Algorithmic Learning Theory
Spring 2017
Lecture 2

Instructor: Farid Alizadeh
Scribe: Chien-Ming Huang
Edit: Yuan Qu

1/25/2017

1. Review Bayes Theory(Lecture 1)
2. Random Variable and Distribution
   (a) Random variable
      i. DRV, discrete random variable
      ii. CRV, continuous random variable
   (b) Distribution function
      i. CDF, cumulative distribution function
      ii. pdf or pmf, probability density(Mass) function
   (c) Discrete distribution
      i. Discrete uniform distribution
      ii. Beunoulli's distribution
      iii. Binomial distribution
   (d) Continuous distribution
      i. Continuous uniform distribution
      ii. Normal distribution
3. Multivariate Distributions
   (a) Random vector
   (b) Discrete multivariate distribution
   (c) Binormal distribution
   (d) Marginal distribution
   (e) Conditional distribution
4. Bayes Classification

Scribe:Chien-Ming Huang Edit:Yuan Qu
Lecture 2 **Date:** 01/25/2017

1 Review Bayes Theory(Lecture 1)

See notes in "Lecture 1".

2 Random Variable and Distribution

2.1 Random Variable

2.1.1 Discrete Random Variable(D.R.V)

$x \to t_1, t_2, \cdots, t_n$

2.1.2 Continuous Random Variable(C.R.V)

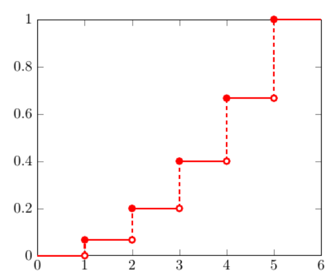
$x \to [a, b]$ a range of value.

2.2 Distribution Function

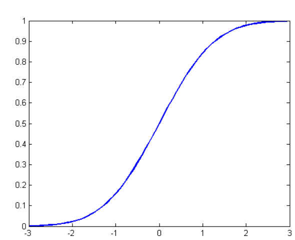
2.2.1 CDF:

Cumulative Distribution Function
$F_x(t) = P[x \leq t]$, probability can only increase

1. For Discrete:



2. For Continuous:

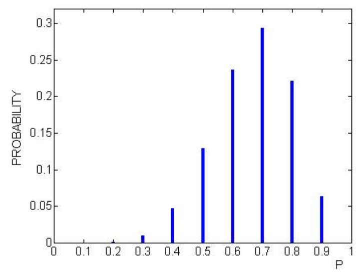


2.2.2 pdf or pmf:

Probability Density(Mass) Function

1. For Discrete:

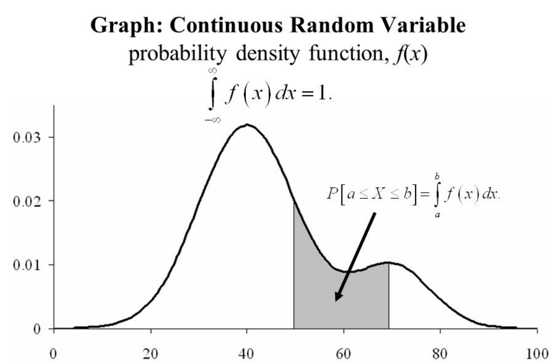
$$x : F_x(t) = P[x = t]$$



2. For Continuous:

$$f_x(t) = \frac{d}{dt}F_x(t), \; F_x(t) = \int_{-\infty}^{t} f_x(t)dt$$

i $f_x(t) \geq 0$
ii $\int_{-\infty}^{\infty} f_x(t)dt = 1$

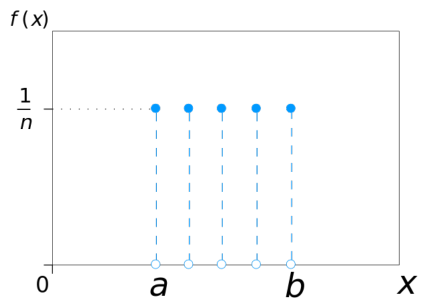


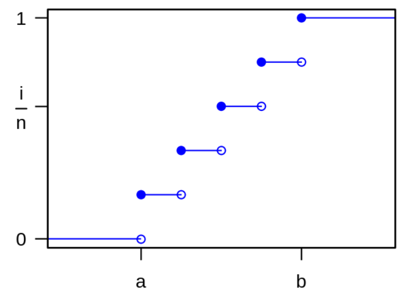
2.3 Discrete Distribution

2.3.1 Discrete Uniform Distribution

$$x : 1, 2, 3, \cdots, k$$

$$\mathrm{pdf} : u_x(t) = \begin{cases} \frac{1}{n}, & \text{if } t = 1, 2, \cdots, n \\ 0, & \text{otherwise} \end{cases}$$





2.3.2 Bernoulli Distribution

$$\text{pdf}: f_x(t) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\text{CDF}: F_x(t) = \begin{cases} 0, & x \leqslant 0 \\ 1-p, & 0 \leqslant x < 1 \\ 0, & x \geqslant 1 \end{cases}$$

2.3.3 Binomial Distribution

numbers of 0's in independent Bernoulli trial with $P[0] = p$

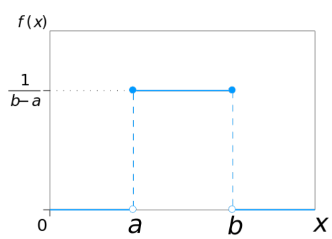
$$\text{pdf: } b(t|p,n) = \binom{n}{t} p^t (1-p)^{n-t}, \ \binom{n}{t} = \frac{n!}{t!(n-t)!}$$

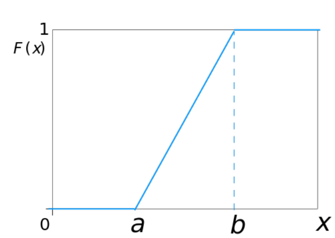
$$\text{CDF: } B(t|p,n) = \sum_{n=0}^{t} b(t|p,n)$$

2.4 Continuous distribution

2.4.1 Continuous uniform distribution

$$u(t|a,b) = \begin{cases} \frac{1}{b-a}, & a \leqslant t \leqslant b \\ 0, & \text{otherwise} \end{cases}$$





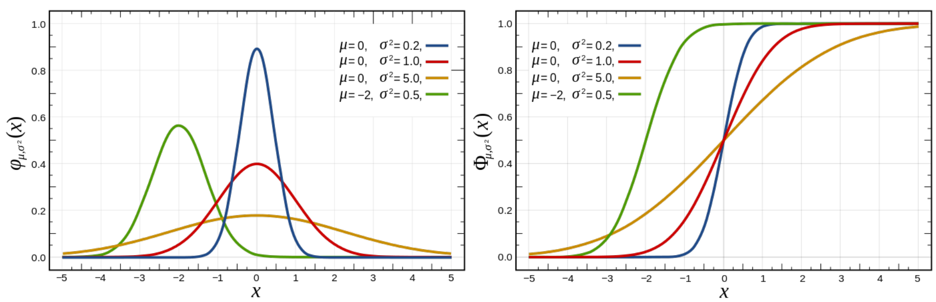
2.4.2 Normal distribution

mean$= \mu$ and std.$= \sigma$

$$\text{pdf: } f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\text{CDF: } \frac{1}{2}[1 + \mathrm{erf}(\frac{x-\mu}{\sigma\sqrt{2}})]$$

$$\mathrm{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} dt$$



3 Multivariate Distributions

3.1 Random Vector

$X_N = (x_1, x_2, \cdots x_N)$ can be continuous or discrete.

$$\text{CDF: } F_x(t_1, t_2, \cdots t_n) = P[x_1 \leqslant t_1, x_2 \leqslant t_2, \cdots, x_n \leqslant t_n]$$

$$\text{pdf: } \begin{cases} \frac{\partial}{\partial t_1 \partial t_2 \partial t_3 \cdots \partial t_n} F(t_1, t_2, \cdots t_n) = f_x(t_1, \cdots, t_n), & \text{all continuous} \\ P[x_1 = t_1, x_2 = t_2, \cdots, x_n = t_n] = f_x(t_1, \cdots, t_n), & \text{all discrete} \end{cases}$$

both are joint distribution R.V. $x_1, \cdots, x_n$

3.2 Discrete Multivariate Distribution

$$Y \to 1, 2, \cdots, r; \quad P[Y = r_1] = P_u; \quad \sum P_u = 1$$

repeat n times, x_k= number of times $Y = k$ occurs

$$\underset{\sim}{x} = (x_1, \cdots, x_n)$$

$$\text{pdf: } f_x(x_1, x_2, \cdots, x_n) = P[x_1 = t_1, x_2 = t_2, \cdots, x_n = t_n] \binom{n}{t_1, t_2, \cdots, t_n} {p_1}^{t_1} {p_2}^{t_2} \cdots {p_n}^{t_n}$$

$$\binom{n}{t_1, t_2, \cdots, t_r} = \frac{n!}{t_1! t_2! \cdots t_n!}$$

3.3 Binormal distribution

$$\underset{\sim}{x} = (x_1, x_2), \text{both continuous}$$

$$\underset{\sim}{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \sigma_1^2 \to x_1, \sigma_2^2 \to x_2, \sigma_{21} \to x_1, x_2$$

$$\text{Covariance Matrix: } \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

$$\phi(t_1, t_2 | \underset{\sim}{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi \cdot \mathrm{Det}(\Sigma)}} \exp[(t - \underset{\sim}{\mu})^T \Sigma^{-1} (t - \underset{\sim}{\mu}), t = \begin{pmatrix} t_1 \\ t_2 \end{pmatrix}$$

$$\phi(t_1) = \int_{-\infty}^{\infty} \phi(t_1, t_2 | \cdots) dt_2$$

Joint pdf: $f_{(x1,x2)}(t_1, t_2)$
In a similar way, for multi-normal distribution,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \cdots & \sigma_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \cdots & \cdots & \sigma_n^n \end{bmatrix}$$

3.4 Marginal Distribution

$$\underset{\sim}{x} = (x_1, x_2, \cdots, x_n) \to f_{\underset{\sim}{x}}(t_1, \cdots, t_n)$$

To calculate pdf,

$$f_{\underset{\sim}{x}}(t_1, t_2, \cdots, t_n) = \int_{t_{k+1}, \cdots, t_n}^{\infty} f_{\underset{\sim}{x}}(t_1, \cdots, t_n) dt_{k+1}, \cdots, dt_n = \sum_{t_{k+1}} \sum_{t_{k+2}} \cdots \sum_{t_n} f_{\underset{\sim}{x}}(t_1, \cdots, t_n)$$

Joint pdf: $f(t_1, t_2)$

3.5 Conditional Distribution (2 Vars)

$$\underset{\sim}{x} = (x_1, x_2) \to \text{ joint:} f(t_1, t_2)$$

Conditional distribution means: $f_{x_1|x_2}(t_1|t_2 = a)$, a is a given constant, x_2 is given, fixed.

$$f_{x_1|x_2}(t_1|t_2) = \frac{f_{\underset{\sim}{x}}(t_1, t_2)}{f_{x_2}(t_2)}$$

By defination, $f_{\underset{\sim}{x}}(t_1, t_2)$ is joint distribution, $f_{x_2}(t_2)$ is marginal distribution. And, we have,

$$f_{\underset{\sim}{x}}(t_1, t_2) = f_{x_1|x_2}(t_1|t_2) f_{x_2}(t_2)$$

$$f_{x_2|x_1}(t_2|t_1) = \frac{f_{\underset{\sim}{x}}(t_1, t_2)}{f_{x_1}(t_1)}$$

$$f_{\underset{\sim}{x}}(t_1, t_2) = f_{x_2|x_1}(t_2|t_1) f_{x_1}(t_1)$$

So,

$$f_{x_1|x_2}(t_1|t_2) = \frac{f_{x_2|x_1}(t_2|t_1) f_{x_1}(t_1)}{f_{x_2}(t_2)} \longrightarrow \text{Bayes Rule}$$

e.g.

x_2 \ x_1	1	2	3	$f_{x_2}(t)$
0	0.1	0.4	0.2	0.7
1	0.2	0.05	0.05	0.3
$f_{x_1}(t)$	0.3	0.45	0.45	

$$P[x_2 = 0] = P[x_2 = 0|x_1 = 1] + P[x_2 = 0|x_1 = 2] + P[x_2 = 0|x_1 = 3] = 0.7$$

$$P[x_2 = 1] = 1 - P[x_2 = 0] = 0.3$$

4 Bayes Classification

$$f_{x_1|x_2}(t_1|t_2) = \frac{f_{x_2|x_1}(t_2|t_1) f_{x_1}(t_1)}{f_{x_2}(t_2)} \longrightarrow \text{Bayes Rule}$$

For discrete: $f_{x_2}(t_2) = \sum_{t_1} f_{x_2|x_1}(t_2|t_1) f_{x_1}(t_1)$

For continuous: $f_{x_2}(t_2) = \int_{-\infty}^{\infty} f_{x_2|x_1}(t_2|t_1) f_{y_1}(t_1) dt$

e.g. Height Supposed that: $x_1 \longrightarrow$ Height, $x_2 \longrightarrow$ Gender, $\begin{cases} 0 & \text{male} \\ 1 & \text{female} \end{cases}$

For $x_2 = 0 \longrightarrow$ Height $\sim N(69, 4.5) \Leftrightarrow f_{x_1|x_2}(t_1|t_2 = 0)\phi(t_1|\phi = 69, \sigma = 4.5)$

For $x_2 = 1 \longrightarrow$ Height $\sim N(65, 4.2) \Leftrightarrow f_{x_1|x_2}(t_1|t_2 = 1)\phi(t_1|\phi = 65, \sigma = 4.2)$

Marginal distribution of height for people:

$$f_{x_1}(t_1) = f_{x_1|x_2}(t_1|t_2 = 0) * f_{x_2}(0) + f_{x_1|x_2}(t_1|t_2 = 1) * f_{x_2}(1)$$
$$= \phi(t_1|69, 4.5) \cdot 0.5 + \phi(t_2|65, 4.2) * 0.5$$
$$= \phi(t|\frac{69 + 65}{2}, \sqrt{\frac{4.5^2 + 4.2^2}{2}})$$

A person has height 6'7", caculate the probability of each gender.

$$f(x_2 = 0|x_1 = 67) = \frac{f_{x_1|x_2}(67|x_2 = 0) f_{x_2}(0)}{f_{x_1}(67)} = \frac{\phi(67|69, 4.5) \cdot 0.5}{f_{x_1}(67)}$$

$$f(x_2 = 1|x_1 = 67) = \frac{f_{x_1|x_2}(67|x_2 = 1)f_{x_2}(1)}{f_{x_1}(67)} = \frac{\phi(67|65, 4.2) \cdot 0.5}{f_{x_1}(67)}$$

$$\text{Loss}(\hat{f}, \underset{\sim}{x}|f), \rightarrow \text{Minimum} E_x : Loss(\hat{f}|f)$$

Misclassification Rate: Loss$(\hat{f}, \underset{\sim}{x}|f) : \begin{cases} 0, & f_x = \hat{f}(x) \\ 1, & f_x \neq \hat{f}(x) \end{cases}$, Risk $= E_x[Loss(\hat{f}, \underset{\sim}{x}|f)]$

Probability of Misclassification: $E_x = \begin{cases} \sum_{t_i} t_i f_x(t_i), & \text{for discrete} \\ \int_0^1 t \int_x (t)dt, & \text{for continuous} \end{cases}$

Bayes Classification Rule for Binary: choose k that $P[y = k|\underset{\sim}{x}]$,

$$k = \underset{k}{\text{argmax}} \frac{P[x|k]P[k]}{P[x]} \propto P[x|k]P[k]$$

For cost matrix C, assumed that

$$C_{ij} = \text{the cost of classification } i \text{by wrong classified in } j$$

$$E_x(loss(j = \hat{f}(x)|i) = \sum_{i=1}^{k} P[i|x]C_{ij} = \sum_{i=1}^{k} \frac{f(x|i)P[i]C_{ij}}{f_x} \cdot C_{ij}$$

$$\propto \sum_{i=1}^{k} f(x|i)P[i]C_{ij}$$

us

Example: In a box $\frac{1}{4}$ of coins are fake, $\frac{3}{4}$ of coins are real
The probability to get fake: $P_r[head] = \frac{1}{3}, P_r[tail] = \frac{2}{3}$
The probability to get real: $P_r[head] = \frac{1}{2}, P_r[tail] = \frac{2}{2}$
Take a random coin selected, n= 20 times, t= 7 heads, what is $P_r[real]$? what is $P_r[false]$?
x_1= number of heads in n= 20 trials
$x_2 = \begin{cases} 0, \text{fake} & f_{x_2}(0) = \frac{1}{4} \\ 1, \text{real} & f_{x_2}(1) = \frac{3}{4} \end{cases}$
$f(t_2 = 0|x_1 = 7, n = 20) = \frac{f_{x_1|x_2}(7|fake, n=20) * f_{x_2}(0)}{f_{x_1}(7)} = \binom{20}{7} \frac{1}{3}^7 \frac{2}{3}^1 3 * 0.25 = 0.45$
$f(t_2 = 1|x_1 = 7, n = 20) = \frac{f_{x_1|x_2}(7|real, n=20) * f_{x_2}(1)}{f_{x_1}(7)} \binom{20}{7} \frac{1}{2}^7 \frac{1}{2}^1 3 * 0.75 = 0.55$

4.1 Bayes Classfication

4.2 Modify Bayes Rule(Uneven Cost)

$c = avgmin \sum_{i=1}^{k} f(x|i)P_r[i]C_{ij}$
Example: Real and Fake: $C = \begin{pmatrix} 0 & 1 \\ 4 & 0 \end{pmatrix}$

real:$P_r[x_2 = 1|x_1 = 7] * C_{11} + P_r[x_2 = 0|x_1 = 7] * C_{12}$
fake:$P_r[x_2 = 1|x_1 = 7] * C_{21} + P_r[x_2 = 0|x_1 = 7] * C_{22}$

$$min : \begin{cases} 0.55 * 0 + 0.45 * 1 = 0.45 \\ 0.55 * 4 + 0.45 * 0 = 1.8 \end{cases}$$