

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

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

Two theorems from the Axioms:

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

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

If A is a random variable that can take on exactly one value out of
 $\{v_1, v_2, \dots, v_k\}$

$$p(A = v_1 \vee A = v_2 \vee \dots \vee A = v_k)$$

$$= \sum_{j=1}^k p(A = v_j) = 1$$

$$p(B \wedge [A = v_1 \vee A = v_2])$$

$$= \sum_{j=1}^2 p(B \wedge A = v_j)$$

$$p(B) = \sum_{j=1}^k p(B \wedge A = v_j)$$

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 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...

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)

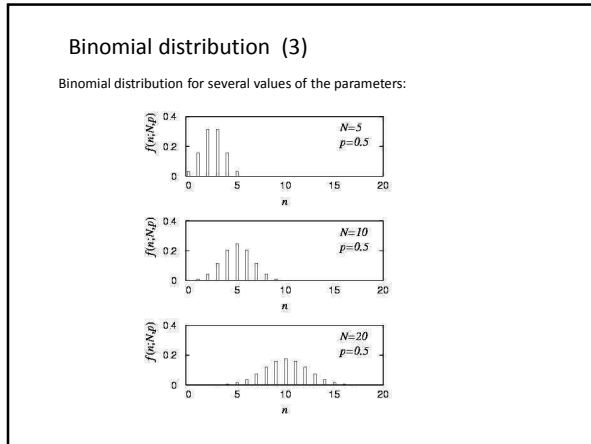
The binomial distribution is therefore

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

For the expectation value and variance we find:

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

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



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

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

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:

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

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

$$\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}; \Sigma = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{bmatrix}$$

Symmetric non-negative

Gaussian Distribution(2)

$$\sigma_{xy} = Cov [x, y] = E[(x - \mu_x)(y - \mu_y)]$$

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

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

Properties:

-Linear transformation
 $x \sim N(\mu_x, \Sigma_x); y = Ax$
 $\Rightarrow y \sim (A\mu_x, A\Sigma_x A^T)$

-Addition
 $x \sim N(\mu_x, \Sigma_x); y \sim N(\mu_y, \Sigma_y)$
 $\Rightarrow x + y \sim (\mu_x + \mu_y, \Sigma_x + \Sigma_y)$

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.

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

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


$B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i)$,
 $\rightarrow P(B) = P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i)$
 $\rightarrow P(B) = \sum_i P(B|A_i)P(A_i)$ 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!

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

Build a classifier

input → **classifier** → categorical output

Eat_veg	Exercise	regular_sleep	Y
X ₁	X ₂	X ₃	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	health
1	1	1	very healthy

Maximum likelihood Estimator

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

e.g. (0,0,0)

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

$$p(Y = \gamma | x_1 = \mu_1; x_2 = \mu_2; \dots; x_m = \mu_m) = \frac{p(x_1 = \mu_1; x_2 = \mu_2; \dots; x_m = \mu_m | Y = \gamma) p(Y = \gamma)}{p(x_1 = \mu_1; x_2 = \mu_2; \dots; x_m = \mu_m)}$$

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

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

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 γ_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

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

Discrete Random Variables

- Multinomial Distribution: each trial has k possible outcomes

If X_1, X_2, \dots, X_n are mutually exclusive events with $p(X_1=x_1)=p_1, \dots, p(X_n=x_n)=p_n$. Then the probability that X_1 occurs x_1 times, ..., X_n occurs x_n times is given by:

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

where

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