited time offer
          *  0.1 -- BODY: Free Offer
          *  0.4 -- BODY: Stop with the offers, coupons, discounts etc!
          *  0.1 -- BODY: HTML font color is red
          *  0.1 -- BODY: Image tag with an ID code to identify you
          *  2.8 -- BODY: Bayesian classifier says spam probability is 80 to 90%
                     [score: 0.8204]
          *  0.8 -- BODY: HTML font color is green
          *  0.3 -- BODY: FONT Size +2 and up or 3 and up
          *  0.1 -- BODY: HTML font color not within safe 6x6x6 palette
          *  0.1 -- BODY: HTML font color is blue
          *  0.3 -- BODY: Message is 70% to 80% HTML
          *  1.2 -- Date: is 6 to 12 hours after Received: date
          ---- End of SpamAssassin results
```

Spamassassin results

```
X-Spam-Report: ---- Start SpamAssassin results
          10.70 points, 4 required;
          *  0.4 -- BODY: Message is 40% to 50% HTML
          *  1.0 -- URI: URL contains username and (optional) password
          *  0.8 -- URI: Uses a username in a URL
          *  1.2 -- RBL: Received via a relay in dnsbl.njabl.org
                    [RBL check: found 78.199.241.24.dnsbl.njabl.org.,]
                    [type: 127.0.0.9]
          *  4.3 -- RBL: Received via a relay in list.dsbl.org
                    [RBL check: found 78.199.241.24.list.dsbl.org.]
          *  0.1 -- Message has X-MSMail-Priority, but no X-MimeOLE
          *  0.1 -- Message only has text/html MIME parts
          *  2.8 -- Forged mail pretending to be from MS Outlook IMO
          ---- End of SpamAssassin results
```

Naive Bayes Classifier

Let $f(\mathbf{x})$ be a target function for classification: $f(\mathbf{x}) \in \{+, -\}$.

Let $\mathbf{x} = \langle a_1, a_2, ..., a_n \rangle$

We want to find the most probable hypothesis given the data $\mathbf{x}$:

$$h_{MAP} \text{ (maximum } a \textit{ posteriori} \text{ hypothesis)}$$

$$= \underset{class \in \{+,-\}}{\text{argmax}} \; P(class \mid \mathbf{x})$$

$$= \underset{class \in \{+,-\}}{\text{argmax}} \; P(class \mid a_1, a_2, ..., a_n)$$

By Bayes Theorem:

$$h_{MAP} = \underset{class \in \{+,-\}}{\text{argmax}} \frac{P(a_1, a_2, ..., a_n \mid class) P(class)}{P(a_1, a_2, ..., a_n)}$$

$$= \underset{class \in \{+,-\}}{\text{argmax}} P(a_1, a_2, ..., a_n \mid class) P(class)$$

$P(class)$ can be estimated from the training data. How?

How about estimating $P(a_1, a_2, ..., a_n \mid class)$ from training data? Would this work?

- Naive Bayes classifier: Assume each attribute is conditionally independent, given *class*.

$$P(a_1, a_2, ..., a_n \mid class) = P(a_1 \mid class) P(a_2 \mid class) \cdots P(a_n \mid class)$$

Given this assumption, here's how to classify an instance $\boldsymbol{x} = \langle a_1, a_2, ..., a_n \rangle$:

For each a_i, estimate $P(a_i \mid class)$ from training data.

For new instance **x:**

$$h_{NB} = \underset{class \in \{+,-\}}{\text{argmax}} P(class) \prod_i P(a_i \mid class)$$

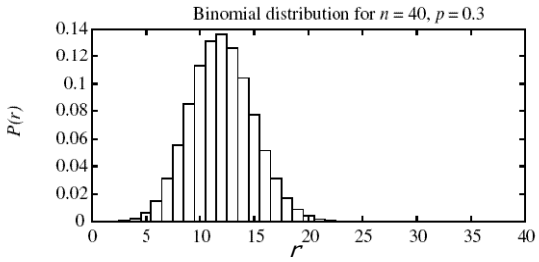
Example

Suppose you have a binary classification problem in which instances **x** have three attributes: **x** = (a_1, a_2, a_3), where $a_i \in \{-,+\}$. Given the training set below, show how a naive Bayes classification algorithm would classify the new instance **x** = (0, 0, 0).

Training set:

$\mathbf{x_1}$ = (1, 0, 0), class = +
$\mathbf{x_2}$ = (0, 1, 1), class = +
$\mathbf{x_3}$ = (1, 1, 0), class = −
$\mathbf{x_4}$ = (0, 0, 1), class = −
$\mathbf{x_5}$ = (1, 1, 1), class = +

Binomial distribution



- *n* identical trials
- Outcome of trial can be one of two possible values, say 0 and 1
- Probability of 1 on a single trial is given by constant *p*.
- Trials are independent of one another.

Details of Binomial Distribution

Suppose n independent experiments yield
$Z_1, Z_2, \ldots Z_n \in \{0,1\}$.

Let

$$R = \sum_{i \text{ such that } Z_i = 1} Z_i$$

Then:

$$\Pr(R = r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

- Expected value of random variable Y that takes on possible values $y_1, y_2, \ldots, y_n$:

$$E[Y] = \sum_{i=1}^{n} y_i \Pr(Y = y_i)$$

- If Y follows a binomial distribution, then

$$E[Y] = np \qquad \text{(why?)}$$

$$E[Y] = \sum_{i=1}^{n} (1 * p) + 0(1-p) = np$$

Variance of Y:

$$Var[Y] = E[(Y - E[Y])^2] = E(Y^2) - (E(Y))^2$$

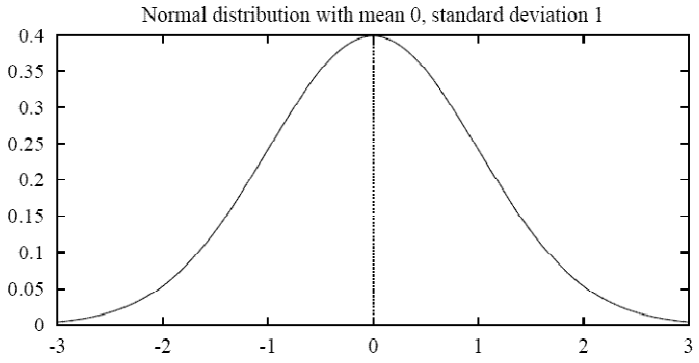
That is, variance = expected squared error in using a single sample of Y to estimate mean $E[Y]$.

If Y is binomially distributed,

$$Var[Y] = np(1-p)$$

(see any statistics book for derivation).

Central Limit Theorem: Sum of a large number of independent, identically distributed (iid) random variables follows a distribution that is approximately Normal (*Gaussian*).



For large n, any binomial distribution is closely approximated by a Normal distribution with the same mean and variance.

Normal distribution

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

for $x \in (-\infty, +\infty)$