(Brief) Intro to probability and Density Estimation

Basic notations

- Random Variable
  - referring to an element/event whose status/value is unknown
- Example A = “it will rain tomorrow”
- Domain: (usually denoted by Ω)
  - A = “CS534 will be canceled on Friday”: binary
  - A = “Your CS534 grade”: categorical (discrete)
  - A = “The amount of time you will spend each week on studying for CS534”: continuous

Axioms of probability (Kolmogorov's axioms)

- A variety of useful facts can be derived from just three axioms:

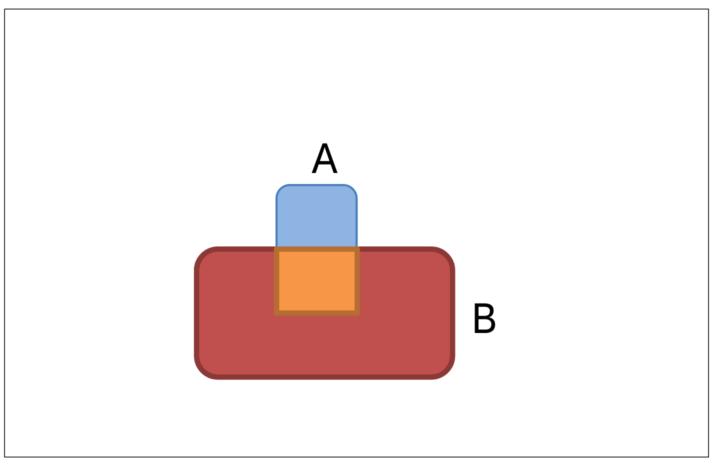
1. $0 \leq P(A) \leq 1$
2. $P(true) = 1$, $P(false) = 0$
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

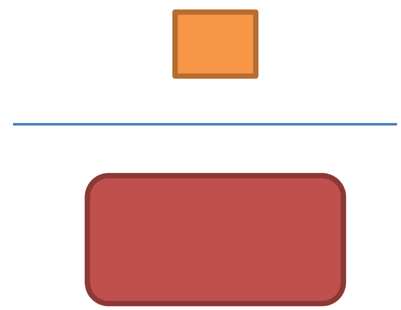
Joint Distribution

- The probability that a set of random variables will take a specific value combination
- Notation: $P(A \wedge B)$ or $P(A, B)$ - probability that both A and B are true
- Example: P(Headache, Flu)
- If two variables are independent then $P(A,B) = P(A)P(B)$

Conditional Probability

- P(A|B) = Fraction of worlds in which B is true that also have A true

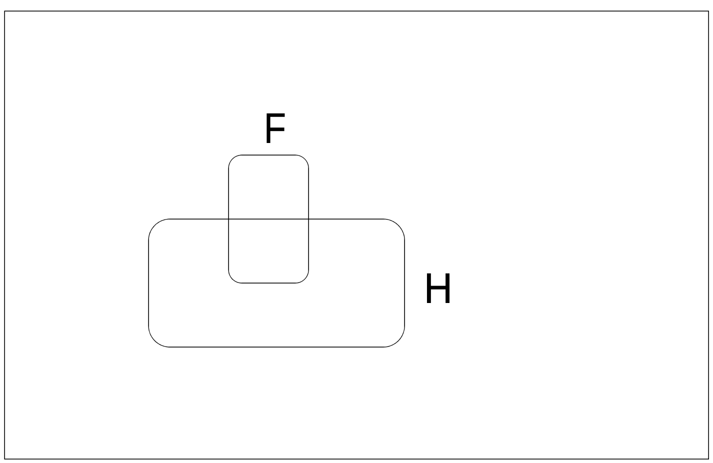




- If A and B are independent, P(A|B)=P(A)

Conditional Probability

- Some times, knowing one or more random variables can improve upon our prior belief of another random variable



H = "Have a headache"
F = "Coming down with Flu"

P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

"Headaches are rare (1/10), but if you're coming down with 'flu there's a 50-50 chance you'll have a headache."

Chain Rule

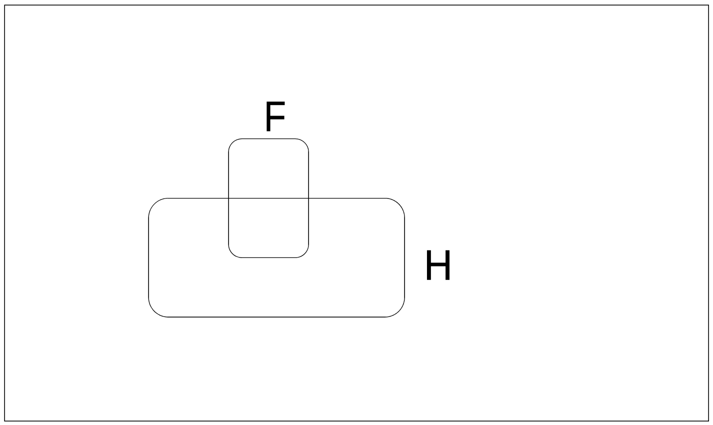
$$P(A \wedge B) = P(A|B)\, P(B)$$

- Chain rule can be used to derive the Bayes rule:

$$P(A \wedge B) = P(A|B)\, P(B) = P(B|A) P(A)$$

$$\Rightarrow \quad P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

Probabilistic Inference



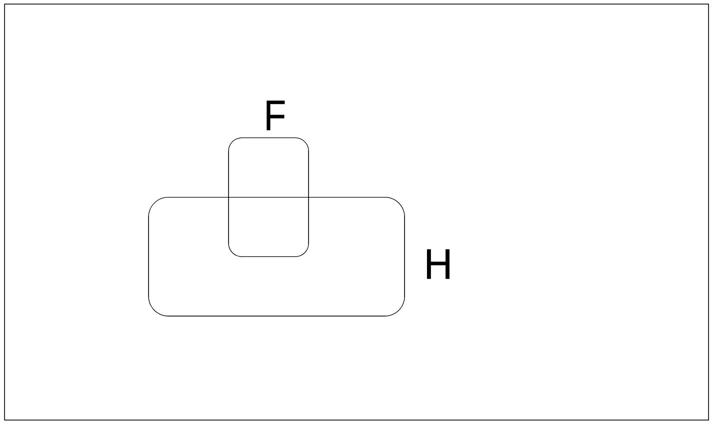
H = “Have a headache”
F = “Coming down with Flu”

$P(H) = 1/10$
$P(F) = 1/40$
$P(H|F) = 1/2$

One day you wake up with a headache. You think: “Drat! 50% of flus are associated with headaches so I must have a 50-50 chance of coming down with flu”

Is this reasoning good?

Probabilistic Inference



H = “Have a headache”
F = “Coming down with Flu”

P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

Prior: the degree of belief in an event in the absence of any other information

$$P(F|H) = \frac{P(F \wedge H)}{P(H)} = \frac{P(H|F)P(F)}{P(H)} = \frac{\frac{1}{40} * \frac{1}{2}}{1/10} = \frac{1}{8}$$

Posterior: the degree of belief in an event after obtaining some evidential information

More General Forms of Bayes Rule

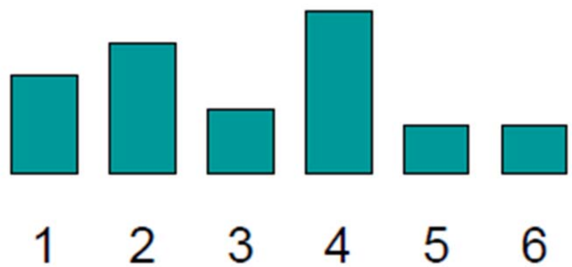
$$P(A|B \wedge X) = \frac{P(B|A \wedge X)P(A \wedge X)}{P(B \wedge X)}$$

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

$$P(A = v_i|B) = \frac{P(B|A = v_i)P(A = v_i)}{\sum_{k=1}^{n_A} P(B|A = v_k)P(A = v_k)}$$

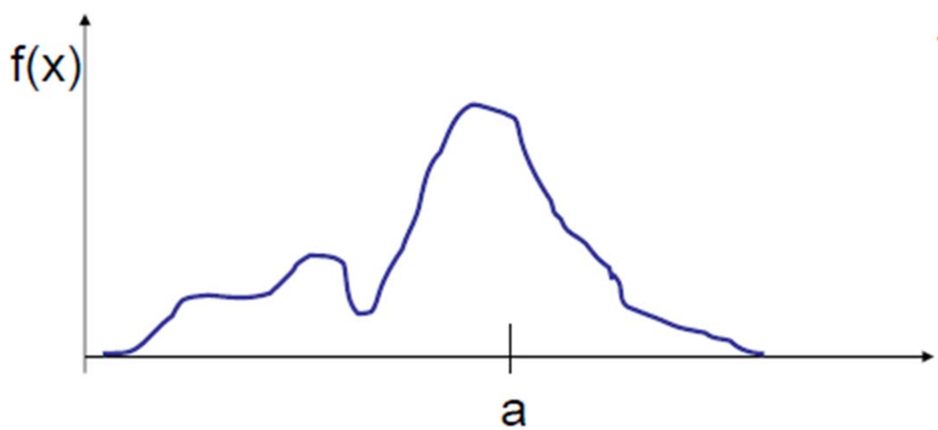
Probability Density Function

- Discrete distribution:



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

- Continuous: Probability density function (PDF) f(x)



Cumulative density function

- Cumulative Density Function F(*x*):

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

- Properties:

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

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

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

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

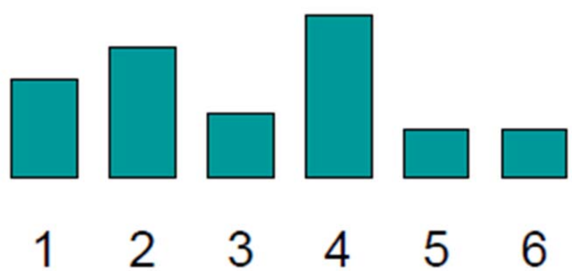
$$F(a) \geq F(b)\ \forall a \geq b$$

Multivariate

- Joint distribution of x and y is described by a **pdf** function f(x, y): $$P((x,y) \in A) = \int\int_A f(x,y)dxdy$$
- Marginal: $$f(x) = \int f(x,y)dy$$
- Conditional: $$f(x|y) = \frac{f(x,y)}{f(y)}$$
- Chain rule: $$f(x,y) = f(x|y)f(y) = f(y|x)f(x)$$
- Bayes rule: $$f(x|y) = \frac{f(y|x)f(x)}{f(y)} = \frac{f(y|x)f(x)}{\int f(y|x)f(x)dx}$$

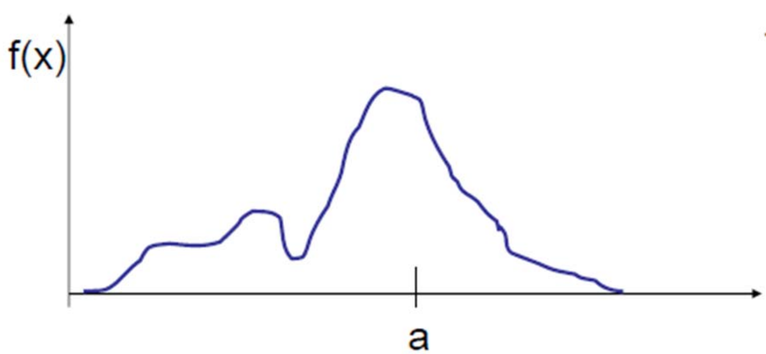
Expectations

- Expectation of a random variable of x is the weighted average of all possible values that x can take
- **Discrete :**



$$\bar{X} = E(X) = \sum_i x_i P(X = x_i)$$

- **Continuous**:



$$\bar{X} = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Variance

- Var(x) describes how far the values of x lie from the expected value of x (mean)

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

$$E[x^2] = \int x^2 f(x) dx$$

$$E[g(x)] = \int g(x) f(x) dx$$

Commonly Used Discrete Distributions

Bernoulli distribution: Ber(p)

$$P(x) = \begin{cases} 1-p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases} \quad \Rightarrow \quad P(x) = p^x (1-p)^{1-x}$$

$$\boxed{\begin{aligned} E(x) &= p \\ Var(x) &= p(1-p) \end{aligned}}$$

Binomial distribution: x ~ Binomial(n , p)

the probability to see x heads out of n flips

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

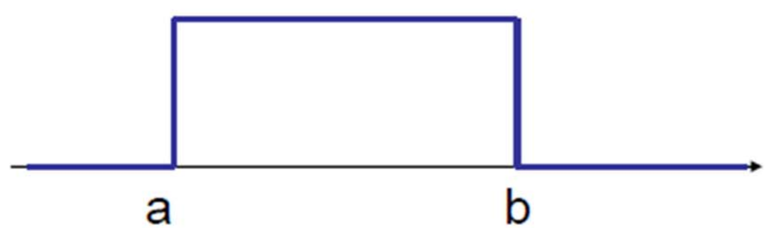
$$\boxed{\begin{aligned} E(x) &= np \\ Var(x) &= np(1-p) \end{aligned}}$$

Continuous Distributions

- Uniform: equal probability within regin [a,b]

$$x \sim U(a,b) \qquad f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & otherwise \end{cases}$$

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



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

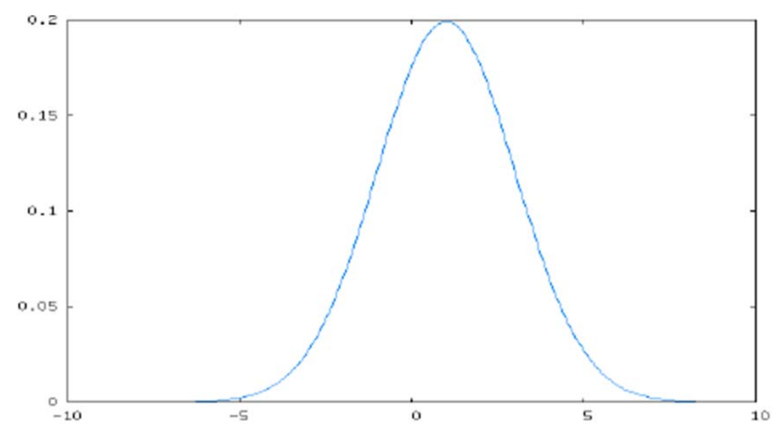
Gaussian (Normal)

- If we look at the height of woman in the US, it will approximately look like Gaussian

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

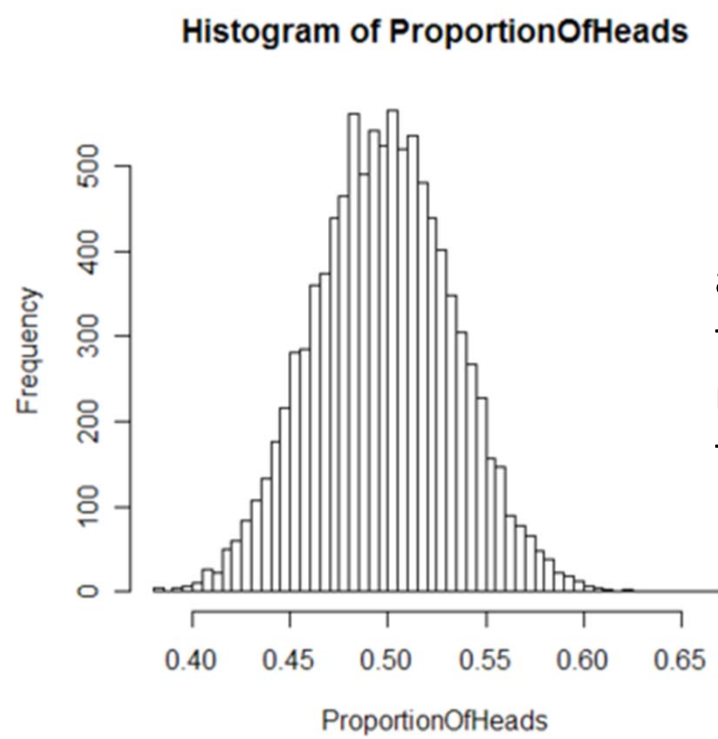
$$E[x] = \mu$$

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



Central Limit Theorem

The sum of a large number of independent random variables is approximately Gaussian



average proportion of heads in a fair coin toss, over a large number of sequences of coin tosses

Multivariate Gaussian

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

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

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

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

$$Cov(x, y) = E((x - \bar{x})(y - \bar{y}))$$

Density Estimation

- Estimate the distribution (or conditional distribution) of a random variable
- Types of variables
  - Binary: coin flip (p)
  - Discrete: dice, grades ($p_i = P(X = x_i)$)
  - Continuous: height, weight, temperature (e.g, $\mu\ and\ \Sigma\ for\ Guassian$)

Maximum Likelihood Principle

We can define the likelihood of the data given the model as follows:

$$\hat{P}(\text{dataset} \mid M) = \hat{P}(x_1 \wedge x_2 \cdots \wedge x_n \mid M) = \prod_{k=1}^{n} \hat{P}(x_k \mid M)$$

M is our model (usually a collection of parameters)

For example M is

- The probability of ‘head’ for a coin flip
- The probabilities of observing 1,2,3,4 and 5 for a dice
- etc.

Maximum Likelihood Principle

$$\hat{P}(\text{dataset} \mid M) = \hat{P}(x_1 \wedge x_2 \cdots \wedge x_n \mid M) = \prod_{k=1}^{n} \hat{P}(x_k \mid M)$$

- Our goal is to determine the values for the parameters in *M*
- We can do this by maximizing the probability of generating the observed samples
- For example, let Θ *be the probabilities for a coin flip*
- Then

$$L(x_1, \ldots, x_n \mid \Theta) = p(x_1 \mid \Theta) \ldots p(x_n \mid \Theta)$$

- The observations (different flips) are assumed to be independent
- For such a coin flip with $P(H)=q$ the best assignment for Θ_h is

  $argmax_q = \#H/\#samples$

- Why?

Maximum Likelihood Principle: Binary variables

- For a binary random variable A with P(A=1)=q
  argmax_q = #1/#samples
- Why?

Data likelihood: $P(D \mid M) = q^{n_1}(1-q)^{n_2}$

We would like to find: $\arg\max_q q^{n_1}(1-q)^{n_2}$

Maximum Likelihood Principle

Data likelihood: $P(D \mid M) = q^{n_1}(1-q)^{n_2}$

We would like to find: $\arg\max_q q^{n_1}(1-q)^{n_2}$

$$\frac{\partial}{\partial q} q^{n_1}(1-q)^{n_2} = n_1 q^{n_1-1}(1-q)^{n_2} - q^{n_1} n_2 (1-q)^{n_2-1}$$

$$\frac{\partial}{\partial q} = 0 \Rightarrow$$

$$n_1 q^{n_1-1}(1-q)^{n_2} - q^{n_1} n_2 (1-q)^{n_2-1} = 0 \Rightarrow$$

$$q^{n_1-1}(1-q)^{n_2-1}(n_1(1-q) - q n_2) = 0 \Rightarrow$$

$$n_1(1-q) - q n_2 = 0 \Rightarrow$$

$$n_1 = n_1 q + n_2 q \Rightarrow$$

$$q = \frac{n_1}{n_1 + n_2}$$

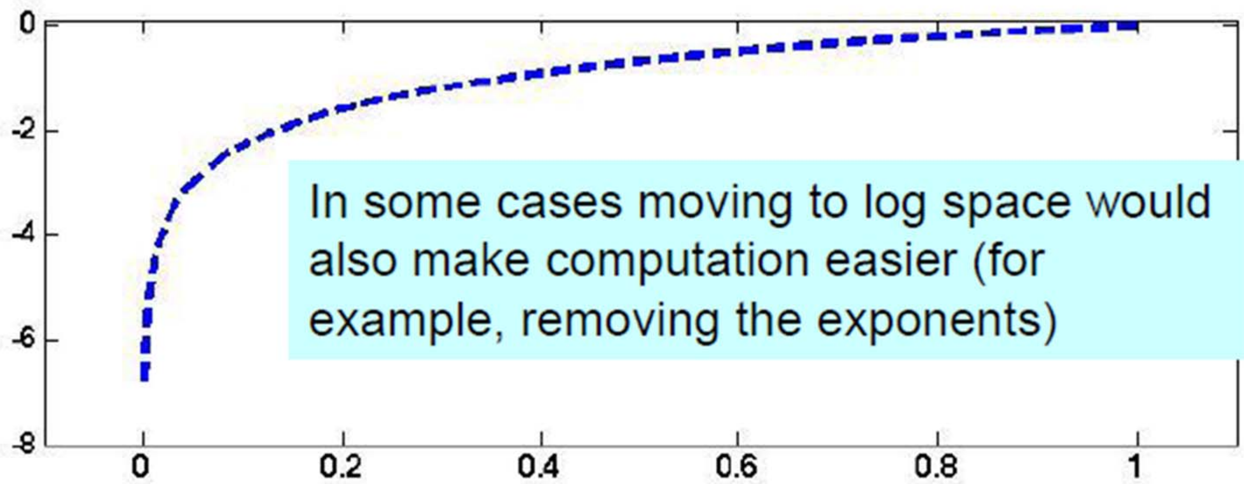
Log Probabilities

When working with products, probabilities of entire datasets often get too small. A possible solution is to use the log of probabilities, often termed 'log likelihood'

$$\log \hat{P}(\text{dataset} \mid M) = \log \prod_{k=1}^{n} \hat{P}(x_k \mid M) = \sum_{k=1}^{n} \log \hat{P}(x_k \mid M)$$

Maximizing this likelihood function is the same as maximizing P(dataset | M)

Log values between 0 and 1



In some cases moving to log space would also make computation easier (for example, removing the exponents)