DATA MINING TECHNIQUES
Review of Probability Theory

Yijun Zhao

Northeastern University

spring 2015

Review of Probability Theory

Based on "Review of Probability Theory" from CS 229 Machine Learning, Stanford University
(Handout posted on the course website)

Elements of Probability

- Sample space Ω: the set of all the outcomes of an experiment
- Event space F: a collection of possible outcomes of an experiment. $F \subseteq \Omega$.
- Probability measure: a function P: $F \rightarrow R$ that satisfies the following properties:
  - $P(A) \geq 0 \ \forall \ A \in F$
  - $P(\Omega) = 1$
  - If $A_1, A_2, \ldots$ are disjoint events, then $P(\cup_i A_i) = \sum_i P(A_i)$

Properties of Probability

- If $A \subseteq B \implies P(A) \leq P(B)$
- $P(A \cap B) \leq \min\ (P(A), P(B))$
- $P(A \cup B) \leq P(A) + P(B)$ (Union Bound)
- $P(\Omega \setminus A) = 1 - P(A)$
- If $A_1, \ldots, A_k$ is a disjoint partition of Ω, then

$$\sum_{i=1}^{k} P(A_k) = 1$$

Conditional Probability

- A conditional probability $P(A|B)$ measures the probability of an event A after observing the occurrence of event B

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

- Two events A and B are independent iff $P(A|B) = P(A)$ or equivalently, $P(A \cap B) = P(A)P(B)$

Conditional Probability Examples

- A math teacher gave her class two tests. 25% of the class passed both tests and 42% of the class passed the first test. What percent of those who passed the first test also passed the second test?
- In New England, 84% of the houses have a garage and 65% of the houses have a garage and a back yard. What is the probability that a house has a backyard given that it has a garage?

Independent Events Examples

- What's the probability of getting a sequence of 1,2,3,4,5,6 if we roll a dice six times?
- A school survey found that 9 out of 10 students like pizza. If three students are chosen at random with replacement, what is the probability that all three students like pizza?

Random Variable

A random variable X is a function that maps a sample space Ω to real values. Formally,

$$X : \Omega \longrightarrow R$$

Examples:

- Rolling one dice
  $X =$ number on the dice at each roll
- Rolling two dice at the same time
  $X =$ sum of the two numbers

Random Variable

A random variable can be continuous. E.g.,

- $X =$ the length of a randomly selected phone call
  (What's the Ω?)
- $X =$ amount of coke left in a can marked 12oz
  (What's the Ω?)

Probability Mass Function

If X is a discrete random variable, we can specify a probability for each of its possible values using the probability mass function (*PMF*). Formally, a *PMF* is a function p: $\Omega \longrightarrow R$ such that

$$p(x) = P(X = x)$$

- Rolling a dice:
  $p(X = i) = \frac{1}{6} \quad i = 1, 2, \ldots, 6$
- Rolling two dice at the same time:
  $X =$ sum of the two numbers
  $p(X = 2) = \frac{1}{36}$

Probability Mass Function

- $X \sim Bernoulli(p),\ p \in [0,1]$

$$p(x) = \begin{cases} p & \text{if} \quad x = 1 \\ 1-p & \text{if} \quad x = 0 \end{cases}$$

- $X \sim Binomial(n,p),\ p \in [0,1]$ and $n \in Z^+$

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

- $X \sim Geometric(p),\ p > 0$

$$p(x) = p(1-p)^{x-1}$$

- $X \sim Poisson(\lambda),\ \lambda > 0$

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

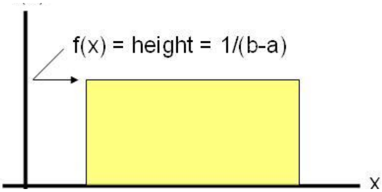
Probability Density Function

- If X is a continuous random variable, we can NOT specify a probability for each of its possible values (why?)
- We use a probability density function *PDF* to describe the relative likelihood for a random variable to take on a given value
- A (*PDF*) specifies the probability of X takes a value within a range. Formally, a *PDF* is a function $f(x)$: $\Omega \longrightarrow R$ such that

$$P(a < X < b) = \int_a^b f(x)dx$$

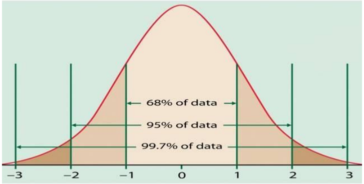
Probability Density Function

- $X \sim$ uniform on $[a, b]$:



$$f(x) = \frac{1}{b-a}$$

- $X \sim N(\mu, \sigma)$:



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

Joint Probability Mass Function

If we have two discrete random variables X, Y, we can define their joint probability mass function (PMF) $p_{XY} : R^2 \longrightarrow [0, 1]$ as:

$$p(x, y) = P(X = x, Y = y)$$

where $p(x, y) \leq 1$ and $\sum_{x \in X} \sum_{y \in Y} p(x, y) = 1$

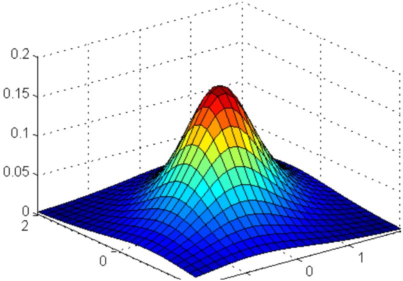
- X, Y: rolling two dice
  $p(x, y) = \frac{1}{36} \quad x, y = 1, 2, \ldots, 6$
- X: rolling one dice $\quad Y$: drawing a colored ball
  $p(6, green) = ? \quad p(5, red) = ?$

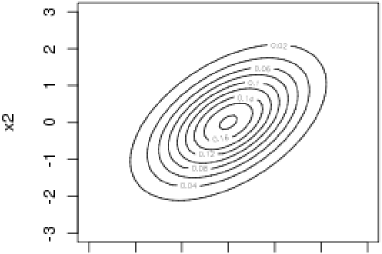
Joint Probability Density Function

If we have two continuous random variables X, Y, we can define their joint probability density function (PDF) $f_{XY} : R^2 \longrightarrow [0, 1]$ as:

$$P(a < X < b, c < Y < d) = \int_c^d \int_a^b f(x, y)dxdy$$

- 2D Gaussian





Marginal Probability Mass Function

How does the joint *PMF* over two discrete variables relate to the *PMF* for each variable separately? It turns out that

$$p(x) = \sum_{y \in Y} p(x, y)$$

- X, Y: rolling two dice

$$p(x, y) = \frac{1}{36} \quad x, y = 1, 2, \ldots, 6$$

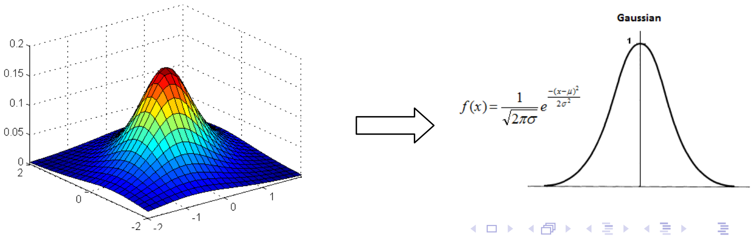
$$p(x) = \sum_{y=1}^{6} p(x, y) = \frac{1}{6}$$

Marginal Probability Density Function

Similarly, we can obtain a marginal *PDF* (also called marginal density) for a continuous random variable from a joint *PDF*:

$$f(x) = \int_{-\infty}^{\infty} f(x, y)dy$$

- Integrating out one variable in the 2D Gaussian gives a 1D Gaussian in either dimension



Conditional Probability Distribution

A conditional probability distribution defines the probability distribution over Y when we know that X must take on a certain value x

- Discrete case: conditional *PMF*

$$p(y|x) = \frac{p(x,y)}{p(x)} \Longleftrightarrow p(x,y) = p(y|x)p(x)$$

- Continuous case: conditional *PDF*

$$f(y|x) = \frac{f(x,y)}{f(x)} \Longleftrightarrow f(x,y) = f(y|x)f(x)$$

Marginal vs. Conditional

- **Marginal probability:**

$i \backslash j$	1	2	3	4	5	6	$p_X(i)$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p_Y(j)$	1/6	1/6	1/6	1/6	1/6	1/6	

- **Conditional probability: probability of rolling a 2**

$i \backslash j$	1	2	3	4	5	6	$p_X(i)$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p_Y(j)$	1/6	1/6	1/6	1/6	1/6	1/6	

Bayes Rule

- We can express the joint probability in two ways:

$$p(x,y) = p(y|x)p(x)$$
$$p(x,y) = p(x|y)p(y)$$

- Bayes rule:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \qquad \text{(discrete)}$$

$$f(y|x) = \frac{f(x|y)f(y)}{f(x)} \qquad \text{(continuous)}$$

Bayes Rule Application

A patient underwent a HIV test and got a positive result. Suppose we know that

- Overall risk of having HIV in the population is 0.1%
- The test can accurately identify 98% of HIV infected patients
- The test can accurately identify 99% of healthy patients

What's the probability the person indeed infected HIV?

Bayes Rule - Application

We have two random variables here:

- $X \in \{+, -\}$: the outcome of the HIV test
- $C \in \{Y, N\}$: the patient has HIV or not

We want to know: $P(C{=}\mathbf{Y}|X{=}+)$?

Apply Bayes rule:

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

$P(X{=}+|C{=}Y) = 0.98$ $\quad$ $P(C{=}Y) = 0.001$

$P(X{=}+) = 0.98 * 0.001 + (1\text{-}0.99) * 0.999 = 0.01097$

Answer: $0.98 * 0.001/0.01097 = 8.9\%$

Bayes Rule Terminology

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

$P(Y)$: prior probability or, simply, prior

$P(X|Y)$: conditional probability or, likelihood

$P(X)$: marginal probability

$P(Y|X)$: posterior probability or, simply, posterior

Independence

Two random variables X and Y are independent iff

- For discrete random variables
  $$p(x, y) = p(x)p(y) \quad \forall x \in X, y \in Y$$
- For discrete random variables
  $$p(y|x) = p(y) \qquad \forall y \in Y \text{ and } p(x) \neq 0$$
- For continuous random variables
  $$f(x, y) = f(x)f(y) \quad \forall x, y \in R$$
- For continuous random variables
  $$f(y|x) = f(y) \qquad \forall y \in R \text{ and } f(x) \neq 0$$

Multiple Random Variables

Extend to multiple random variables :

- Joint Distribution (discrete):

$$p(x_1, \ldots, x_n) = P(X1 = x_1, \ldots, X_n = x_n)$$

- Conditional Distribution (chain rule - discrete)

$$p(x_1, \ldots, x_n) = p(x_n|x_1, \ldots, x_{n-1})p(x_1, \ldots, x_{n-1})$$

$$= p(x_n|x_1, \ldots, x_{n-1})p(x_{n-1}|x_1, \ldots, x_{n-2})p(x_1, \ldots, x_{n-2})$$

$$= p(x_1) \prod_{i=2}^{n} p(x_i|x_1, \ldots, x_{i-1})$$

(continuous case can be defined similarly using PDF)

Multiple Random Variables

- Independence:

  Discrete case: $X_1, \ldots, X_n$ are independent iff

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$$

  Continuous case: $X_1, \ldots, X_n$ are independent iff

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$$

Multiple Random Variables

- Bayes rule:

Discrete case:

$$p(x_n|x_1,\ldots,x_{n-1}) = \frac{p(x_1,\ldots,x_{n-1}|x_n)p(x_n)}{p(x_1,\ldots,x_{n-1})}$$

Continuous case:

$$f(x_n|x_1,\ldots,x_{n-1}) = \frac{f(x_1,\ldots,x_{n-1}|x_n)f(x_n)}{f(x_1,\ldots,x_{n-1})}$$

Probabilistic View of a Dataset

What about a dataset $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$?

- We can view S as $d+1$ random variables where d is the number of attributes in $\mathbf{x}$, i.e.

$$X_1,\ X_2,\ \ldots,\ X_d,\ Y$$

- Uncover(model) $p(x_1, x_2, \ldots, x_d, y)$ from the training data
- For ANY $(x_1, x_2, \ldots, x_n)$, we will compute:

$$P(y = 0 | x_1, x_2, \ldots, x_n)\ ?$$

$$P(y = 1 | x_1, x_2, \ldots, x_n)\ ?$$

That is predicting y from $\mathbf{x}$!