Review of Probability Theory

CSE 4334 / 5334 Data Mining
Spring 2019

Won Hwa Kim

Contents borrowed from Prof. Mark Craven at UW-Madison

Probability Theory

Frequentist interpretation: the probability of an event from a random experiment is the proportion of the time events of same kind will occur in the long run, when the experiment is repeated

Examples

- the probability my flight to Chicago will be on time
- the probability this ticket will win the lottery
- the probability it will rain tomorrow

Always a number in the **interval [0,1]**

0 means "never occurs"

1 means "always occurs"

Sample space

Uncertainty

- Probability theory is the study of uncertainty.

Sample Space

- a set of possible outcomes for some event

Examples

– flight to Chicago: {on time, late}
– lottery: {ticket 1 wins, ticket 2 wins,…,ticket n wins}
– weather tomorrow:
{rain, not rain} or
{sun, rain, snow} or
{sun, clouds, rain, snow, sleet} or…

Random variables

Random Variable

- A variable whose possible values are numerical outcomes of a random phenomenon.
- E.g., flipping a coin, an outcome from a dice, Will it rain tomorrow?

Example

- *X* represents the outcome of my flight to Chicago
- we write the probability of my flight being on time as $P(X = \text{on-time})$
- or when it's clear which variable we're referring to, we may use the shorthand $P(\text{on-time})$

Notation

- uppercase letters and capitalized words denote random variables
- lowercase letters and uncapitalized words denote values
- we'll denote a particular value for a variable as follows

$$P(X = x) \qquad P(Fever = true)$$

- we'll also use the shorthand form

$$P(x) \quad \text{for} \quad P(X = x)$$

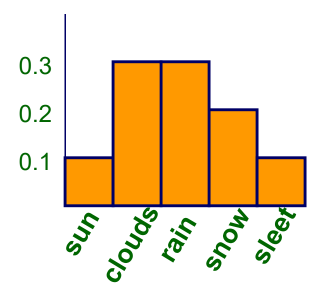
- for Boolean random variables, we'll use the shorthand

$$P(fever) \quad \text{for} \quad P(Fever = true)$$
$$P(\neg fever) \quad \text{for} \quad P(Fever = false)$$

Probability distribution

- if X is a random variable, the function given by $P(X = x)$ for each x is the *probability distribution* of X
- requirements:

$$P(x) \geq 0 \quad \text{for every } x$$
$$\sum_x P(x) = 1$$



Joint distribution

- *joint probability distribution*: the function given by $P(X = x, Y = y)$
- read "X equals x <u>and</u> Y equals y"
- example

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

probability that it's sunny and my flight is on time

Marginal distribution

- The *marginal distribution* of X is defined by

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

"the distribution of X ignoring other variables"

- This definition generalizes to more than two variables, e.g.

$$P(x) = \sum_y \sum_z P(x, y, z)$$

- Also known as sum rule

Marginal distribution example

joint distribution

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

marginal distribution for *X*

x	$P(X = x)$
sun	0.3
rain	0.5
snow	0.2

Conditional distribution

- the *conditional distribution* of *X* given *Y* is defined as:

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

"the distribution of *X* given that we know the value of *Y*"

Conditional distribution example

joint distribution

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

conditional distribution for X given Y=on-time

x	$P(X = x \mid Y = \textit{on-time})$
sun	0.20/0.45 = 0.444
rain	0.20/0.45 = 0.444
snow	0.05/0.45 = 0.111

The product rule

- rearranging the definition of the conditional distribution

$$P(x \mid y) = \frac{P(x,y)}{P(y)}$$

- leads to the product rule

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

The chain rule

- by repeated application of the product rule, a joint distribution can be expressed as

$$P(x_1, x_2, \ldots, x_n) = P(x_1) \prod_{i=2}^{n} P(x_i \mid x_1, \ldots, x_{i-1})$$

- permits the calculation of the joint distribution of a set of random variables using only conditional probabilities

- important idea for Bayesian networks

Independence

- two random variables, *X* and *Y*, are *independent* if

$$P(x, y) = P(x) \times P(y) \quad \text{for all } x \text{ and } y$$

- equivalently

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

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

$$P(x, y \mid z) = P(x \mid z) \times P(y \mid z) \quad \text{for all } x, y \text{ and } z$$

Independence example

joint distribution

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

marginal distributions

x	$P(X = x)$
sun	0.3
rain	0.5
snow	0.2

y	$P(Y = y)$
on-time	0.45
late	0.55

Are X and Y independent here? NO.

Independence example

joint distribution

x, y	$P(X = x, Y = y)$
sun, fly-United	0.27
rain, fly-United	0.45
snow, fly-United	0.18
sun, fly-Delta	0.03
rain, fly-Delta	0.05
snow, fly-Delta	0.02

marginal distributions

x	$P(X = x)$
sun	0.3
rain	0.5
snow	0.2

y	$P(Y = y)$
fly-United	0.9
fly-Delta	0.1

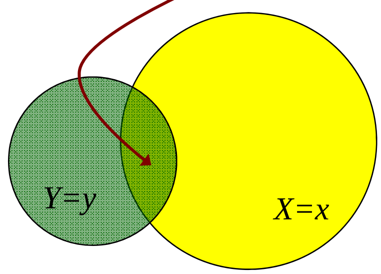
Are X and Y independent here? YES.

Probability of union of events

- the probability of the union of two events is given by:

$$P(x \vee y) = P(x) + P(y) - P(x, y)$$

this term needed to avoid double counting



Bayes rule (or theorem)

recall the product rule

$$\begin{aligned} P(x, y) &= P(x|y)P(y) \\ &= P(y|x)P(x) \end{aligned}$$

dividing both expressions on the right by $P(y)$

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)} = \frac{P(y \mid x)P(x)}{\sum_{x'} P(y \mid x')P(x')}$$

Bayes rule example

- $P(\text{stiff}-\text{neck}|\text{meningitis}) = 0.5$
- $P(\text{meningitis}) = \frac{1}{50{,}000}$
- $P(\text{stiff}-\text{neck}) = \frac{1}{20}$

$$P(\text{meningitis}|\text{stiff}-\text{neck}) = \frac{P(\text{stiff}-\text{neck}|\text{meningitis})P(\text{meningitis})}{P(\text{stiff}-\text{neck})}$$

$$= \frac{0.5 \times \frac{1}{50{,}000}}{\frac{1}{20}} = 0.0002$$

Why use Bayes rule?

- Causal knowledge such as $P(\text{stiff}-\text{neck}|\text{meningitis})$ is often more reliably estimated than diagnostic knowledge such as $P(\text{meningitis}|\text{stiff}-\text{neck})$
- Bayes' rule lets us use causal knowledge to make diagnostic inferences

Expected values

- the *expected value* of a random variable that takes on numerical values is defined as:

$$E[X] = \sum_x x \times P(x)$$

this is the same thing as the *mean*

- we can also talk about the expected value of a function of a random variable

$$E[g(X)] = \sum_x g(x) \times P(x)$$

Expected values

$$E[Shoesize] =$$
$$5 \times P(Shoesize = 5) + ... + 14 \times P(Shoesize = 14)$$

- Suppose each lottery ticket costs $1 and the winning ticket pays out $100. The probability that a particular ticket is the winning ticket is 0.001.

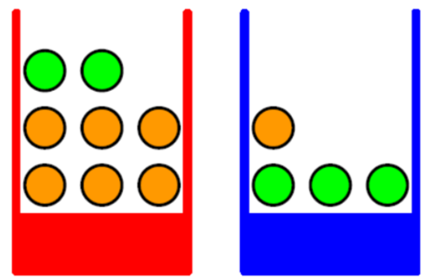
$$E[gain(Lottery)] =$$
$$gain(\text{winning})P(\text{winning}) + gain(\text{losing})P(\text{losing}) =$$
$$(\$100 - \$1) \times 0.001 - \$1 \times 0.999 =$$
$$-\$0.90$$

Probability Theory

Simple example

- Red box: 2 apples and 6 oranges
- Blue box: 3 apples and 1 orange
- Chance of selecting red / blue box: 40% / 60%
- What is the probability that we pick an apple?

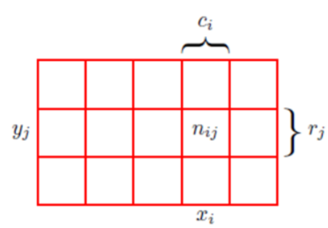
  - B: random variable for box selection
  - $p(B = r) = 4/10$, $p(B = b) = 6/10$



Probability Theory

Understanding probability

- $p(X = x_i) = \frac{c_i}{N}$
- Joint probability: $p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$
- Conditional proability: $p(Y = y_i | X = x_i) = \frac{n_{ij}}{c_i}$

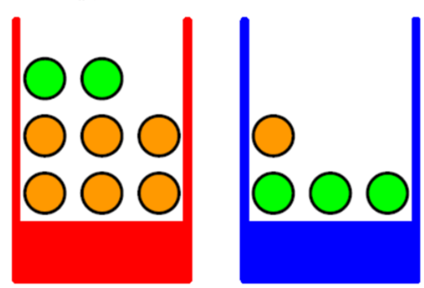


Sum rule

- Marginal probability: $p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)$

Product rule

- $p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \frac{c_i}{N}$
  $= p(X = x_i | Y = y_j) p(X = x_i)$



Probability Theory

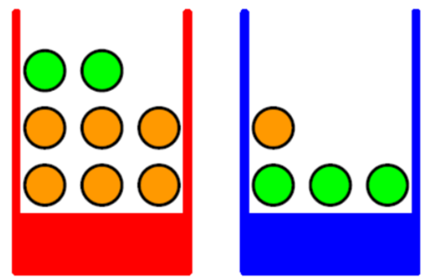
Bayes Theorem

- From the product rule and symmetry of joint probability,

$$p(Y|X) = \frac{P(X|Y)p(Y)}{P(X)}$$

Back to the example…

- $p(B = r) = 4/10$
- $p(B = b) = 6/10$
- $p(F = a|B = r) = 1/4$
- $p(F = o|B = r) = 3/4$
- $p(F = a|B = b) = 3/4$
- $p(F = o|B = b) = 1/4$



Probability Theory

Rules

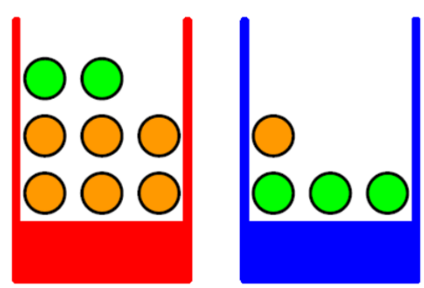
- $p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)$
- $p(X = x_i, Y = y_j) = p(X = x_i|Y = y_j)p(X = x_i)$
- $p(Y|X) = \frac{P(X|Y)p(Y)}{P(X)}$

You pick an apple

$$p(F = a) = p(F = a|B = r)p(B = r) + p(F = a|B = b)P(B = b)$$
$$= \frac{1}{4}\frac{4}{10} + \frac{3}{4}\frac{6}{10} = \frac{11}{20}$$

You pick an orange… which bag?

$$p(B = r|F = o) = \frac{p(F = o|B = r)p(B = r)}{p(F = o)}$$
$$= \frac{3}{4}\frac{4}{10}\frac{20}{9} = \frac{2}{3}$$



Probability Theory

Expectation

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

Conditional Expectation

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

Variance

$$\begin{aligned} var[f] &= \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] \\ &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \end{aligned}$$

Covariance

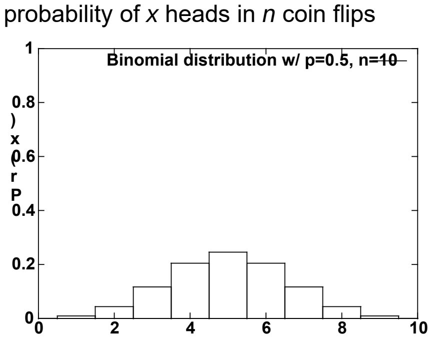
$$\begin{aligned} cov[x, y] &= \mathbb{E}_{x,y}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned}$$

Binomial distribution

- distribution over the number of successes in a fixed number *n* of independent trials (with same probability of success *p* in each)

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- e.g. the probability of *x* heads in *n* coin flips



Probability Theory

Probability density

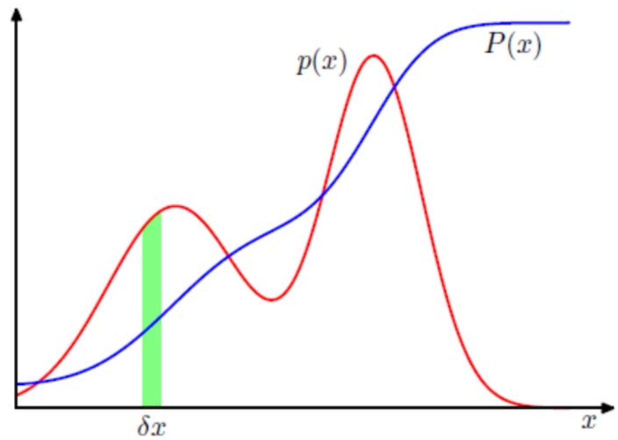
- Probability density function (PDF)

$$p(x) \geq 0$$

$$\int p(x) = 1$$

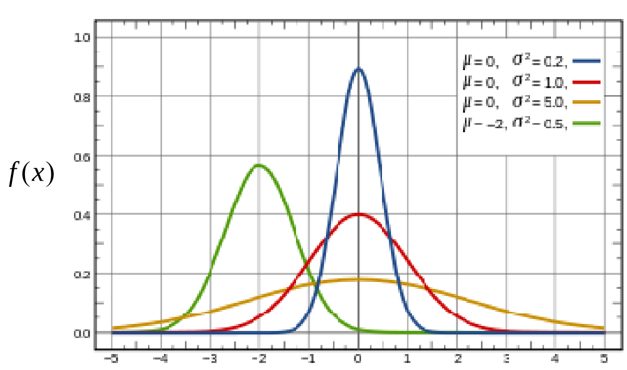
- Cumulative density function (CDF)

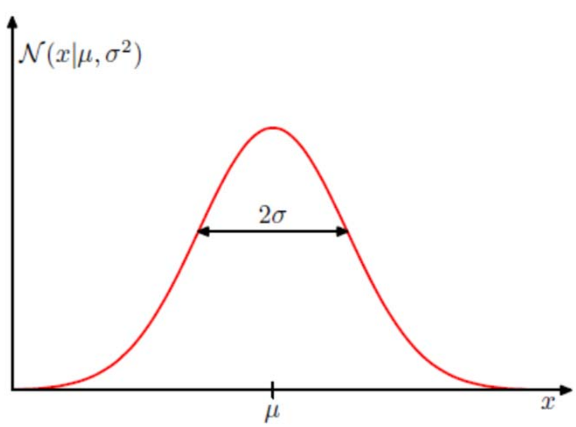
$$P(z) = \int_{-\infty}^{z} p(x)dx$$



Gaussian Distribution

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

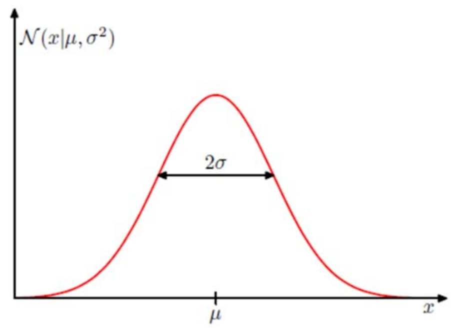




Probability Theory

Gaussian Distribution

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



Properties of the Gaussian Distribution

- $N(x|\mu,\sigma^2) \geq 0$
- $\int N(x|\mu,\sigma^2)dx = 1$
- $\mathbb{E}[x] = \int N(x|\mu,\sigma^2)xdx = \mu$
- $\mathbb{E}[x^2] = \int N(x|\mu,\sigma^2)x^2dx = \mu + \sigma^2$
- $var[x] = \sigma^2$ using that $var[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$