Computer Science

CSC380: Principles of Data Science

Probability 4

Xinchen Yu

Recap

- PDF of a transformation of a continuous RV
  - $X + b$ has a PDF that is a translation of X's PDF by b units
  - aX's PDF is X's PDF stretched by a factor of a horizontally
- Mean
  - $\mathrm{E}[\mathrm{X}] = \int x\, f(x)\, dx$
  - $\mathrm{E}[r(X)] = \int r(x) f(x)\, dx$
- Variance
  - $\mathrm{Var}(X) = \sigma^2 = \mathrm{E}[(X-\mu)^2] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$
- Properties
  - $\mathrm{E}[aX] = a\,\mathrm{E}[X]$
  - $\mathrm{Var}(aX) = a^2 \mathrm{Var}(X)$
  - $\mathrm{E}[X+b] = \mathrm{E}[X] + b$
  - $\mathrm{Var}(X+b) = \mathrm{Var}(X)$

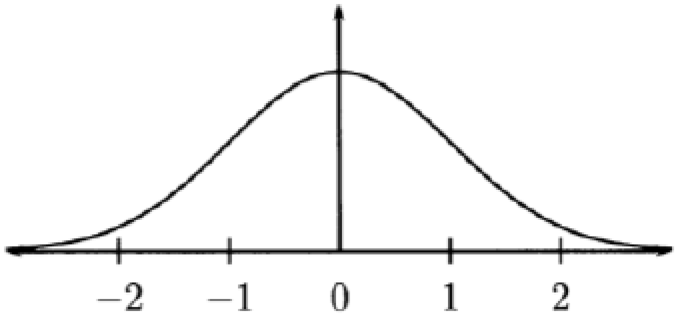
Outline

- Calculating probabilities about Gaussians

- Multivariate Random Variables
    - Joint distribution vs. Marginal distribution

The standard Gaussian distribution

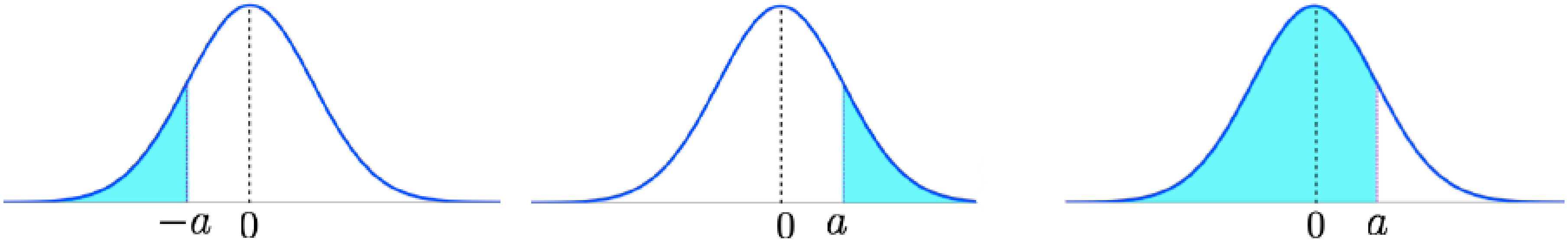
- Gaussian distribution with $\mu = 0$ and $\sigma^2 = 1$



- Denoted by $Z \sim N(0,1)$
- Its PDF denoted by $\phi(z)$, and CDF denoted by $\Phi(z)$

Calculating probabilities about Gaussians

- Symmetry of ϕ => $\Phi(-a) = 1 - \Phi(a)$



$\Phi(-a) = P(Z \leq -a)$ $= P(Z \geq a)$ $= 1 - P(Z \leq a) = 1 - \Phi(a)$

Calculating probabilities about Gaussians

- Suppose $X \sim N(5, 2^2)$, how can I calculate $P(1 < X < 8)$?

- Transform X into standard normal Z:
  - $X \sim N(\mu, \sigma^2)$

  $\Rightarrow X - \mu \sim N(0, \sigma^2)$

  $\Rightarrow Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

$$\mathrm{E}[aX] = a\,\mathrm{E}[X]$$
$$\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$$
$$\mathrm{E}[X+b] = \mathrm{E}[X] + b$$
$$\mathrm{Var}(X+b) = \mathrm{Var}(X)$$

- We can write $\mathrm{P}(a < X < b)$ using $\mathrm{P}(c < Z < d)$, which in turn can be written in Φ.

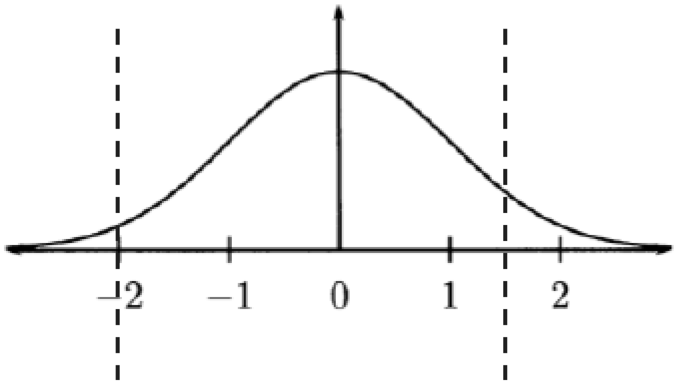
Calculating probabilities about Gaussians

- $P(a < X < b)$

$$= P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right)$$

$$= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$



Example Suppose $X \sim N(5, 2^2)$, calculate $P(1 < X < 8)$

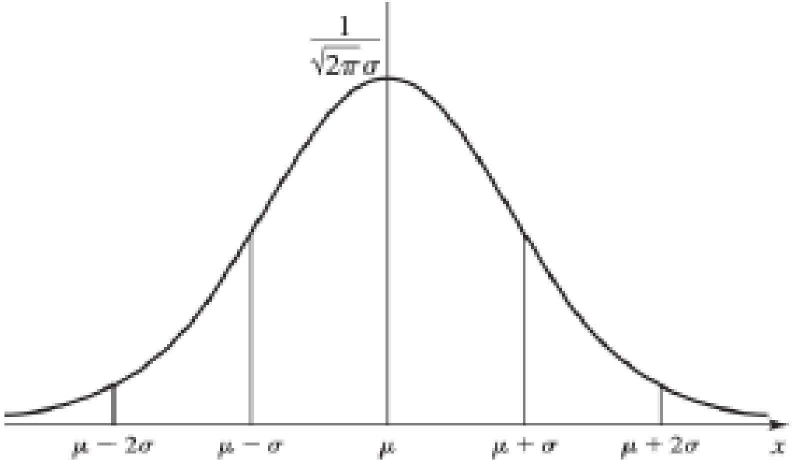
This is $\Phi\left(\frac{8-5}{2}\right) - \Phi\left(\frac{1-5}{2}\right) = \Phi(1.5) - \Phi(-2) = \Phi(1.5) - (1 - \Phi(2))$

```
from scipy.stats import norm
print(norm.cdf(1.5)-(1-norm.cdf(2)))
```

```
0.9104426667829627
```

Calculating probabilities about Gaussians

- What is the probability that a Gaussian RV X is within k ($k = 1, 2, \ldots$) std of its mean?



- $\mathrm{P}(\mu - k\sigma \le X < \mu + k\sigma)$

Calculating probabilities about Gaussians

- $p_k = \mathrm{P}(\mu - k\sigma \leq X < \mu + k\sigma)$

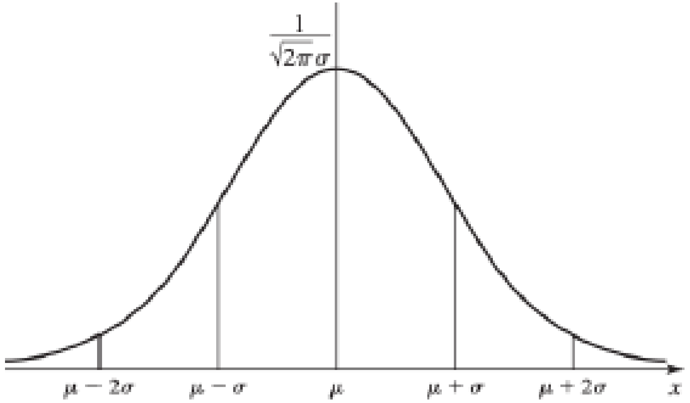
$= P\left(-k < \frac{X-\mu}{\sigma} < k\right)$

$= P(-k < Z < k)$

$= \Phi(k) - (1 - (\Phi(k))$

$= 2\Phi(k) - 1$

k	p_k
1	0.6826
2	0.9544
3	0.9974
4	0.99994

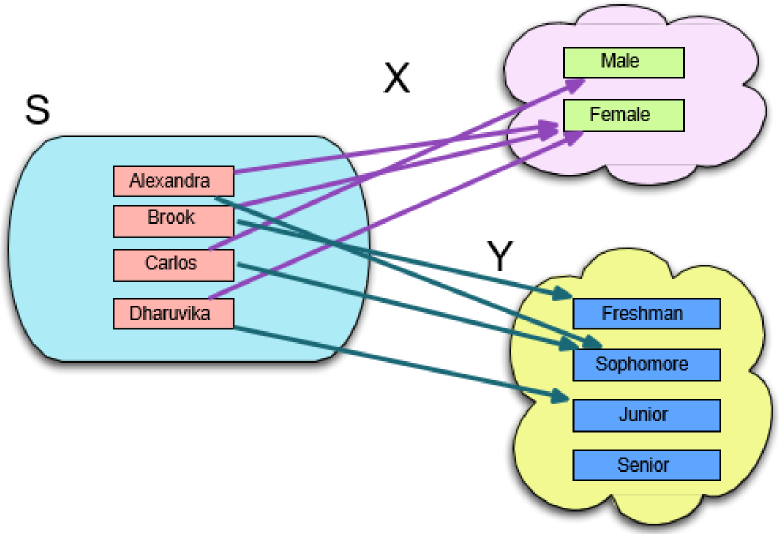


In words,

- With probability about 95%, X is within 2 std of its mean
- With overwhelming prob. (99.7%), X within 3 std of mean

Multivariate Random Variables

Multivariate RVs: example



- X: people -> their genders
- Y: people -> their class year
- We'd like to answer questions such as: does X and Y have a correlation?
  - I.e., is a student in higher class year more likely to be male?
- We call (X, Y) a multivariate RV, and will study its *joint* distribution

Joint distribution of discrete RVs

- The joint PMF (probability mass function) of discrete random variables X, Y:

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

Examples

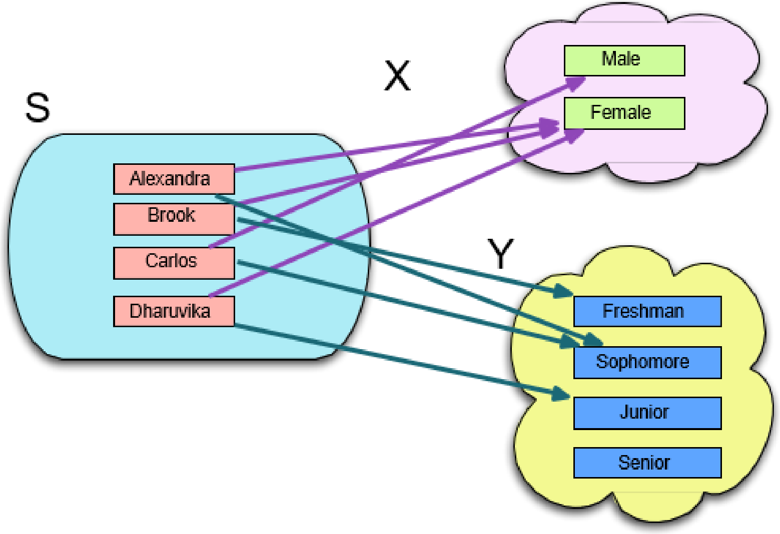
Alexandra

$$P(X = \mathrm{Fem}, Y = \mathrm{Soph}) = \frac{1}{4}$$

Dharuvika

$$P(X = \mathrm{Fem}, Y = \mathrm{Jun}) = \frac{1}{4}$$

…



Marginal distributions

Given joint distribution of (X, Y), need distribution of one of them, say X.

- Named the *marginal distribution* of X.
- How to find $P(X = x)$?
- Using law of total probability:

$$f_1(x) = \sum_y f(x, y)$$

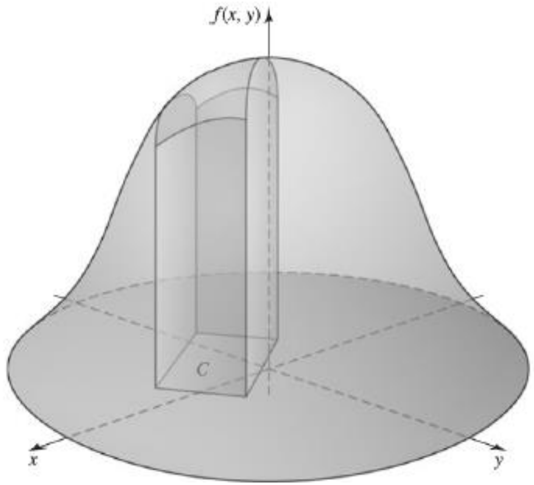
	y			
x	1	2	3	4
1	0.1	0	0.1	0
2	0.3	0	0.1	0.2
3	0	0.2	0	0

- This operation is called *marginalization* ('marginalizing out variable Y', or variable elimination)

Joint distribution of continuous RVs

- Any continuous random vector (X,Y) has a *joint probability density function* (PDF) $f(x,y)$, such that for all C,

$$P\big((X,Y) \in C\big) = \iint_C f(x,y)\,dx\,dy$$



$f(x,y)$: represent a 2D surface

double integral: the *volume* under the surface

Properties:

- f is nonnegative
- $\iint_{R^2} f(x,y)\,dx\,dy = 1$ ($R^2 =$ the whole x-y plane)
  - $P\big((X,Y) \in R^2\big) = 1$

Marginal distribution of continuous RV

Given joint distribution of continuous RV (X, Y), how to find X's PDF f_1?

Fact (marginalization) $f_1(x) = \int_R f(x, y)\, dy$

Replacing summation with integration in the continuous case ('marginalizing / integrating out variable Y')

How about Y's PDF f_2?

- Marginalize out X

Joint distribution of more than 3 RVs

- We can consider the joint distribution of more than 3 random variables,
  - E.g. (A,B,C), A = gender, B = class year, C = blood type
- Discrete RVs: can still define joint PMFs

a	b	c	$P(A=a, B=b, C=c)$
0	0	0	0.06
0	0	1	0.09
0	1	0	0.08
0	1	1	0.12
1	0	0	0.06
1	0	1	0.24
1	1	0	0.10
1	1	1	0.25

Marginalization

a	b	c	$P(A=a,B=b,C=c)$
0	0	0	0.06
0	0	1	0.09
0	1	0	0.08
0	1	1	0.12
1	0	0	0.06
1	0	1	0.24
1	1	0	0.10
1	1	1	0.25

Given the joint distribution of (A, B, C)

- What is the distribution of A?
  - Need to find $P(A=0)$ and $P(A=1)$

$$P(A=0) = \sum_{b,c} P(A=0, B=b, C=c)$$

Marginalization: summing over irrelevant variables

- What is the joint distribution of (A, B)?
  - Need to find $P(A=0, B=0), \ldots, P(A=1, B=1)$

$$P(A=0, B=0) = \sum_{c} P(A=0, B=0, C=c)$$

Marginalization for continuous RVs

Suppose joint PDF of (A, B, C) is $f(a, b, c)$

- What is the PDF of A?

$$f_A(a) = \iint_{R^2} f(a, b, c)\ db\, dc$$

- What is the joint PDF of (A, B)?

$$f_{A,B}(a, b) = \int_R f(a, b, c) dc$$

Marginalization: summing over irrelevant variables

- These operations generalize to joint PDFs of more RVs..