

## Probability Theory

Chen Yu  
Indiana University

## Probability

$P(A)$  as the fraction of possible worlds in which  $A$  is true.

The axioms of probability

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

$$(2) \quad p(A \vee B) = p(A) + P(B) - P(A \wedge B)$$

Two theorems from the Axioms:

$$p(\neg A) = 1 - P(A)$$

$$p(A) = p(A \wedge B) + p(A \wedge \neg B)$$

## Example

- The results of a race are captured by a random variable  $R$ , with sample space {win,lose}. A horse Harry won 20 of 100 races, then we could say  
 $P(R=\text{win}) = 20/100=0.2$ ;  
 $P(R=\text{lose})=20/100=0.8$ ;
- To represent richer situations, we might care what the weather was like at each race (with a random variable  $W$  with values rain or shine). To model how these two variables interact, we need a **joint probability distribution** that combines the two random variables  $R$  and  $W$ :  
 $\langle \text{win, rain} \rangle \langle \text{lose, rain} \rangle \langle \text{win, shine} \rangle \langle \text{lose, shine} \rangle$

## Example

- It rained in 30 of Harry's races, and he won 15 of them:  
 $P(R=\text{win}, W=\text{rain})=15/100=0.15$ ;  
 $P(R=\text{win}, W=\text{shine}) = 5/100 = 0.05$ ;  
 $P(R=\text{lose}, W=\text{rain})=15/100=0.15$ ;  
 $P(R=\text{lose}, W=\text{shine})=65/100=0.65$ ;
- Given a joint distribution, we can derive the probability distributions for each single random variable (called **marginal probability**):  
 $P(R=\text{win}) = P(R=\text{win}, W \in \{\text{rain, shine}\})$   
 $= P(R=\text{win}, W=\text{rain}) + P(R=\text{win}, W=\text{shine})$

### Conditional Probability

- If it is raining, what is the probability that Harry will win?

$$P(R=win | W=rain) = P(R=win, W=rain) / P(W=rain)$$

$$= 15/30 = 0.5$$

- The conditional probability is written as P(A|B)  
 $P(A|B) = P(A, B) / P(B)$
- Almost all our knowledge about the world is in the form of conditional probabilities, it is relative to a certain context.

### Conditional Probability

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

$$p(A \wedge B) = p(A | B) p(B)$$

### The Berkeley Restaurant Project (BeRP)

eat on	.16	eat Thai	.03
eat some	.06	eat breakfast	.03
eat lunch	.06	eat in	.02
eat dinner	.05	eat Chinese	.02
eat at	.04	eat Mexican	.02
eat a	.04	eat tomorrow	.01
eat Indian	.04	eat dessert	.007
eat today	.03	eat British	.001

<start> I	.25	Want some	.04
<start> I'd	.06	Want Thai	.01
<start> Tell	.04	To eat	.26
<start> I'm	.02	To have	.14
I want	.32	To spend	.09
I would	.29	To be	.02
I don't	.08	British food	.60
I have	.04	British restaurant	.15
Want to	.65	British cuisine	.01
Want a	.05	British lunch	.01

## What are calculated?

- What's being captured with ...
  - $P(\text{want} | I) = .32$
  - $P(\text{to} | \text{want}) = .65$
  - $P(\text{eat} | \text{to}) = .26$
  - $P(\text{food} | \text{Chinese}) = .56$
  - $P(\text{lunch} | \text{eat}) = .055$
- What about...
  - $P(I | I) = .0023$
  - $P(I | \text{want}) = .0025$
  - $P(I | \text{food}) = .013$

- $P(I | I) = .0023$  I I I want
- $P(I | \text{want}) = .0025$  I want I want
- $P(I | \text{food}) = .013$  the kind of food I want is ...

## Applications

- Why do we want to predict a word, given some preceding words?
  - Rank the **likelihood** of sequences containing various alternative hypotheses,  
Theatre owners say %?#corn sales have doubled...
  - Theatre owners say popcorn/unicorn sales have doubled...

## More Realistic

- To make the above example more realistic, we should include many other variables as we use to represent the world. For instance, we might represent what Harry had for lunch, or what Harry's current health is, or how Harry did in his last race.
- The key issues to formulate any problem is selecting and identifying what random variables to use, and when we formulate knowledge, knowing what the dependencies are.

## Independence

- The variables, such as when we got up in the morning, and how we got to the raceback, should not affect how Harry performs in the race, unless a few of us believe in lucky omens. This is the notion of independence.
- Two random variables are independent if the probability distribution of one isn't affected by the values of the other.  
e.g. G states whether we drove or walked to the racetrack. We believe:  
 $P(R|G)=P(R)$
- Making independence assumptions is critical in building probabilistic models for it allows us to ignore certain information that is not relevant.

## Independence

- An equivalent way to define independence relates the joint distribution to the marginal distributions. In particular, random variables X and Y are independent if  
 $P(X,Y)=P(X)*P(Y)$
- A new random variable indicates whether I attend the race or not (at home). I go to the track 60 times, and Harry wins 12 races. This means that the probability that Harry wins if I attend the race is  $12/60=.2$ . The same as if I don't attend, and the same probability over all races. Thus we can prove that the random variable R is independent of the random variable A  
 $P(R=win|A=track) = 12/60 = .2$   
 $P(R=win|A=home)=8/40 = .2$   
 $P(R|A) = P(R)$

## Bayes' theorem

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

but  $P(A \cap B) = P(B \cap A)$ , so

Bayes' theorem

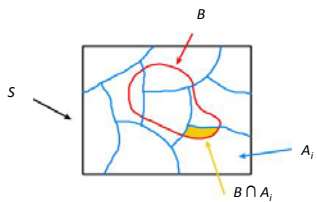
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



## Bayes Rule

- $P(A|B)=(P(B|A)*P(A))/P(B)$
- Let's assume that we do know the probability that he wins any races(.2), the probability that he wins when its raining (.5), and the probability that it rains (.3). We can calculate the probability that it rained on a day when Harry won a race:

$$P(\text{rain} | \text{win}) = (P(\text{win} | \text{rain}) * P(\text{rain})) / P(\text{win}) \\ = .5 * .3 / .2 = .75$$



- $B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i)$ ,
- $P(B) = P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i)$
- $P(B) = \sum_i P(B|A_i)P(A_i)$  law of total probability

Bayes' theorem becomes

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

### Bayes rule (1763)

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

More general forms of Bayes rule

$$(1) p(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)}$$

$$(2) p(A|B \wedge C) = \frac{P(B|A \wedge C)P(A \wedge C)}{P(B \wedge C)}$$

$$(3) p(A = v_i | B) = \frac{P(B|A = v_i)P(A = v_i)}{\sum_{j=1}^k p(B|A = v_j)P(A = v_j)}$$

### An example using Bayes' theorem

Suppose the probability (for anyone) to have AIDS is:

$$P(\text{AIDS}) = 0.001$$

$$P(\text{no AIDS}) = 0.999$$

Consider an AIDS test: result is + or -

$$P(+|\text{AIDS}) = 0.98$$

$$P(-|\text{AIDS}) = 0.02$$

$$P(+|\text{no AIDS}) = 0.03$$

$$P(-|\text{no AIDS}) = 0.97$$

Suppose your result is +. How worried should you be?

### Bayes' theorem example (cont.)

The probability to have AIDS given a + result is

$$P(\text{AIDS}|+) = \frac{P(+|\text{AIDS})P(\text{AIDS})}{P(+|\text{AIDS})P(\text{AIDS}) + P(+|\text{no AIDS})P(\text{no AIDS})}$$

$$= \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.03 \times 0.999}$$

$$= 0.032$$

You're probably OK!

### Discrete Random Variables

- Bernoulli: the distribution of a single binary variable

e.g. flip a coin, head  $p$ ; tail  $(1-p)$

- Binomial:  $n$  independent Bernoulli Trials

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

### Binomial distribution

Consider  $N$  independent experiments (Bernoulli trials):

Define  $n$  = number of heads ( $0 \leq n \leq N$ ).

Probability of a specific outcome (in order), e.g. 'hhtht' is

$$pp(1-p)p(1-p) = p^n (1-p)^{N-n}$$

But order not important; there are  $\frac{N!}{n!(N-n)!}$

ways (permutations) to get  $n$  successes in  $N$  trials, total probability for  $n$  is sum of probabilities for each permutation.

### Binomial distribution (2)

The binomial distribution is therefore

$$f(n; N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

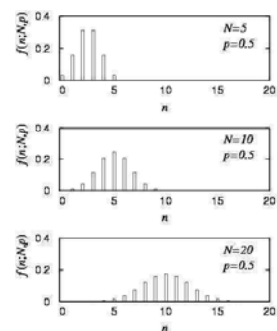
For the expectation value and variance we find:

$$E[n] = \sum_{n=0}^N n f(n; N, p) = Np$$

$$V[n] = E[n^2] - (E[n])^2 = Np(1-p)$$

### Binomial distribution (3)

Binomial distribution for several values of the parameters:



### Multinomial distribution

Like binomial but now  $n$  outcomes instead of two:

$$p_1 + p_2 + \dots + p_n = 1$$

For  $N$  trials we want the probability to obtain:

- $x_1$  of outcome 1,
- $x_2$  of outcome 2,
- ...
- $x_n$  of outcome  $n$ .

This is the multinomial distribution for

$$p_N(x_1, x_2, \dots, x_n) = \frac{N!}{x_1! \dots x_n!} p_1^{x_1} p_2^{x_2} \dots p_n^{x_n}$$

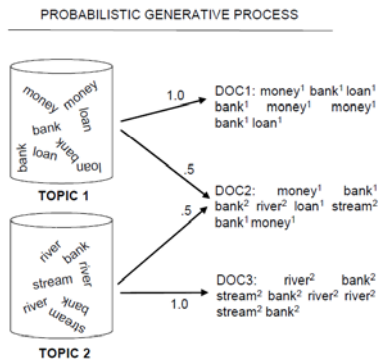
where

$$\sum_{i=1}^n x_i = N ; \sum_{i=1}^n p_i = 1.$$

### Text Mining

word	prob.	word	prob.	word	prob.	word	prob.
DRUGS	.069	RED	.202	MIND	.081	DOCTOR	.074
DRUG	.060	BLUE	.099	THOUGHT	.066	DR.	.063
MEDICINE	.027	GREEN	.096	REMEMBER	.064	PATIENT	.061
EFFECTS	.026	YELLOW	.073	MEMORY	.037	HOSPITAL	.049
BODY	.023	WHITE	.048	THINKING	.050	CARE	.046
MEDICINES	.019	COLOR	.048	PROFESSOR	.028	MEDICAL	.042
PAIN	.016	BRIGHT	.030	FELT	.025	NURSE	.031
PERSON	.016	COLORS	.029	REMEMBERED	.022	PATIENTS	.029
MARIJUANA	.014	ORANGE	.027	THOUGHTS	.020	DOCTORS	.028
LABEL	.012	BROWN	.027	FORGOTTEN	.020	HEALTH	.025
ALCOHOL	.012	PINK	.017	MOMENT	.020	MEDICINE	.017
DANGEROUS	.011	LOOK	.017	THINK	.019	NURSING	.017
ABUSE	.009	BLACK	.016	THING	.016	DENTAL	.015
EFFECT	.009	PURPLE	.015	WONDER	.014	NURSES	.013
KNOWN	.008	CROSS	.011	FORGET	.012	PHYSICIAN	.012
PILLS	.008	COLORED	.009	RECALL	.012	HOSPITALS	.011

Documents are mixtures of topics, where a topic is a multinomial probability distribution over words.



### Gaussian Distribution

1-dimensional:

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

2-dimensional:

$$z = \begin{bmatrix} x \\ y \end{bmatrix}$$

$$p(z) = \frac{1}{2\pi \Sigma^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(z-\mu)^T \Sigma^{-1}(z-\mu)\right)$$

$$\mu = \begin{bmatrix} u_x \\ u_y \end{bmatrix}; \Sigma = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix}$$

Symmetric non-negative

**Gaussian Distribution(2)**

$$\sigma_{xy} = \text{Cov} [x, y] = E [(x - u_x)(y - u_y)]$$

$$\sigma_{xx} = \text{Var} [x] = E [(x - u_x)^2]$$

$$\sigma_{yy} = \text{Var} [y] = E [(y - u_y)^2]$$

Properties:

-Linear transformation

$$x \propto N(\mu_x, \Sigma_x); y = Ax$$

$$\Rightarrow y \propto (A\mu_x, A\Sigma_x A^T)$$

-Addition

$$x \propto N(\mu_x, \Sigma_x); y \propto N(\mu_y, \Sigma_y)$$

$$\Rightarrow x + y \propto (\mu_x + \mu_y, \Sigma_x + \Sigma_y)$$

**Gaussian Distribution(3)**

Popular because:

1. This distribution is very tractable analytically.
2. The distribution has the familiar symmetric bell shape.
3. There is the central limit theorem which shows that under mild conditions, the normal distribution can be used to approximate a large variety of other distributions in large samples.

**Joint Distribution Table**

Eat_veg	Exercise	regular_sleep	P
0	0	0	0.12
0	0	1	0.24
0	1	0	0.04
0	1	1	0.20
1	0	0	0.05
1	0	1	0.15
1	1	0	0.10
1	1	1	0.10

**Joint Distribution Table**

1. Assume that if we have the table, we can ask for the probability of any logical expression

$$P(E) = \sum p(\text{row})$$

e.g.

$$p(\text{exercise}) =$$

$$P(\text{exercise} \cap \text{regular\_sleep}) =$$

### Joint Distribution Table

#### 2. Inference

compute the probability of an event given some evidence

$$p(E_1 | E_2) = \frac{p(E_1 \wedge E_2)}{p(E_2)} = \frac{\sum_{\text{matching } - E_1 \text{ and } - E_2} p(\text{rows})}{\sum_{\text{matching } - E_2} p(\text{rows})}$$

e.g.  
 $p(\text{eat\_veg}|\text{exercise}) =$

### How to obtain the Table

Eat_veg	Exercise	regular_sleep	P
0	0	0	?
0	0	1	?
0	1	0	?
0	1	1	?
1	0	0	?
1	0	1	?
1	1	0	?
1	1	1	?

-Made up by experts  
 -Learn from the data

$p(\text{row}) = \frac{\text{records matching the row}}{\text{\# of records}}$

### Build a classifier

input → classifier → categorical output

Eat_veg	Exercise	regular_sleep	
X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	Y
0	0	0	not healthy
0	0	1	healthy
0	0	1	healthy
1	1	1	very healthy
1	0	0	not healthy
1	0	1	very healthy
1	1	0	healthy
1	1	1	very healthy

### Maximum likelihood Estimator (ML)

e.g. (0,0,0)

$$\hat{Y} = \arg \max_{\gamma} p(x_1 = 0; x_2 = 0; x_3 = 0 | Y = \gamma)$$

**Maximum A-Posteriori Estimator (MAP)**

e.g. (0,0,0)

$$\hat{Y} = \arg \max_{\gamma} p(Y = \gamma | x_1 = 0; x_2 = 0; x_3 = 0)$$

$$\begin{aligned} & p(Y = \gamma | x_1 = \mu_1; x_2 = \mu_2; \dots; x_m = \mu_m) \\ &= \frac{p(x_1 = \mu_1; x_2 = \mu_2; \dots; x_m = \mu_m | Y = \gamma) p(Y = \gamma)}{p(x_1 = \mu_1; x_2 = \mu_2; \dots; x_m = \mu_m)} \\ &= \frac{p(x_1 = \mu_1; x_2 = \mu_2; \dots; x_m = \mu_m | Y = \gamma) p(Y = \gamma)}{\sum_{j=1}^N p(x_1 = \mu_1; x_2 = \mu_2; \dots; x_m = \mu_m | Y = \gamma_j) p(Y = \gamma_j)} \end{aligned}$$

$$\hat{Y} = \arg \max_{\gamma} p(x_1 = \mu_1, \dots, x_m = \mu_m | Y = \gamma) p(Y = \gamma)$$

**Naïve Bayes Classifier**

$$\hat{Y} = \arg \max_{\gamma} p(Y = \gamma) \prod_{j=1}^N p(x_j = \mu_j | Y = \gamma)$$

**Building a classifier**Step 1: for each category  $\gamma_j$ 

$$p(x_1, x_2, \dots, x_m | Y = \gamma_j)$$

Step 2: estimate

$$p(Y = \gamma_j) = \frac{\text{records labeled as } \gamma_j}{\text{\# of records}}$$

Step 3: given a new feature vector

$$\hat{Y} = \arg \max_{\gamma} p(x_1 = \mu_1, \dots, x_m = \mu_m | Y = \gamma) p(Y = \gamma)$$