10601

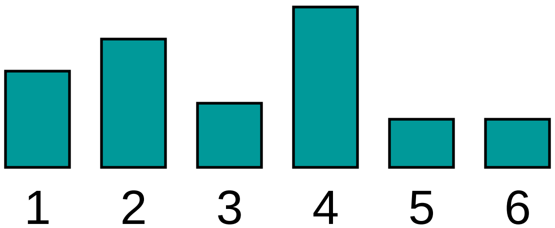
Probabilistic Reasoning and Inference:
Statistics and distributions

Outline

- Continuous distributions
  - Probability density functions, Cumulative density functions
  - Recap on the probability rules
- Gaussian distribution, multivariate Gaussian
- Density estimation example
  - Joint density estimation
  - Naïve density estimation
- Preview of Bayesian networks

Probability Density Function

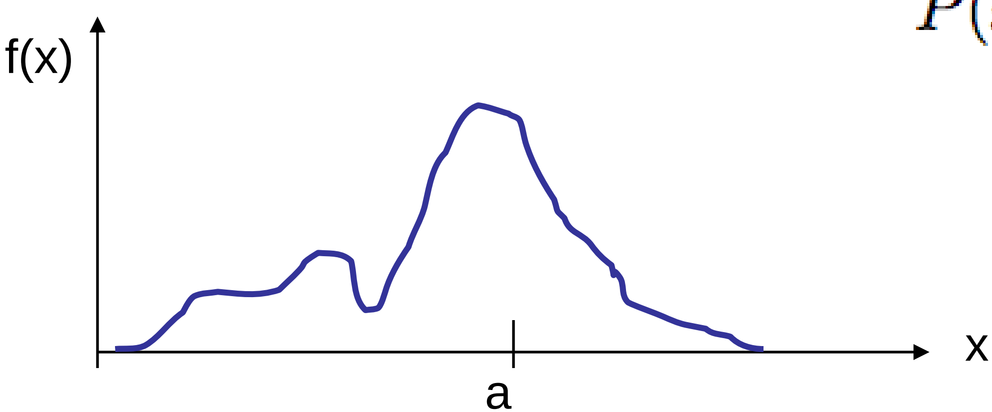
- Discrete distributions



X is the event space

$$\sum_i P(X = x_i) = 1$$

- Continuous: Cumulative Density Function (CDF): *F(a)*



$$P(x \leq a) = \int_{-\infty}^{a} f(\tau) d\tau$$

Cumulative Density Functions

- Total probability

$$P(\Omega) = \int_{-\infty}^{\infty} f(x)dx = 1$$

- Probability Density Function (PDF)

$$\frac{d}{dx}F(x) = f(x)$$

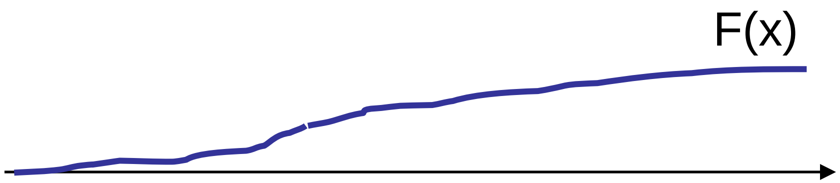
- Properties:

$$P(a \leq x \leq b) = \int_{b}^{a} f(x)dx = F(b) - F(a)$$

$$\lim_{x \to -\infty} F(x) = 0$$

$$\lim_{x \to \infty} F(x) = 1$$

$$F(a) \geq F(b) \; \forall a \geq b$$



Expectations

- Mean/Expected Value: $$E[x] = \bar{x} = \int x f(x) dx$$
- Variance:
  - Note: $$Var(x) = E[(x - \bar{x})^2] = E[x^2] - (\bar{x})^2$$
- In general: $$E[x^2] = \int x^2 f(x) dx$$ $$E[g(x)] = \int g(x) f(x) dx$$

Multivariate

- Joint for (x,y)

$$P((x,y) \in A) = \int \int_A f(x,y) dx dy$$

- Marginal:

$$f(x) = \int f(x,y) dy$$

- Conditionals:

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

- Chain rule:

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

Bayes Rule

- Standard form:

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

- Replacing the bottom:

$$f(x|y) = \frac{f(y|x)f(x)}{\int f(y|x)f(x)dx}$$

Binomial

- Distribution: a <u>discrete </u>probability distribution of the number of successes in a sequence of *n* <u>independent</u> yes/no experiments.
  p is the probability of success
- Mean/Var: $x \sim Binomial(p, n)$

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

$$E[x] = np$$

$$Var(x) = np(1 - p)$$

Uniform

- Anything is equally likely in the region [a,b]

- Distribution: $x \sim U(a,b)$

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & otherwise \end{cases}$$

- Mean/Var

$$E[x] = \frac{a+b}{2}$$

$$Var(x) = \frac{a^2+ab+b^2}{3}$$

Gaussian (Normal)

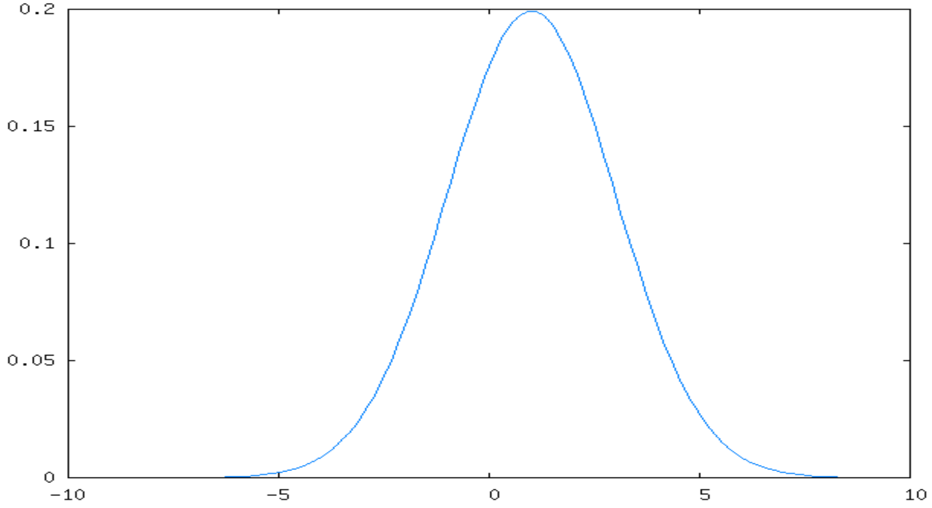
- If I look at the height of women in country xx, it will look approximately Gaussian
- Distribution:

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

- Mean/var $x \sim N(\mu, \sigma^2)$

$$E[x] = \mu$$

$$Var(x) = \sigma^2$$



Why Do People Use Gaussians

- Central Limit Theorem: (loosely)

  Sum of a large number of independent and identically distributed (IID) random variables is approximately Gaussian

Multivariate Gaussians

- Distribution for vector x

$$x = (x_1, \ldots, x_N)^T, \quad x \sim N(\mu, \Sigma)$$

- PDF:

$$f(x) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$$E[x] = \mu = (E[x_1], \ldots, E[x_N])^T$$

$$Var(x) \rightarrow \Sigma = \begin{pmatrix} Var(x_1) & Cov(x_1, x_2) & \ldots & Cov(x_1, x_N) \\ Cov(x_2, x_1) & Var(x_2) & \ldots & Cov(x_2, x_N) \\ \vdots & & \ddots & \vdots \\ Cov(x_N, x_1) & Cov(x_N, x_2) & \ldots & Var(x_N) \end{pmatrix}$$

Multivariate Gaussians

$$f(x) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

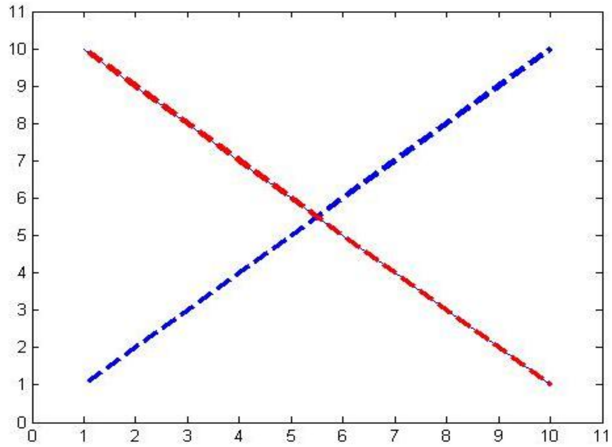
$$E[x] = \mu = (E[x_1], \ldots, E[x_N])^T$$

$$Var(x) \to \Sigma = \begin{pmatrix} Var(x_1) & Cov(x_1, x_2) & \ldots & Cov(x_1, x_N) \\ Cov(x_2, x_1) & Var(x_2) & \ldots & Cov(x_2, x_N) \\ \vdots & & \ddots & \vdots \\ Cov(x_N, x_1) & Cov(x_N, x_2) & \ldots & Var(x_N) \end{pmatrix}$$

$$\mathrm{cov}(\chi_1, \chi_2) = \frac{1}{n} \sum_{i=1}^{n} (x_{1,i} - \mu_1)(x_{2,i} - \mu_2)$$

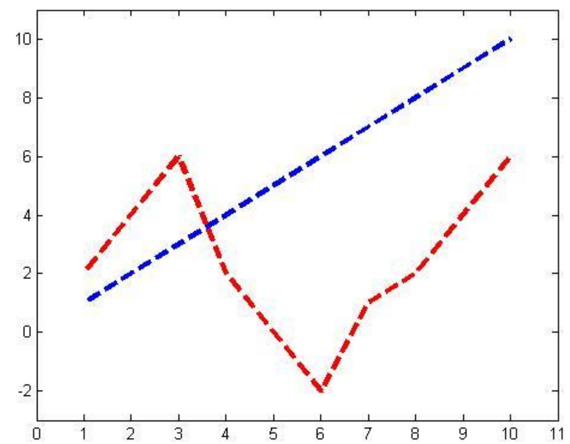
Covariance examples

Anti-correlated



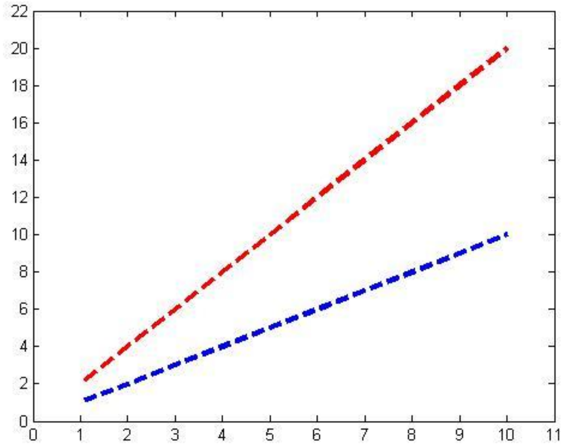
Covariance: -9.2

Independent (almost)



Covariance: 0.6

Correlated



Covariance: 18.33

Sum of Gaussians

- The sum of two Gaussians is a Gaussian:

$$x \sim N(\mu, \sigma^2) \quad y \sim N(\mu_y, \sigma_y^2)$$

$$ax + b \sim N(a\mu + b, (a\sigma)^2)$$

$$x + y \sim N(\mu + \mu_y, \sigma^2 + \sigma_y^2)$$

Independence

- In some cases the additional information does not help

P(slept) = 0.5

P(slept | rain = 1) = 0.5

- In this case, the extra knowledge about rain does not change our prediction
- Slept and rain are independent!

Liked movie	Slept	raining	P
1	1	1	0.05
1	0	1	0.1
0	0	1	0.025
0	1	1	0.075
1	1	0	0.15
1	0	0	0.3
0	0	0	0.075
0	1	0	0.225

Independence (cont.)

- Notation: $P(S \mid R) = P(S)$
- Using this we can derive the following:

  - $P(\neg S \mid R) = P(\neg S)$

  - $P(S,R) = P(S)P(R)$

  - $P(R \mid S) = P(R)$

Independence

- Independence allows for easier models, learning and inference
- For our example:
  - P(raining, slept movie) = P(raining)P(slept movie)
  - Instead of 4 by 2 table (4 parameters), only 2 are required
  - The saving is even greater if we have many more variables …
- In many cases it would be useful to assume independence, even if its not the case

Conditional independence

- Two dependent random variables may become independent when conditioned on a third variable:

  P(A,B | C) = P(A | C) P(B | C)

- Example

  P(liked movie) = 0.5

  P(slept) = 0.4

  P(liked movie, slept) = 0.1

  P(liked movie | long) = 0.4

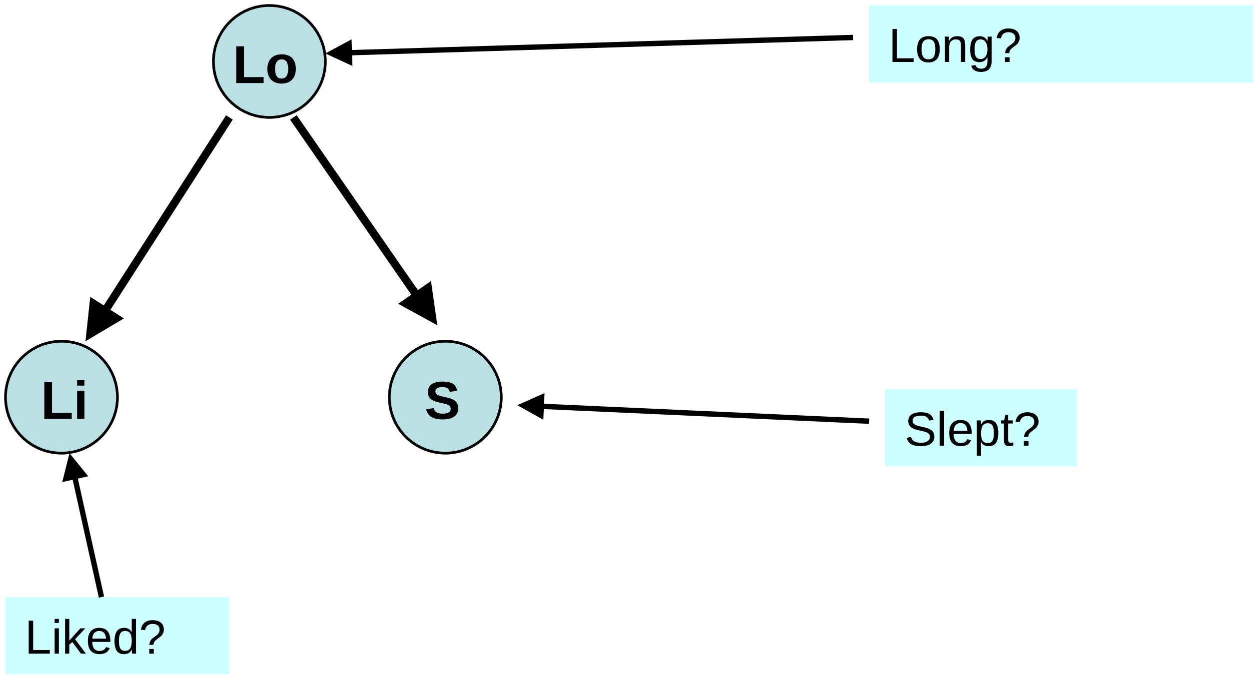
  P(slept | long) 0.6

  P(slept, like movie | long) = 0.24

Given knowledge of length, the two other variables become independent

Bayesian networks

- Bayesian networks are *directed graphs* with nodes representing *random variables* and edges representing *dependency assumptions*



What you should know

- Thoroughly understand:
  - Probability theory
  - The different distributions