Review of probability

Computer Sciences 760
Spring 2014

http://pages.cs.wisc.edu/~dpage/cs760/

Goals for the lecture

you should understand the following concepts
• definition of probability
• random variables
• joint distributions
• conditional distributions
• chain rule
• independence
• union rule
• expected values, variance, covariance
• uncorrelated vs. independent
• Bayes theorem
• multinomial distribution
• probability density function
• normal distribution
Definition of probability

• *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 spaces

• *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 representing the outcome of an event

- 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
  $$ P(X = x) \quad P(\text{Fever} = \text{true}) $$
- 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
  $$ P(\text{fever}) \quad \text{for} \quad P(\text{Fever} = \text{true}) $$
  $$ P(\neg \text{fever}) \quad \text{for} \quad P(\text{Fever} = \text{false}) $$
Probability distributions

- 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 distributions

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

<table>
<thead>
<tr>
<th>$x, y$</th>
<th>$P(X = x, Y = y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun, on-time</td>
<td>0.20</td>
</tr>
<tr>
<td>rain, on-time</td>
<td>0.20</td>
</tr>
<tr>
<td>snow, on-time</td>
<td>0.05</td>
</tr>
<tr>
<td>sun, late</td>
<td>0.10</td>
</tr>
<tr>
<td>rain, late</td>
<td>0.30</td>
</tr>
<tr>
<td>snow, late</td>
<td>0.15</td>
</tr>
</tbody>
</table>

probability that it’s sunny and my flight is on time
Marginal distributions

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

---

Marginal distribution example

<table>
<thead>
<tr>
<th>joint distribution</th>
<th>marginal distribution for $X$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x, y$</td>
<td>$P(X = x, Y = y)$</td>
</tr>
<tr>
<td>sun, on-time</td>
<td>0.20</td>
</tr>
<tr>
<td>rain, on-time</td>
<td>0.20</td>
</tr>
<tr>
<td>snow, on-time</td>
<td>0.05</td>
</tr>
<tr>
<td>sun, late</td>
<td>0.10</td>
</tr>
<tr>
<td>rain, late</td>
<td>0.30</td>
</tr>
<tr>
<td>snow, late</td>
<td>0.15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$x$</th>
<th>$P(X = x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>0.3</td>
</tr>
<tr>
<td>rain</td>
<td>0.5</td>
</tr>
<tr>
<td>snow</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Conditional distributions

- 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

<table>
<thead>
<tr>
<th>joint distribution</th>
<th>$P(X = x, Y = y)$</th>
<th>conditional distribution for $X$ given $Y$=on-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x, y$</td>
<td></td>
<td>$x$</td>
</tr>
<tr>
<td>sun, on-time</td>
<td>0.20</td>
<td>sun</td>
</tr>
<tr>
<td>rain, on-time</td>
<td>0.20</td>
<td>rain</td>
</tr>
<tr>
<td>snow, on-time</td>
<td>0.05</td>
<td>snow</td>
</tr>
<tr>
<td>sun, late</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>rain, late</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>snow, late</td>
<td>0.15</td>
<td></td>
</tr>
</tbody>
</table>

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
Chain rule

- A joint probability distribution can be expressed as

\[ P(X_1, \ldots, X_n) = P(X_1) \prod_{i=2}^{n} P(X_i|X_1, \ldots, X_{i-1}) \]

- Permits the calculation of the joint distribution of a set of random variables using only conditional probabilities.
- Useful for Bayesian network.

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,

\[ P(X \mid Y) = P(X) \]

- Equivalently,

\[ P(Y \mid X) = P(Y) \]
Conditional independence

- 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 #1

<table>
<thead>
<tr>
<th>joint distribution</th>
<th>marginal distributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x, y$</td>
<td>$x$</td>
</tr>
<tr>
<td>sun, on-time</td>
<td>$P(X = x, Y = y)$</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
</tr>
<tr>
<td>rain, on-time</td>
<td>0.20</td>
</tr>
<tr>
<td>snow, on-time</td>
<td>0.05</td>
</tr>
<tr>
<td>sun, late</td>
<td>0.10</td>
</tr>
<tr>
<td>rain, late</td>
<td>0.30</td>
</tr>
<tr>
<td>snow, late</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Are $X$ and $Y$ independent here? NO.
Independence example #2

<table>
<thead>
<tr>
<th></th>
<th>$P(X = x, Y = y)$</th>
<th></th>
<th>$P(X = x)$</th>
<th>$P(Y = y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun, fly-United</td>
<td>0.27</td>
<td>sun</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>rain, fly-United</td>
<td>0.45</td>
<td>rain</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>snow, fly-United</td>
<td>0.18</td>
<td>snow</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>sun, fly-Delta</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rain, fly-Delta</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>snow, fly-Delta</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Are $X$ and $Y$ independent here? YES.

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 value examples

\[ E[\text{Shoesize}] = 5 \times P(\text{Shoesize} = 5) + \ldots + 14 \times P(\text{Shoesize} = 14) \]

- 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.

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

More examples

- the expected value of a function of a random variable

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

- Example 1 (k-th moment): \( g(X) = X^k \)
- Example 2 (entropy): \( g(X) = -\log P(X) \)
- Example 3 (variance): \( g(X) = (X - E(X))^2 \)
Variance

- The variance of a random variable $X$ is its second central moment, i.e. the expected value of the squared deviation from the mean:

$$Var(X) = E\{[X - E(X)]^2\} = E(X^2) - \{E(X)\}^2$$

- Properties:
  - (1) $c$ is a constant, then $Var(cX) = c^2Var(X)$
  - (2) If $X$ and $Y$ are independent, then

$$Var(X + Y) = Var(X) + Var(Y)$$

Covariance

- covariance is a measure of how much two random variables change together:

$$Cov(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$$

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

- Correlation coefficient: $\rho_{xy} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$

- Properties:
  - (1) $Cov(aX,bY) = abCov(X,Y)$
  - (2) $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$
Uncorrelated vs. Independent

• $X$ and $Y$ are **uncorrelated** if $\text{Cov}(X, Y) = 0$
  \[ E(XY) = E(X)E(Y) \]

• $X$ and $Y$ are **independent** if
  \[ P(X, Y) = P(X)P(Y) \]

• Are they equivalent?
  - independence $\rightarrow$ uncorrelated, but not the other way around
  - Equivalent for Gaussian variables

Probability of union of events

• the probability of the union of two events is given by:
  \[ P(x \lor y) = P(x) + P(y) - P(x, y) \]

• Example:
  - $x$: taking CS540 this semester
  - $y$: taking CS760 this semester

this term needed to avoid double counting
Bayes theorem

\[ P(x|y) = \frac{P(y|x)P(x)}{P(y)} = \frac{P(y|x)P(x)}{\sum_{x'} P(y|x')P(x')} \]

- this theorem is extremely useful
- there are many cases when it is hard to estimate \( P(x \mid y) \) directly, but it’s not too hard to estimate \( P(y \mid x) \) and \( P(x) \)

Bayes theorem example

- MDs usually aren’t good at estimating \( P(\text{Disorder} \mid \text{Symptom}) \)
- they’re usually better at estimating \( P(\text{Symptom} \mid \text{Disorder}) \)
- if we can estimate \( P(\text{Fever} \mid \text{Flu}) \) and \( P(\text{Flu}) \) we can use Bayes’ Theorem to do diagnosis

\[ P(\text{flu} \mid \text{fever}) = \frac{P(\text{fever} \mid \text{flu})P(\text{flu})}{P(\text{fever} \mid \text{flu})P(\text{flu}) + P(\text{fever} \mid \neg \text{flu})P(\neg \text{flu})} \]
The 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

The geometric distribution

- distribution over the number of trials before the first failure (with same probability of success $p$ in each)

\[ P(x) = (1-p) p^x \]

- e.g. the probability of $x$ heads before the first tail
The multinomial distribution

- $k$ possible outcomes on each trial
- probability $p_i$ for outcome $x_i$ in each trial
- distribution over the number of occurrences $x_i$ for each outcome in a fixed number $n$ of independent trials

$$P(x) = \frac{n!}{\prod_i (x_i!)} \prod_i p_i^{x_i}$$

- e.g. with $k=6$ (a six-sided die) and $n=30$

$$P([7,3,0,8,10,2]) = \frac{30!}{7! \times 3! \times 0! \times 8! \times 10! \times 2!} \left( p_1^7 p_2^3 p_3^0 p_4^8 p_5^{10} p_6^2 \right)$$

Continuous random variables

- up to now, we’ve considered only discrete random variables, but we can have RVs describing continuous variables too (weight, temperature, etc.)

- a continuous random variable is described by a probability density function (p.d.f.)
Probability density functions

- A continuous random variable is described by a probability density function $f(x)$

\[
\forall x \ f(x) \geq 0
\]

\[
P[a \leq X \leq b] = \int_{a}^{b} f(x) \, dx
\]

\[
\int_{a}^{b} f(x) \, dx = 1
\]

The normal (Gaussian) distribution

\[
f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]

Standard normal: $Z \sim N(0,1)$

Let $X \sim N(\mu, \sigma^2)$, then $X = \mu + \sigma \cdot Z$
The normal (Gaussian) distribution

\( X_1, X_2, \ldots, X_n \) are i.i.d. \( N(\mu, \sigma^2) \)

Let \( \bar{X} \) be the sample mean and \( S^2 \) be the sample variance. Then

\[
\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1)
\]

\[
\frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t(n-1)
\]

Some p.d.f.’s