Probability Theory Primer plus Naïve Bayes all ways

James G. Shanahan [1,2]

[1]Church and Duncan Group,

[2]School of Informatics, Computing and Engineering, Indiana University

EMAIL: James_DOT_Shanahan_AT_gmail_DOT_com

References and Resources

- **Manning, Raghavan, Schutz, IRBook**
- **Sebastian Raschka - Naive Bayes & Text Classification**
  - http://sebastianraschka.com/Articles/2014_naive_bayes_1.html
  - https://onlinecourses.science.psu.edu/stat200/node/42

Probability Basics → Naïve Bayes Models

- **Introduction**
- **Probability Basics**
  - Probability Axioms
  - Conditional probabilities
  - Product Rule, Chain Rule, Bayes Rule
- **Bayes Nets And Naïve Bayes**
  - Learning
  - Independence
  - Conditional independence
  - Naïve Bayes derivation (discrete case plus smoothing)
- **Naïve Bayes Flavors**
  - Continuous input variables
  - Discrete input variables (2 flavors: Bernoulli, multinomial)
- **Case Study: Spam detector in Naïve Bayes**
- **Summary**

0-1 Loss function for Naïve Bayes

- The 0-1 loss function penalizes misclassification, i.e. it assigns the smallest loss to the solution that has greatest number of correct classifications. So in both cases we are talking about estimating mode. Recall that mode is the *most common value* in the dataset, or the *most probable value*, so both maximizing the posterior probability and minimizing the 0-1 loss leads to estimating the mode.
- If you need a formal proof, the one is given in the *Introduction to Bayesian Decision Theory* paper by Angela J. Yu:

https://stats.stackexchange.com/questions/296014/why-is-the-naive-bayes-classifier-optimal-for-0-1-loss

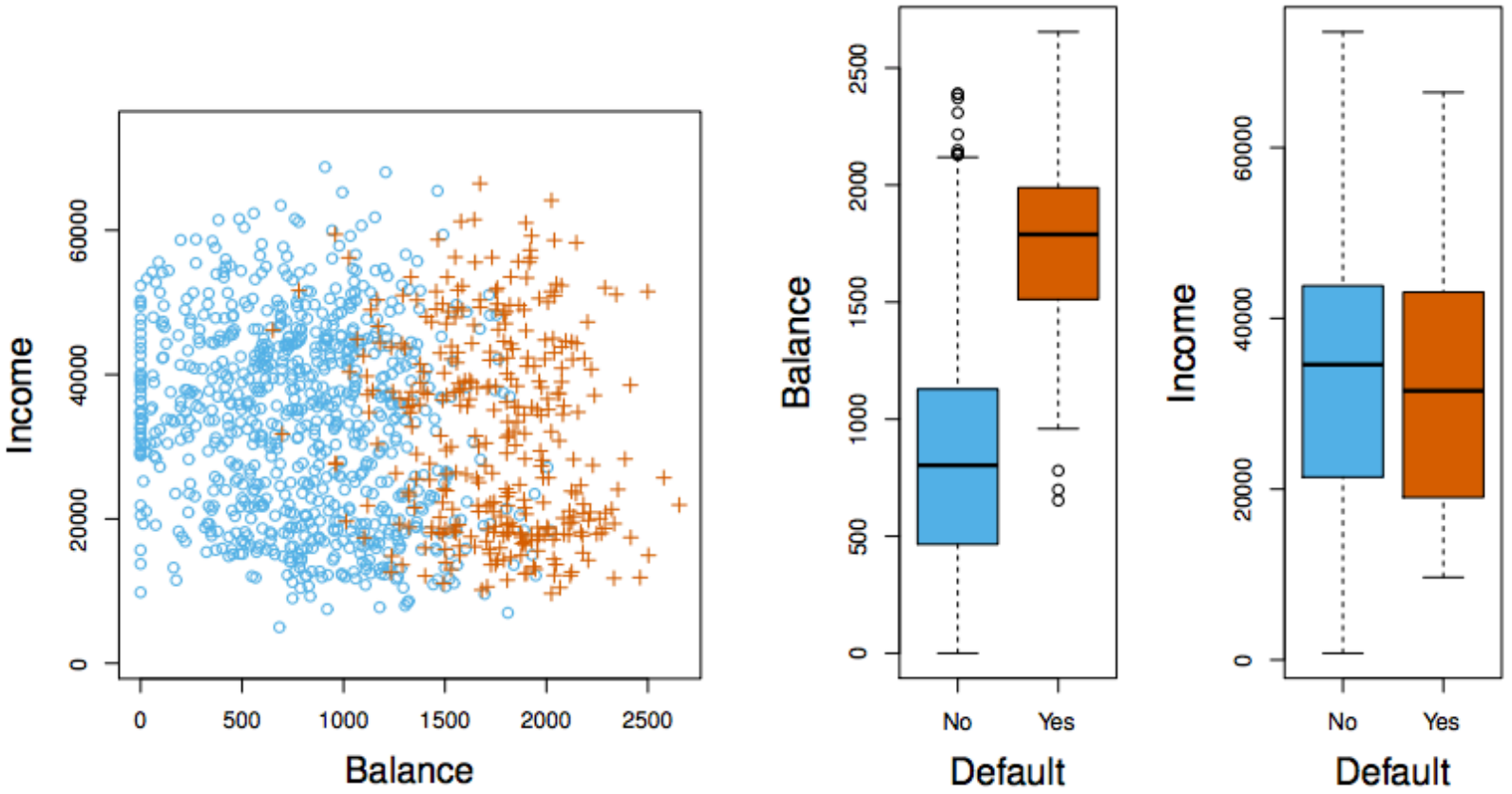
Supervised ML: Build Classifiers

- Qualitative variables take values in an unordered set $\mathcal{C}$, such as:
  $\texttt{eye color} \in \{\texttt{brown}, \texttt{blue}, \texttt{green}\}$
  $\texttt{email} \in \{\texttt{spam}, \texttt{ham}\}$.
- Given a feature vector X and a qualitative response Y taking values in the set $\mathcal{C}$, the classification task is to build a function $C(X)$ that takes as input the feature vector X and predicts its value for Y; i.e. $C(X) \in \mathcal{C}$.
- Often we are more interested in estimating the *probabilities* that X belongs to each category in $\mathcal{C}$.

For example the SPAMiness of an email

Exploratory Data Analysis

Example: Credit Card Defualt



From complex class conditional joint probability distributions of the order 2^n to n

- **From complex class conditional joint probability distributions of the order 2^n to n**

- **2^n Possibilities -> n combinations that we have to estimate class conditional probabilities for**

Summary

- Bayesian prediction:
  - requires solving density estimation problems.
  - often difficult to estimate $\Pr[\mathbf{x} \mid y]$ for $\mathbf{x} \in \mathbb{R}^N$.
  - but, simple and easy to apply; widely used.
- Naive Bayes:
  - strong assumption.
  - straightforward estimation problem.
  - specific linear classifier.
  - sometimes surprisingly good performance.

Decompose Bayesian Networks

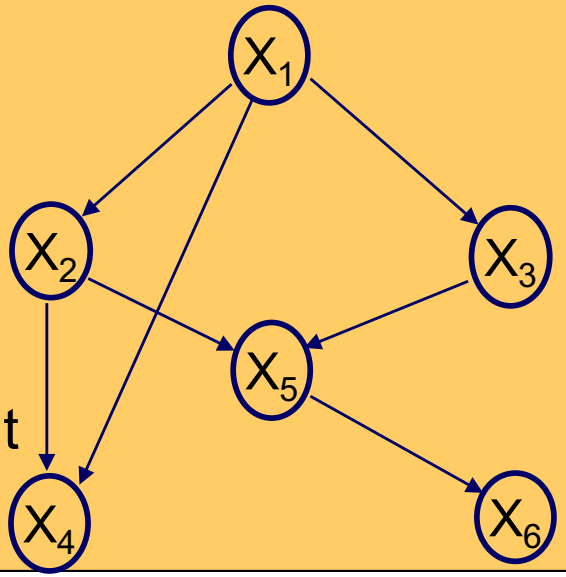
GOAL: 2^N Possibilities -> $1N$

P: Joint Probability Distribution

#	X1	X2	X3	X4	X5	X6	Pr(X1,X2, X3, X4, X5, X6)
1	1						
2	1						
..	1						
4..	1						
5	1						
6	1						
7	1						
64							

Factorize the joint

G: Directed Acyclic Graph



$p(x_1, x_2, x_3, x_4, x_5, x_6)$ **1. Partial Order**

2. By Chain Rule

$= p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1, x_2)\, p(x_4 \mid x_1, x_2, x_3)\, p(x_5 \mid x_4,x_3,x_2,x_1)\, p(x_6 \mid x_5,x_4,x_3,x_2,x_1)$

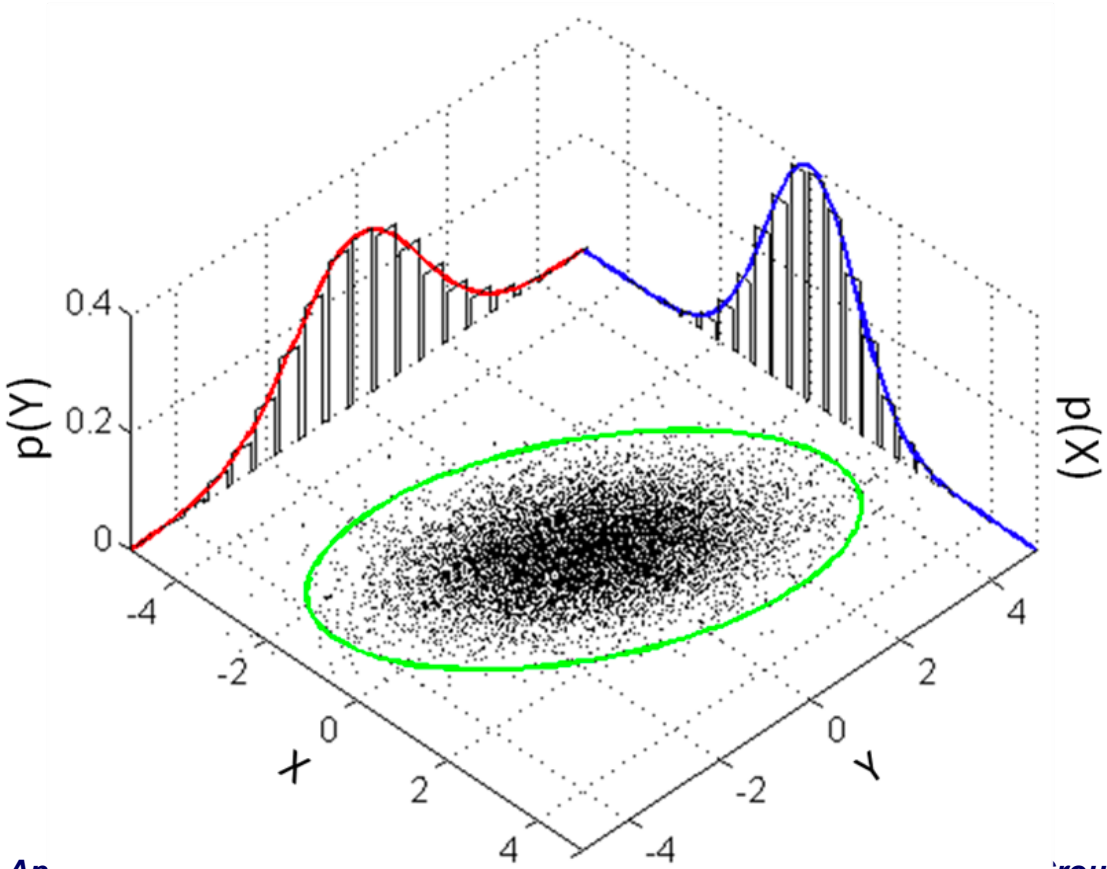
$= p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1)\, p(x_4 \mid x_2, x_1)\, p(x_5 \mid x_3, x_2)\, p(x_6 \mid x_5)$ **3. Markov Property**

$= p(x_1)\, p(x_2)\, p(x_3)\, p(x_4)\, p(x_5)\, p(x_6)$ **4: Independence (see next section)**

$P(y|x_6, x_5,x_4,x_3,x_2,x_1) = p(x_1|y)\, p(x_2|y)\, p(x_3|y)\, p(x_4|y)\, p(x_5|y)\, p(x_6|y)$ **5: Naïve Bayes via Cond. Independence**

Decompose large joint distributions

- **Decompose large joint distributions of many variables into production many univariate distributions**



Probability Basics → Naïve Bayes Models

- **Introduction**
- **Probability Basics**
  - Probability Axioms
  - Conditional probabilities
  - Product Rule, Chain Rule, Bayes Rule
- **Bayes Nets And Naïve Bayes**
  - Learning
  - Independence
  - Conditional independence
  - Naïve Bayes derivation (discrete case plus smoothing)
- **Naïve Bayes Flavors**
  - Continuous input variables
  - Discrete input variables (2 flavors: Bernoulli, multinomial)
- **Case Study: Spam detector in Naïve Bayes**
- **Summary**

Probability theory 1/3

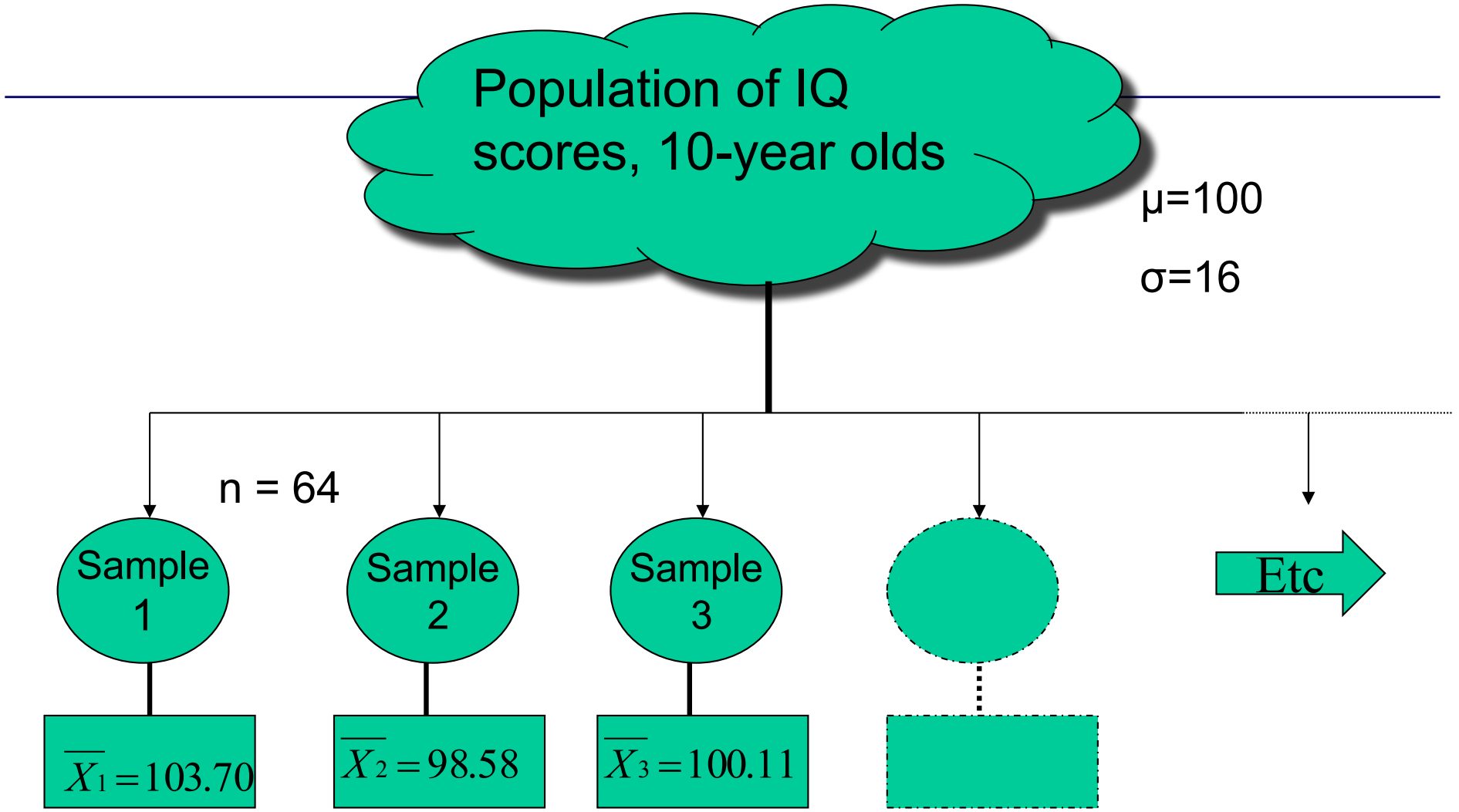
- **Probability theory is the branch of mathematics concerned with probability, the analysis of random phenomena.**
- **The central objects of probability theory are random variables, stochastic processes, and events:**
  - mathematical abstractions of non-deterministic events or measured quantities that may either be single occurrences or evolve over time in an apparently random fashion.

Probability theory 2/3

- **It is not possible to predict precisely results of random events.**
- **However, if a sequence of individual events, such as coin flipping or the roll of dice, is influenced by other factors, such as friction, it will exhibit certain patterns, which can be studied and predicted.**
- **Two representative mathematical results describing such patterns are the**
  - (1) law of large numbers (LoLN) and the (see next slide)
  - (2) central limit theorem. (central tendencies) (see next slide)

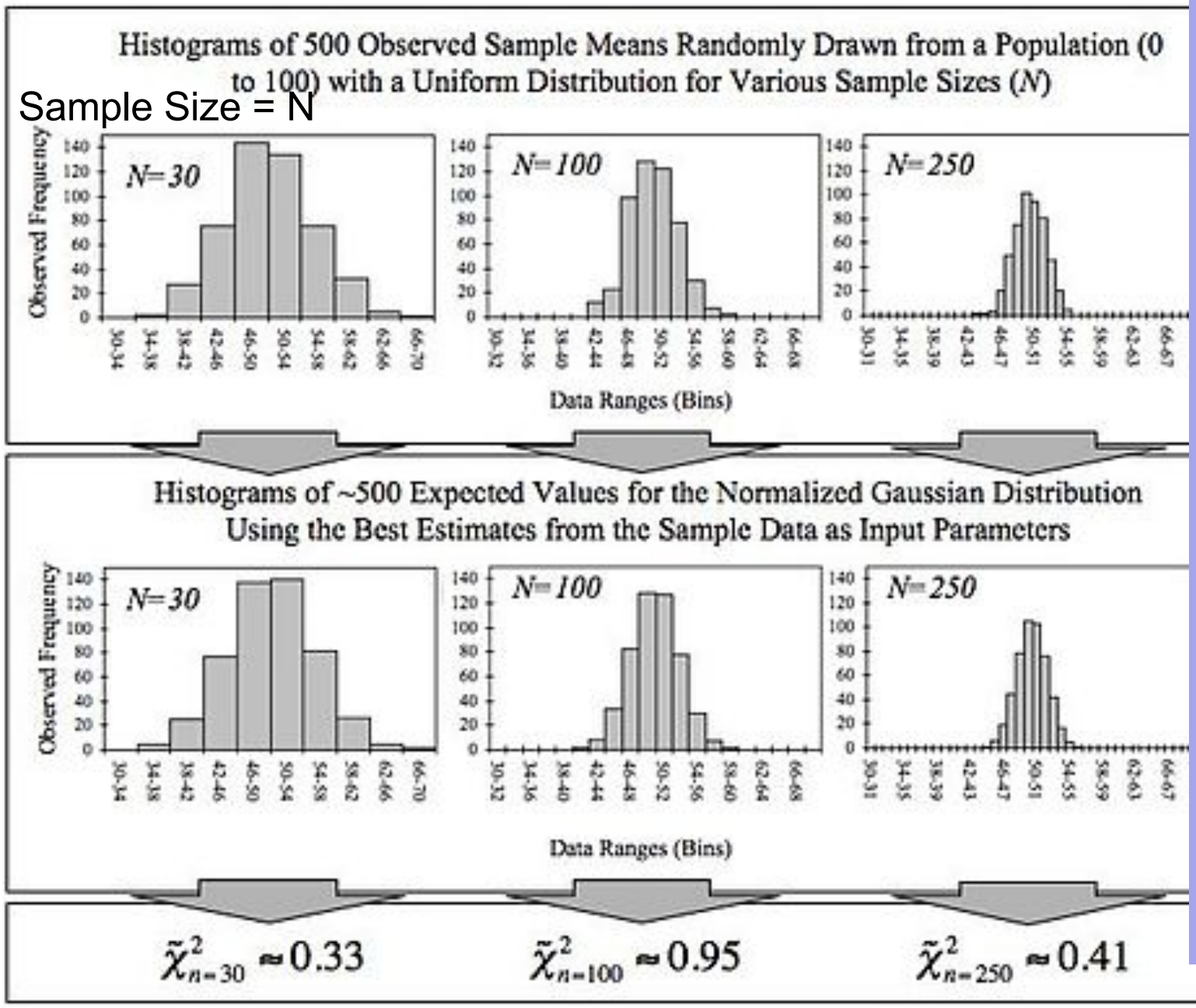
CLT and LoLN

- **Both the law of large numbers and central limit theorem are about many independent samples from same distribution**
- **CLT is about shape: shape is bell-like the more samples we draw**
  - The Central Limit Theorem tell us that as the sample size tends to infinity, the of the distribution of sample means approaches the normal distribution.
  - This is a statement about the SHAPE of the distribution. A normal distribution is bell shaped so the shape of the distribution of sample means begins to look bell shaped as the sample size increases.
- **The Law of Large Numbers tells us where the center (maximum point) of the bell is located.**
  - Again, as the sample size approaches infinity the center of the distribution of the sample means becomes very close to the population mean.



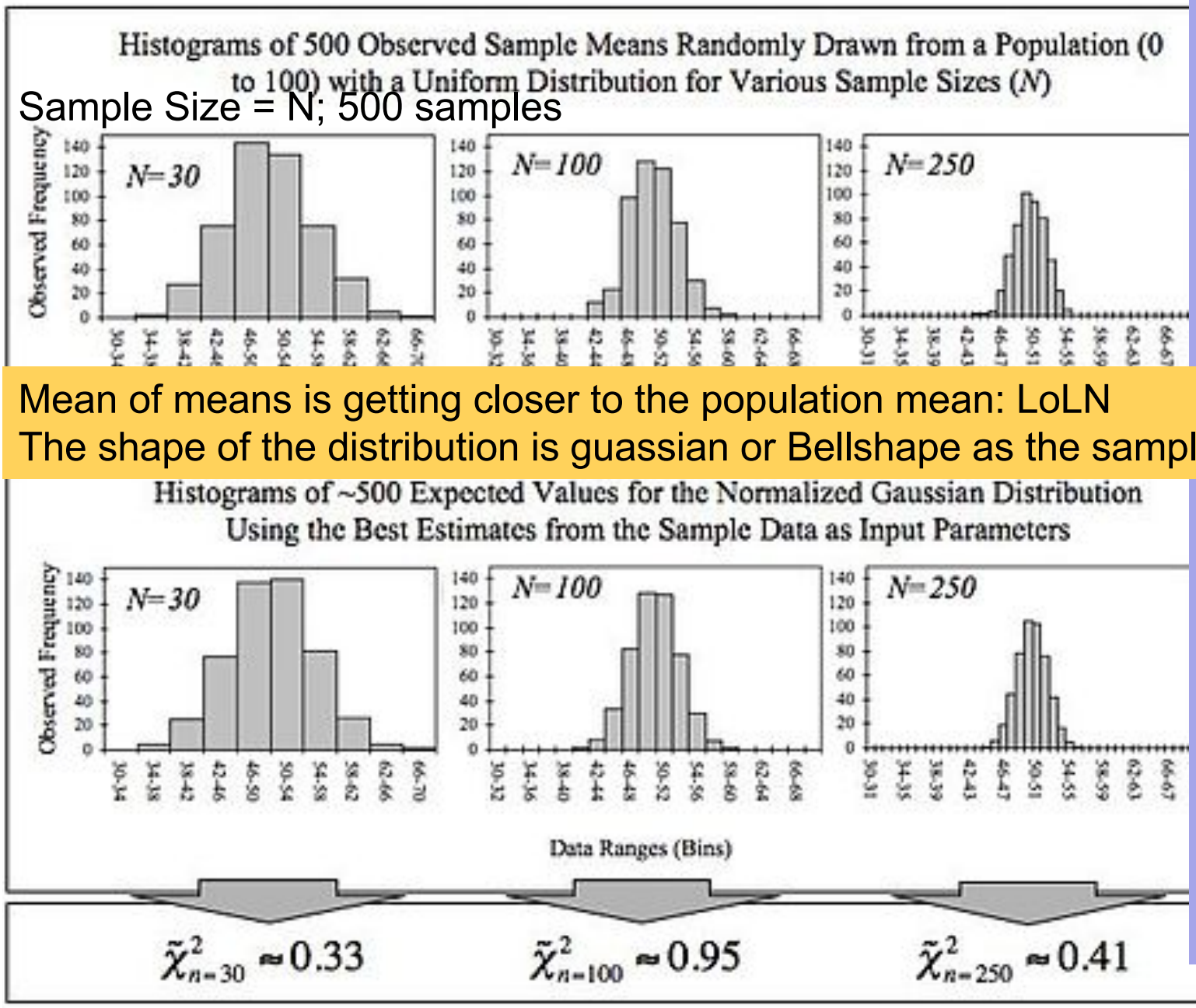
Is sample 2 a likely representation of our population?

Uniform distribution [0,100], Pop Mean is 50. vs. Gaussian based population



This figure demonstrates the central limit theorem. The sample means are generated using a random number generator, which draws numbers between 0 and 10 from a uniform probability distribution. It illustrates that increasing sample sizes result in the 500 measured sample means being more closely distributed about the population mean (50 in this case). It also compares the observed distributions with the distributions that would be expected for a normalized Gaussian distribution, and shows the chi-squared values that quantify the goodness of the fit (th fit is good if the reduced chi-squared value is less than or approximately equal to one). The input into the normalized Gaussia function is the mean of sample means (~50) and the mean samp standard deviation divided by the square root of the sample size (~28.87/√n), which is called the standard deviation of the mean (since it refers to the spread of sample means).

Uniform distribution [0,100], Pop Mean is 50. vs. Gaussian based population



Sample Size = N; 500 samples

Mean of means is getting closer to the population mean: LoLN

The shape of the distribution is guassian or Bellshape as the sample size N increases

This figure demonstrates the central limit theorem. The sample means are generated using a random number generator, which draws numbers between 0 and 10 from a uniform probability distribution. It illustrates that increasing sample sizes result in the 500 measured sample means being more closely distributed about the population mean (50 in this case). It also compares the Gaussian distribution, and shows the chi-squared values that quantify the goodness of the fit (th fit is good if the reduced chi-squared value is less than or approximately equal to one). The input into the normalized Gaussia function is the mean of sample means (~50) and the mean samp standard deviation divided by the square root of the sample size (~28.87/√n), which is called the standard deviation of the mean (since it refers to the spread of sample means).

Probability theory 3/3: many apps

- **As a mathematical foundation for statistics, probability theory is essential to many human activities that involve quantitative analysis of large sets of data.**
- **Many applications**
  - Machine learning
  - statistical mechanics: Methods of probability theory also apply to descriptions of complex systems given only partial knowledge of their state, as in statistical mechanics.
  - quantum mechanics: A great discovery of twentieth century physics was the probabilistic nature of physical phenomena at atomic scales, described in quantum mechanics.

law of large numbers

- In probability theory, the law of large numbers (LLN) is a theorem that describes the result of performing the same experiment a large number of times.
- According to the law, the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.

CLT: mean will be approximately normally distributed

- In probability theory, the central limit theorem (CLT) states that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution.[1][2]

- To illustrate what this means, suppose that a sample is obtained containing a large number of observations, each observation being randomly generated in a way that does not depend on the values of the other observations, and that the arithmetic average of the observed values is computed.
- If this procedure is performed many times, the central limit theorem says that the computed values of the average will be distributed according to the normal distribution (commonly known as a "bell curve").
- A simple example of this is that if one flips a coin many times, the probability of getting a given number of heads should follow a normal curve, with mean equal to half the total number of flips.

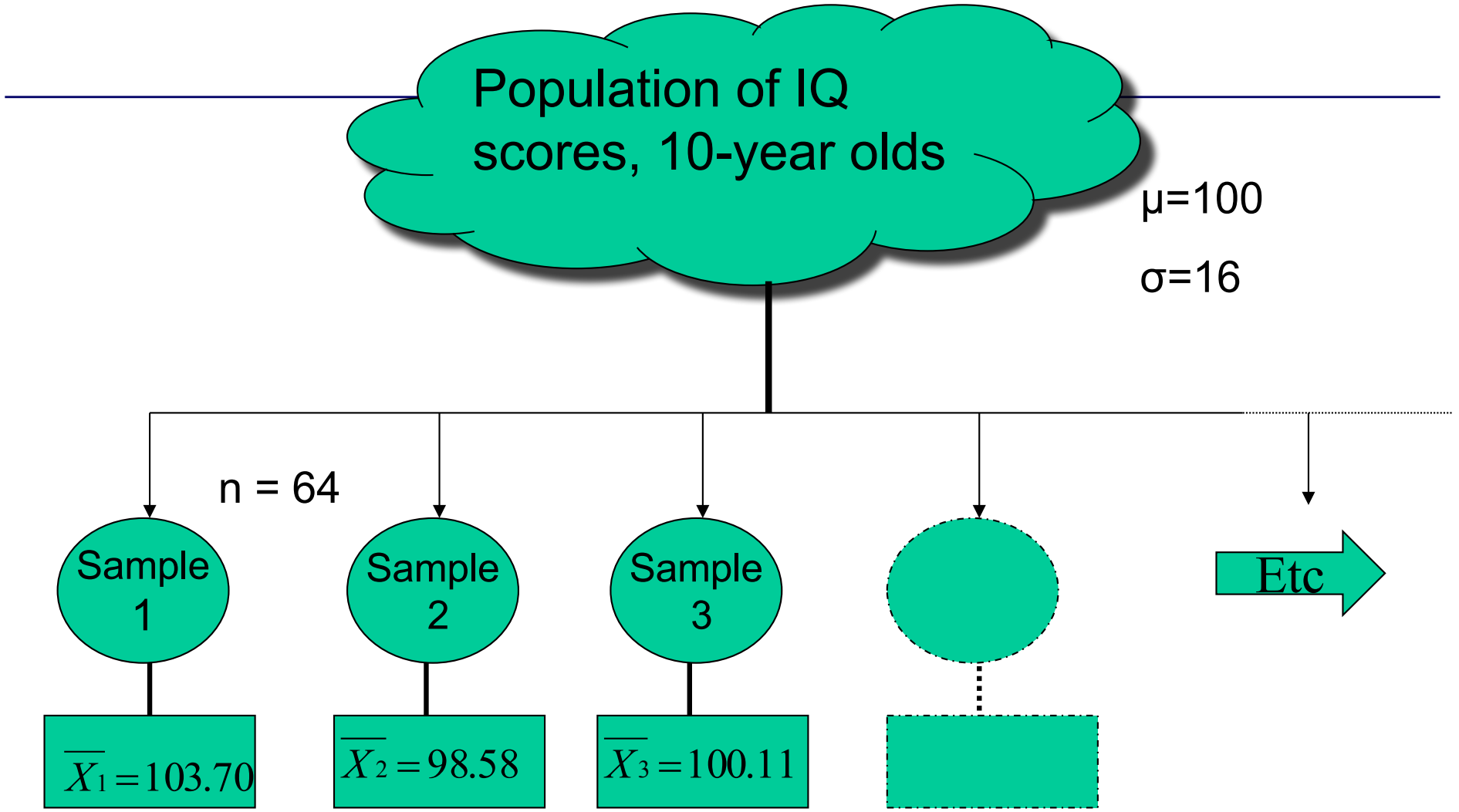
Sampling Distributions of Sample Statistics

- **Sample Statistics**
  - Two common statistics are the sample proportion, $\hat{p}$ ("p-hat"), and the sample mean, $\bar{x}$ ("x-bar").
  - Sample statistics are random variables because they vary from sample to sample.
- **Sampling distribution**
  - As a result, sample statistics also have a distribution called the sampling distribution.
  - These sampling distributions, similar to distributions discussed previously, have a mean and standard deviation.
- **Standard error**
  - We refer to the standard deviation of a sampling distribution as the standard error. Thus, the standard error is simply the standard deviation of a sampling distribution. Often times statisticians will interchange these two terms

https://onlinecourses.science.psu.edu/stat200/node/42

Standard Error

- The standard error is the standard deviation of the sampling distribution of a statistic.[1]
- The term may also be used to refer to an estimate of that standard deviation, derived from a particular sample used to compute the estimate.
- For example, the sample mean is the usual estimator of a population mean. However, different samples drawn from that same population would in general have different values of the sample mean. The standard error of the mean (i.e., of using the sample mean as a method of estimating the population mean) is the standard deviation of those sample means over all possible samples (of a given size) drawn from the population. Secondly, the standard error of the mean can refer to an estimate of that standard deviation, computed from the sample of data being analyzed at the time.



Is sample 2 a likely representation of our population?

Distribution of Sample Means

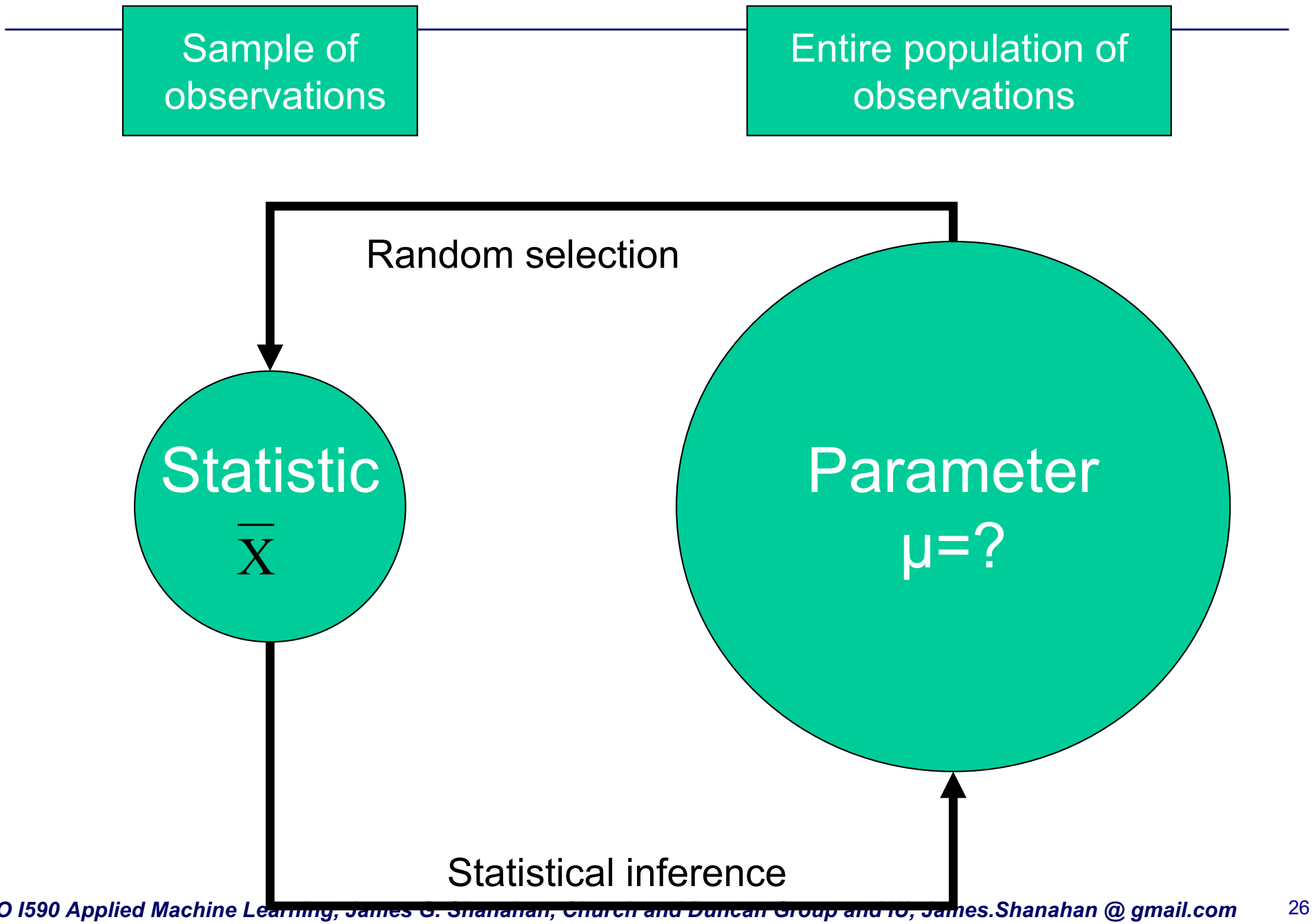
1. The mean of a sampling distribution is identical to mean of raw scores in the population (μ)
2. If the population is Normal, the distribution of sample means is also Normal
3. If the population is not Normal, the distribution of sample means approaches Normal distribution as the size of sample on which it is based gets larger

Standard Error of the Mean

- The standard deviation of means in a sampling distribution is known as the standard error of the mean.
- It can be calculated from the standard deviation of observat

$$S_{\bar{x}} = \frac{s}{\sqrt{n}} \qquad \text{Standard Error of } \bar{X} : \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

3. The larger our sample size, the smaller our standard error


Sample of observations
Entire population of observations
Random selection
Statistic
$\bar{x}$
Parameter
μ=?
Statistical inference

STANDARD ERROR CALCULATION

Procedure:

Step 1: Calculate the mean (Total of all samples divided by the number of samples).

Step 2: Calculate each measurement's deviation from the mean (Mean minus the individual measurement).

Step 3: Square each deviation from mean. Squared negatives become positive.

Step 4: Sum the squared deviations (Add up the numbers from step 3).

Step 5: Divide that sum from step 4 by one less than the sample size (n-1, that is, the number of measurements minus one)

Step 6: Take the square root of the number in step 5. That gives you the "standard deviation (S.D.)."

Step 7: Divide the standard deviation by the square root of the sample size (n). That gives you the "standard error".

Step 8: Subtract the standard error from the mean and record that number. Then add the standard error to the mean and record that number. You have plotted mean± 1 standard error (S. E.), the distance from 1 standard error below the mean to 1 standard error above the mean

Draw a sample of size N from the population.

In the SE calculation N is the sample size (not the number of samples drawn)

Example:

Name	Height to nearest 0.5 cm	**2** Deviations (m-i)	**3** Squared deviations $(m-i)^2$
1. Waldo	150.5	11.9	141.61
2. Finn	170.0	-7.6	57.76
3. Henry	160.0	2.4	5.76
4. Alfie	161.0	1.4	1.96
5. Shane	170.5	-8.1	65.61
n= 5	**1 Mean m** = 162.4 cm		**4** Sum of squared deviations $\sum(m-i)^2 = 272.70$

5 Divide by number of measurements-1. $\sum (m-i)^2 / (n-1) = 272.70 / 4 = 68.175$

6 **Standard deviation** = square root of $\sum (m-i)^2/n-1 = \sqrt{68.175} = 8.257$

7 **Standard error** = Standard deviation/$\sqrt{n}$ = 8.257/2.236 = 3.69

8 **m ± 1SE** = 162 ± 3.7 or 159cm to 166cm for the men (162.4 - 3.7 to 162.4 + 3.7).

Estimation Procedures

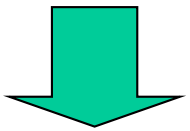
- **Point estimates**
  - For example mean of a sample of 25 patients
    - No information regarding probability of accuracy
  - Interval estimates
  - Estimate a range of values that is likely
    - Confidence interval between two limit values
      - The degree of confidence depends on the probability of including the population mean

$$95\%\ \mathrm{CI} = \bar{X} \pm 1.96\,\sigma_{\bar{x}}$$

$$99\%\ \mathrm{CI} = \bar{X} \pm 2.58\,\sigma_{\bar{x}}$$

When Sample size is small …

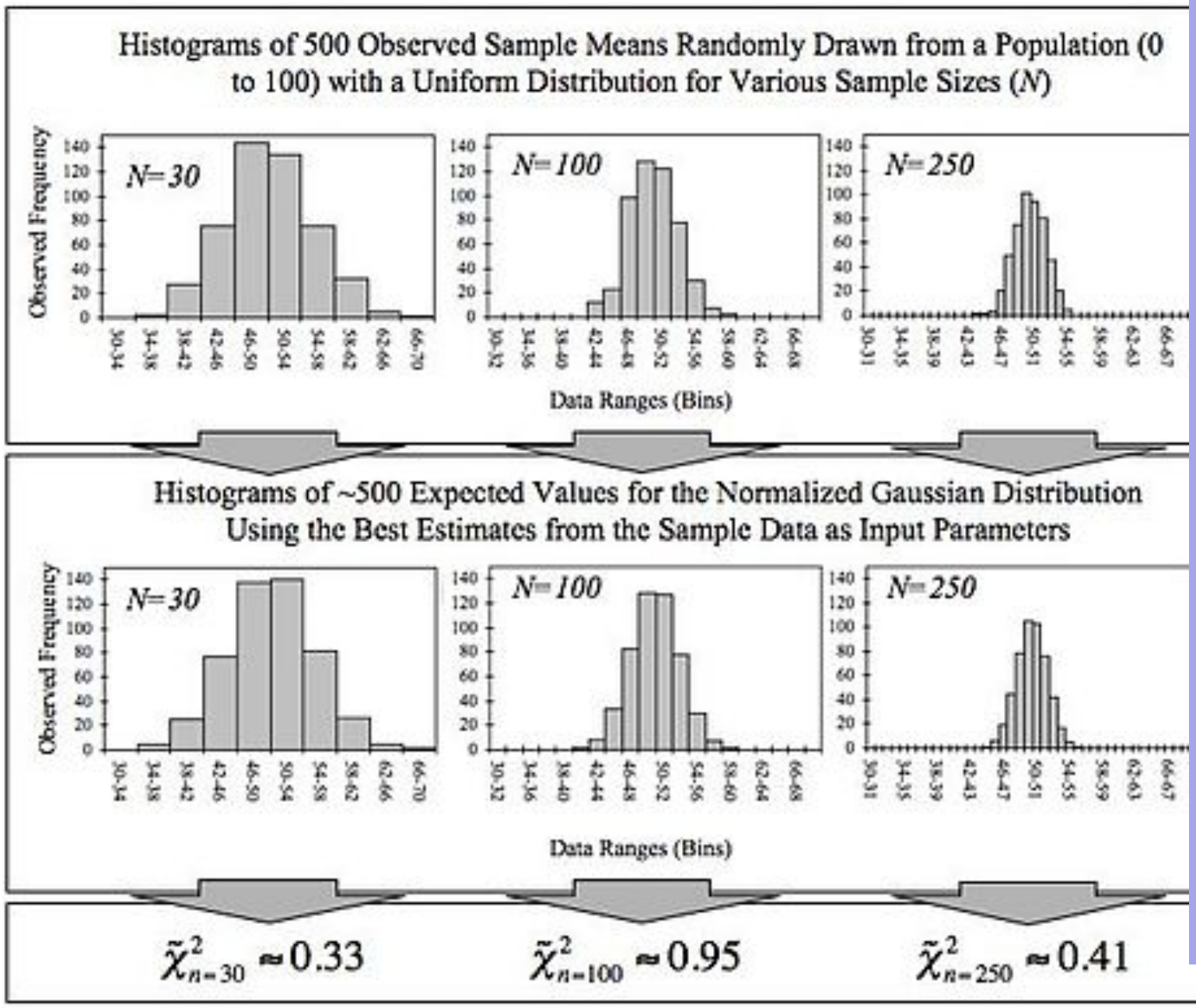
$$95\%\ \mathrm{CI} = \bar{X} \pm 1.96\ \sigma_{\bar{x}}$$



$$95\%\ \mathrm{CI} = \bar{X} \pm t\ S_{\bar{x}}$$

A constant from
Student t Distribution
that depends on confidence
interval and sample size

Uniform distribution [0,100], Pop Mean is 50. vs. Gaussian based population



This figure demonstrates the central limit theorem. The sample means are generated using a random number generator, which draws numbers between 0 and 10 from a uniform probability distribution. It illustrates that increasing sample sizes result in the 500 measured sample means being more closely distributed about the population mean (50 in this case). It also compares the observed distributions with the distributions that would be expected for a normalized Gaussian distribution, and shows the chi-squared values that quantify the goodness of the fit (th fit is good if the reduced chi-squared value is less than or approximately equal to one). The input into the normalized Gaussia function is the mean of sample means (~50) and the mean samp standard deviation divided by the square root of the sample size (~28.87/√n), which is called the standard deviation of the mean (since it refers to the spread of sample means).

Notation

- Proposition - statement or assertion about a state of the world
- Variable X is a set of mutually exclusive propositions x_i
- Variables – upper-case
- Propositions – lowercase
  - **Example (X=x, Y=y, Z=z)**
  - **Shortened: (x,y,z)**
- Sets of variables – bold
  - Example: (**X, Y, Z**)
- Latent/Hidden variable – states are inferred but never observed directly

From Axioms: deduce theorems and propositions

- **Math**
  - One strategy in mathematics is to start with a few statements, then build up more mathematics from these statements.
  - The beginning statements are known as axioms.
    - An axiom is typically something that is mathematically self evident.
    - From a relatively short list of axioms, deductive logic is used to prove other statements, called theorems or propositions.
- **Probability Theory has 3 axioms → theorems or propositions**
  - The area of mathematics known as probability is no different.
  - Underlying probability is a handful of axioms from which we can derive all sorts of results. But what are these probability axioms?
  - Probability can be reduced to three axioms.
  - It presupposes that we have a set of outcomes called the sample space S comprised of subsets called events $E_1, E_2, \ldots, E_n$ and a way of assigning a probability to any event E. The probability of the event E is denoted by P(E).

Axioms of Probabilities

http://statistics.about.com/od/Mathstat/a/what-is-the-power-set.htm

Axiom 1 : $0 \le P(A) \le 1$

Axiom 2: $P(Sure\ Proposition) = 1$

Axiom 3: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Marginal probability: $P(A) = P(A, B) + P(A, \neg B)$

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

The first axiom of probability is that the probability of any event is a nonnegative real number. This means that the smallest that a probability can ever be is zero, and that it cannot be infinite.

The third axiom of probability deals with mutually exclusive events. If E1 and E2 are mutually exclusive, meaning that they have an empty intersection and we use U to denote the union, then P(E1 U E2) = P(E1) + P(E2).

Axiom Applications: Pr(an impossible event)

The three axioms set an upper bound for the probability of any event. We denote the complement of the event E by EC. From set theory E and EC have empty intersection and are mutually exclusive. Furthermore E U EC = S, the entire sample space.

These facts, combined with the axioms give us:

$1 = P(S) = P(E \cup E^C) = P(E) + P(E^C)$.

We rearrange the above equation and see that $P(E) = 1 - P(E^C)$. Since we know that probabilities must be nonnegative, we now have that an upper bound for the probability of any event is 1.

By rearranging the formula again we have $P(E^C) = 1 - P(E)$. We also can deduce from this formula that the probability of an event not occurring is one minus the probability that it does occur.

The above equation also provides us a way to calculate the probability of the impossible event, denoted by the empty set. To see this, recall that the empty set is the complement of the universal set, in this case S^C. Since $1 = P(S) + P(S^C) = 1 + P(S^C)$, by algebra we have $P(S^C) = 0$.

Complement space

2nd Axiom: states that the probability of all the events, i.e., the probability of the entire sample space is 1.
Mathematically, if S represents the Sample space, then P(S)=1.
This means that there are no events outside the sample space and it includes all possible events in it.

$$P(A) + P(\neg A) = 1$$

$$P(A \mid B)$$

Pr(A) → P(A|B)

- Belief in A under the assumption that B is known with absolute certainty
- A is conditioned on B

Probability Basics → Naïve Bayes Models

- **Introduction**
- **Probability Basics**
  - Probability Axioms
  - Conditional probabilities
  - Product Rule, Chain Rule, Bayes Rule
- **Bayes Nets And Naïve Bayes**
  - Learning
  - Independence
  - Conditional independence
  - Naïve Bayes derivation (discrete case plus smoothing)
- **Naïve Bayes Flavors**
  - Continuous input variables
  - Discrete input variables (2 flavors: Bernoulli, multinomial)
- **Case Study: Spam detector in Naïve Bayes**
- **Summary**

Sample Space: possible outcomes

- **In probability theory, the sample space of an experiment or random trial is the set of all possible outcomes or results of that experiment.**

A sample space is usually denoted using set notation, and the possible outcomes are listed as elements in the set. It is common to refer to a sample space by the labels S, Ω, or U (for "universal set").

- **E.g.,**
  - For example, if the experiment is tossing a coin, the sample space is typically the set {head, tail}.
  - For tossing two coins, the corresponding sample space would be {(head,head), (head,tail), (tail,head), (tail,tail)}.
  - For tossing a single six-sided die, the typical sample space is {1, 2, 3, 4, 5, 6} (in which the result of interest is the number of pips facing up).

{head, tail}

{(head,head), (head,tail), (tail,head), (tail,tail)}

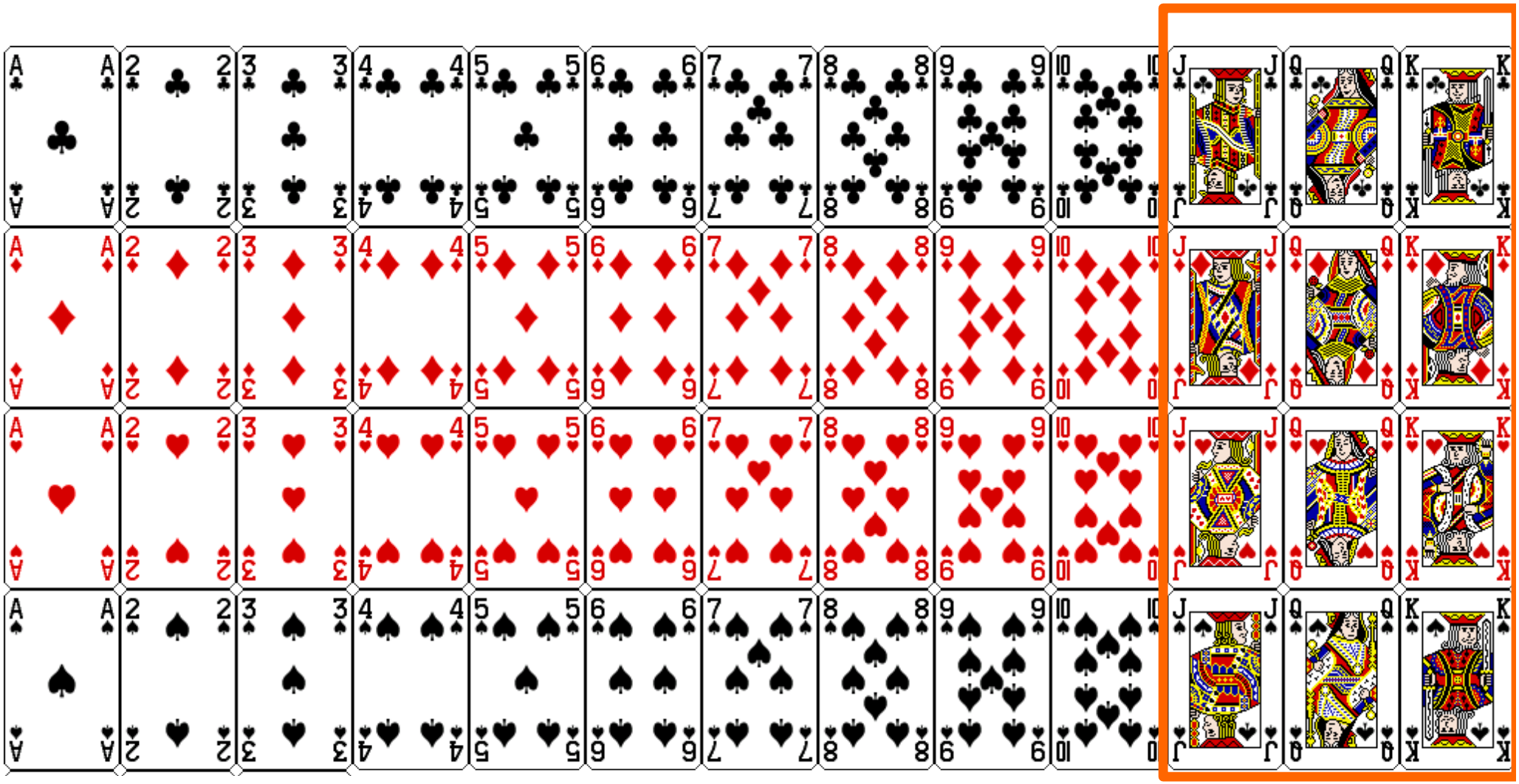
Multiple sample spaces (views)

- **For many experiments, there may be more than one plausible sample space available, depending on what result is of interest to the experimenter.**
- **For example, when drawing a card from a standard deck of fifty-two playing cards,**
  - **Rank sample space (picture cards)**
    - one possibility for the sample space could be the various ranks (Ace through King),
  - **Suits sample space**
    - while another could be the suits (clubs, diamonds, hearts, or spades).
  - Cartesian Sample Space
    - A more complete description of outcomes, however, could specify both the denomination and the suit, and a sample space describing each individual card can be constructed as the Cartesian product of the two sample spaces noted above (this space would contain fifty-two equally likely outcomes).
  - Still other sample spaces are possible, such as {right-side up, up-side down} if some cards have been flipped when shuffling.

PictureCards X Suit = 4 X 4 = 16

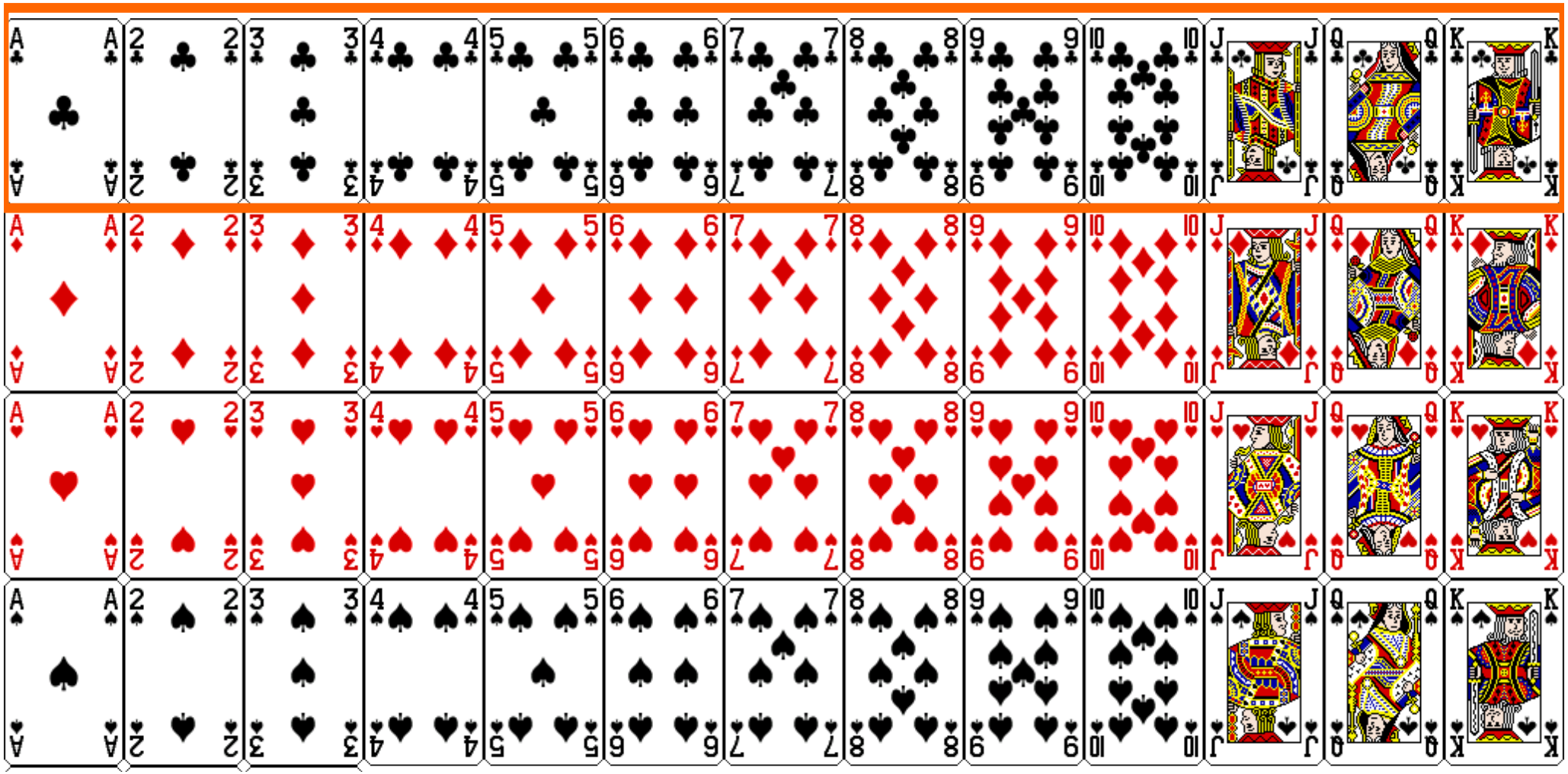
Sample space

Royal Cards = {Jack, Queen, King}



Sample space

Clubs

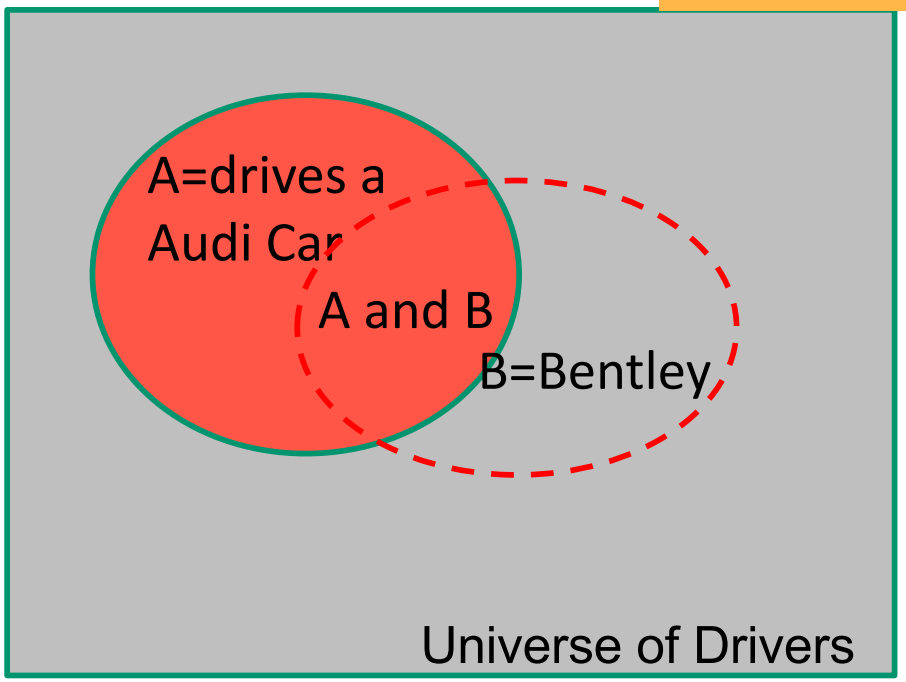


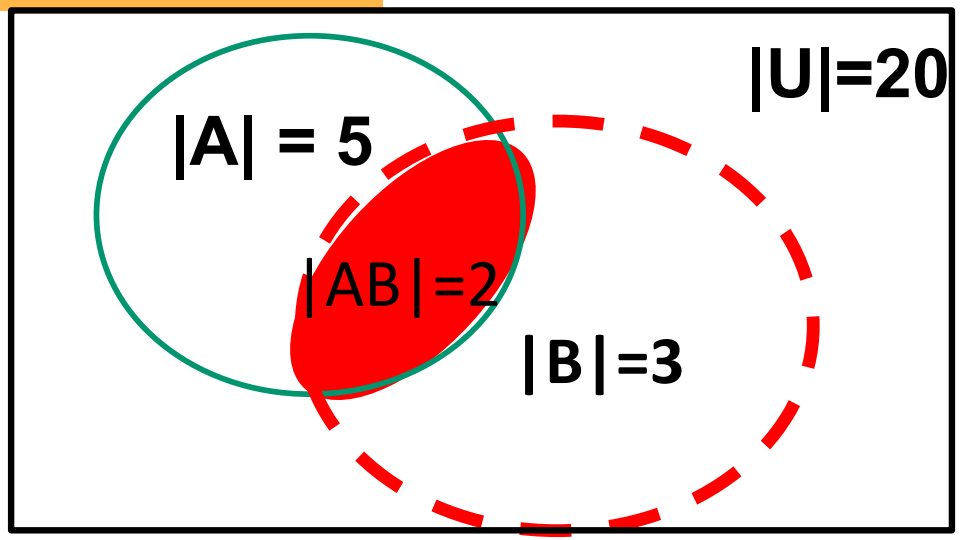
Conditional Probability as probability wrt a reduced sample space

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

Conditional probability can be seen to be the probability with respect to a reduced sample space. We can illustrate the conditional probability with the Venn diagram.

Pr(A) → P(A|B)





A = 5, B = 5, A and B = 2, universe = 20

Pr(A|B) = 2/20 / 5/20 = 2/5

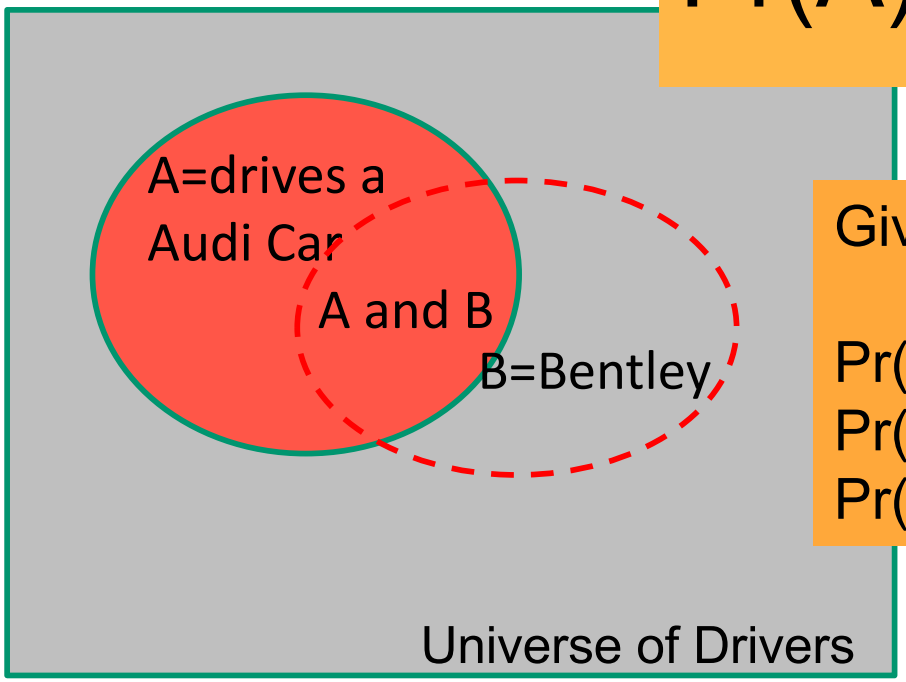
Pr(A|not B) = 3/20 / 15/20 = 3/5

Calculating Conditional Probability

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

Conditional probability can be seen to be the probability with respect to a reduced sample space. We can illustrate the conditional probability with the Venn diagram.

$$\Pr(A) \rightarrow P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2}{5}$$

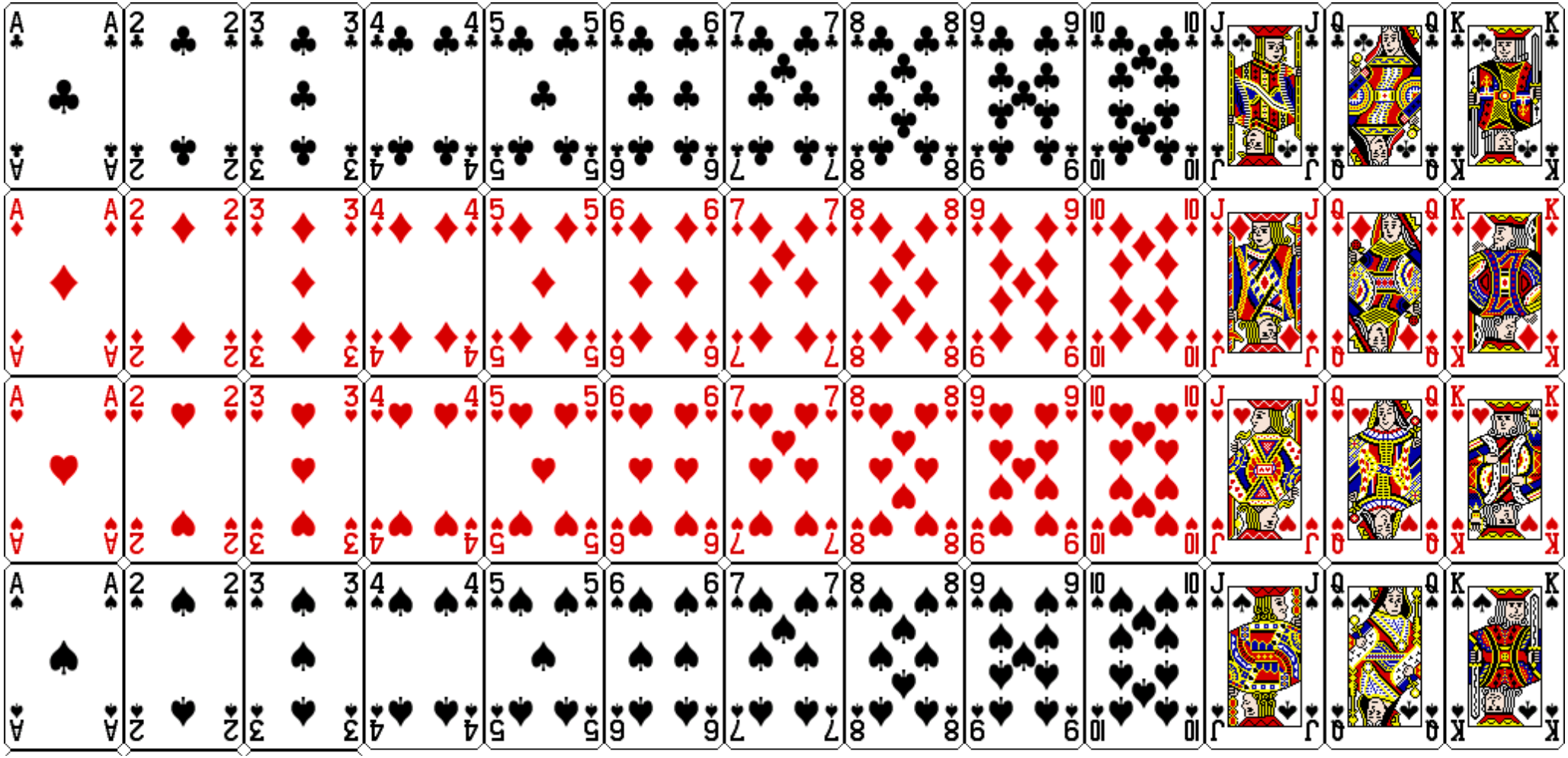


Given: A = 5 B = 5, A and B = 2, universe = 20

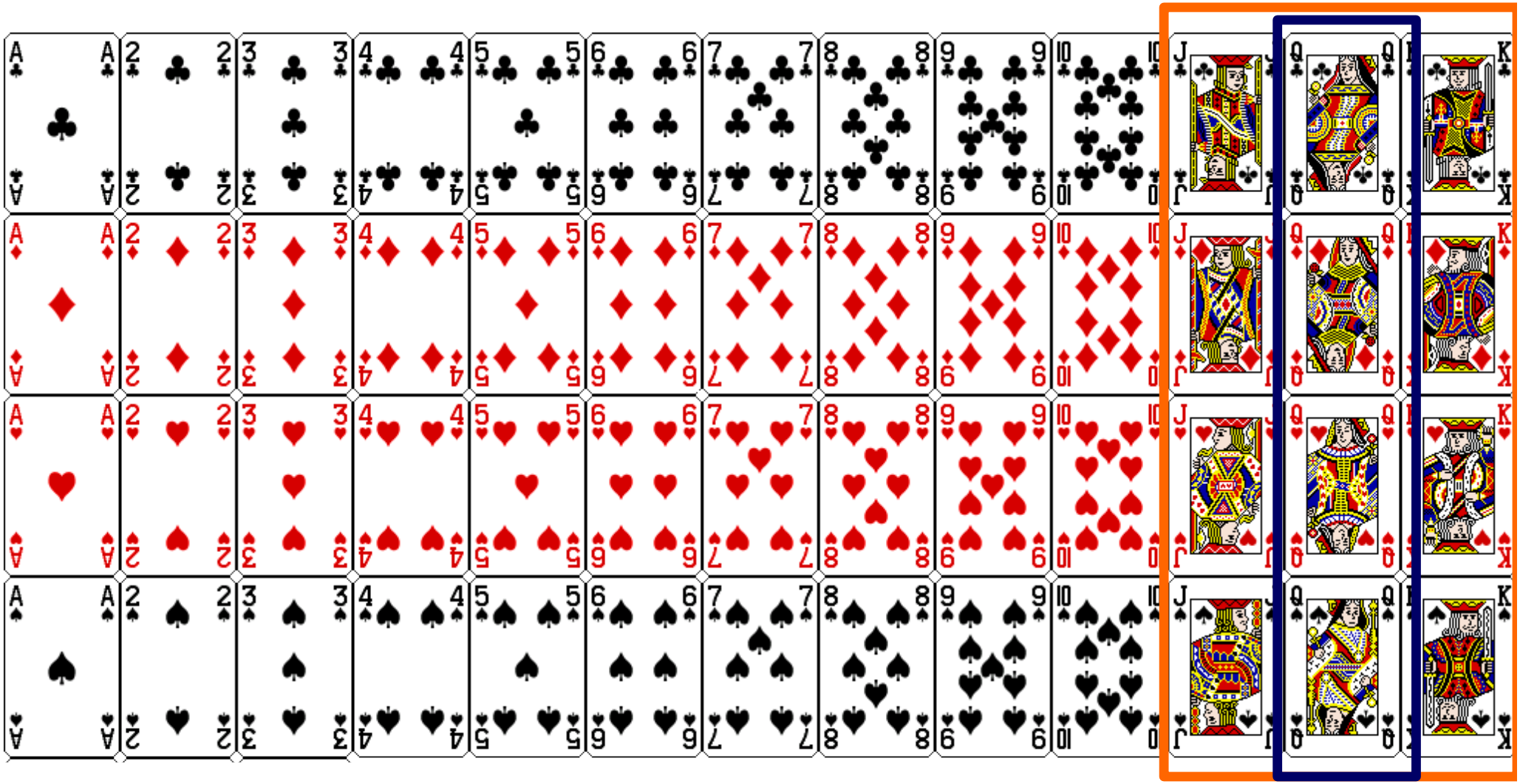
Pr(A|B) = 2/20 / 5/20 = 2/5
Pr(A|not B) = 3/20 / 15/20 = 3/5
Pr(A) = Pr(A|B) + Pr(A|not B) = 2/5 + 3/5 = 1

Conditional probability example

Royal Cards = {Jack, Queen, King}

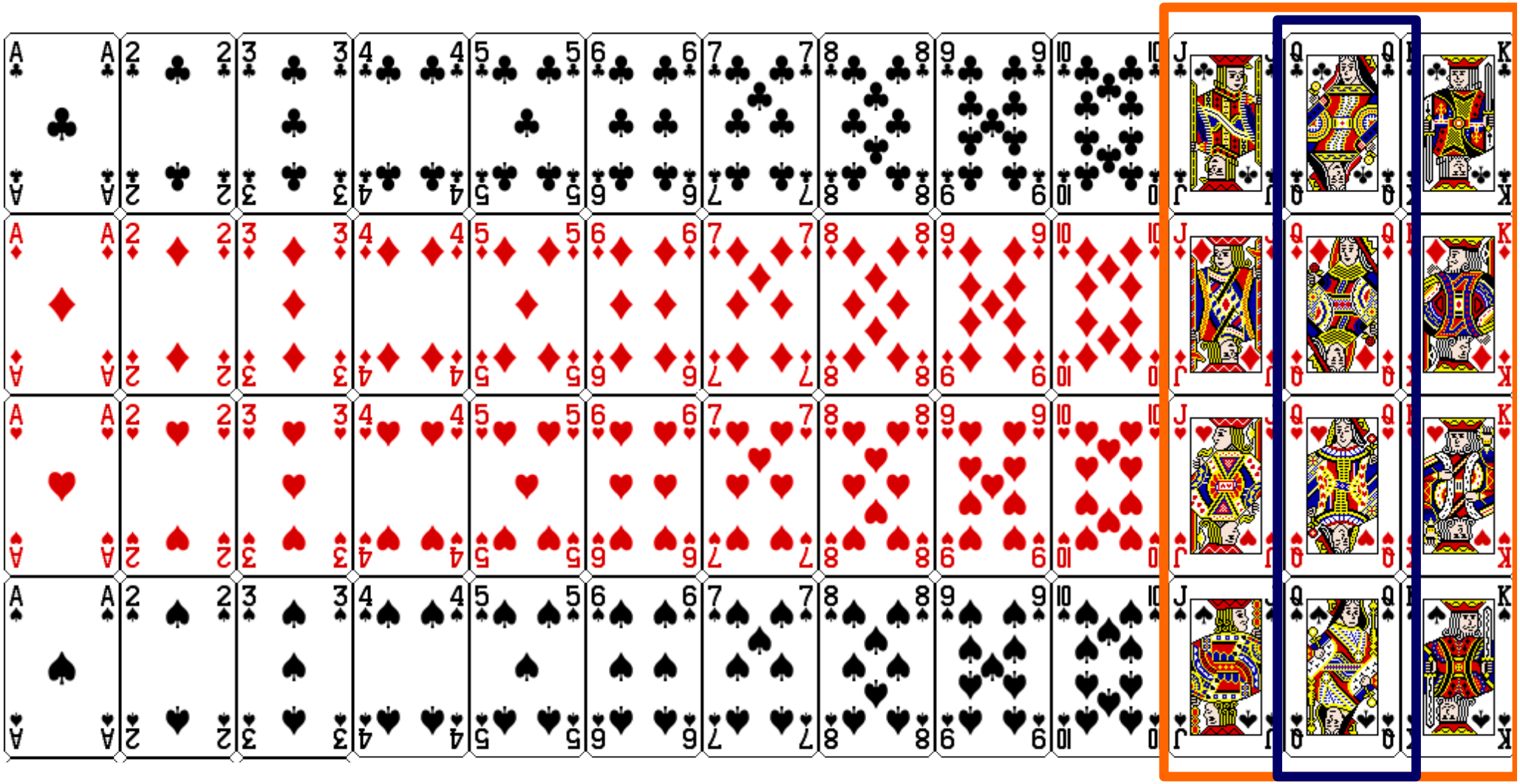


Conditional probability example



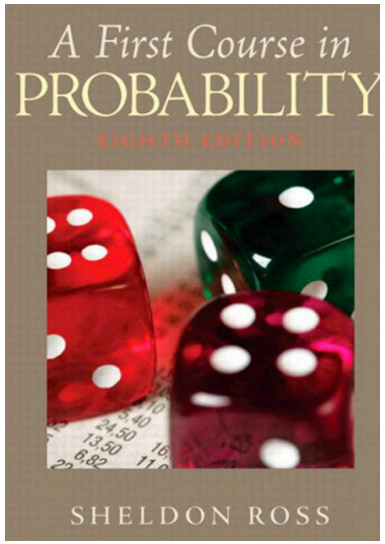
$$P(\text{Queen}|\text{RoyalCard}) = \frac{P(\text{Queen} \cap \text{RoyalCard})}{P(\text{RoyalCard})} = \frac{1/13}{3/13} = \frac{1}{3}$$

Pr(Queen|Club) is independent?



$$P(\text{Queen}|\text{RoyalCard}) = \frac{P(\text{Queen} \cap \text{RoyalCard})}{P(\text{RoyalCard})} = \frac{1/13}{3/13} = \frac{1}{3}$$

Conditional Probabilities Example



3.2 CONDITIONAL PROBABILITIES

Suppose that we toss 2 dice, and suppose that each of the 36 possible outcomes is equally likely to occur and hence has probability $\frac{1}{36}$. Suppose further that we observe that the first die is a 3. Then, given this information, what is the probability that the sum of the 2 dice equals 8? To calculate this probability, we reason as follows: Given that the initial die is a 3, there can be at most 6 possible outcomes of our experiment, namely, (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), and (3, 6). Since each of these outcomes originally had the same probability of occurring, the outcomes should still have equal probabilities. That is, given that the first die is a 3, the (conditional) probability of each of the outcomes (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), and (3, 6) is $\frac{1}{6}$, whereas the (conditional) probability of the other 30 points in the sample space is 0. Hence, the desired probability will be $\frac{1}{6}$.

If we let E and F denote, respectively, the event that the sum of the dice is 8 and the event that the first die is a 3, then the probability just obtained is called the *conditional probability that E occurs given that F has occurred* and is denoted by

$$P(E|F)$$

A general formula for $P(E|F)$ that is valid for all events E and F is derived in the same manner: If the event F occurs, then, in order for E to occur, it is necessary that the actual occurrence be a point both in E and in F; that is, it must be in EF. Now, since we know that F has occurred, it follows that F becomes our new, or reduced, sample space; hence, the probability that the event EF occurs will equal the probability of EF relative to the probability of F. That is, we have the following definition.

58

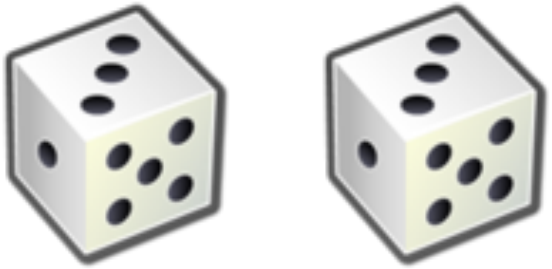
Definition

If $P(F) > 0$, then

$$P(E|F) = \frac{P(EF)}{P(F)} \tag{2.1}$$

Quiz: Conditional Probabilities Puzzle

Roll a die twice



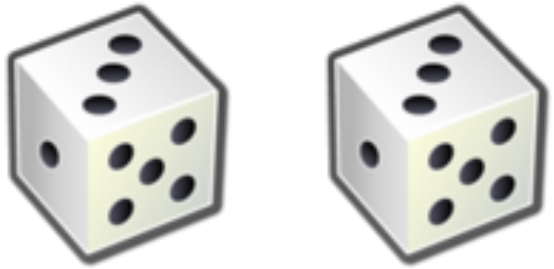
Q1

2 Questions

Roll a die twice. What is the probability that the total we get is 3? Q2 Given the information that the first number is 1, what is probability that the total we get is 3?

Conditional Probabilities Puzzle

Roll a die twice



Question Roll a die twice. What is the probability that the total we get is 3? Given the information that the first number is 1, what is probability that the total we get is 3?

Answer: 2/36 and 1/6.

2 of 6x6 = 36 equilikely events: 1+2, 2+1

Since 1st number =1, 2nd number must =2 to get a sum of 3. There are 6 possible 2nd outcomes, 1 of which is to get a 2, hence 1/6

Conditional Probabilities Example 2

Example 1: If the probability that a research project will be well planned is 0.8 and the probability that is will be well planned and well executed is 0.72, what is the probability that a well planned research project will be well executed?

Conditional Probabilities Examples

Example 1: If the probability that a research project will be well planned is 0.8 and the probability that is will be well planned and well executed is 0.72, what is the probability that a well planned research project will be well executed?

Answer: $0.72/0.8 = 0.9$. P[Exec|Plan]=P[Exec&Plan]/P[Plan]

Conditional probability as an easier path to an answer

Sometimes the conditional probability can be determined easily, so we can actually use the conditional probability to calculate probability.

The multiplication rule: $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$.

Example: There are 3 red balls and 2 blue balls in a box. Randomly take 2 balls from the box. What is the probability that both are red?

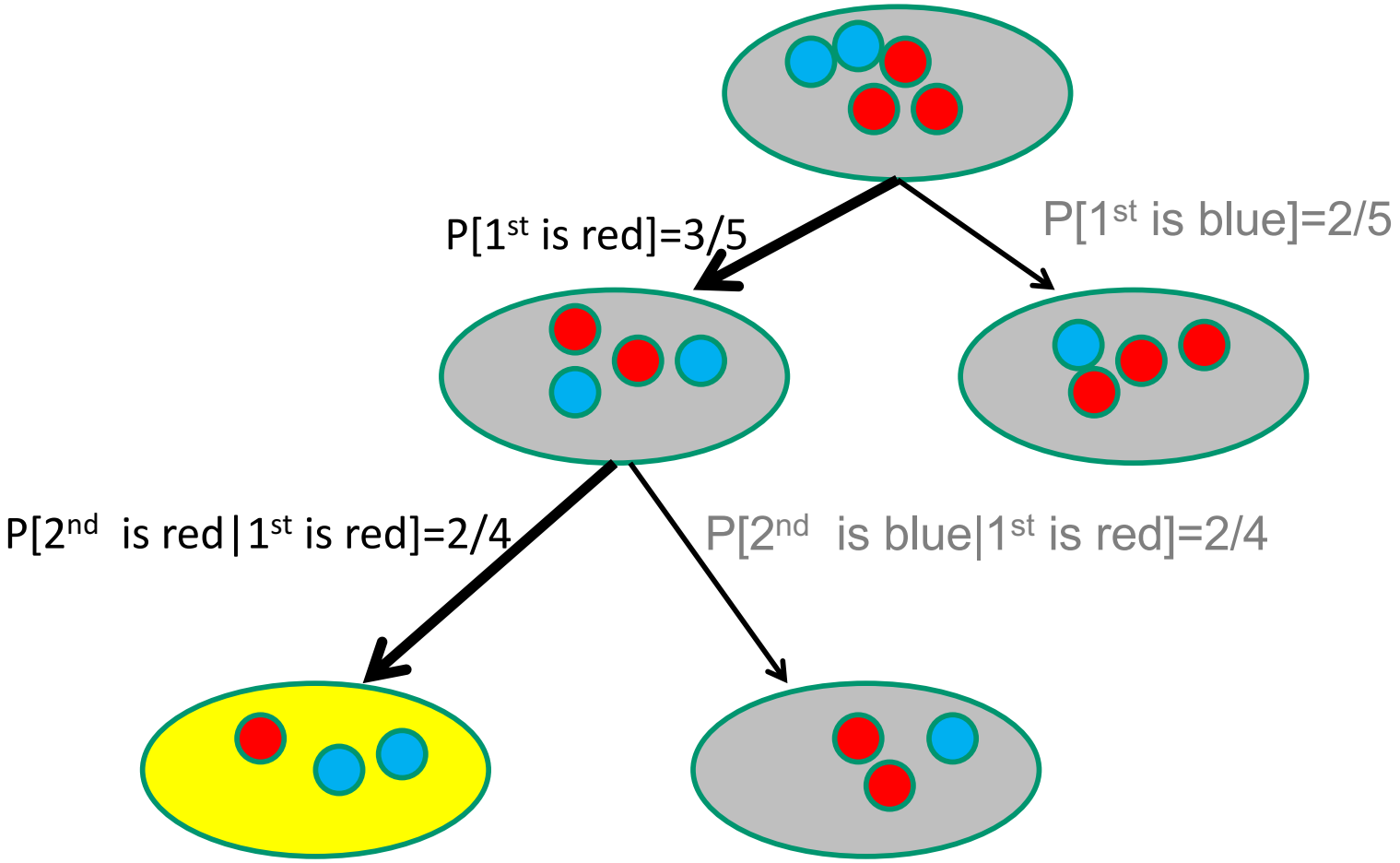
Answer: $P(R_1 \cap R_2) = P(R_1)P(R_2|R_1) = (3/5)(2/4) = 0.3$.

3 of the 5 are red

With 1 red ball gone, 2 of the remaining 4 balls are red

Alternative solution: P[2 red in 2 tries] = Comb(3,2)*Comb(2,0)/Comb(5,2) = 0.3

Solution via Tree Diagram



P[both are red] = p[2nd is red|1st is red] x P[1st is red] = 2/4 x 3/5 = 6/20 = 0.3

Conditional Probability Example

The initial intuition for conditional probability comes from considering probabilities that are ratios. In the case of ratios, $P(\mathsf{E}|\mathsf{F})$, as defined above, is the fraction of items in F that are also in E. We show this as follows. Let n be the number of items in the sample space, n_{F} be the number of items in F, and n_{EF} be the number of items in $\mathsf{E} \cap \mathsf{F}$. Then

$$\frac{P(\mathsf{E} \cap \mathsf{F})}{P(\mathsf{F})} = \frac{n_{\mathsf{EF}}/n}{n_{\mathsf{F}}/n} = \frac{n_{\mathsf{EF}}}{n_{\mathsf{F}}},$$

which is the fraction of items in F that are also in E. As far as meaning, $P(\mathsf{E}|\mathsf{F})$ means the probability of E occurring given that we know F has occurred.

Example 1.6 *Again consider drawing the top card from a deck of cards, let* Queen *be the set of the* 4 *queens,* RoyalCard *be the set of the* 12 *royal cards, and* Spade *be the set of the* 13 *spades. Then*

$$P(\mathsf{Queen}) = \frac{1}{13}$$

$$P(\mathsf{Queen}|\mathsf{RoyalCard}) = \frac{P(\mathsf{Queen} \cap \mathsf{RoyalCard})}{P(\mathsf{RoyalCard})} = \frac{1/13}{3/13} = \frac{1}{3}$$

$$P(\mathsf{Queen}|\mathsf{Spade}) = \frac{P(\mathsf{Queen} \cap \mathsf{Spade})}{P(\mathsf{Spade})} = \frac{1/52}{1/4} = \frac{1}{13}.$$

Probability Basics → Naïve Bayes Models

- **Introduction**
- **Probability Basics**
  - Probability Axioms
  - Conditional probabilities
  - Product Rule, Chain Rule, Bayes Rule
- **Bayes Nets And Naïve Bayes**
  - Learning
  - Independence
  - Conditional independence
  - Naïve Bayes derivation (discrete case plus smoothing)
- **Naïve Bayes Flavors**
  - Continuous input variables
  - Discrete input variables (2 flavors: Bernoulli, multinomial)
- **Case Study: Spam detector in Naïve Bayes**
- **Summary**

Probability Theory Rules

- **cond[al] probability** $P(A \mid B) = \frac{P(A,B)}{P(B)}$
- **PRODUCT RULE and Chain Rule:**

$$P(A,B) = P(A \mid B)P(B)$$

$$\mathrm{P}(A,B,C,D) = \mathrm{P}(A \mid B,C,D)\mathrm{P}(B \mid C,D)\mathrm{P}(C \mid D)\mathrm{P}(D)$$

- **Total Probabilities :**
  - Pr(A)=Pr(A) = Pr(A|B) + Pr(A|not B)
- **Bayes Rules**

Posterior **Likelihood** **Prior**

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

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{\sum_{i=1}^{n} P(B \mid A = a_i)P(A = a_i)}$$

Product Rule is Fundamental and follows from cond[al] probability

$$P(Queen \mid C\,\text{lub}) = \frac{P(Queen, C\,\text{lub})}{P(C\,\text{lub})} = \frac{\frac{1}{52}}{\frac{13}{52}} = \frac{1}{13}$$

Chain Rule, Bayes Rule follow from the Product Rule

Cond[al] Prob:

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

WHERE $P(A, B)$ is the belief in the joint event of A and B (joint probability)

$P(B)$ is the marginal probability of *B*

Rewrite the cond[al] probability to give us the product rule

PRODUCT RULE :

$$P(A, B) = P(A \mid B)P(B)$$

Divide and conquer

Product Rule Part 2

$$P(Queen \mid C\,\text{lub}) = \frac{P(Queen, C\,\text{lub})}{P(C\,\text{lub})} = \frac{\frac{1}{52}}{\frac{13}{52}} = \frac{1}{13}$$

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

Chain Rule, Bayes Rule follow from the Product Rule

PRODUCT RULE :

$$P(A,B) = P(A \mid B)P(B)$$

$$P(A \mid B) = \frac{P(A,B)}{P(B)} \; and \; P(B \mid A) = \frac{P(A,B)}{P(A)}$$

Product Rule Part 2!

(Often left to your imaginations)

$$P(A,B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

Law of Total Probabilities: Event and partition

Boole's inequality

Venn diagram · Tree diagram

The law of total probability is[1] the proposition that if $\{B_n : n = 1, 2, 3, \ldots\}$ is a finite or countably infinite partition of a sample space (in other words, a set of pairwise disjoint events whose union is the entire sample space) and each event B_n is measurable, then for any event A of the same probability space:

$$\Pr(A) = \sum_n \Pr(A \cap B_n)$$

or, alternatively,[1]

$$\Pr(A) = \sum_n \Pr(A \mid B_n) \Pr(B_n),$$

where, for any n for which $\Pr(B_n) = 0$ these terms are simply omitted from the summation, because $\Pr(A \mid B_n)$ is finite.

The summation can be interpreted as a weighted average, and consequently the marginal probability, $\Pr(A)$, is sometimes called "average probability";[2] "overall probability" is sometimes used in less formal writings.[3]

The law of total probability can also be stated for conditional probabilities. Taking the B_n as above, and assuming C is an event independent with any of the B_n:

$$\Pr(A \mid C) = \sum_n \Pr(A \mid C \cap B_n) \Pr(B_n \mid C) = \sum_n \Pr(A \mid C \cap B_n) \Pr(B_n)$$

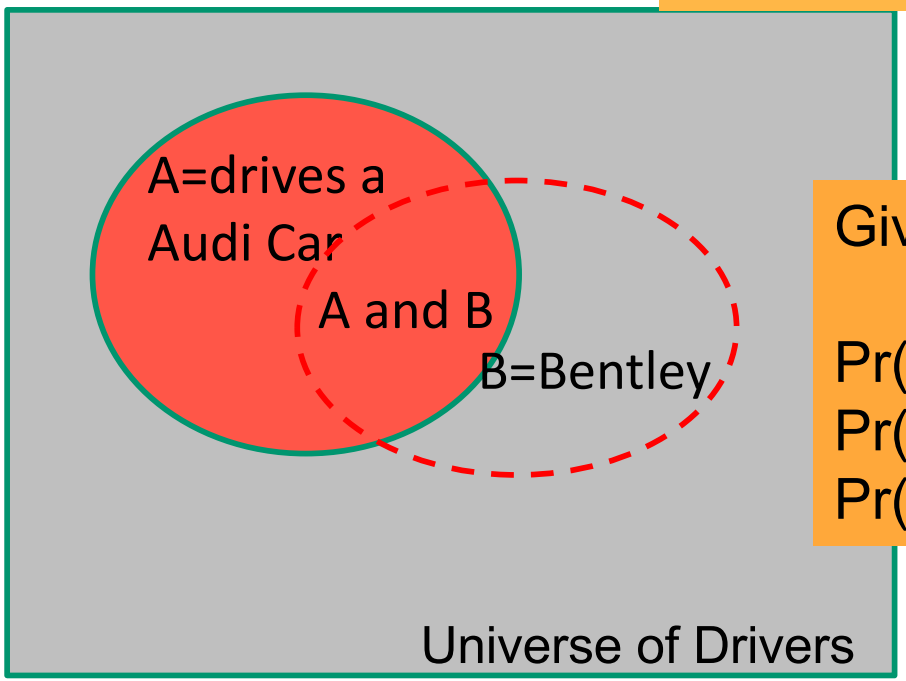
Given an sample space S, event A and partition of S, B
For any event A we can compute its marginal probability with respect to a partition B as the sum of Pr(A and B_n)

Total Probabilities Example: Pr(A)=Pr(A) = Pr(A|B) + Pr(A|not B)

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

Conditional probability can be seen to be the probability with respect to a reduced sample space. We can illustrate the conditional probability with the Venn diagram.

Pr(A) → P(A|B)



Given: A = 5 B = 5, A and B = 2, universe = 20

Pr(A|B) = 2/20 / 5/20 = 2/5
Pr(A|not B) = 3/20 / 15/20 = 3/5
Pr(A) = Pr(A|B) + Pr(A|not B) = 2/5 + 3/5 = 1

Marginalize via Conditional Probability

Simpler calculation: get marginal from joint; decompose joint using product rule

FROM $P(A) = \sum_i P(A, B_i)$ Marginalize A from joint A,B

AND PRODUCT RULE $P(A, B) = P(A \mid B)P(B)$

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

the belief in any event *A* is a weighted sum over the beliefs in all the distinct ways that *A* might be realized.

Chain Rule is a generalization of the Product Rule

Decompose joint probabilities into a product of simpler conditionals

PRODUCT RULE: $P(A,B) = P(A \mid B)P(B)$

1. Chain Rule = Generalization of PRODUCT RULE.
2. It permits the calculation of any member of the joint distribution of a set of random variables using only conditional probabilities

$$P(E_1, E_2 \ldots E_n) = P(E_n \mid E_{n-1}, \ldots, E_2, E_1)P(E_{n-1}, \ldots, E_2, E_1)$$

let's further decompose!

$$\ldots\ldots\ldots\ldots\ldots\ldots = P(E_n \mid E_{n-1}, \ldots, E_2, E_1) \ldots P(E_2 \mid E_1)P(E_1)$$

- **For example, derive using the chain rule**

$$\mathrm{P}(A,B,C,D) = \mathrm{P}(A \mid B,C,D)\mathrm{P}(B \mid C,D)\mathrm{P}(C \mid D)\mathrm{P}(D)$$

Chain Rule Example: Use to calculate joint prob.

- The rule is useful in the study of Bayesian networks, which describe a probability distribution in terms of conditional probabilities.
- Assume Urn 1 has 1 black balls and 2 white balls and Urn 2 has 1 black ball and 3 white balls. Suppose we pick an urn at random and then select a ball from that urn.
- Let event A be choosing the first urn: P(A) = P(~A) = 1/2. Let event B be the chance we choose a white ball. Chance of choosing a white ball, given that we've chose the first urn, is P(B|A) = 2/3. Chance of choosing a white ball, given that we've chosen the second urn is P(B|~A) = 3/4.
- Event A, B would be their intersection; choosing the first urn and a white ball from it. The probability can be found by the chain rule for probability:

$$\mathrm{P}(A, B) = \mathrm{P}(B \mid A)\mathrm{P}(A) = 2/3 \times 1/2 = 1/3$$

[Russell, Stuart J.; Norvig, Peter (2003), *Artificial Intelligence: A Modern Approach* (2nd ed.), Upper Saddle River, NJ: Prentice Hall, ISBN 0-13-790395-2, http://aima.cs.berkeley.edu/ , p. 496.]

Bayes Rule: 1, 2, 3

1. $P(A \mid B) = \frac{P(A,B)}{P(B)}$ $\quad P(B \mid A) = \frac{P(A,B)}{P(A)}$ Cond^{al} Prob

2. $P(A,B) = P(A \mid B)P(B)$ $\quad P(A,B) = P(B \mid A)P(A)$ Product Rule

3. $P(A \mid B)P(B) = P(B \mid A)P(A)$

Prior Probability

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

Posterior probability

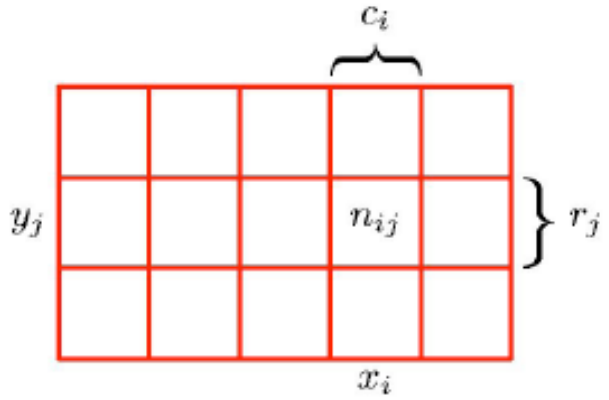
Posterior $\quad$ Likelihood $\quad$ Prior

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{\sum_{i=1}^{n} P(B \mid A = a_i)P(A = a_i)}$$

Marginalize B
Total probabilities

Rules of Probability

- Given random variables X and Y
- Sum Rule gives Marginal Probability



$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j) = \frac{c_i}{N}$$

- Product Rule: joint probability in terms of conditional and marginal

$$p(X,Y) = \frac{n_{ij}}{N} = p(Y \mid X)p(X) = \frac{n_{ij}}{c_i} \times \frac{c_i}{N}$$

- Combining we get Bayes Rule

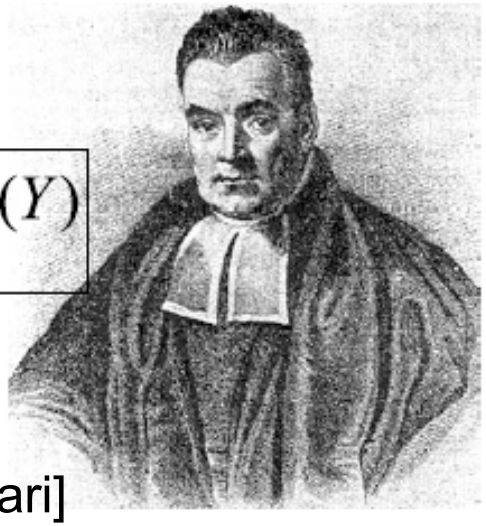
$$p(Y \mid X) = \frac{p(X \mid Y)p(Y)}{p(X)}$$

where

$$p(X) = \sum_{Y} p(X \mid Y)p(Y)$$

Viewed as

Posterior α likelihood x prior



[Srihari]

Probability Theory Summary

1. **cond[al] probability** $P(A \mid B) = \frac{P(A,B)}{P(B)}$
2. **PRODUCT RULE and Chain Rule:**

$$P(A,B) = P(A \mid B)P(B)$$

$$\mathrm{P}(A,B,C,D) = \mathrm{P}(A \mid B,C,D)\mathrm{P}(B \mid C,D)\mathrm{P}(C \mid D)\mathrm{P}(D)$$

3. **Total Probabilities :**
   1. Pr(A)=Pr(A) = Pr(A|B) + Pr(A|not B)
4. **Bayes Rules**

Posterior **Likelihood** **Prior**

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

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{\sum_{i=1}^{n} P(B \mid A = a_i)P(A = a_i)}$$

Maximum a Posteriori (MAP)

- **Any such maximally probable hypothesis is called a *maximum a posteriori (MAP) hypothesis***

Bayes Rule Puzzle

- **Great example to test your understanding of Bayes Rule in practice**
- **Please test yourself**

Bayes Rule Puzzle

test with two possible outcomes: ⊕ (positive) and ⊖ (negative). We have prior knowledge that over the entire population of people only .008 have this disease. Furthermore, the lab test is only an imperfect indicator of the disease. The test returns a correct positive result in only 98% of the cases in which the disease is actually present and a correct negative result in only 97% of the cases in which the disease is not present. In other cases, the test returns the opposite result. The above situation can be summarized by the following probabilities:

PUZZLE

Fill in the probabilities based on the above text

Pr(Cancer) = ??
Pr(+|Cancer) =??
Pr(??????|????)=???

Bayes Rule Example

test with two possible outcomes: $\oplus$ (positive) and $\ominus$ (negative). We have prior knowledge that over the entire population of people only .008 have this disease. Furthermore, the lab test is only an imperfect indicator of the disease. The test returns a correct positive result in only 98% of the cases in which the disease is actually present and a correct negative result in only 97% of the cases in which the disease is not present. In other cases, the test returns the opposite result. The above situation can be summarized by the following probabilities:

$$P(cancer) = .008, \quad P(\neg cancer) = .992$$
$$P(\oplus|cancer) = .98, \quad P(\ominus|cancer) = .02$$
$$P(\oplus|\neg cancer) = .03, \quad P(\ominus|\neg cancer) = .97$$

Patient has cancer

PUZZLE:

Observe a new patient for whom the lab tests return a positive result

Should we diagnose the patient as having cancer or not?

Bayes Rule Example

test with two possible outcomes: ⊕ (positive) and ⊖ (negative). We have prior knowledge that over the entire population of people only .008 have this disease. Furthermore, the lab test is only an imperfect indicator of the disease. The test returns a correct positive result in only 98% of the cases in which the disease is actually present and a correct negative result in only 97% of the cases in which the disease is not present. In other cases, the test returns the opposite result. The above situation can be summarized by the following probabilities:

$$P(cancer) = .008, \quad P(\neg cancer) = .992$$
$$P(\oplus|cancer) = .98, \quad P(\ominus|cancer) = .02$$
$$P(\oplus|\neg cancer) = .03, \quad P(\ominus|\neg cancer) = .97$$

Patient has cancer

PUZZLE:

Observe a new patient for whom the lab tests return a positive result

Should we diagnose the patient as having cancer or not?

Use Bayes theorem to calculate
Given Pr(Cancer|+) = Pr(+|Cancer) X Pr(Cancer)/ Pr(+)
Then…

Bayes Rule Example

test with two possible outcomes: ⊕ (positive) and ⊖ (negative). We have prior knowledge that over the entire population of people only .008 have this disease. Furthermore, the lab test is only an imperfect indicator of the disease. The test returns a correct positive result in only 98% of the cases in which the disease is actually present and a correct negative result in only 97% of the cases in which the disease is not present. In other cases, the test returns the opposite result. The above situation can be summarized by the following probabilities:

$$P(cancer) = .008, \quad P(\neg cancer) = .992$$
$$P(\oplus|cancer) = .98, \quad P(\ominus|cancer) = .02$$
$$P(\oplus|\neg cancer) = .03, \quad P(\ominus|\neg cancer) = .97$$

Patient has cancer

Suppose we now observe a new patient for whom the lab test returns a positive result. Should we diagnose the patient as having cancer or not? The maximum a posteriori hypothesis can be found using Equation (6.2):

$$P(Cancer \mid Test = +) = \frac{P(Test = + \mid Cancer)P(Can}{P(Test = +)}$$

Numerator for P(cancer|+)

$$P(\oplus|cancer)\,P(cancer) = (.98).008 = .0078$$

$$P(\oplus|\neg cancer)\,P(\neg cancer) = (.03).992 = .0298$$

Numerator for P(NOTcancer|+)

Thus, $h_{MAP} = \neg cancer$. The exact posterior probabilities can also be determined by normalizing the above quantities so that they sum to 1 (e.g., $P(cancer|\oplus) = \frac{.0078}{.0078+.0298} = .21$). This step is warranted because Bayes theorem states that the posterior probabilities are just the above quantities divided by the probability of the data, $P(\oplus)$. Although $P(\oplus)$ was not provided directly as part of the problem statement, we can calculate it in this fashion because we know that $P(cancer|\oplus)$ and $P(\neg cancer|\oplus)$ must sum to 1 (i.e., either the patient has cancer or they do not). Notice that while the posterior probability of *cancer* is significantly higher than its prior probability, the most probable hypothesis is still that the patient does

Probability Basics → Naïve Bayes Models

- **Introduction**
- **Probability Basics**
  - Probability Axioms
  - Conditional probabilities
  - Product Rule, Chain Rule, Bayes Rule
- **Bayes Nets And Naïve Bayes**
  - Learning
  - Independence
  - Conditional independence
  - Naïve Bayes derivation (discrete case plus smoothing)
- **Naïve Bayes Flavors**
  - Continuous input variables
  - Discrete input variables (2 flavors: Bernoulli, multinomial)
- **Case Study: Spam detector in Naïve Bayes**
- **Summary**

Digression

- **Digress to cover Confidence intervals**

Confidence interval for a binomial distribution

- In statistics, a binomial proportion confidence interval is a confidence interval for a proportion in a statistical population. It uses the proportion estimated in a statistical sample and allows for sampling error. There are several formulas for a binomial confidence interval, but all of them rely on the assumption of a binomial distribution.
- In general, a binomial distribution applies when an experiment is repeated a fixed number of times, each trial of the experiment has two possible outcomes (labeled arbitrarily success and failure), the probability of success is the same for each trial, and the trials are statistically independent.
- A simple example of a binomial distribution is the set of various possible outcomes, and their probabilities, for the number of heads observed when a (not necessarily fair) coin is flipped ten times.
- The observed binomial proportion is the fraction of the flips which turn out to be heads.
- Given this observed proportion, the confidence interval for the true proportion innate in that coin is a range of possible proportions which may contain the true proportion.
- A 95% confidence interval for the proportion, for instance, will contain the true proportion 95% of the times that the procedure for constructing the confidence interval is employed.
- Note that this does not mean that a calculated 95% confidence interval will contain the true proportion with 95% probability. Instead, one should interpret it as follows: the process of drawing a random sample and calculating an accompanying 95% confidence interval will generate a confidence interval that contains the true proportion in 95% of all cases.
- There are several ways to compute a confidence interval for a binomial proportion. The normal approximation interval is the simplest formula, and the one introduced in most basic Statistics classes and textbooks. This formula, however, is based on an approximation that does not always work well.

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n}\hat{p}\left(1-\hat{p}\right)}$$

https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval

Confidence interval for a binomial distribution

- **In statistics, a binomial proportion confidence interval is a confidence interval for a proportion in a statistical population.**
  - The number of people who voted for Donald Trump in a poll
- **It uses the proportion estimated in a statistical sample and allows for sampling error.**
- **There are several formulas for a binomial confidence interval, but all of them rely on the assumption of a binomial distribution.**

Confidence interval for a binomial distribution

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n}\hat{p}\left(1-\hat{p}\right)}$$

Confidence interval for a binomial distribution

https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval

Confidence interval for a binomial distribution

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n}\hat{p}\,(1-\hat{p})}$$

$$\hat{p} \pm \frac{z^*}{2\sqrt{n}}$$

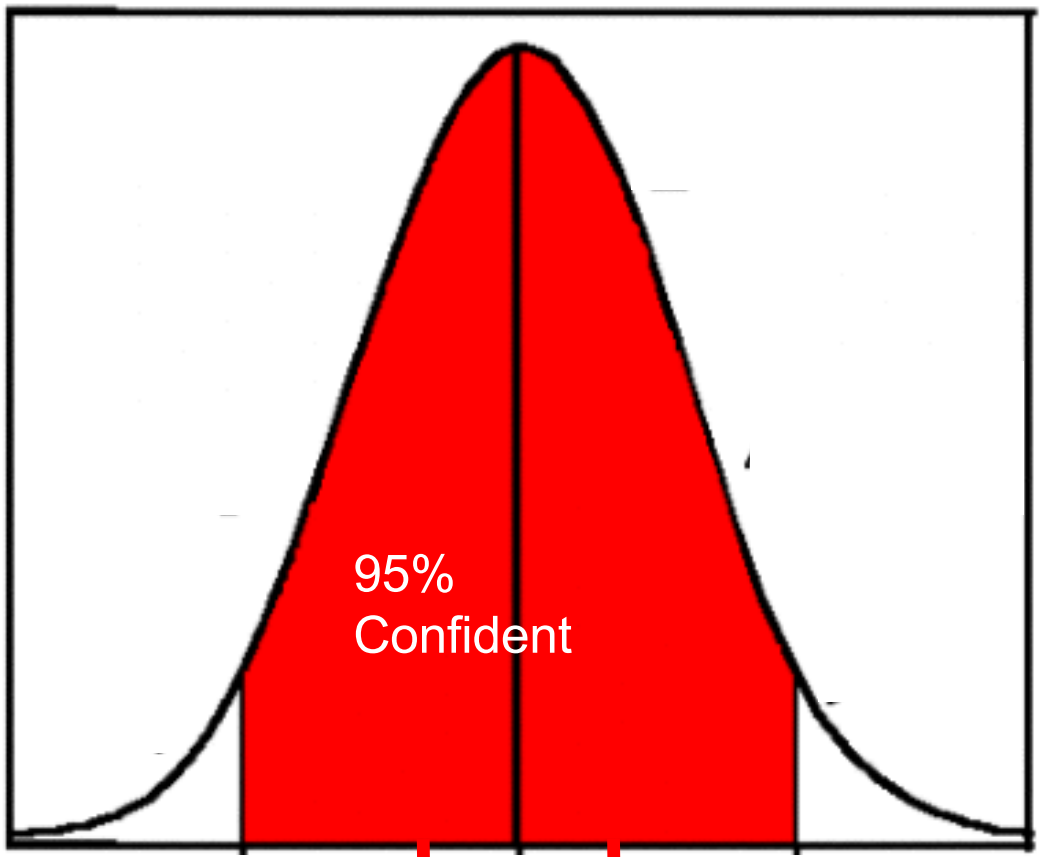
Example: Poll of 1,000 people
2*1/(2*sqrt(n)), e.g., 1/sqrt(1000)= ± 3%

Election Poll of 1000 people (N=1000)
Assume 45% (P=0.45, Q=1-P=55%) favor candidate A (Donald) versus 49% for Candidate B (Hillary)
Standard error about the mean = SQRT(PQ/N) = SQRT(0.45*0.55/1000)=0.015=1.5%
So the 95% confidence interval surrounding Donald's support of 45% is 45%+/- 2*1.5 = [42,…48]

In machine learning:
Pr(Y=Business) = 100/1000 = 0.1
SQRT(0.1*0.9/1000) =0.01 = 1%
yielding Pr(Y=Business) Confidence interval = 0.1+/- 0.01*2
[0.08,0.12] Note we need 1,000 examples to get this tight CI;

How much data do we need to estimate Pr(Y|X) with confidence?

Estimating Reliable Probabilities



1.6% 2.6% 3.6% CTR (after 1,000 impressions)

2.3% 2.9% (after 10,000 impressions)

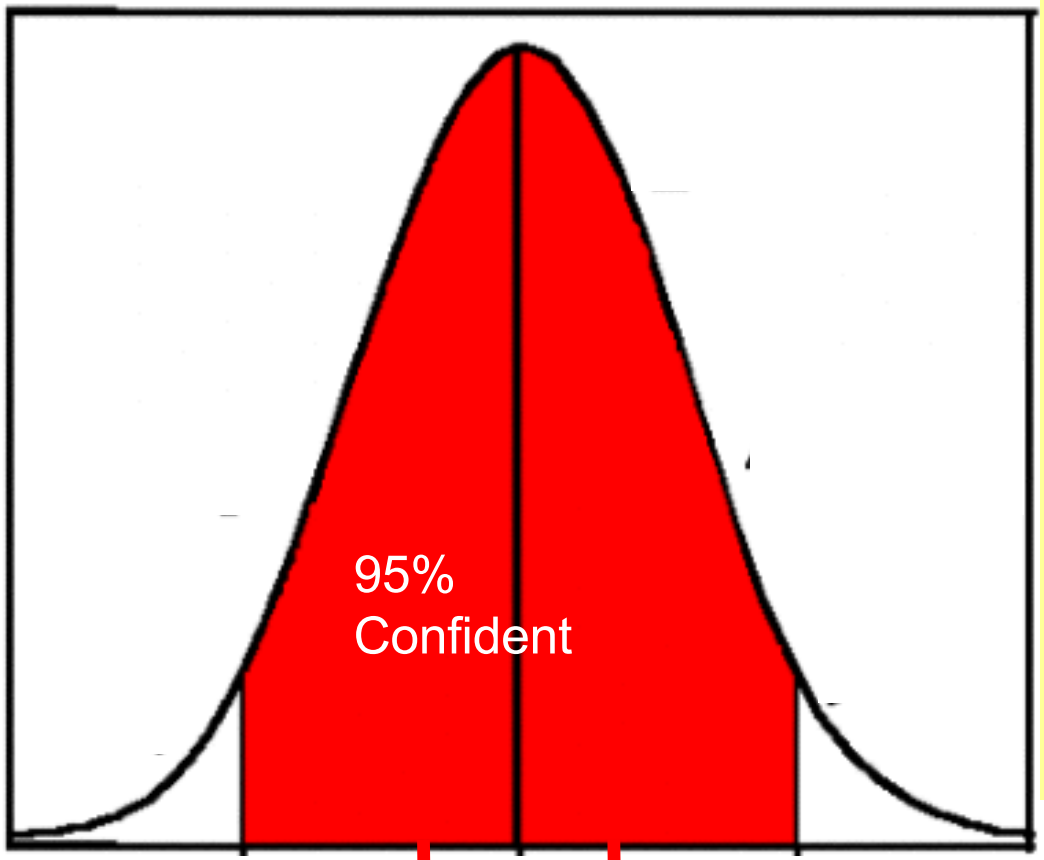
Estimate using Binomial
MLE Estimates
I.e., #Clicks/#Impression

$40/1,000 @CPC of $1.60
$400/10,000

Standard error of the mean of one sample is the estimate of the standard deviation that would be obtained from the means of a large number of samples drawn from that population.

StdErr= sqrt(.026*(1-0.026)/1000) * 1.96=±1%

Estimating CTR (and later AR)



For a network of
~10^9 target pages,
~10^6 ads
~10^7 users
……
- Cannot afford this evaluation/auditioning
- Borrow strength, marginalize
- CoD (curse of dim^{ality})

1.6% 2.6% 3.6% CTR (after 1,000 impressions)

2.3% 2.9% (after 10,000 impressions)

Confidence Intervals in a Nutshell

- **The assumptions required for CI for a population proportion to be valid:**
  - the sample size *n* is large enough (check: $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$)
  - the data are a *random sample* from that population
- **General Confidence Interval for the Population Proportion *p*:**

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- **Approximate 95% Confidence Interval for the Population Proportion *p*:**

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

```
> 2*sqrt(.4*.6/1000)  #president's satisfaction rating
[1] 0.03098387
```

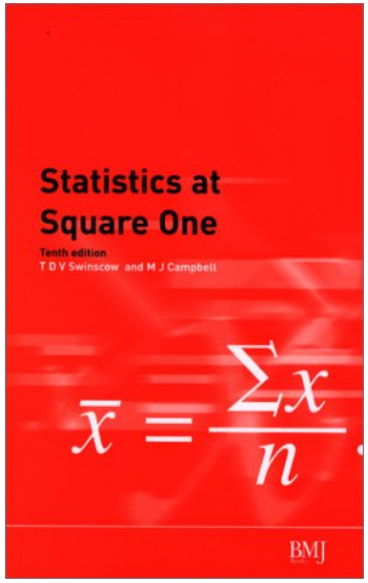
  - Note: standard error of is which is largest when $\hat{p} = \frac{1}{2}$
- **Conservative Confidence Interval for the Population Proportion *p*:**

$$\hat{p} \pm \frac{z^*}{2\sqrt{n}}$$

2*1/(2*sqrt(n)), e.g., 1/sqrt(1000)= ± 3%

Sample Size Needed

- **Sample Size Needed for Desired Confidence Level and Error Margin where *m* is the desired margin of error.**



$$n = \left(\frac{z^*}{2m} \right)^2$$

SE= =1/sqrt(n)

Measure CTRs confidently p± 0.001
> (1.96/(2*.001))^2
[1] 960,400 sample size

[http://www.bmj.com/statsbk/ Statistics at Square One (Ninth Edition), T D V Swinscow]

Bayes Theorem for Machine Learning

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- P(h) = Prior probability of hypothesis
- P(D) = Prior probability of training data D.
- P(h|D) = Probability of h given D.
- P(D|h) = Probability of D given h.

- **Bayesian Learning via**
  - Max Likelihood inference
  - Hierarchical Bayesian inference leading to the MAP hypothesis

Bayesian Learning: Discrete input/output variables

- **Here we consider the relationship between supervised learning, or function approximation problems, and Bayesian reasoning.**

- **We begin by considering how to design learning algorithms based on Bayes rule.**
  - Consider a supervised learning problem in which we wish to approximate an unknown target function
    - $f : X \rightarrow Y$, or equivalently $P(Y|X)$.
  - To begin, we will assume Y is a boolean-valued random variable, and X is a vector containing n boolean attributes.
  - In other words, X =<X1,X2 ...,Xn>, where Xi is the boolean random variable denoting the ith attribute of X.

Probability Basics → Naïve Bayes Models

- **Introduction**
- **Probability Basics**
  - Probability Axioms
  - Conditional probabilities
  - Product Rule, Chain Rule, Bayes Rule
- **Bayes Nets And Naïve Bayes**
  - Learning
  - Independence
  - Conditional independence
  - Naïve Bayes derivation (discrete case plus smoothing)
- **Naïve Bayes Flavors**
  - Continuous input variables
  - Discrete input variables (2 flavors: Bernoulli, multinomial)
- **Case Study: Spam detector in Naïve Bayes**
- **Summary**

Supervised Learning Via Bayes Rule

- **Consider a supervised learning problem in which we wish to approximate an unknown target function *f : X →Y, or equivalently P(Y_j | X).***
- ***For pedagogical reasons assume discrete binary variables for both and input (X)and output (Y)***
  - *To begin, we will* assume *Y is a boolean-valued random variable, and X is a vector containing n* boolean attributes. In other words, *X = hX1;X2 : : : ;Xni, where Xi is the boolean* random variable denoting the *ith attribute of X.*
- **Applying Bayes rule, we see that *P(Y = y_i | X) can be represented as***

$$P(Y = y_i | X = x_k) = \frac{P(X = x_k | Y = y_i) P(Y = y_i)}{\sum_j P(X = x_k | Y = y_j) P(Y = y_j)}$$

  - where y_m *represents the* m^{tn} *possible value for Y, and where the summation in the* denominator is over all legal values of the random variable *Y*

Goal is to learn Class conditionals Pr(X|Y)

Applying Bayes rule, we see that $P(Y = y_i|X)$ can be represented as

$$P(Y = y_i|X = x_k) = \frac{P(X = x_k|Y = y_i)P(Y = y_i)}{\sum_j P(X = x_k|Y = y_j)P(Y = y_j)}$$

- **Then use to classify X**

Learning Classifiers based on Bayes Rule

Posterior | Likelihood | Prior

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

Assume
Y is a boolean-valued random variable,
X is a vector containing n boolean attributes.
In other words, X =<X1,X2 ...,Xn>, where Xi is the boolean random variable denoting the ith attribute of X

Applying Bayes rule, we see that $P(Y = y_i|X)$ can be represented as

$$P(Y = y_i|X = x_k) = \frac{P(X = x_k|Y = y_i)P(Y = y_i)}{\sum_j P(X = x_k|Y = y_j)P(Y = y_j)}$$

LHS RHS

where y_m denotes the mth possible value for Y, x_k denotes the kth possible vector value for X, and where the summation in the denominator is over all legal values of the random variable Y.

One way to learn $P(Y|X)$ is to use the training data to estimate $P(X|Y)$ and $P(Y)$. We can then use these estimates, together with Bayes rule above, to determine $P(Y|X = x_k)$ for any new instance x_k.

Confidence interval for a binomial distribution

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n}\hat{p}\,(1-\hat{p})}$$

Estimating P(Y), P(X|Y) from data

- **If we are going to train a Bayes classifier by estimating P(X|Y) and P(Y), then it is reasonable to ask how much training data will be required to obtain reliable estimates of these distributions.**
- **Let us assume training examples are generated by drawing instances at random from an unknown underlying distribution P(X), then allowing a teacher to label this example with its Y value.**
- **A 1,000 independently drawn training examples will usually suffice to obtain a maximum likelihood estimate of P(Y) that is within a few percent of its correct value when Y is a boolean variable.**

Note for a simple proposition the Standard error:
SE(P(Y))=SQRT(0.1*0.9/1000)=0.01 for business labeled docs; assume 100 out 1000

- **However, accurately estimating P(X|Y) typically requires many more examples**

Assume n binary variables: 2^n propositions X 1000 training examples
See example in a moment

Impractical: so what to do?

Learning a (FULL) Bayesian Model

- **One way to learn *P(Y | X)***
  1. *Estimate the joint probability distribution P(X |Y) and P(Y) from the training data.*
  2. *Use these estimates, together with Bayes rule above, to determine P(Y | X = x_k) for any new instance x_k.*
- ***A maximum likelihood estimate of P(Y) can be accomplished with just a few hundred examples***
  - *#ExamplesWithLabel/#TotalNumberOfExamples. E.g., 45 examples are in class 1 and 55 are in class2 then Pr(Y=Class1) = 45/100=0.45.*
- ***However, estimating the joint probability distribution P(X |Y) requires an exponential amount training examples (even with this assumption of binary input and output variable)***
  - *Assume n input attributes X_i take 2 discrete values and Y has 2 possible class values; $2^n * 2$ possible states of the world (parameters)*

Estimating the Joint Prob. Directly?

- **$2^n * 2$ possible states of the world (parameter estimates)**
  - Assume n input attributes X_i take 2 discrete values and Y has 2 possible class values; $2^n * 2$ possible states of the world (parameters) that we need to estimate from data
    - $\theta_{iJ} = P(X = x_i | Y = y_J)$; 2^n possible states of the input world 2 possible output states
- **Can reduce $2^n - 1 * 2$ parameters by exploiting the sum-to-1**
  - Class conditional multinomial needs to sum to 1 so we exploit this and can infer one of the class conditional probs from the rest
  - Each state is a complex combination of feature values
- **Require 200k examples (or more) for reliable estimates of 10 binary variable problem**
  - So for 10 input variables and one output variable we have to estimate 2046 states. To estimate probabilities requiring at least 204,600 examples for reliable estimates.
- **So not very realistic, even in these WWW times**

Learning a Bayesian Model

- *$2^n * 2$ possible states of the world*
  - *Assume n input attributes X_i take 2 discrete values and Y has 2 possible class values; $2^n * 2$ possible states of the world (parameters) that* we need to estimate from data
    - $\theta_{iJ} = P(X = x_i | Y = y_J)$; *$2^n$ possible states of the input world 2 possible output states*

Index $X=X^i$	X1 $X=x_1^i$)	X2 $X=x_2^i$)	Y $Y=y_j$	$\theta_{ij}=P(X=x^i \mid y=y_j)$
1	1	1	1	θ_{11}
2	0	1	1	θ_{21}
3	1	0	1	θ_{31}
4	0	0	1	θ_{41}
5	1	1	0	θ_{50}
6	0	1	0	θ_{60}
7	1	0	0	θ_{70}
8	0	0	0	θ_{80}

Sum to 1

Cant Estimate Joint Probability

- **Look for ways to combat the intractable data needs for learning a Bayesian Classifier**
  - Leverage the chain rule and other assumptions (Markov, Naïve Bayes); this leads to Bayesian Networks
  - Make a conditional independence assumption; this leads to a Naïve Bayes classifier
    - Reduces the number of parameters from $2^n - 1 * 2$ parameters to 2n

Bayesian Networks

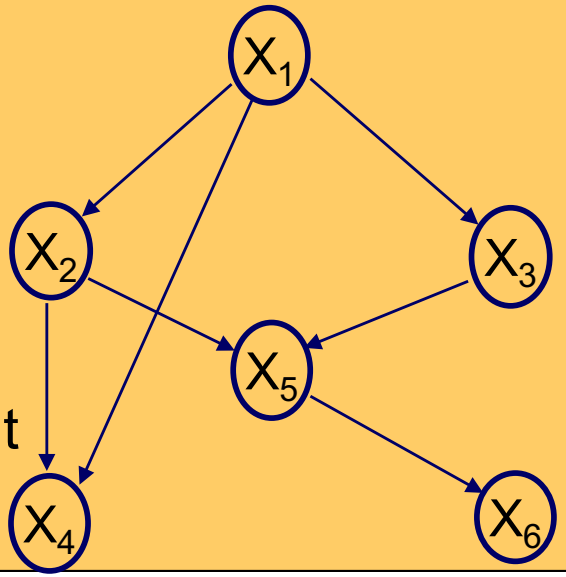
GOAL: 2^N Possibilities -> 2N

P: Joint Probability Distribution

#	X1	X2	X3	X4	X5	X6	Pr(X1,X2, X3, X4, X5, X6)
1	1						
2	1						
..	1						
4..	1						
5	1						
6	1						
7	1						
64							

Factorize the joint

G: Directed Acyclic Graph



$p(x_1, x_2, x_3, x_4, x_5, x_6)$ **1. Partial Order**

2. By Chain Rule

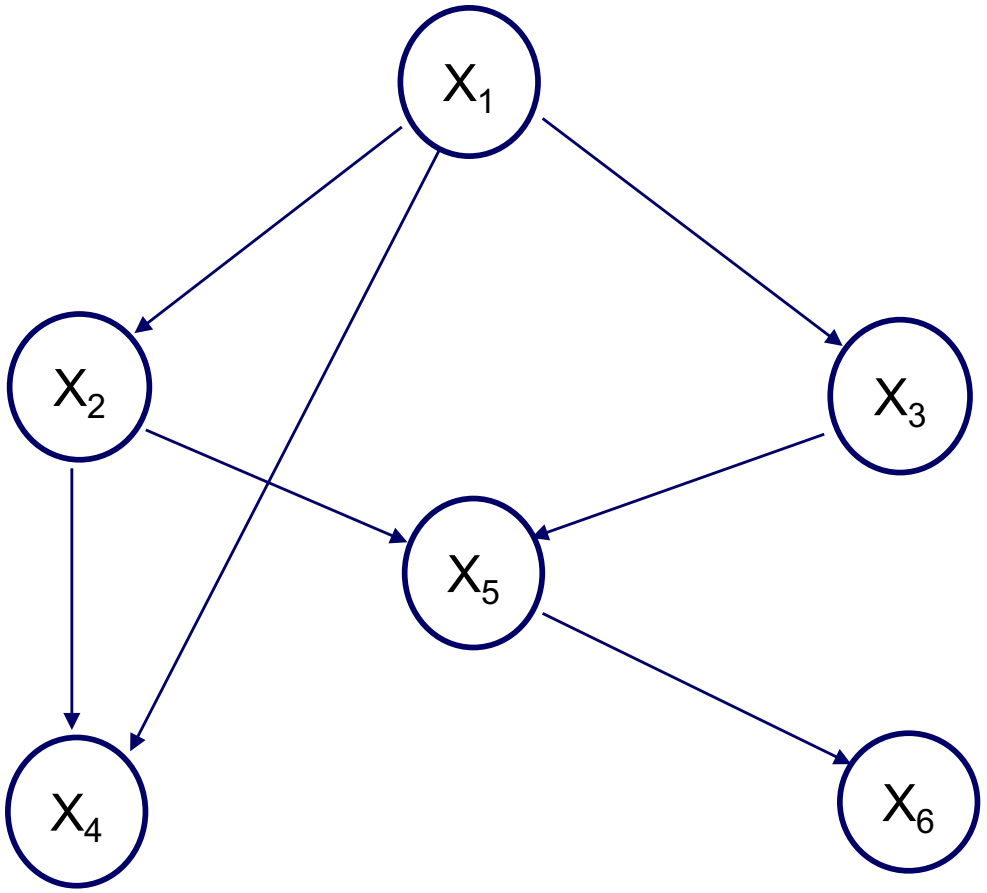
$= p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1, x_2)\, p(x_4 \mid x_1, x_2, x_3)\, p(x_5 \mid x_4,x_3,x_2,x_1)\, p(x_6 \mid x_5,x_4,x_3,x_2,x_1)$

$= p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1)\, p(x_4 \mid x_2, x_1)\, p(x_5 \mid x_3, x_2)\, p(x_6 \mid x_5)$ **3. Markov Property**

$= p(x_1)\, p(x_2)\, p(x_3)\, p(x_4)\, p(x_5)\, p(x_6)$ **4: Independence (see next section)**

$P(y|x_6, x_5,x_4,x_3,x_2,x_1) = p(x_1|y)\, p(x_2|y)\, p(x_3|y)\, p(x_4|y)\, p(x_5|y)\, p(x_6|y)$ **5: Naïve Bayes via Cond. Independence**

1, 2, 3 Example of Factorization



2. By Chain Rule

$p(x_1, x_2, x_3, x_4, x_5, x_6)$ 1. Partial Order

$= p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1, x_2)\, p(x_4 \mid x_1, x_2, x_3)\, p(x_5 \mid x_4, x_3, x_2, x_1)\, p(x_6 \mid x_5, x_4, x_3, x_2, x_1)$

$= p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1)\, p(x_4 \mid x_2, x_1)\, p(x_5 \mid x_3, x_2)\, p(x_6 \mid x_5)$ 3. Markov Property

Factorization

- Given a DAG G, topologically sort the variables:

  $X_1, \ldots, X_n$ (s.t. if i<j, then X_j is not an ancestor of X_i)

- For any joint distribution P that is Markov to G, factorize it as follows:

  $P(X_1, \ldots, X_n)$

  $= P(X_1)\, P(X_2|X_1) \ldots P(X_n|X_1, \ldots, X_{n-1})$ By Chain Rule

  $= \prod_i P(X_i \mid Pa(X_i))$ Exploit Local Markov Property

- Markov property: every variable is independent of its non-descendants given its parents.
- Markov Condition requires that every conditional independence in the graph is in the joint probability distribution

Probability Basics → Naïve Bayes Models

- **Introduction**
- **Probability Basics**
  - Probability Axioms
  - Conditional probabilities
  - Product Rule, Chain Rule, Bayes Rule
- **Bayes Nets And Naïve Bayes**
  - Learning
  - Independence
  - Conditional independence
  - Naïve Bayes derivation (discrete case plus smoothing)
- **Naïve Bayes Flavors**
  - Continuous input variables
  - Discrete input variables (2 flavors: Bernoulli, multinomial)
- **Case Study: Spam detector in Naïve Bayes**
- **Summary**

Independence

- **In probability theory, two events are independent, statistically independent, or stochastically independent if the occurrence of one does not affect the probability of the other.**
  - P(A|B) = Pr(A)
  - As a result P(A and B) = Pr(A) X Pr(B);

Event

- **Similarly, two random variables are independent if the realization of one does not affect the probability distribution of the other.**
  - The concept of independence extends to dealing with collections of more than two events or random variables, in which case the events are pairwise independent if each pair are independent of each other, and the events are mutually independent if each event is independent of each other combination of events.

Variables

https://en.wikipedia.org/wiki/Independence_(probability_theory)

P(A and B) = Pr(A) X Pr(B)

Two events [edit]

Two events *A* and *B* are **independent** (often written as $A \perp B$ or $A \perp\!\!\!\perp B$) if and only if their joint probability equals the product of their probabilities:

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B).$$

Why this defines independence is made clear by rewriting with conditional probabilities:

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B) \Leftrightarrow \mathrm{P}(A) = \frac{\mathrm{P}(A)\mathrm{P}(B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \mathrm{P}(A \mid B)$$

and similarly

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B) \Leftrightarrow \mathrm{P}(B) = \mathrm{P}(B \mid A).$$

Thus, the occurrence of *B* does not affect the probability of *A*, and vice versa. Although the derived expressions may seem more intuitive, they are not the preferred definition, as the conditional probabilities may be undefined if *P*(*A*) or *P*(*B*) are 0. Furthermore, the preferred definition makes clear by symmetry that when *A* is independent of *B*, *B* is also independent of *A*.

More than two events [edit]

A finite set of events $\{A_i\}$ is **pairwise independent** if and only if every pair of events is independent[2]—that is, if and only if for all distinct pairs of indices *m*, *k*,

$$\mathrm{P}(A_m \cap A_k) = \mathrm{P}(A_m)\mathrm{P}(A_k).$$

A finite set of events is **mutually independent** if and only if every event is independent of any intersection of the other events[2]—that is, if and only if for every *n*-element subset $\{A_i\}$,

$$\mathrm{P}\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \mathrm{P}(A_i).$$

This is called the *multiplication rule* for independent events. Note that it is not a single condition involving only the product of all the probabilities of all single events (see below for a counterexample); it must hold true for all subset of events.

For more than two events, a mutually independent set of events is (by definition) pairwise independent; but the converse is not necessarily true (see below for a counterexample).

Two events are independent if…

Definition 1.3 *Two events* E *and* F *are independent if one of the following hold:*

1. $P(\mathsf{E}|\mathsf{F}) = P(\mathsf{E})$ *and* $P(\mathsf{E}) \neq 0,\ P(\mathsf{F}) \neq 0.$
2. $P(\mathsf{E}) = 0$ *or* $P(\mathsf{F}) = 0.$

Notice that the definition states that the two events are independent even though it is based on the conditional probability of E given F. The reason is that independence is symmetric. That is, if $P(\mathsf{E}) \neq 0$ and $P(\mathsf{F}) \neq 0$, then $P(\mathsf{E}|\mathsf{F}) = P(\mathsf{E})$ if and only if $P(\mathsf{F}|\mathsf{E}) = P(\mathsf{F})$. It is straightforward to prove that E and F are independent if and only if $P(\mathsf{E} \cap \mathsf{F}) = P(\mathsf{E})P(\mathsf{F})$.

Independence is symmetric

Independence Example
Pr(Queen) = Pr(Queen|Spade)

Example 1.6 *Again consider drawing the top card from a deck of cards, let* Queen *be the set of the* 4 *queens,* RoyalCard *be the set of the* 12 *royal cards, and* Spade *be the set of the* 13 *spades. Then*

$$P(\text{Queen}) = \frac{1}{13}$$

$$P(\text{Queen}|\text{RoyalCard}) = \frac{P(\text{Queen} \cap \text{RoyalCard})}{P(\text{RoyalCard})} = \frac{1/13}{3/13} = \frac{1}{3}$$

$$P(\text{Queen}|\text{Spade}) = \frac{P(\text{Queen} \cap \text{Spade})}{P(\text{Spade})} = \frac{1/52}{1/4} = \frac{1}{13}.$$

Notice in the previous example that $P(\text{Queen}|\text{Spade}) = P(\text{Queen})$. This means that finding out the card is a spade does not make it more or less probable that it is a queen. That is, the knowledge of whether it is a spade is irrelevant to whether it is a queen. We say that the two events are independent in this case, which is formalized in the following definition.

[Learning Bayesian Networks, Prentice-Hall, Richard E. Neapolitan]

Independence more generally on sets: Set A and B are independent

Definition 1.6 *Suppose we have a probability space (Ω, P), and two sets* A *and* B *containing random variables defined on* Ω. *Then the sets* A *and* B *are said to be* ***independent*** *if, for all values of the variables in the sets* a *and* b, *the events* $\mathsf{A} = \mathsf{a}$ *and* $\mathsf{B} = \mathsf{b}$ *are independent. That is, either* $P(\mathsf{a}) = 0$ *or* $P(\mathsf{b}) = 0$ *or*

$$P(\mathsf{a}|\mathsf{b}) = P(\mathsf{a}).$$

When this is the case, we write

$$I_P(\mathsf{A}, \mathsf{B}),$$

where I_P *stands for independent in* P.

Independence Example for Random Variables R, T, S

Example 1.18 *Let Ω be the set of all cards in an ordinary deck, and let P assign $1/52$ to each card. Define random variables as follows:*

Variable	Value	Outcomes Mapped to this Value
R	$r1$	*All royal cards*
	$r2$	*All nonroyal cards*
T	$t1$	*All tens and jacks*
	$t2$	*All cards that are neither tens nor jacks*
S	$s1$	*All spades*
	$s2$	*All nonspades*

Then we maintain the sets $\{R, T\}$ and $\{S\}$ are independent. That is,

$$I_P(\{R, T\}, \{S\}).$$

To show this, we need show for all values of r, t, and s that

$$P(r, t|s) = P(r, t).$$

(Note that it we do not show brackets to denote sets in our probabilistic expression because in such an expression a set represents the members of the set. See the discussion following Example 1.14.) The following table shows this is the case:

Independence Example for Random Variables R, T, S

s	r	t	$P(r,t\|s)$	$P(r,t)$
$s1$	$r1$	$t1$	$1/13$	$4/52 = 1/13$
$s1$	$r1$	$t2$	$2/13$	$8/52 = 2/13$
$s1$	$r2$	$t1$	$1/13$	$4/52 = 1/13$
$s1$	$r2$	$t2$	$9/13$	$36/52 = 9/13$
$s2$	$r1$	$t1$	$3/39 = 1/13$	$4/52 = 1/13$
$s2$	$r1$	$t2$	$6/39 = 2/13$	$8/52 = 2/13$
$s2$	$r2$	$t1$	$3/39 = 1/13$	$4/52 = 1/13$
$s2$	$r2$	$t2$	$27/39 = 9/13$	$36/52 = 9/13$

Definition 1.7 *Suppose we have a probability space* (Ω, P), *and three sets* A, B, *and* C *containing random variable defined on* Ω. *Then the sets* A *and* B *are said to be conditionally independent given the set* C *if, for all values of the variables in the sets* a, b, and c, *whenever* $P(\mathsf{c}) \neq 0$, *the events* $\mathsf{A} = \mathsf{a}$ *and* $\mathsf{B} = \mathsf{b}$ *are conditionally independent given the event* $\mathsf{C} = \mathsf{c}$. *That is, either* $P(\mathsf{a}|\mathsf{c}) = 0$ *or* $P(\mathsf{b}|\mathsf{c}) = 0$ *or*

$$P(\mathsf{a}|\mathsf{b},\mathsf{c}) = P(\mathsf{a}|\mathsf{c}).$$

When this is the case, we write

$$I_P(\mathsf{A},\mathsf{B}|\mathsf{C}).$$

Probability Basics → Naïve Bayes Models

- **Introduction**
- **Probability Basics**
  - Probability Axioms
  - Conditional probabilities
  - Product Rule, Chain Rule, Bayes Rule
- **Bayes Nets And Naïve Bayes**
  - Learning
  - Independence
  - Conditional independence
  - Naïve Bayes derivation (discrete case plus smoothing)
- **Naïve Bayes Flavors**
  - Continuous input variables
  - Discrete input variables (2 flavors: Bernoulli, multinomial)
- **Case Study: Spam detector in Naïve Bayes**
- **Summary**

Conditional Independence

Definition: Given random variables X, Y and Z, we say X is **conditionally independent** of Y given Z, if and only if the probability distribution governing X is independent of the value of Y given Z; that is

Knowlng Y given Z tells us nothing about X

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

As an example, consider three boolean random variables to describe the current weather: *Rain*, *Thunder* and *Lightning*. We might reasonably assert that *Thunder* is independent of *Rain* given *Lightning*. Because we know *Lightning* causes *Thunder*, once we know whether or not there is *Lightning*, no additional information about *Thunder* is provided by the value of *Rain*. Of course there is a clear dependence of *Thunder* on *Rain* in general, but there is no *conditional* dependence once we know the value of *Lightning*.

Pr(Rain | Lightning, Thunder) == Pr(Rain | Lightning)

P(R|L, T)= P(R|L)

Conditional independence notation

- **V is** conditionally independent **of set *Vi* given set *Vj***
  - if p(V | ***Vi***, ***Vj***) = p(V | ***Vj***)
  - notation: I(V, ***Vi*** | ***Vj***) or V ⊥ ***Vi*** | ***Vj***

- **Intuition**
  - if I(V, ***Vi*** | ***Vj***) then knowing ***Vi*** & ***Vj*** tells nothing more about V than knowing ***Vj*** alone
  - if we know ***Vj*** we can ignore ***Vi***

$$P(Queen \mid C\,\text{lub}) = \frac{P(Queen, C\,\text{lub})}{P(C\,\text{lub})} = \frac{\frac{1}{52}}{\frac{13}{52}} = \frac{1}{13}$$

$$P(Queen) = \frac{4}{52} = \frac{1}{13}$$

Conditionally Independent

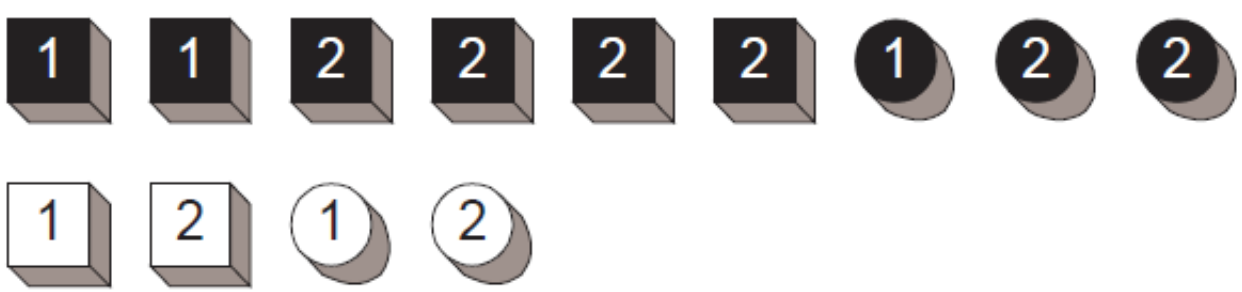


Figure 1.2: Containing a '1' and being a square are not independent, but they are conditionally independent given the object is black and given it is white.

Example 1.19 *Let Ω be the set of all objects in Figure 1.2, and let P assign $1/13$ to each object. Define random variables S (for shape), V (for value), and C (for color) as follows:*

Variable	Value	Outcomes Mapped to this Value
V	$v1$	*All objects containing a '1'*
	$v2$	*All objects containing a '2'*
S	$s1$	*All square objects*
	$s2$	*All round objects*
C	$c1$	*All black objects*
	$c2$	*All white objects*

Then we maintain that $\{V\}$ and $\{S\}$ are conditionally independent given $\{C\}$. That is,

$$I_P(\{V\}, \{S\}|\{C\}).$$

To show this, we need show for all values of v, s, and c that

$$P(v|s, c) = P(v|c).$$

The results in Example 1.8 show $P(v1|s1, c1) = P(v1|c1)$ and $P(v1|s1, c2) = P(v1|c2)$. The table that follows shows the equality holds for the other values of the variables too:

c	s	v	$P(v\|s,c)$	$P(v\|c)$
$c1$	$s1$	$v1$	$2/6 = 1/3$	$3/9 = 1/3$
$c1$	$s1$	$v2$	$4/6 = 2/3$	$6/9 = 2/3$
$c1$	$s2$	$v1$	$1/3$	$3/9 = 1/3$
$c1$	$s2$	$v2$	$2/3$	$6/9 = 2/3$
$c2$	$s1$	$v1$	$1/2$	$2/4 = 1/2$
$c2$	$s1$	$v2$	$1/2$	$2/4 = 1/2$
$c2$	$s2$	$v1$	$1/2$	$2/4 = 1/2$
$c2$	$s2$	$v2$	$1/2$	$2/4 = 1/2$

[Learning Bayesian Networks, Prentice-Hall, Richard E. Neapolitan]

Conditional Independence (Wiki)

Conditional independence [edit]

Main article: Conditional independence

Intuitively, two random variables *X* and *Y* are conditionally independent given *Z* if, once *Z* is known, the value of *Y* does not add any additional information about *X*. For instance, two measurements *X* and *Y* of the same underlying quantity *Z* are not independent, but they are **conditionally independent given *Z*** (unless the errors in the two measurements are somehow connected).

The formal definition of conditional independence is based on the idea of conditional distributions. If *X*, *Y*, and *Z* are discrete random variables, then we define *X* and *Y* to be *conditionally independent given Z* if

$$\mathrm{P}(X \leq x, Y \leq y \mid Z = z) = \mathrm{P}(X \leq x \mid Z = z) \cdot \mathrm{P}(Y \leq y \mid Z = z)$$

for all *x*, *y* and *z* such that $\mathrm{P}(Z = z) > 0$. On the other hand, if the random variables are continuous and have a joint probability density function *p*, then *X* and *Y* are conditionally independent given *Z* if

$$p_{XY|Z}(x, y|z) = p_{X|Z}(x|z) \cdot p_{Y|Z}(y|z)$$

for all real numbers *x*, *y* and *z* such that $p_Z(z) > 0$.

If *X* and *Y* are conditionally independent given *Z*, then

$$\mathrm{P}(X = x|Y = y, Z = z) = \mathrm{P}(X = x|Z = z)$$

for any *x*, *y* and *z* with $\mathrm{P}(Z = z) > 0$. That is, the conditional distribution for *X* given *Y* and *Z* is the same as that given *Z* alone. A similar equation holds for the conditional probability density functions in the continuous case.

Independence can be seen as a special kind of conditional independence, since probability can be seen as a kind of conditional probability given no events.

https://en.wikipedia.org/wiki/Conditional_independence

Conditional independence notation

- **V is** conditionally independent **of set *Vi* given set *Vj***
  - if p(V | ***Vi***, ***Vj***) = p(V | ***Vj***)
  - notation: I(V, ***Vi*** | ***Vj***) or V ⊥ ***Vi*** | ***Vj***
- **Intuition**
  - if I(V, ***Vi*** | ***Vj***) then knowing ***Vi*** & ***Vj*** tells nothing more about V than knowing ***Vj*** alone
  - if we know ***Vj*** we can ignore ***Vi***

$$P(Queen \mid C\,\text{lub}) = \frac{P(Queen, C\,\text{lub})}{P(C\,\text{lub})} = \frac{\frac{1}{52}}{\frac{13}{52}} = \frac{1}{13}$$

$$P(Queen) = \frac{4}{52} = \frac{1}{13}$$

Pairwise Independence

- **Single Vi conditionally independent of single Vj given *V***
  - that is, I(Vi, Vj | ***V***)
- **From definitions we have that**
  - p(Vi | Vj,**V**) = p(Vi | ***V***) and
  - p(Vi,Vj | ***V***) = p(Vi | Vj,**V**) p(Vj | ***V***)
    (Product Rule: $P(A,B) = P(A \mid B)P(B)$)
- **Thus**
  - p(Vi,Vj | ***V***) = p(Vi | ***V***) p(Vj | ***V***)
  - Vi and Vj is **pairwise independent**

Independence: Pairwise and Conditional

- **Pairwise Independence**
  - P(A|C) = Pr(A)
  - P(A,B|C)=P(A|C)P(B|C)
    - Since P(A,B|C) = P(A|B,C)P(B|C) [Chain rule]
    - and P(A|B, C)= P(A|C)
- **Conditional Independence**
  - P(A|B)=P(A)
  - P(A|B, C)= P(A|C)
  - P(*A, B*)=P(*A*)P(*B*)

Decompose joint probs using chain rule

Independence: Pairwise and Conditional

- **Pairwise Independence**
  - P(A|C) = Pr(A)
  - P(A,B|C)=P(A|C)P(B|C)
    - Since P(A,B|C) = P(A|B,C)P(B|C) [Chain rule]
    - and P(A|B, C)= P(A|C)
- **Conditional Independence**
  - P(A|B)=P(A)
  - P(A|B, C)= P(A|C)
  - P(*A, B*)=P(*A*)P(*B*)

Decompose joint probs using chain rule

Only a skip away from Naïve Bayes

Probability Basics → Naïve Bayes Models

- **Introduction**
- **Probability Basics**
  - Probability Axioms
  - Conditional probabilities
  - Product Rule, Chain Rule, Bayes Rule
- **Bayes Nets And Naïve Bayes**
  - Learning
  - Independence
  - Conditional independence
  - Naïve Bayes derivation (discrete case plus smoothing)
- **Naïve Bayes Flavors**
  - Continuous input variables
  - Discrete input variables (2 flavors: Bernoulli, multinomial)
- **Case Study: Spam detector in Naïve Bayes**
- **Summary**

Bayes Classifier: Simplify RHS

$$P(Y = y_k|X_1, X_2, ..., X_N) = \frac{P(Y=y_k)P(X_1,X_2,..,X_N|Y=y_k)}{\sum_j P(Y=y_j)P(X_1,X_2,..,X_N|Y=y_j)}$$

$$P(Y = y_k|X_1, X_2, ..., X_N) = \frac{P(Y=y_k)P(X_1,X_2,..,X_N|Y=y_k)}{\sum_j P(Y=y_j)P(X_1,X_2,..,X_N|Y=y_j)}$$

- **The denominator is common to all classes y_k so set to 1**

$$Y \leftarrow argmax_{y_k} P(Y = y_k) P(Y=y_k)P(X_1,X_2,..,X_N|Y=y_k)$$

argmax_yk means find the value of yk that maximises the expression

Bayes Classifier: Simplify RHS

$$P(Y = y_k | X_1, X_2, ..., X_N) = \frac{P(Y=y_k) P(X_1, X_2, .., X_N | Y=y_k)}{\sum_j P(Y=y_j) P(X_1, X_2, .., X_N | Y=y_j)}$$

$$P(Y = y_k | X_1, X_2, ..., X_N) = \frac{P(Y=y_k) P(X_1, X_2, .., X_N | Y=y_k)}{\sum_j P(Y=y_j) P(X_1, X_2, .., X_N | Y=y_j)}$$

- **The denominator is common to all classes y_k so set to 1**

$$Y \leftarrow argmax_{y_k} P(Y = y_k) P(X_1, X_2, .., X_N | Y=y_k)$$

BUT $P(X_1, X_2, .., X_N | Y=y_k)$ **is still complex**

Simplify the class conditional term
using the product rule and conditional independence

$$Y \leftarrow argmax_{y_k} P(Y = y_k) P(X_1, X_2, .., X_N | Y = y_k)$$

$$P(X_1, X_2, .., X_N | Y = y_k)$$

X is vector of <X1, X2>

Using the Product Rule
Simplify the Class Cond^{l} term

$$P(X|Y) = P(X_1, X_2|Y)$$

$$P(A, B) = P(A \mid B)P(B)$$

$$= P(X_1|X_2, Y)P(X_2|Y)$$

Assume cond^{al} independence.

$$= P(X_1|Y)P(X_2|Y)$$

Naïve Bayes

Independence
P(X|Y)=P(X)
P(*X*, *Y*)=P(*X*)P(*Y*)
$P(X_1 | X_2, Y) = P(X_1 | Y)$

Where
- the second line follows from a general property of probabilities (product Rule),
- the third line follows directly from our above definition of conditional independence.

Key Slide

Derive NB Algorithm

$$P(Y = y_k | X_1, X_2, ..., X_N) = \frac{P(Y=y_k) P(X_1, X_2, .., X_N | Y=y_k)}{\sum_j P(Y=y_j) P(X_1, X_2, .., X_N | Y=y_j)}$$

$$= \frac{P(Y=y_k) \Pi_i P(X_i | Y=y_k)}{\sum_j P(Y=y_j) \Pi_i P(X_i | Y=y_j)}$$

The Naive Bayes algorithm is a classification algorithm based on Bayes rule, that assumes the attributes $X_1 \ldots X_n$ are all conditionally independent of one another, given Y. The value of this assumption is that it dramatically simplifies the representation of $P(X|Y)$, and the problem of estimating it from the training data. Consider, for example, the case where $X = \langle X_1, X_2 \rangle$. In this case

Independence
P(X|Y)=P(X)
P(*X*, *Y*)=P(*X*)P(*Y*)
$P(X_1 | X_2, Y) = P(X_1 | Y)$

X is vector of <X1, X2>

$$P(X|Y) = P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y) = P(X_1|Y)P(X_2|Y)$$

Use Product Rule $P(A, B) = P(A \mid B)P(B$

Naïve Bayes

Where
- the second line follows from a general property of probabilities (product Rule),
- the third line follows directly from our above definition of conditional independence.

Bayesian Classifier → Naïve Bayes

$$Y \leftarrow argmax_{y_k} P(Y = y_k) \frac{P(X_1, X_2, .., X_N | Y = y_k)}{}$$

Multivariate Joint Prob → Production of univariate class conditionals

2^n → n. Probabilities to be learned from data for each class

$$Y \leftarrow argmax_{y_k} P(Y = y_k) \Pi_i P(X_i | Y = y_k)$$

argmax_yk means find the value of yk that maximises the expression

Conditional independence dramatically reduces model complexity: to n parameters from 2^n

More generally, when X contains n attributes which are conditionally independent of one another given Y, we have

$$P(X_1 \ldots X_n|Y) = \prod_{i=1}^{n} P(X_i|Y) \quad (1)$$

Notice that when Y and the X_i are boolean variables, we need only $2n$ parameters to define $P(X_i = x_{ik}|Y = y_j)$ for the necessary i, j, k. This is a dramatic reduction compared to the $2(2^n - 1)$ parameters needed to characterize $P(X|Y)$ if we make no conditional independence assumption.

- **High Bias**
- **Assume no interactions between the variables**

Naïve Bayes: summary so far

- **A generative, parametric model**
- **Computes the conditional a-posterior probabilities of a categorical class variable given independent predictor variables using the Bayes rule.**
  - So far focused on discrete input variables
  - See below how to model continous input variables

Probability Basics → Naïve Bayes Models

- **Introduction**
- **Probability Basics**
  - Probability Axioms
  - Conditional probabilities
  - Product Rule, Chain Rule, Bayes Rule
- **Bayes Nets And Naïve Bayes**
  - Learning
  - Independence
  - Conditional independence
  - Naïve Bayes derivation (discrete case plus smoothing)
- **Naïve Bayes Flavors**
  - Discrete input variables (2 flavors: Bernoulli, multinomial)
  - Continuous input variables
- **Case Study: Spam detector in Naïve Bayes**
- **Summary**

Probability Basics → Naïve Bayes Models

- **Introduction**
- **Probability Basics**
  - Probability Axioms
  - Conditional probabilities
  - Product Rule, Chain Rule, Bayes Rule
- **Bayes Nets And Naïve Bayes**
  - Learning
  - Independence
  - Conditional independence
  - Naïve Bayes derivation (discrete case plus smoothing)
- **Naïve Bayes Flavors**
  - Discrete input variables (2 flavors: Bernoulli, multinomial)
  - Continuous input variables
- **Case Study: Spam detector in Naïve Bayes**
- **Summary**

Parameter estimation for Naïve Bayes

See IRBook [Manning et al. 2008] Chapter 13 for more info

- **Multivariate Bernoulli model:**

X_w = t if word present document

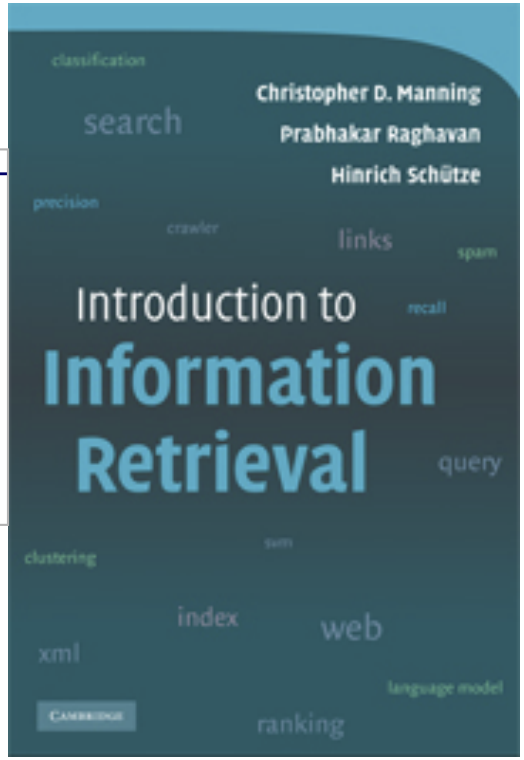
$$\hat{P}(X_w = t \mid c_j) = \text{fraction of documents of topic } c_j \text{ in which word } w \text{ appears}$$

- **Multinomial model:**

$$\hat{P}(X_i = w \mid c_j) = \text{fraction of times in which word } w \text{ appears across all documents of topic } c_j$$

  - Can create a mega-document for topic *j* by concatenating all documents on this topic
  - Use frequency of *w* in mega-document

Naïve Bayes for text



PDF and HTML versions of Chapter 13 are available here
[Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.] http://nlp.stanford.edu/IR-book/

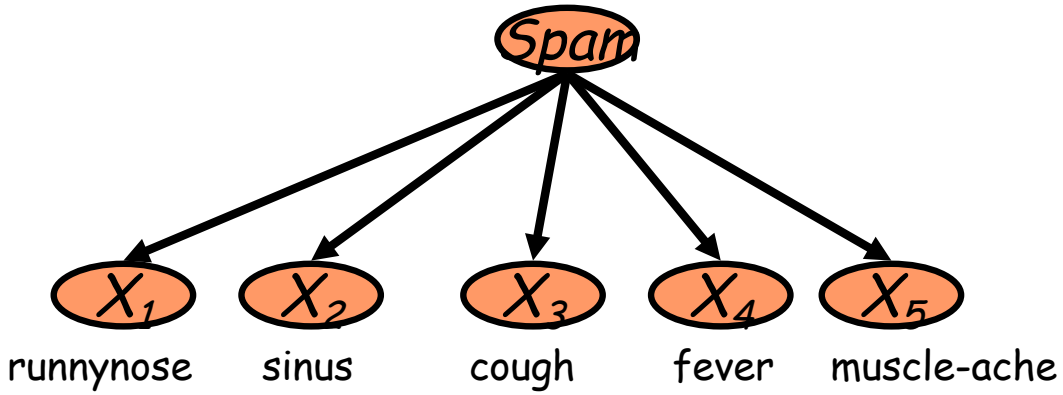
http://www.cs.cmu.edu/~epxing/Class/10701-10s/Lecture/lecture5.pdf
http://www.cs.cmu.edu/~tom/10701_sp11/lectures.shtml

Reading material

- **PDF and HTML versions of Chapter 13 are available here**
  - [Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.] http://nlp.stanford.edu/IR-book/
  - http://www.cs.cmu.edu/~epxing/Class/10701-10s/Lecture/lecture5.pdf
  - http://www.cs.cmu.edu/~tom/10701_sp11/lectures.shtml

The Naïve Bayes Classifier

IRBook 13.3



- **Conditional Independence Assumption: Features (term presence) are *independent* of each other given the class:**

$$P(X_1,\ldots,X_5 \mid C) = P(X_1 \mid C) \bullet P(X_2 \mid C) \bullet \cdots \bullet P(X_5 \mid C)$$

Learn NB Model as a dict with two columns

▶ **Table 13.1** Data for parameter estimation examples.

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

6 is the vocabulary size

Word	Pr(Word|Class)	Pr(Word|NotClass)
CLASSPRIORs	¾ = 0.75	¼ = 0.25
Tokyo	0/(8) =0	1/3
Chinese	5/8	1/3
…..		

wordcount = {}

Use model to predict class of document

▶ **Table 13.1** Data for parameter estimation examples.

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

6 is the vocabulary size

Word	Pr(Word|Class)	Pr(Word|NotClass)
CLASSPRIORs	¾ = 0.75	¼ = 0.25
Tokyo	0/(8) =0	1/3
Chinese	5/8	1/3
…..		

Classiify Document={Tokyo, Chinese}
Pr(Tokyo|Class) = 0/8

$$Y \leftarrow argmax_{y_k} P(Y = y_k) \Pi_i P(X_i | Y = y_k)$$

wordcount = {}

Model file can be a dict with two columns

▶ **Table 13.1** Data for parameter estimation examples.

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

6 is the vocabulary size

Word	Pr(Word|Class)	Pr(Word|NotClass)
CLASSPRIORs	¾ = 0.75	¼ = 0.25
Tokyo	0/(8) =0	1/3
Chinese	5/8	1/3
…..		

Classiify Document={Tokyo, Chinese}
Pr(Tokyo|Class) = 0/8

$$Y \leftarrow argmax_{y_k} P(Y = y_k) \Pi_i P(X_i | Y = y_k)$$

Run into trouble during classification if we have no estimate θ_{ijk} for variable X_i with value j for class k. The whole expression go to 0 for that class k.

wordcount = {}

Model file can be a dict with two columns

► Table 13.1 Data for parameter estimation examples.

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

Plus 1 smoothing

+6 is the vocabulary size
Prior 1/6 for each word
For both classes

Word	Pr(Word|Class)	Pr(Word|NotClass)
CLASSPRIORs	¾ = 0.75	¼ = 0.25
Tokyo	0+1/(8+6) =1/14	1+1/(3+6) = 2/9
Chinese	5+1/(8+6) = 6/14	1+1/(3+6) = 2/9
…..		

Pr(Tokyo|Class) = 0/8 → 0+1/(8+6)

Laplace Smoothing

▶ **Table 13.1** Data for parameter estimation examples.

	docID	words in document	in c = *China*?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

Plus 1 smoothing

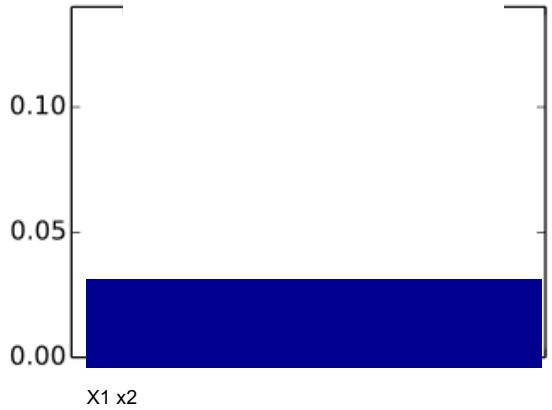
+6 is the vocabulary size
Prior 1/6 for each word
For both classes

Word	Pr(Word|Class)	Pr(Word|NotClass)
CLASSPRIORs	¾ = 0.75	¼ = 0.25
Tokyo	0+1/(8+6) =1/14	1+1/(3+6) = 2/9
Chinese	5+1/(8+6) = 6/14	1+1/(3+6) = 2/9
…..		

MAP Estimates
Learning a NB Via Bayesian hierarchical model

Pr(Tokyo|Class) = 0/8 → 0+1/(8+6)

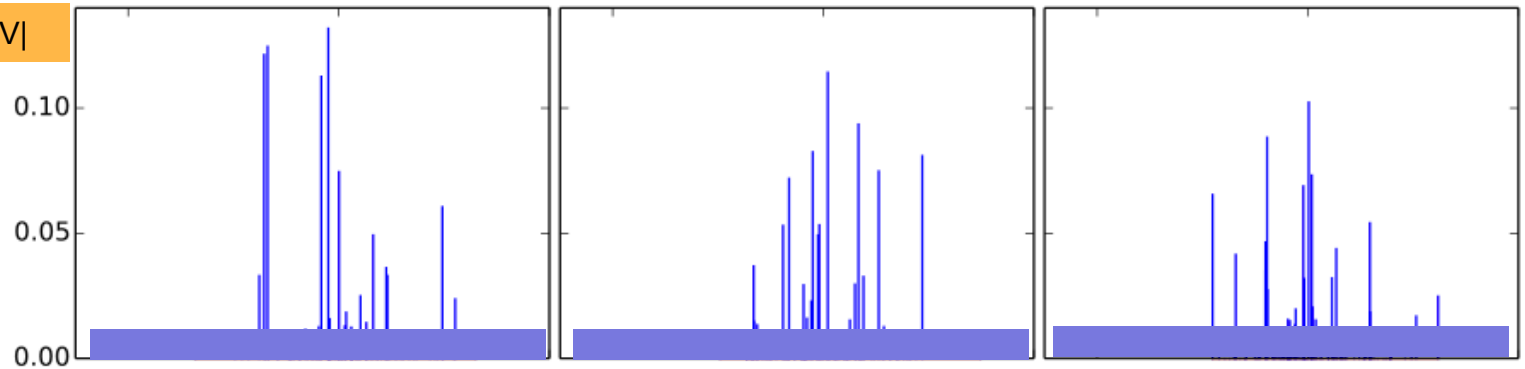
MAP Estimates: Smoothed Probabilities



Uniform Prior: Pr(X=xj|Y=yk) = 1/|V|

+

Training data

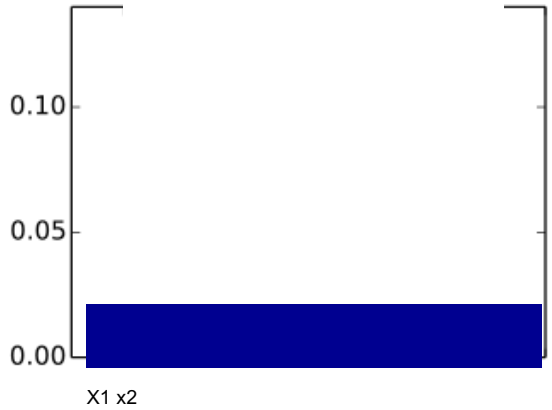


Pr(X=xj|Y=yk) = 1/|V|

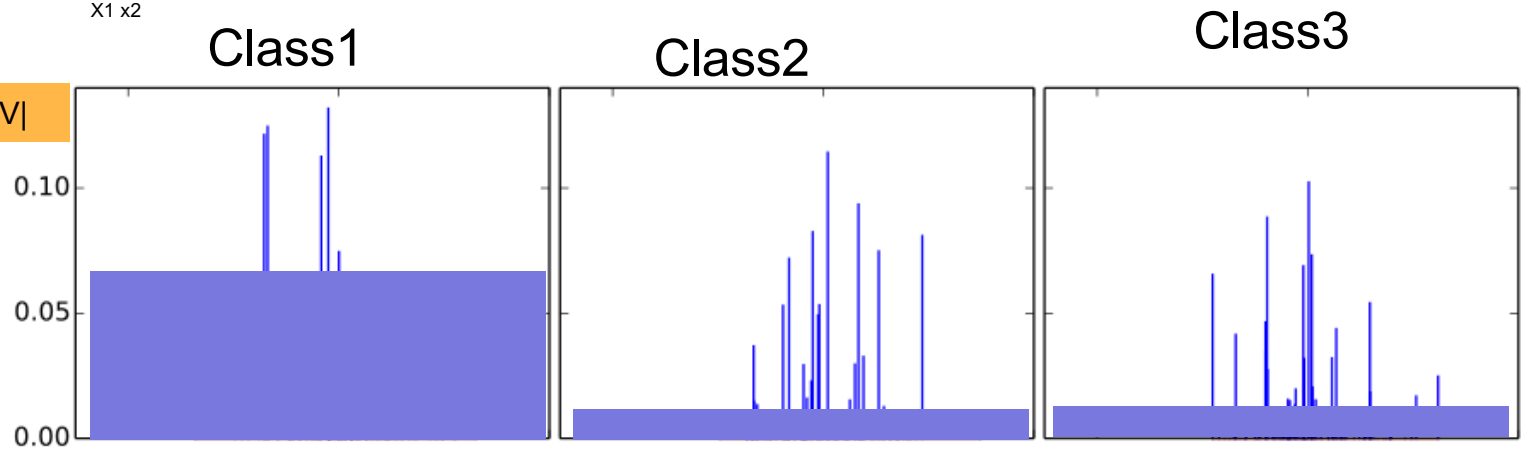
MAP Estimates

Huge Probability mass assigned to background model:
Bias versus variance?

MAP Estimates: Smoothed Probabilities



Uniform Prior: Pr(X=xj|Y=yk) = 1/|V|
+
Training data

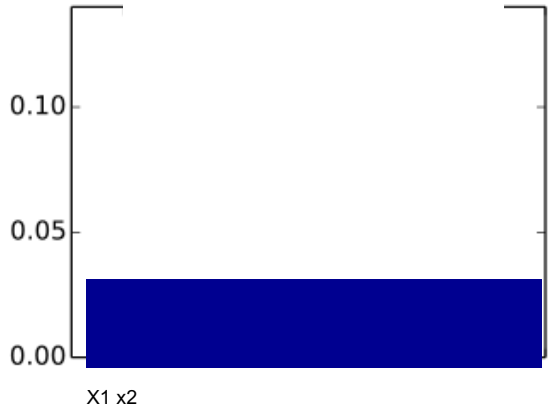


Pr(X=xj|Y=yk) = 1/|V|

MAP Estimates

Huge Probability mass assigned to background model:
Bias versus variance?

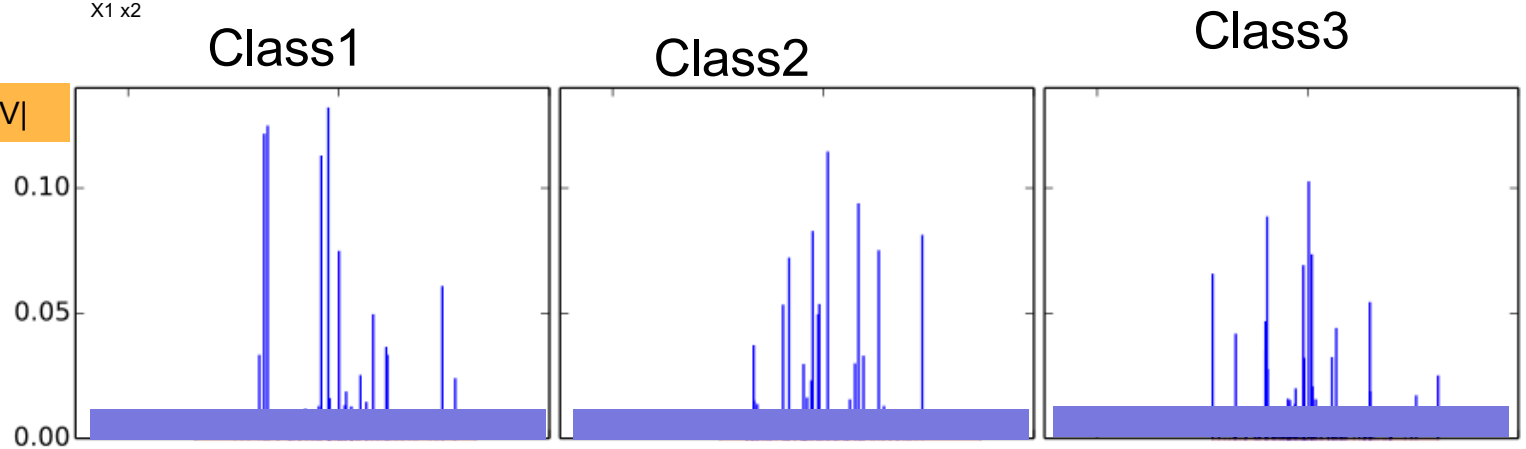
MAP Estimates: Smoothed Probabilities



Uniform Prior: Pr(X=xj|Y=yk) = 1/|V|
+
Training data

Pr(X=xj|Y=yk) = 1/|V|

MAP Estimates



Huge Probability mass assigned to background model:
Bias versus variance? High Bias!

Class Prior Estimates

Maximum likelihood estimates for π_k are

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|} \tag{8}$$

where $|D|$ denotes the number of elements in the training set D.

Alternatively, we can obtain a smoothed estimate, or equivalently a MAP estimate based on a Dirichlet prior over the π_k parameters assuming equal priors on each π_k, by using the following expression

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + l}{|D| + lK} \tag{9}$$

where K is the number of distinct values Y can take on, and l again determines the strength of the prior assumptions relative to the observed data D.

Probability Basics → Naïve Bayes Models

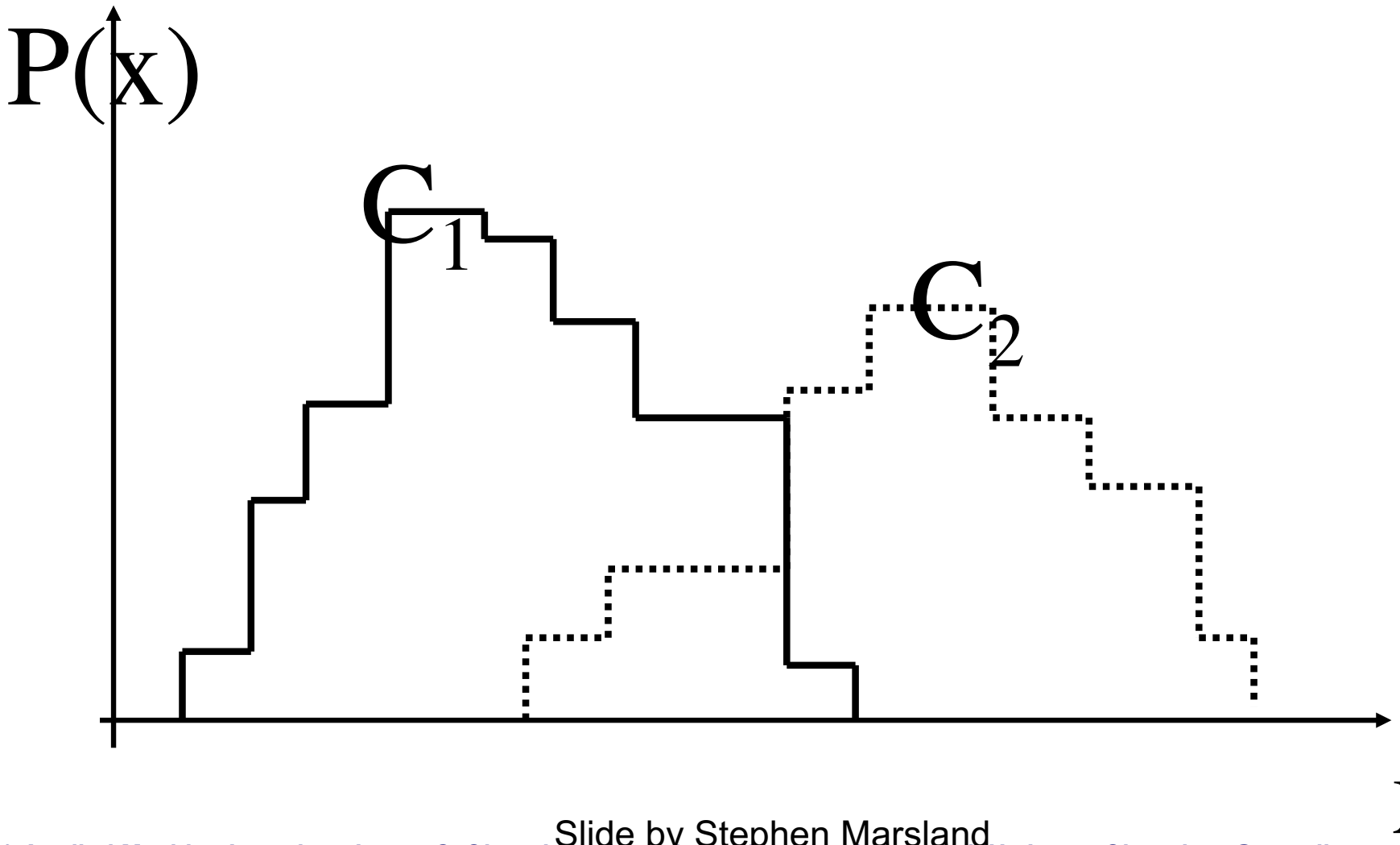
- **Introduction**
- **Probability Basics**
  - Probability Axioms
  - Conditional probabilities
  - Product Rule, Chain Rule, Bayes Rule
- **Bayes Nets And Naïve Bayes**
  - Learning
  - Independence
  - Conditional independence
  - Naïve Bayes derivation (discrete case plus smoothing)
- **Naïve Bayes Flavors**
  - Discrete input variables (2 flavors: Bernoulli, multinomial)
  - Continuous input variables
- **Case Study: Spam detector in Naïve Bayes**
- **Summary**

Naive Bayes for Continuous Inputs

- **When the Xi are continuous we must choose some other way to represent the distributions**
  - P(Xi|Y)=Gaussian(Mu, Sigma).
- **Mix discrete variables with continuous variables**
  - One common approach is to assume that for each possible discrete value yk of Y, the distribution of each continuous Xi is Gaussian, and is defined by a mean and standard deviation specific to Xi and yk.
- **In order to train such a Naïve Bayes classifier we must therefore estimate the mean and standard deviation of each of these Gaussians:**

$$\mu_{ik} = E[X_i | Y = y_k]$$
$$\sigma_{ik}^2 = E[(X_i - \mu_{ik})^2 | Y = y_k]$$

Feature Histograms



Slide by Stephen Marsland

MLE-based Estimates for μ and σ

Continuous Inputs: Pr(Y|X) ~ N(μ, σ)

Again, we can use either maximum likelihood estimates (MLE) or maximum a posteriori (MAP) estimates for these parameters. The maximum likelihood estimator for μ_{ik} is

$$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k) \quad (13)$$

where the superscript j refers to the jth training example, and where $\delta(Y = y_k)$ is 1 if $Y = y_k$ and 0 otherwise. Note the role of δ here is to select only those training examples for which $Y = y_k$.

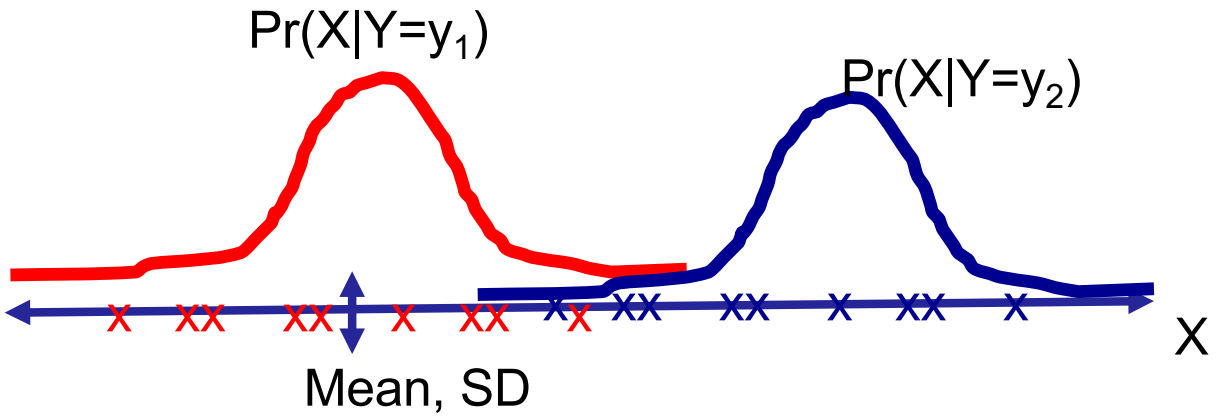
The maximum likelihood estimator for σ_{ik}^2 is

$$\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k) \quad (14)$$

Estimate μ, σ from data

$$\mu_{ik} = E[X_i | Y = y_k]$$

$$\sigma_{ik}^2 = E[(X_i - \mu_{ik})^2 | Y = y_k]$$



Continuous Inputs: Pr(Y|X) ~ N(μ, σ)

If X is Normally distributed with mean μ and standard deviation σ, we write

$$\boxed{X \sim N(\mu, \sigma^2)}$$

μ and σ are the **parameters** of the distribution.

The probability density of the Normal distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-(x-\mu)^2/2\sigma^2}$$

For the purposes of this course we do not need to use this expression. It is included here for future reference.

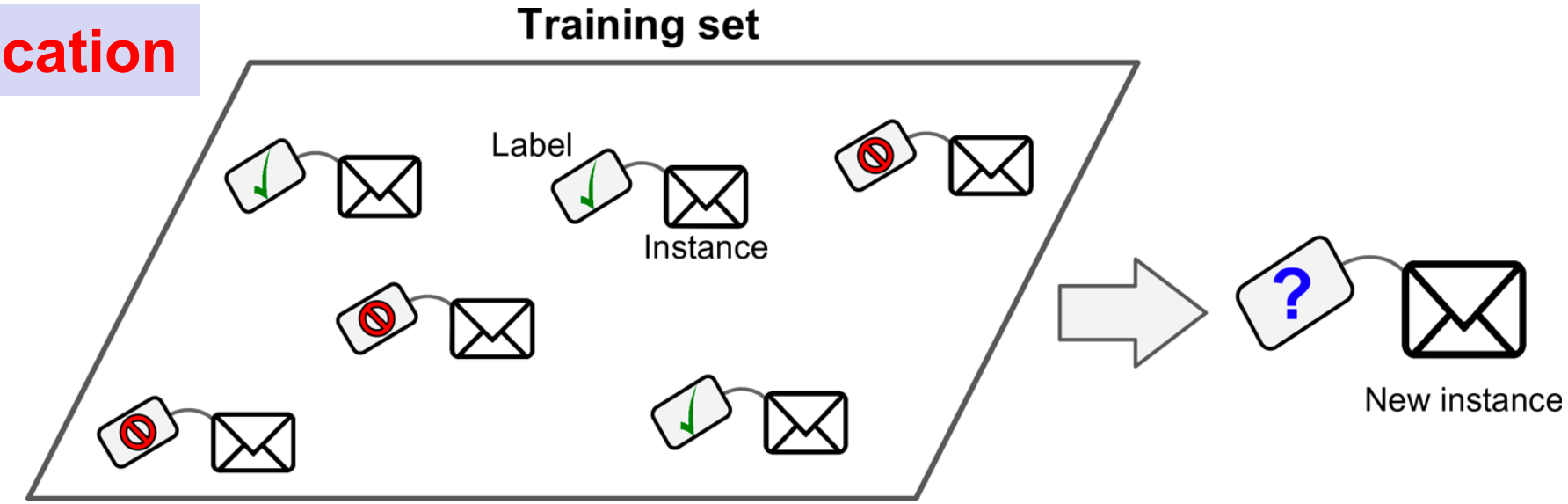
Naïve Bayes

- **Combine discrete input variables with continuous input variables?**
  - Naïve Bayes, Decision trees
- **YES**
- **Whereas these requires feature transformations**
  - Logistic regression: one hot-encoding

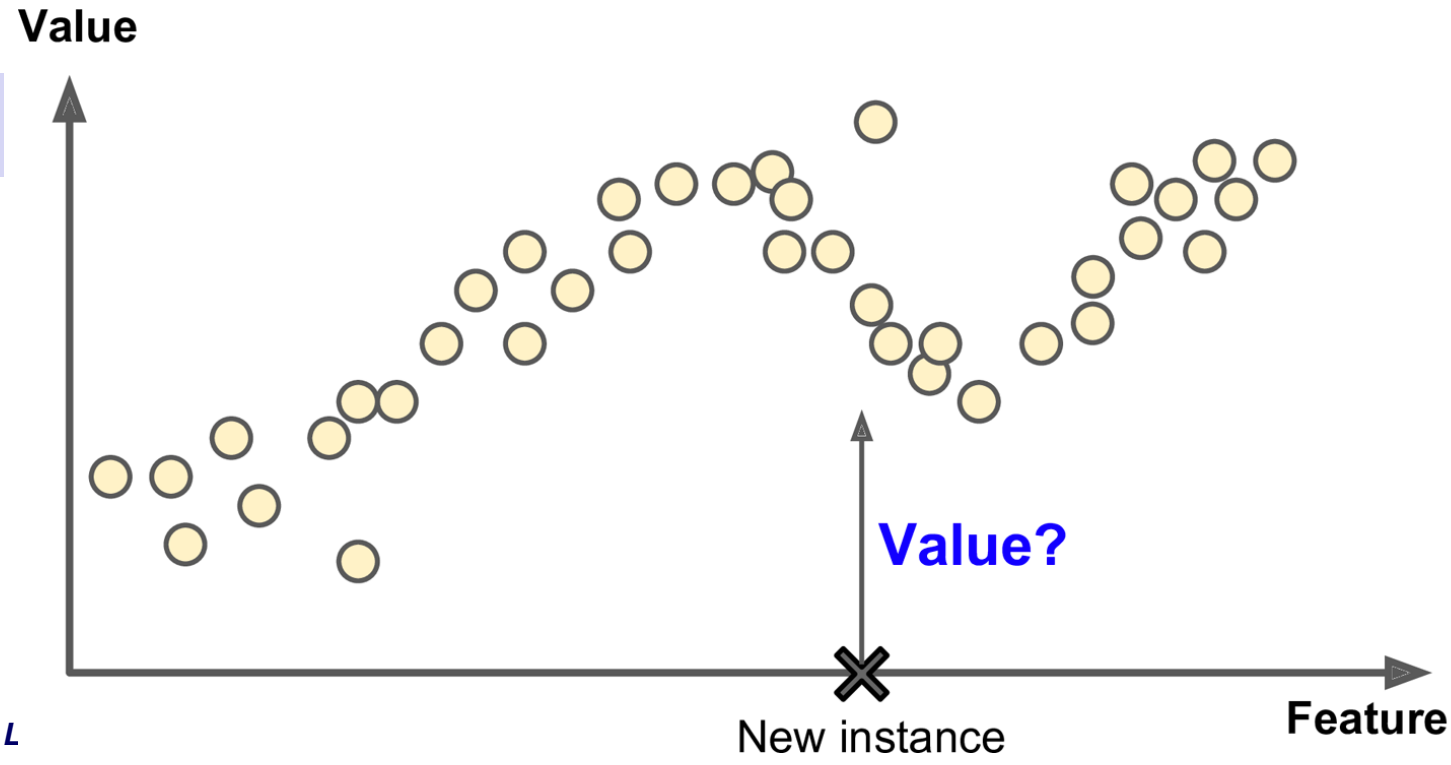
Probability Basics → Naïve Bayes Models

- **Introduction**
- **Probability Basics**
  - Probability Axioms
  - Conditional probabilities
  - Product Rule, Chain Rule, Bayes Rule
- **Bayes Nets And Naïve Bayes**
  - Learning
  - Independence
  - Conditional independence
  - Naïve Bayes derivation (discrete case plus smoothing)
- **Naïve Bayes Flavors**
  - Discrete input variables (2 flavors: Bernoulli, multinomial)
  - Continuous input variables
- **Case Study: Spam detector in Naïve Bayes**
- **Summary**

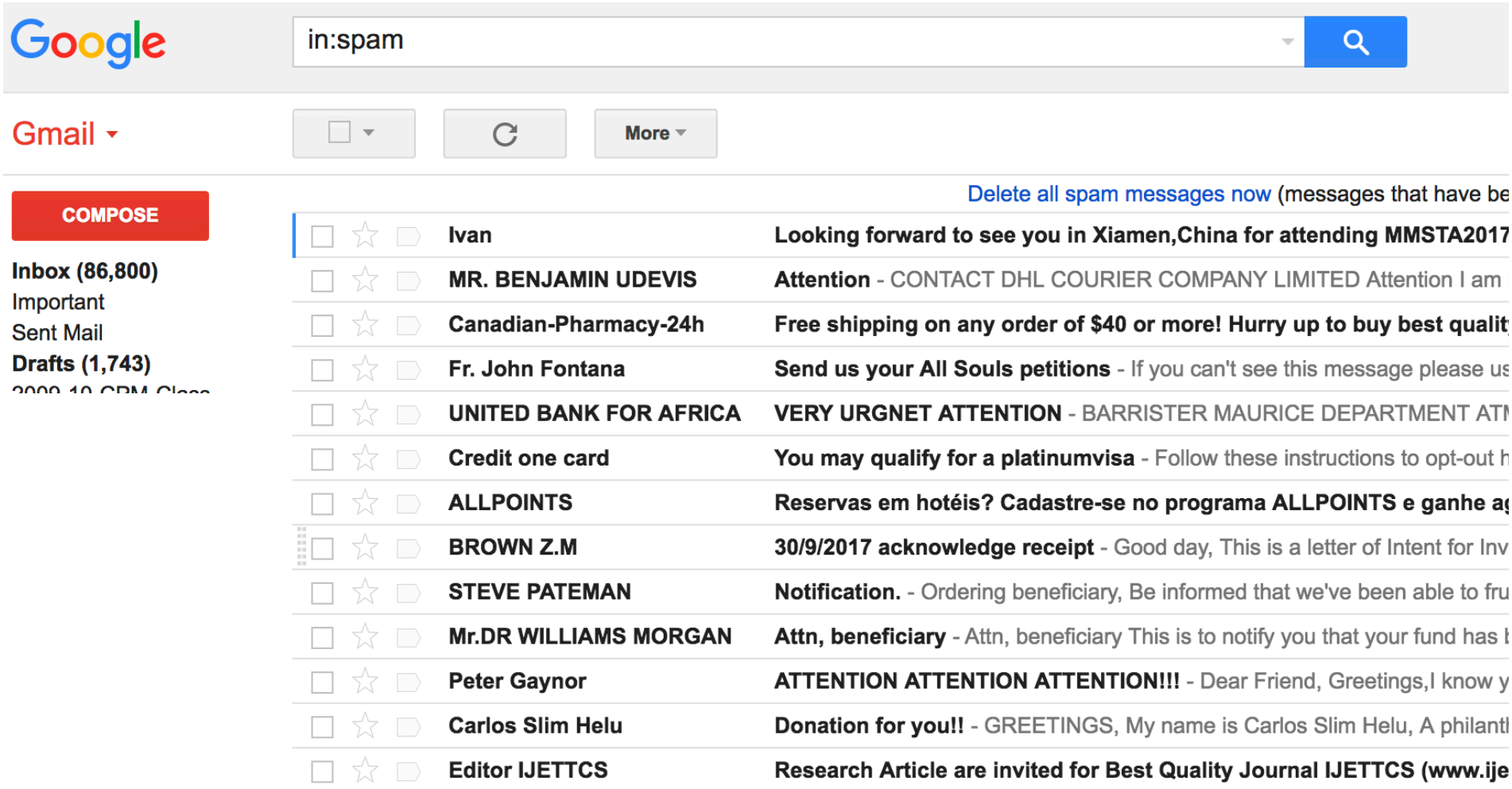
Classification



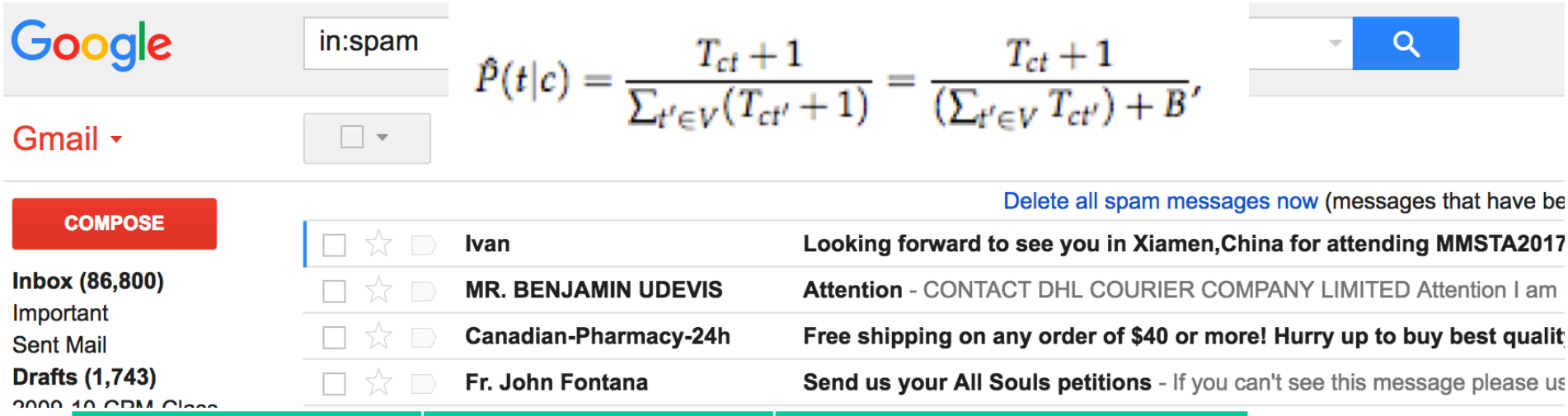
Regression



Email: SPAM versus Ham



Email: SPAM versus Ham



$$\hat{P}(t|c) = \frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'}+1)} = \frac{T_{ct}+1}{(\sum_{t' \in V} T_{ct'}) + B'},$$

Word	Pr(Word|Class)	Pr(Word|NotClass)
CLASSPRIORs	¾ = 0.75	¼ = 0.25
Tokyo	0+1/(8+6) =1/14	1+1/(3+6) = 2/9
Chinese	5+1/(8+6) = 6/14	1+1/(3+6) = 2/9
…..		

Probability Basics → Naïve Bayes Models

- **Introduction**
- **Probability Basics**
  - Probability Axioms
  - Conditional probabilities
  - Product Rule, Chain Rule, Bayes Rule
- **Bayes Nets And Naïve Bayes**
  - Learning
  - Independence
  - Conditional independence
  - Naïve Bayes derivation (discrete case plus smoothing)
- **Naïve Bayes Flavors**
  - Discrete input variables (2 flavors: Bernoulli, multinomial)
  - Continuous input variables
- **Case Study: Spam detector in Naïve Bayes**
- **Summary**

Summary

- Bayesian prediction:
  - requires solving density estimation problems.
  - often difficult to estimate $\Pr[\mathbf{x} \mid y]$ for $\mathbf{x} \in \mathbb{R}^N$.
  - but, simple and easy to apply; widely used.
- Naive Bayes:
  - strong assumption.
  - straightforward estimation problem.
  - specific linear classifier.
  - sometimes surprisingly good performance.

From complex class conditional joint probability distributions of the order 2^n to n

- **From complex class conditional joint probability distributions of the order 2^n to n**

- **2^n Possibilities -> n combinations that we have to estimate class conditional probabilities for**

Decompose Bayesian Networks

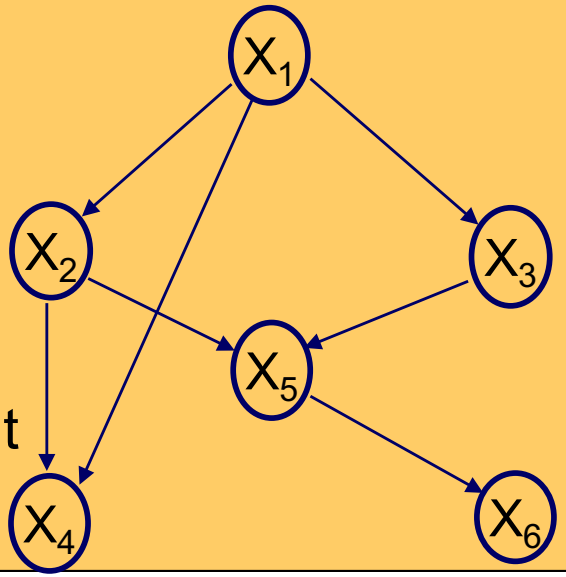
GOAL: 2^n Possibilities -> n

P: Joint Probability Distribution

#	X1	X2	X3	X4	X5	X6	Pr(X1,X2, X3, X4, X5, X6)
1	1						
2	1						
..	1						
4..	1						
5	1						
6	1						
7	1						
64							

Factorize the joint

G: Directed Acyclic Graph



$p(x_1, x_2, x_3, x_4, x_5, x_6)$ **1. Partial Order**

2. By Chain Rule

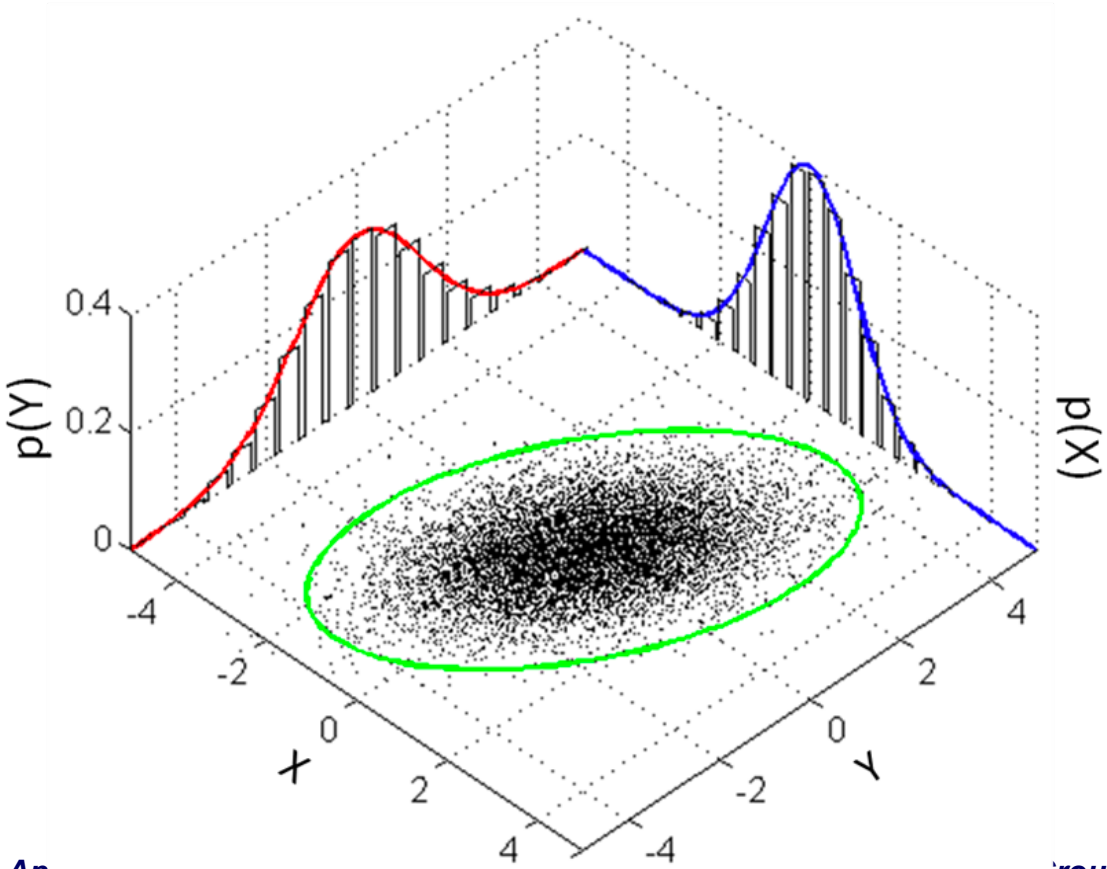
$= p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1, x_2)\, p(x_4 \mid x_1, x_2, x_3)\, p(x_5 \mid x_4,x_3,x_2,x_1)\, p(x_6 \mid x_5,x_4,x_3,x_2,x_1)$

$= p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1)\, p(x_4 \mid x_2, x_1)\, p(x_5 \mid x_3, x_2)\, p(x_6 \mid x_5)$ **3. Markov Property**

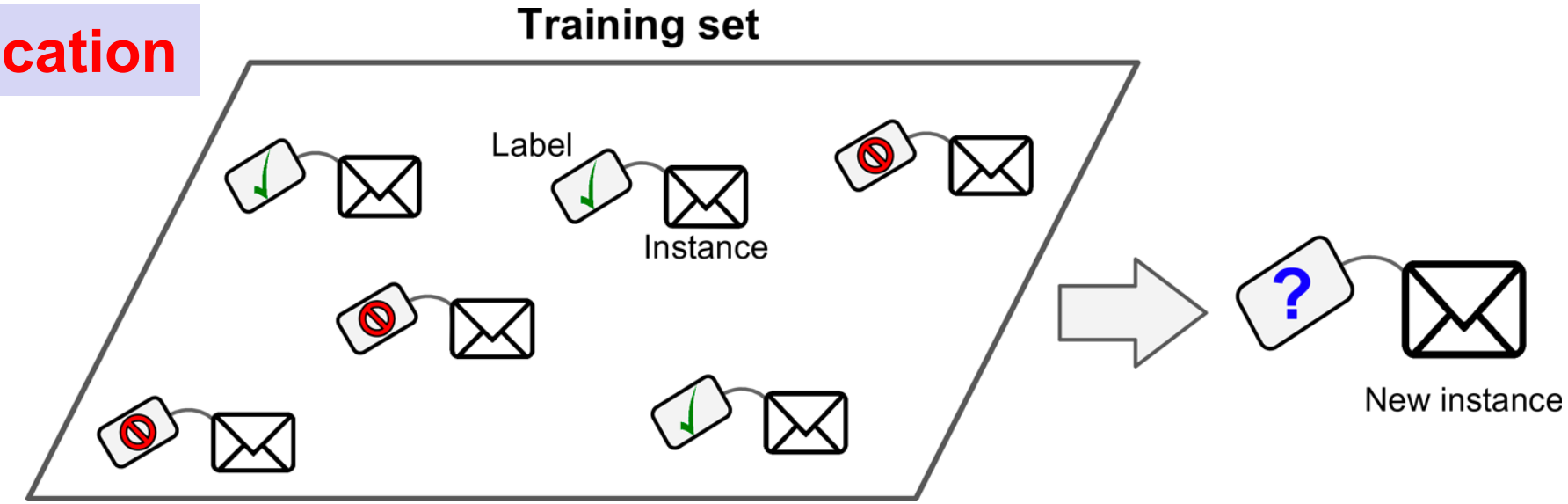
$= p(x_1)\, p(x_2)\, p(x_3)\, p(x_4)\, p(x_5)\, p(x_6)$ **4: Independence (see next section)**

$P(y|x_6, x_5,x_4,x_3,x_2,x_1) = p(x_1|y)\, p(x_2|y)\, p(x_3|y)\, p(x_4|y)\, p(x_5|y)\, p(x_6|y)$ **5: Naïve Bayes via Cond. Independence**

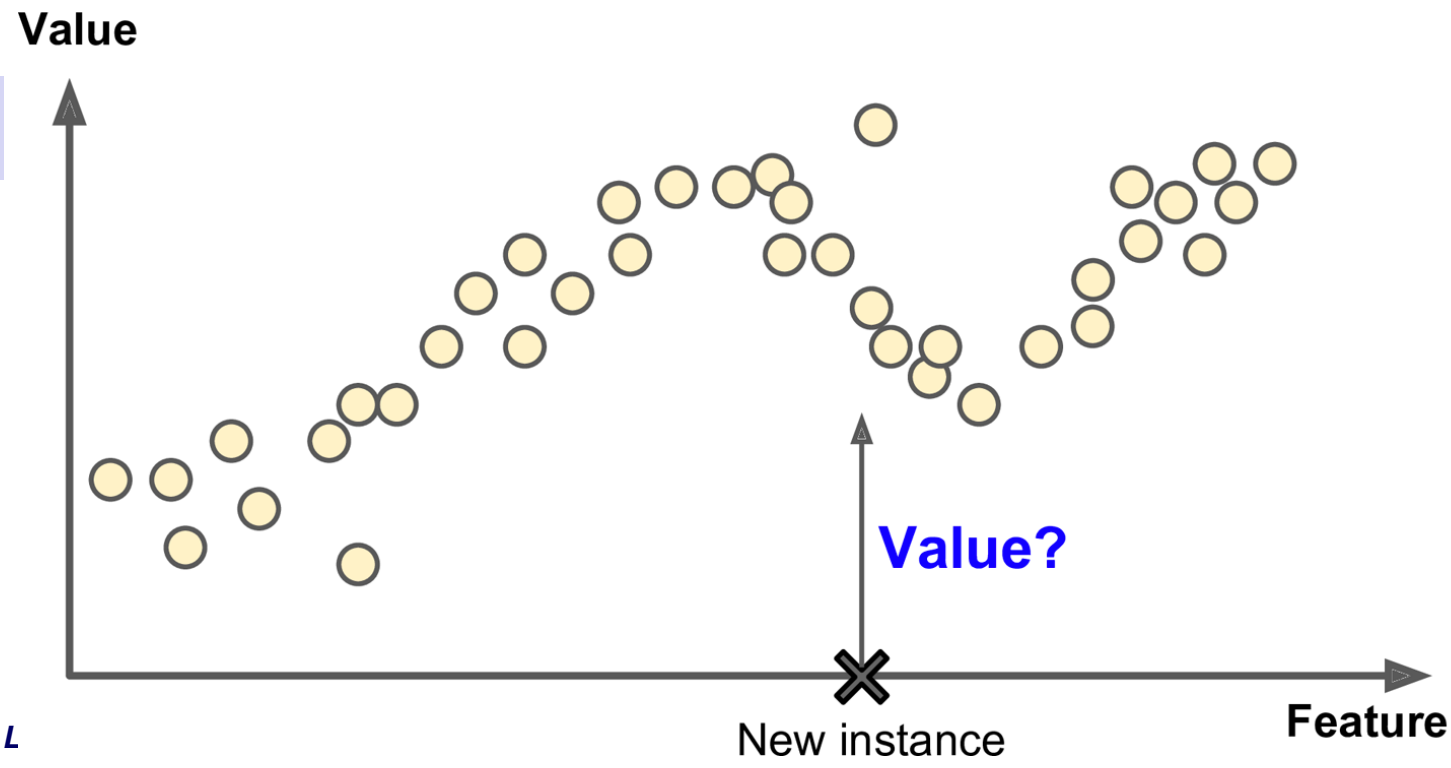
- **Decompose large joint distributions of many variables into production many univariate distributions**



Classification



Regression



Comparison

assuming x in {0 1}

	Learning Objective	Training	Inference
Naïve Bayes	$\text{maximize} \sum_i \left[\sum_j \log P(x_{ij} \mid y_i; \theta_j) + \log P(y_i; \theta_0) \right]$	$\theta_{kj} = \frac{\sum_i \delta(x_{ij} = 1 \wedge y_i = k) + r}{\sum_i \delta(y_i = k) + Kr}$	$\theta_1{}^T \mathbf{x} + \theta_0{}^T (1 - \mathbf{x}) > 0$ where $\theta_{1j} = \log \frac{P(x_j = 1 \mid y = 1)}{P(x_j = 1 \mid y = 0)}$, $\theta_{0j} = \log \frac{P(x_j = 0 \mid y = 1)}{P(x_j = 0 \mid y = 0)}$
Logistic Regression	$\text{maximize} \sum_i \log(P(y_i \mid \mathbf{x}, \boldsymbol{\theta})) + \lambda \lVert \boldsymbol{\theta} \rVert$ where $P(y_i \mid \mathbf{x}, \boldsymbol{\theta}) = 1/(1 + \exp(-y_i \boldsymbol{\theta}^T \mathbf{x}))$	Gradient ascent	$\boldsymbol{\theta}^T \mathbf{x} > 0$
Linear SVM	$\text{minimize } \lambda \sum_i \xi_i + \frac{1}{2} \lVert \boldsymbol{\theta} \rVert$ such that $y_i \boldsymbol{\theta}^T \mathbf{x} \geq 1 - \xi_i \quad \forall i$	Linear programming	$\boldsymbol{\theta}^T \mathbf{x} > 0$
Kernelized SVM	complicated to write	Quadratic programming	$\sum_i y_i \alpha_i K(\hat{\mathbf{x}}_i, \mathbf{x}) > 0$
Nearest Neighbor	most similar features → same label	Record data	y_i where $i = \underset{i}{\text{argmin}} \; K(\hat{\mathbf{x}}_i, \mathbf{x})$

Slide credit: D. Hoiem

End of lecture