Probability& Statistics

Machine learning, 2021

Mansoor Rezghi

Department of Computer science, TMU

Ref:KM, CB

References

•K. Murphy Machine Learning: A Probabilistic Perspective, MIT Press, 2012.(KM)-chap2

•C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.(CB)-chap2

Probability: Discrete random variable

Discrete random variable *X*, which

can take on any value from a finite or countable infinite set χ.

X=x event

P(X=x) or P(x)

$0 \leq P(x) \leq 1$

$$\sum_{x \in \mathcal{X}} p(x) = 1$$

Continues random variable X:
Uncertain continues quantity

$$p(x \in (a, b)) = \int_a^b p(x)\,\mathrm{d}x.$$

$$\begin{aligned} p(x) &\geqslant 0 \\ \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x &= 1. \end{aligned}$$

Basic rules

$$p(A \vee B) = p(A) + p(B) - p(A \wedge B)$$

Joint probability

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

Marginal distribution

$$p(A) = \sum_b p(A, B) = \sum_b p(A|B = b)p(B = b)$$

Conditional Probability

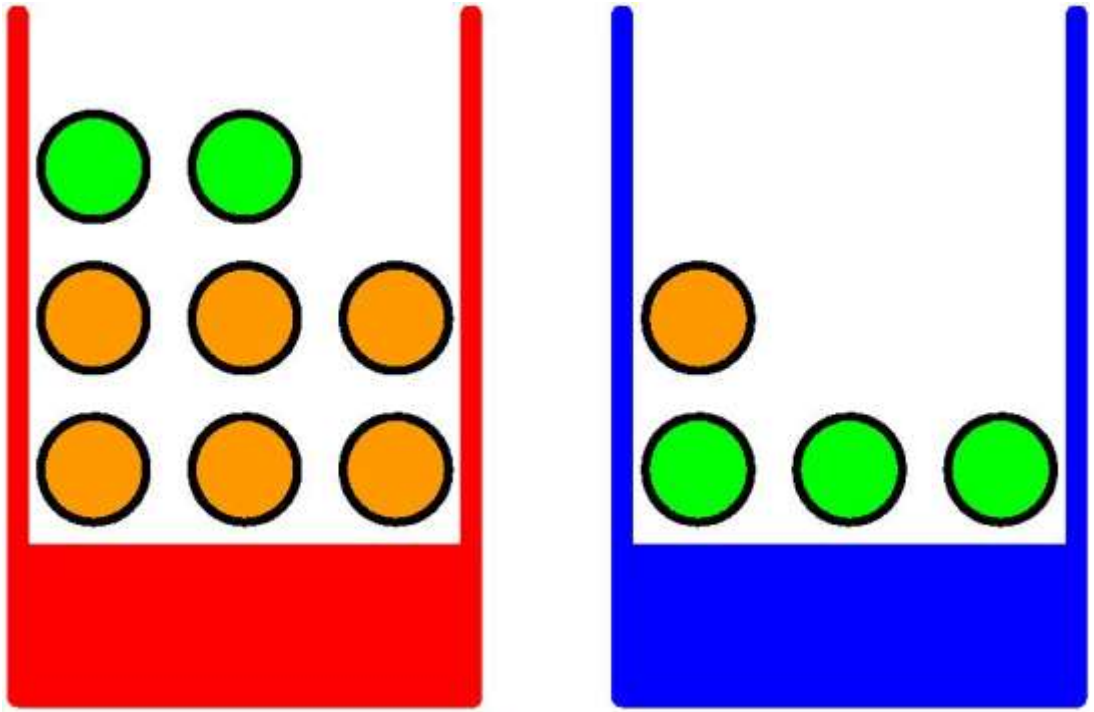
$$p(A|B) = \frac{p(A, B)}{p(B)} \text{ if } p(B) > 0$$

Bayes rule

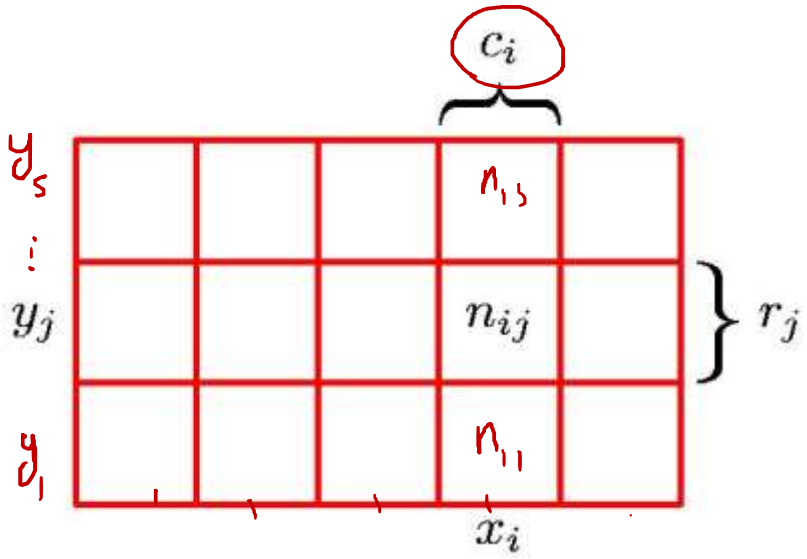
$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{p(X = x)p(Y = y|X = x)}{\sum_{x'} p(X = x')p(Y = y|X = x')}$$

Probability Theory

Apples and Oranges



Probability Theory



Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

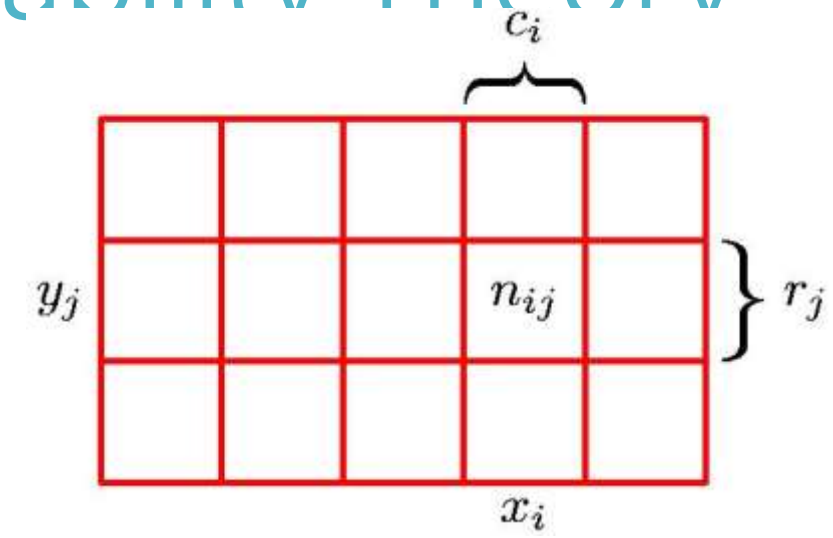
Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Probability Theory



Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

$$= p(Y = y_j | X = x_i) p(X = x_i)$$

$p(x,y) = p(y|x)\, p(x)$ ✓

The Rules of Probability

Sum Rule

$$p(X) = \sum_{Y} p(X, Y)$$

Product Rule

$$p(X, Y) = p(Y|X)p(X)$$

Bayes Theorem

$p(Y|X)$

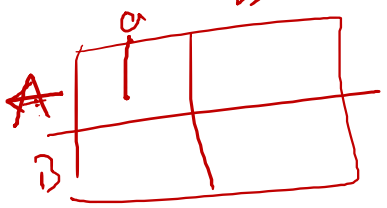
$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(X) = \sum_Y p(X|Y)p(Y)$$

posterior $\propto$ likelihood $\times$ prior

$p(X) = \sum_Y p(X,Y) = \sum_Y p(X|Y)p(Y)$

Bayes rule: Example



Y: you have cancer or NOT,	X: Results of a test about your cancer
Y=1: You have cancer	X=1: test shows your cancer

the test has a **sensitivity** of 80%, which means, if you have cancer, the test will be positive with probability 0.8.

$$p(x=1|y=1)=0.8$$

If $p(y=1)=0.004$ and $p(x=1|y=0)=0.1$ **false positive** or **false alarm.**

$$p(y=1|x=1)=\frac{p(x=1|y=1)p(y=1)}{p(x=1|y=1)p(y=1)+p(x=1|y=0)p(y=0)}$$

$$=\frac{0.8\times 0.004}{0.8\times 0.004+0.1\times 0.996}=0.031$$

Independence

Independence and conditional independence

We say X and Y are **unconditionally independent** or **marginally independent**, denoted $X \perp Y$, if we can represent the joint as the product of the two marginals (see Figure 2.2), i.e.,

$$X \perp Y \iff p(X,Y) = p(X)p(Y) \tag{2.14}$$

$$p(X,Y) = p(X|Y)p(Y)$$

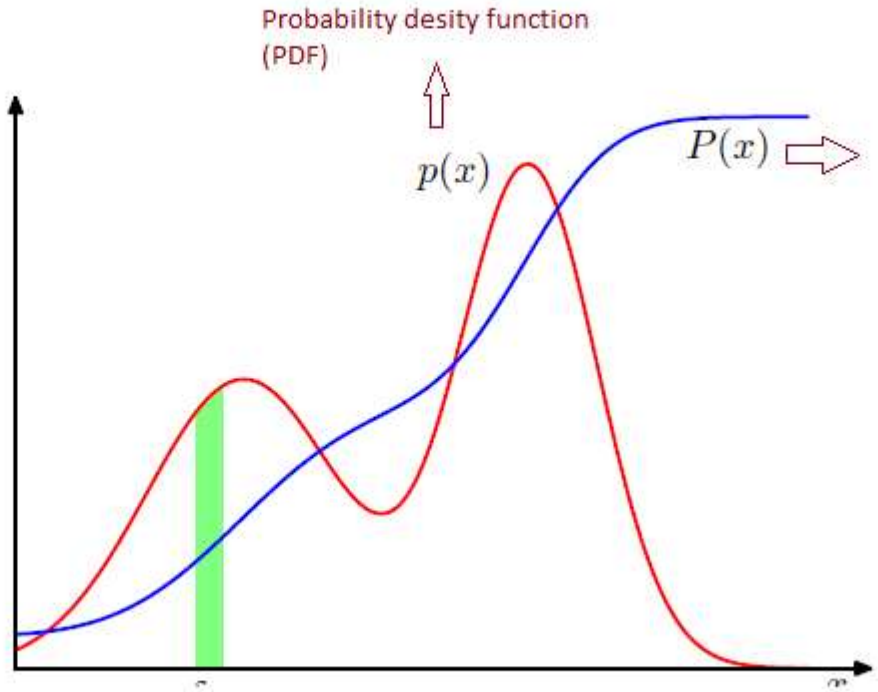
$$p(X) = \frac{p(X|Y)p(Y)}{p(X)}$$

conditionally independent

$$X \perp Y|Z \iff p(X,Y|Z) = p(X|Z)p(Y|Z)$$

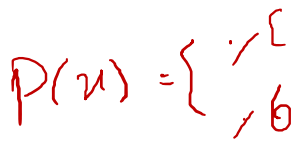
Continues Random Variable

The concept of probability for discrete variables can be extended to that of a probability density $p(x)$ over a continuous variable x and is such that the probability of x lying in the interval $(x, x+\delta x)$ is given by $p(x)\delta x$ for $\delta x \rightarrow 0$. The probability density can be expressed as the derivative of a cumulative distribution function $P(x)$.



cumulative
distribution function (CDF)

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$



Important concepts

Expectations & Variance

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

$$\mathbb{E}[f] = \int p(x) f(x)\, \mathrm{d}x$$

Conditional Expectation (discrete)

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

Variance

$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

$X \longrightarrow p(x) \longrightarrow E(x) = \mu$

۰	۱۰	۲۰
۱۰	۱۰	۱۰

$(X-\mu)^2 \rightarrow p(x)$

$E\left((X-\mu)^2\right) = \sum (x-\mu)^2 p(x)$

$= \sum (x-E(x))^2 p(x)$

$p(x)$

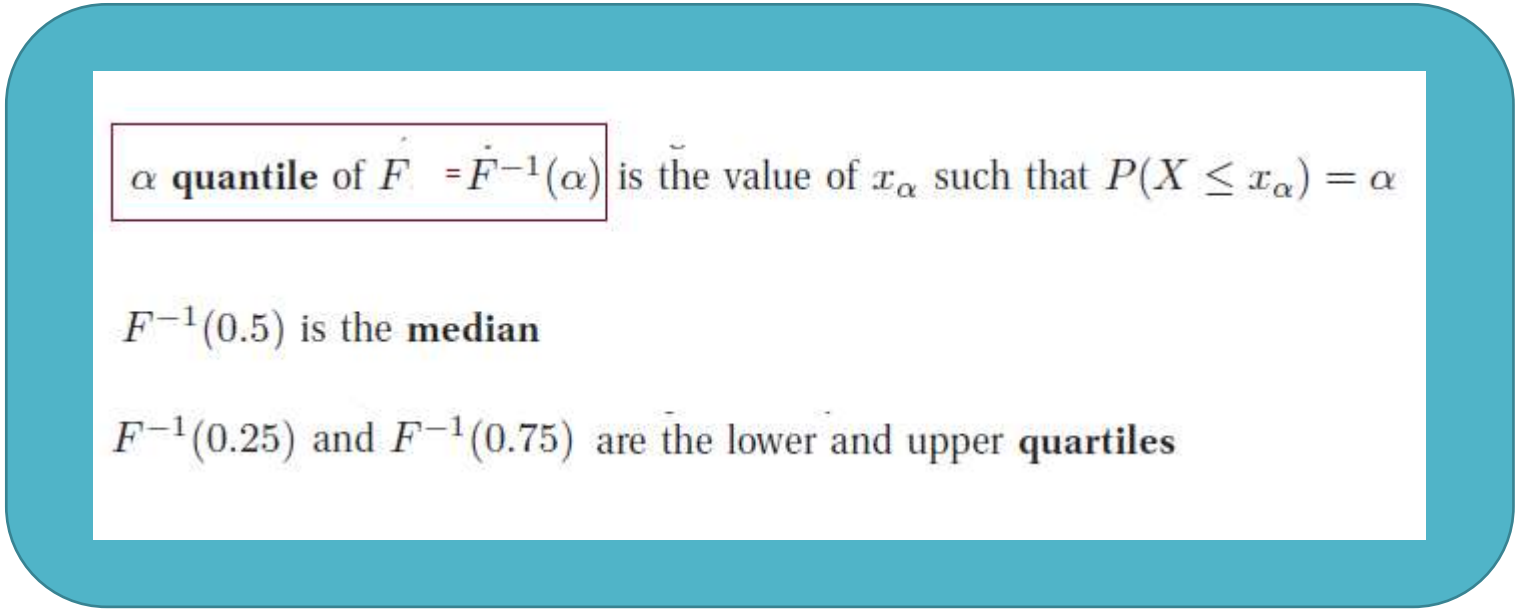
Quintile

$P(X \le x_\alpha) = \alpha$ $\quad 0.5$

α $\quad x_{.5}$ $\quad \alpha = .5$

$x_{.5}$

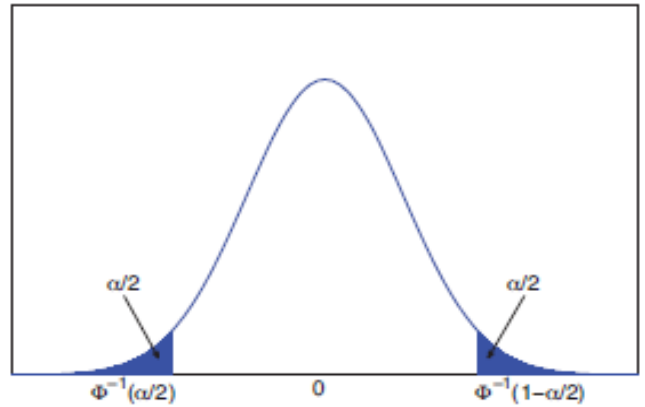
$P(X \le q) = \sum_{x \le q} p(x)$



α **quantile** of F $= F^{-1}(\alpha)$ is the value of x_α such that $P(X \le x_\alpha) = \alpha$

$F^{-1}(0.5)$ is the **median**

$F^{-1}(0.25)$ and $F^{-1}(0.75)$ are the lower and upper **quartiles**



Discrete Distributions

$P(X=1) = \theta$

$P(X=0) = 1-\theta$

$P(X=x) = \theta^{x}(1-\theta)^{1-x}$

The binomial distribution

Suppose we toss a coin n times. Let $X \in \{0, \ldots, n\}$ be the number of heads. If the probability of heads is θ, then we say X has a **binomial** distribution, written as $X \sim \text{Bin}(n, \theta)$. The pmf is given by

$$\text{Bin}(k|n,\theta) \triangleq \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

$E(x) = \sum x\, p(x)$

$$\text{mean} = \theta, \quad \text{var} = n\theta(1-\theta)$$

The Bernoulli distribution $X \in \{0,1\}$

$X \sim \text{Ber}(\theta)$,

$$\text{Ber}(x|\theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)}$$

$p(x)$

$k = 1$

$$Bin(k|n\theta) = \binom{n}{k}\theta^{k}(1-\theta)^{n-k}$$

$E(x) = 1 \times \theta + 0 \times (1-\theta) = \theta$

$E(x) = \sum x\, p(x)$ $\quad x_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ $\quad x_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix}$ $\quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix}$ $\quad 1 \le x_i \le n$ $\quad \sum x_i = n$

The multinomial distribution

let $\mathbf{x} = (x_1, \ldots, x_K)$ be a random vector, where x_j is the number of times side j of the die occurs. Then $\mathbf{x}$ has the following pmf:

$$\mathrm{Mu}(\mathbf{x}|n, \boldsymbol{\theta}) \triangleq \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j} \qquad n = \sum_{k=1}^{K} x_k$$

The Poisson distribution

We say that $X \in \{0, 1, 2, \ldots\}$ has a **Poisson** distribution with parameter $\lambda > 0$, written $X \sim \mathrm{Poi}(\lambda)$, if its pmf is

$$\mathrm{Poi}(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

The first term is just the normalization constant, required to ensure the distribution sums to 1.

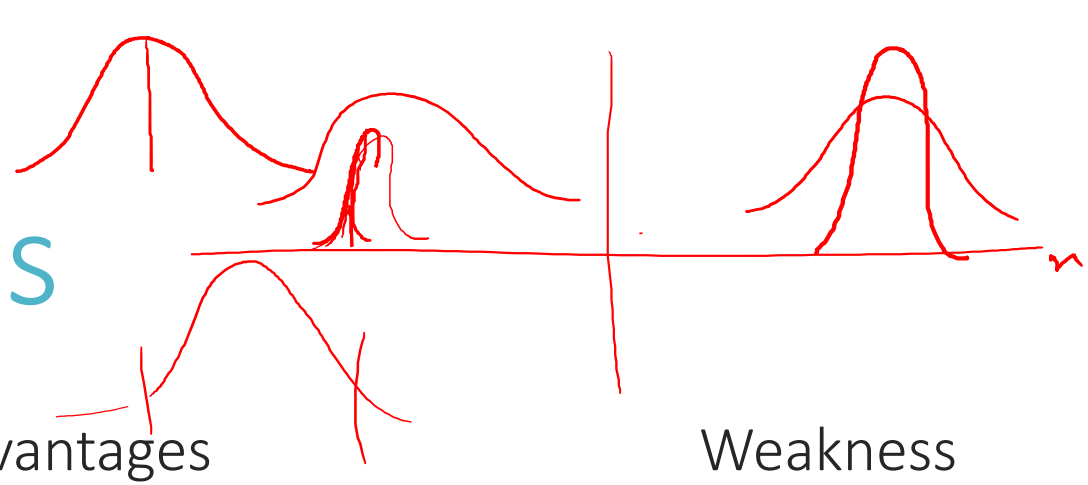
The Poisson distribution is often used as a model for counts of rare events like radioactive decay and traffic accidents.

$x = \begin{pmatrix} 4 \\ 4 \\ 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix}$ $\quad \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \\ \theta_6 \end{pmatrix}$ $\quad x =$

$\frac{n!}{x_1! \cdots x_6!}$

$\frac{1}{6} \quad \frac{1}{6} \quad \cdots \quad \frac{1}{6}$

Continues distributions



Gaussian distribution

$$\mathcal{N}(x|\mu,\sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$\mu = \mathbb{E}[X] \qquad \sigma^2 = \mathrm{var}[X]$$

Precision: $\lambda = 1/\sigma^2$

Advantages

It has two parameters which are easy to interpert

mean and variance.

Central limit Theorem

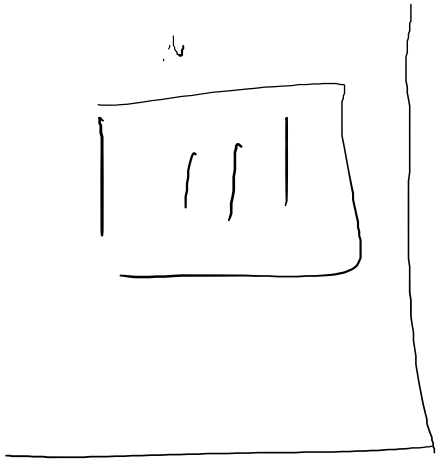
sums of independent random variables have an approximately Gaussian distribution,

Maximum Entropy

Gaussian distribution makes the least number of assumptions

Weakness

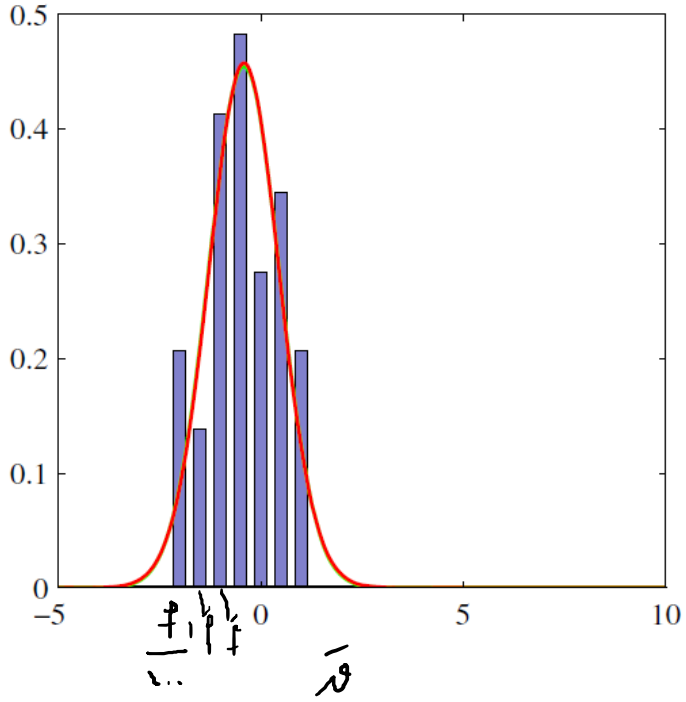
Sensitive to outlier

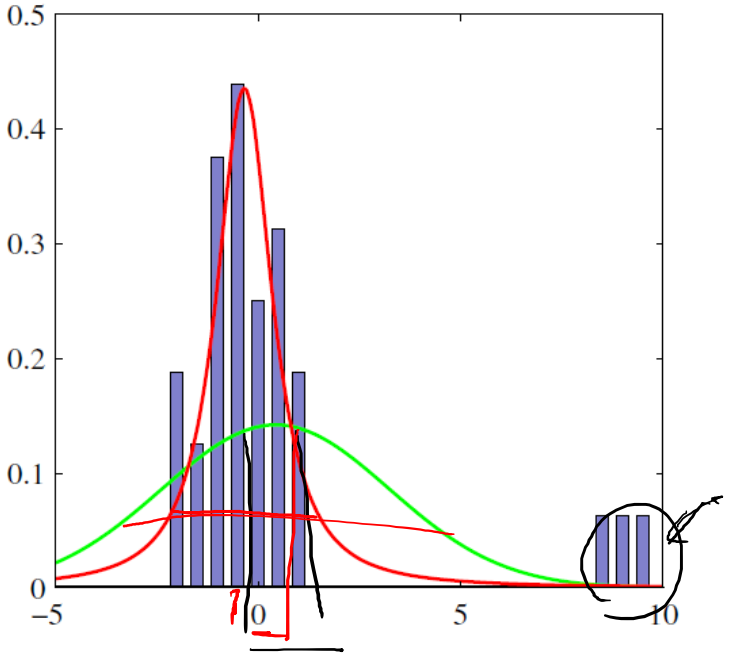


Gaussian and outlier

Red: T-student

Green: Gaussian





T-student

degree of freadom

$$\mathcal{T}(x|\mu,\sigma^2,\nu) \propto \left[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-(\frac{\nu+1}{2})}$$

Scale Parametr

$$\text{mean}=\mu, \text{mode}=\mu, \text{ var}=\frac{\nu\sigma^2}{(\nu-2)}$$

$\nu = 1$, the t-distribution reduces to the *Cauchy* distribution.

$\nu \to \infty$ the t-distribution $\text{St}(x|\mu,\lambda,\nu)$ becomes a Gaussian $\mathcal{N}(x|\mu,\lambda^{-1})$ with mean μ and precision λ

Able to handle out layer

Laplace distribution

$$\text{Lap}(x|\mu, b) \triangleq \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

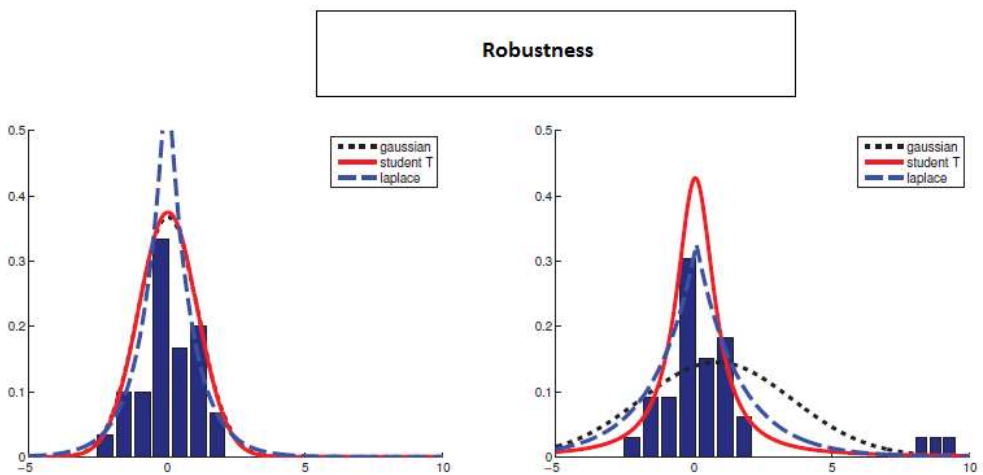
Here μ is a location parameter and $b > 0$ is a scale parameter.

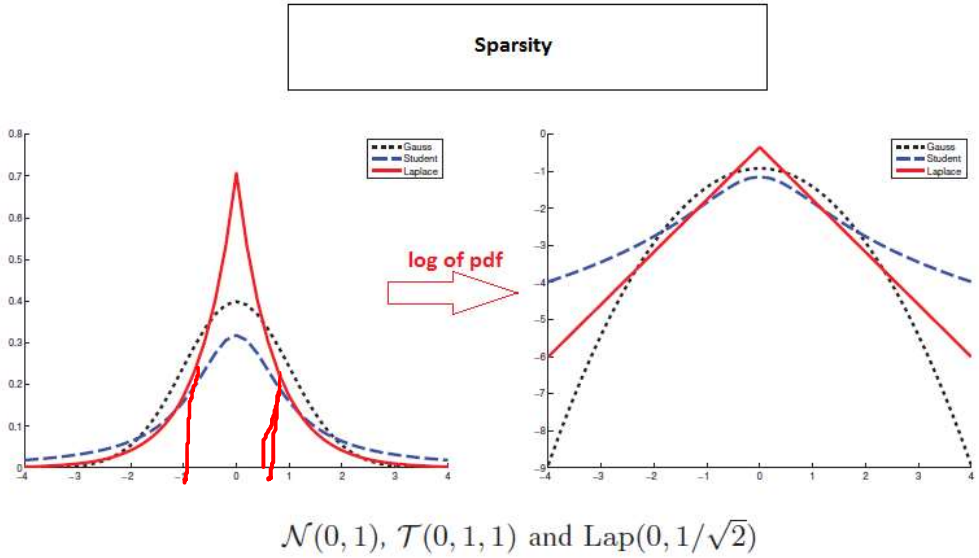
$$\text{mean} = \mu, \ \text{mode} = \mu, \ \text{var} = 2b^2$$

Robust undr outlier

Sparse distribution

Laplace : Robustness and sparsity





$\mathcal{N}(0,1)$, $\mathcal{T}(0,1,1)$ and $\mathrm{Lap}(0,1/\sqrt{2})$

Laplace distribution, which is always log-concave

Multivariate

x

$X = \begin{pmatrix} \mu \\ \sim \end{pmatrix}$ $x \in \mathbb{R}$

$X = \begin{pmatrix} x_1 \\ x_d \end{pmatrix}$ $x \in \mathbb{R}^d$

$P(X = x)$

$P(X = x) =$

$X = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$ $P(X) = P(x_1, \dots, x_N) = \prod_{i=1}^{N} P(x_i)$

$x_i \sim N(x \mid \mu_i, \sigma_i^2)$

$N(x \mid \mu_i, \sigma_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}}$

$(x-\mu)^T \begin{pmatrix} \frac{1}{\sigma_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_d^2} \end{pmatrix} (x-\mu)$

$P(X) = \prod_{i=1}^{d} N(x \mid \mu_i, \sigma_i^2)$?

$= \frac{1}{(2\pi)^{\frac{d}{2}} \prod_{i=1}^{d} \sqrt{\sigma_i^2}} e^{\sum -\frac{(x_i - \mu_i)^2}{2\sigma_i^2}}$

$= \left(\prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\sigma_i^2}} \right) e^{\sum -\frac{(x_i - \mu_i)^2}{2\sigma_i^2}}$

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

(covariance matrix)

$\Sigma = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix}$

$x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$ $\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix}$

Multivariate

A **joint probability distribution** has the form $p(x_1, \ldots, x_D)$ for a set of $D > 1$ variables, and models the (stochastic) relationships between the variables.

$$\text{cov}[X, Y] \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

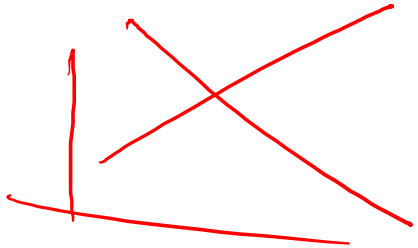
If $\mathbf{x}$ is a d-dimensional random vector, its **covariance matrix** is defined to be the following symmetric, positive definite matrix:

$$\text{cov}[\mathbf{x}] \triangleq \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\right]$$

$$= \begin{pmatrix} \text{var}[X_1] & \text{cov}[X_1, X_2] & \cdots & \text{cov}[X_1, X_d] \\ \text{cov}[X_2, X_1] & \text{var}[X_2] & \cdots & \text{cov}[X_2, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}[X_d, X_1] & \text{cov}[X_d, X_2] & \cdots & \text{var}[X_d] \end{pmatrix}$$

Correlation

$$\text{corr}\,[X,Y] \triangleq \frac{\text{cov}\,[X,Y]}{\sqrt{\text{var}\,[X]\,\text{var}\,[Y]}}$$



$$-1 \leq \text{corr}\,[X,Y] \leq 1$$

A **correlation matrix** has the form

$$\mathbf{R} = \begin{pmatrix} \text{corr}\,[X_1,X_1] & \text{corr}\,[X_1,X_2] & \cdots & \text{corr}\,[X_1,X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \text{corr}\,[X_d,X_1] & \text{corr}\,[X_d,X_2] & \cdots & \text{corr}\,[X_d,X_d] \end{pmatrix}$$

$\text{corr}\,[X,Y] = 1$ if and only if $Y = aX + b$ **a>0**

$\therefore X, Y$

$Cov(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}$

$Cov(X,Y) = E\left(\frac{(X-\mu_x)}{\sqrt{Var(X)}}\,\frac{(Y-\mu_y)}{\sqrt{Var(Y)}}\right) = \sum_{x,y} (x-\mu_x)(y-\mu_y)P(X,Y)$

$-1 \le Cor(X,Y) \le 1$

$Cor(X,Y) = 1 \quad Y = aX + b \quad a > 0$

$Cor(X,Y) = -1 \quad Y = aY + b \quad a < 0$

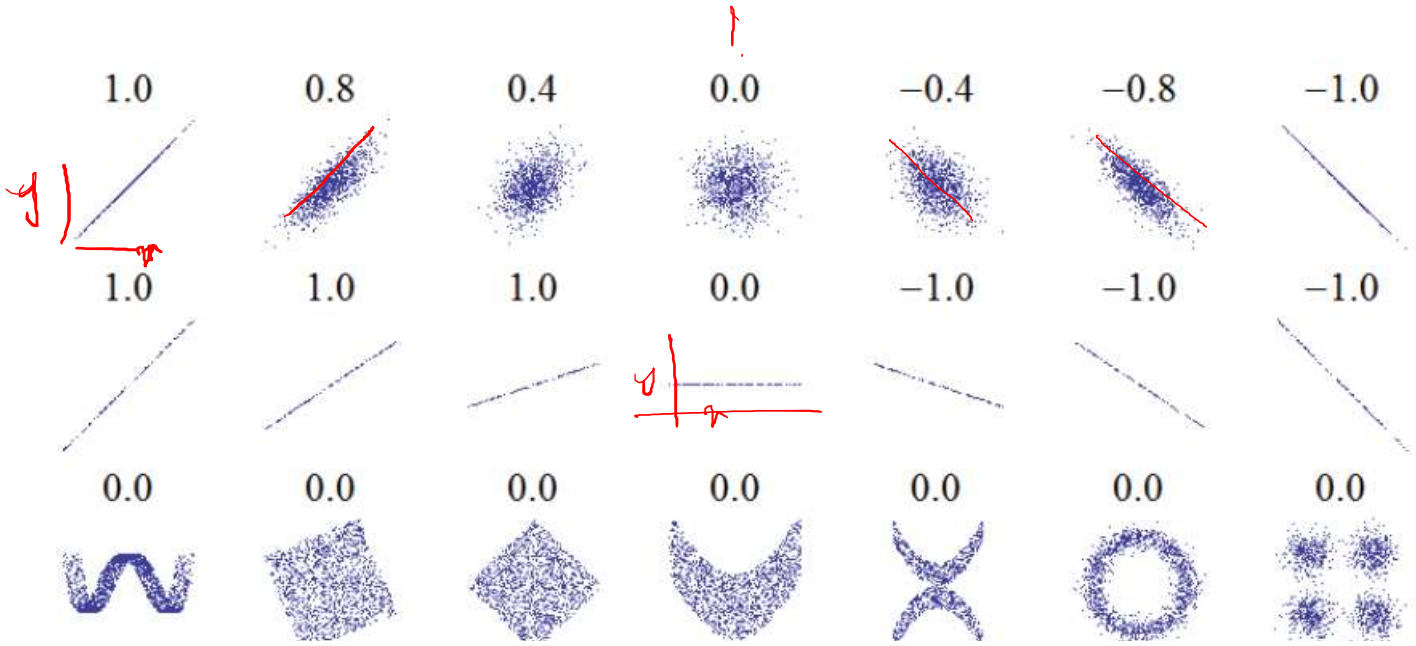


Figure 2.12 Several sets of (x, y) points, with the correlation coefficient of x and y for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero. Source: `http://en.wikipedia.org/wiki/File:Correlation_examples.png`

$$Cov(X,Y) = \sum_{x,y} (x-\mu_x)(y-\mu_y)\frac{p(x,y)}{p(x)p(y)} = \left[\sum_x (x-\mu_x)p(x)\right]\left[\sum_y (y-\mu_y)p(y)\right]$$

Dependence Vs Correlation

$$\sum x p(x) - \mu_x \sum p(x)$$

$$\mu_x - \mu_x = 0$$

If X and Y are independent, meaning $p(X,Y) = p(X)p(Y)$

$\Downarrow$

$\text{cov}[X,Y] = 0$, and hence $\text{corr}[X,Y] = 0$

so they are uncorrelated.

But $\text{corr}[X,Y] = 0$.

Clearly Y is dependent on X

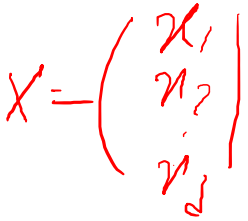
For example, let $X \sim U(-1,1)$ and $Y = X^2$.

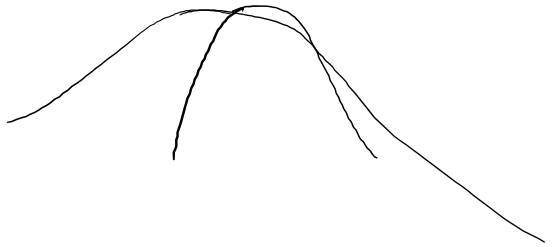
uncorrelated does not imply independent.

Indep $\longrightarrow$ Uncorr

Uncorr $\not\longrightarrow$ Indep

Multivariate Distributions





The multivariate Gaussian

The **multivariate Gaussian** or **multivariate normal** (**MVN**) is the most widely used joint probability density function for continuous variables. We discuss MVNs in detail in Chapter 4; here we just give some definitions and plots.

The pdf of the MVN in D dimensions is defined by the following:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{2.70}$$

where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}] \in \mathbb{R}^D$ is the mean vector, and $\boldsymbol{\Sigma} = \text{cov}[\mathbf{x}]$ is the $D \times D$ covariance matrix. Sometimes we will work in terms of the **precision matrix** or **concentration matrix** instead. This is just the inverse covariance matrix, $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. The normalization constant $(2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{1/2}$ just ensures that the pdf integrates to 1 (see Exercise 4.5).

Figure 2.13 plots some MVN densities in 2d for three different kinds of covariance matrices. A full covariance matrix has $D(D+1)/2$ parameters (we divide by 2 since $\boldsymbol{\Sigma}$ is symmetric). A diagonal covariance matrix has D parameters, and has 0s in the off-diagonal terms. A **spherical** or **isotropic** covariance, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_D$, has one free parameter.

$\Sigma \in R^{d \times d}$ → $\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{pmatrix}$

$y = f(x_1, x_2)$

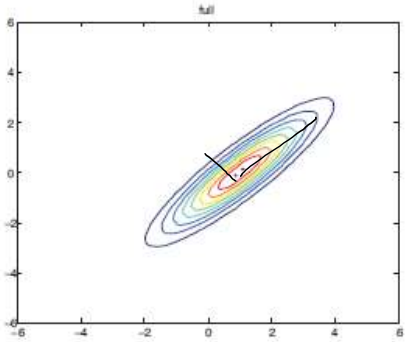
$f(x_1, x_2) = 2$

$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

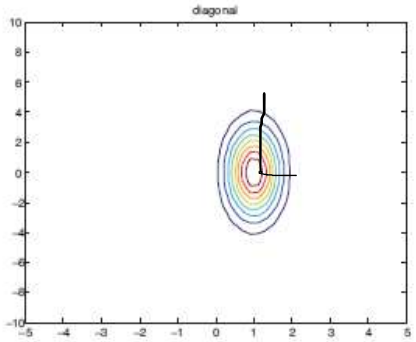
$(x_1 \quad x_2)$

$\Sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$

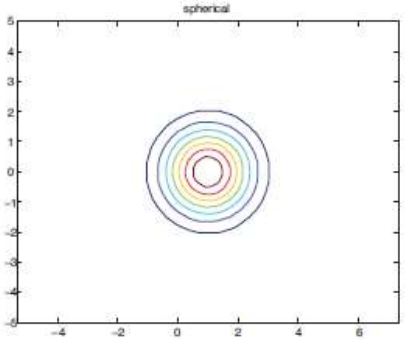
$\Sigma = \sigma I$



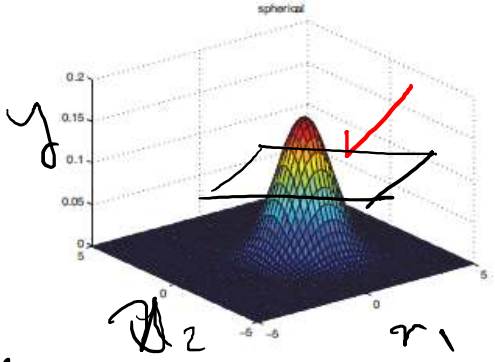
(a)



(b)



(c)



(d)

Figure 2.13 We show the level sets for 2d Gaussians.
(a) A full covariance matrix has elliptical contours.
(b) A diagonal covariance matrix is an **axis aligned** ellipse.
(c) A spherical covariance matrix has a circular shape.
(d) Surface plot for the spherical Gaussian in (c).

$x \in R^2 \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$-\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \begin{pmatrix} \alpha & \beta \\ \gamma & \varepsilon \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} = f'$

e

$-\begin{bmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{bmatrix} \begin{pmatrix} \alpha & \beta \\ \gamma & \varepsilon \end{pmatrix} \begin{pmatrix} x_1 - \mu \\ \cdot \end{pmatrix} = f''$

$a x_1^2 + b x_2^2 + c\, x_1 x_2 = ?$

Multivariate Student t distribution

A more robust alternative to the MVN is the **multivariate Student t** distribution, whose pdf is given by

$$\mathcal{T}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\nu/2 + D/2)}{\Gamma(\nu/2)} \frac{|\boldsymbol{\Sigma}|^{-1/2}}{\nu^{D/2}\pi^{D/2}} \times \left[1 + \frac{1}{\nu}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]^{-(\frac{\nu+D}{2})} \quad (2.71)$$

$$= \frac{\Gamma(\nu/2 + D/2)}{\Gamma(\nu/2)} |\pi\mathbf{V}|^{-1/2} \times \left[1 + (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]^{-(\frac{\nu+D}{2})} \quad (2.72)$$

where $\boldsymbol{\Sigma}$ is called the scale matrix (since it is not exactly the covariance matrix) and $\mathbf{V} = \nu\boldsymbol{\Sigma}$. This has fatter tails than a Gaussian. The smaller ν is, the fatter the tails. As $\nu \to \infty$, the

Mixture Models

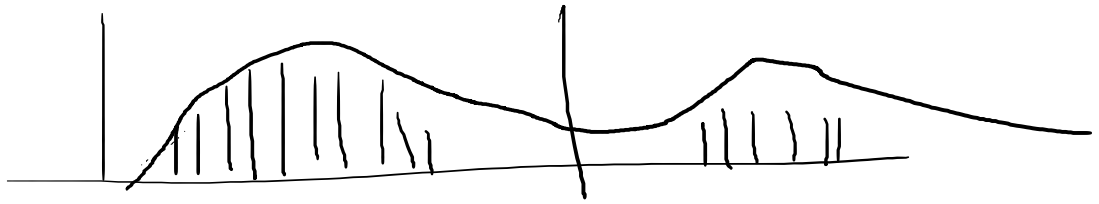
$f(x, \theta_1, \theta_2, \pi_1, \pi_2) = \pi_1 f(x, \theta_1) + \pi_2 f(x, \theta_2)$

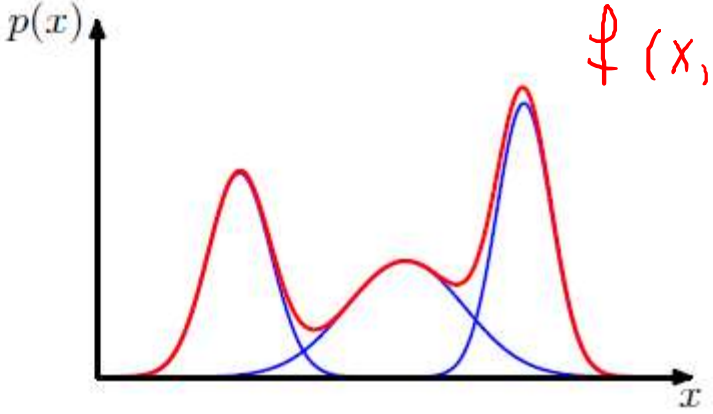
$f(x) \geq 0$

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

Mixtures of Gaussians

Example of a Gaussian mixture distribution in one dimension showing three Gaussians (each scaled by a coefficient) in blue and their sum in red.





$f(x, \theta_1)$

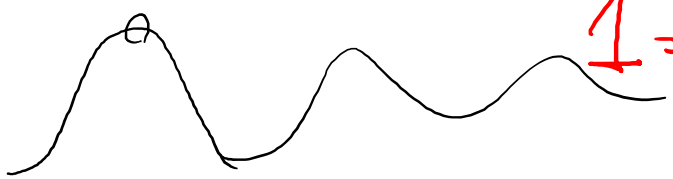
$f(x, \theta_2)$

$\pi_1 + \pi_2 = 1$

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\sum_{k=1}^{K} \pi_k = 1.$$

$0 \leqslant \pi_k \leqslant 1.$



$$1 = \int_{-\infty}^{\infty} f(x, \theta_1, \theta_2, \pi_1, \pi_2)\, dx$$

$$= \pi_1 \int f(x, \theta_1)\, dx + \pi_2 \int f(x, \theta_2)\, dx$$

$$= \pi_1 + \pi_2$$

Transformation of Random Variables

Transformation

Linear

General

$$\mathbf{y} = f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$$

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

$$\operatorname{cov}[\mathbf{y}] = \operatorname{cov}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$$

$$p_y(y) = \sum_{x:f(x)=y} p_x(x)$$

$$p_y(\mathbf{y}) = p_x(\mathbf{x}) \left| \det \left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right) \right|$$

Example

For example, suppose $X \sim U(-1, 1)$, and $Y = X^2$. Then $p_y(y) = \frac{1}{2} y^{-\frac{1}{2}}$

As a simple example, consider transforming a density from Cartesian coordinates $\mathbf{x} = (x_1, x_2)$ to polar coordinates $\mathbf{y} = (r, \theta)$, where $x_1 = r\cos\theta$ and $x_2 = r\sin\theta$. Then

$$\mathbf{J}_{\mathbf{y}\to\mathbf{x}} = \begin{pmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \theta} \\ \frac{\partial x_2}{\partial r} & \frac{\partial x_2}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$

and

$$|\det \mathbf{J}| = |r\cos^2\theta + r\sin^2\theta| = |r|$$

Hence

$$\begin{aligned} p_{\mathbf{y}}(\mathbf{y}) &= p_{\mathbf{x}}(\mathbf{x})|\det \mathbf{J}| \\ p_{r,\theta}(r,\theta) &= p_{x_1,x_2}(x_1,x_2)r = p_{x_1,x_2}(r\cos\theta, r\sin\theta)r \end{aligned}$$

Central limit theorem

Now consider N random variables with pdf's (not necessarily Gaussian) $p(x_i)$, each with mean μ and variance σ^2. We assume each variable is **independent and identically distributed** or **iid** for short. Let $S_N = \sum_{i=1}^{N} X_i$ be the sum of the rv's. This is a simple but widely used transformation of rv's. One can show that, as N increases, the distribution of this sum approaches

$$p(S_N = s) = \frac{1}{\sqrt{2\pi N\sigma^2}} \exp\left(-\frac{(s - N\mu)^2}{2N\sigma^2}\right) \tag{2.96}$$

Hence the distribution of the quantity

$$Z_N \triangleq \frac{S_N - N\mu}{\sigma\sqrt{N}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{N}} \tag{2.97}$$

converges to the standard normal, where $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the sample mean. This is called the **central limit theorem**. See e.g., (Jaynes 2003, p222) or (Rice 1995, p169) for a proof.

In Figure 2.17 we give an example in which we compute the mean of rv's drawn from a beta distribution. We see that the sampling distribution of the mean value rapidly converges to a Gaussian distribution.

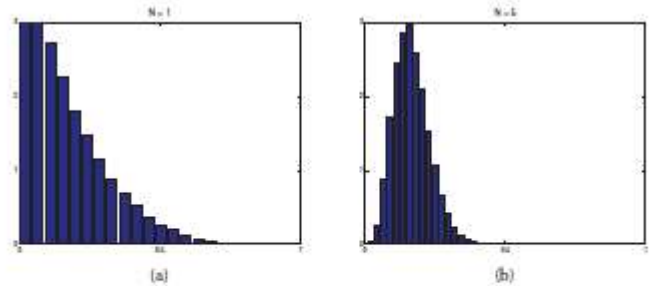
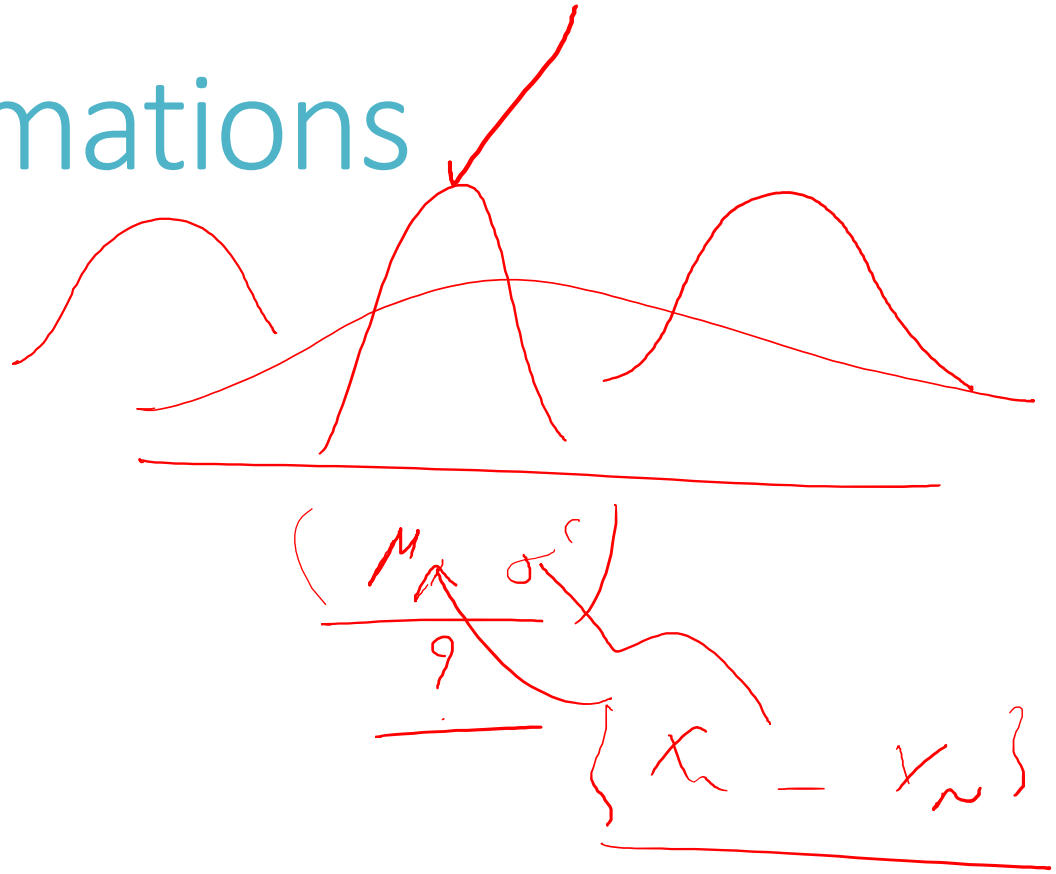


Figure 2.17 The central limit theorem in pictures. We plot a histogram of $\frac{1}{N}\sum_{i=1}^{N} x_{ij}$, where $x_{ij} \sim \text{Beta}(1, 5)$, for $j = 1 : 10000$. As $N \to \infty$, the distribution tends towards a Gaussian. (a) $N = 1$. (b) $N = 5$. Based on Figure 2.6 of (Bishop 2006a). Figure generated by `centralLimitDemo`.

Parameter Estimation

Parameter estimations

- Maximum likelihood
- Maximum a posteriori
- Expected Maximization



ML,MAP,EM

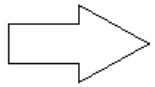
ML: P(X1=x1,...,Xn=xn)=P(X1=x1)...P(Xn=xn)

xi: samples
X=(X1,...,Xn): iid

$$\theta_{MLE} = \arg\max_{\theta} P(X|\theta) = \arg\max_{\theta} \prod_i P(x_i|\theta)$$

MAP: Prior & posterior

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \propto P(X|\theta)P(\theta)$$



$$\theta_{MAP} = \arg\max_{\theta} P(X|\theta)P(\theta) = \arg\max_{\theta} \log P(X|\theta)P(\theta)$$

EM: usually Mixture models

$p_{data}(x)$
Unknow

$\mathbb{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ Samples drawn independently from the true but Unknown data generating distributions $p_{\text{data}}(\mathbf{x})$

$p_{\text{model}}(\mathbf{x}; \boldsymbol{\theta})$: Parametric Model

$$p_{\text{model}}(\mathbb{X}; \boldsymbol{\theta}) = \prod_{i=1}^{m} p_{\text{model}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

Maximum Likelihood

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} p_{\text{model}}(\mathbb{X}; \boldsymbol{\theta})$$
$$= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{\text{model}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

ML

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log p_{\text{model}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

Maximum likelihood

$$\mathcal{X} = \{x^t\}_{t=1}^N$$

We assume that $x^t \sim p(x|\theta)$ and x^t are independent.

We want to find θ that makes sampling x^t from $p(x|\theta)$ as likely as possible

$$\max_\theta \; l(\theta|\mathcal{X}) \equiv p(\mathcal{X}|\theta) = \prod_{t=1}^N p(x^t|\theta)$$

$$\max_\theta \; \mathcal{L}(\theta|\mathcal{X}) \equiv \log l(\theta|\mathcal{X}) = \sum_{t=1}^N \log p(x^t|\theta)$$

ML for Bernoulli

$$P(x) = p^x(1-p)^{1-x}, x \in \{0,1\}$$

iid sample $\mathcal{X} = \{x^t\}_{t=1}^N$, where $x^t \in \{0,1\}$

$$\begin{aligned}\mathcal{L}(p|\mathcal{X}) &= \log \prod_{t=1}^{N} p^{(x^t)}(1-p)^{(1-x^t)} \\ &= \sum_t x^t \log p + \left(N - \sum_t x^t\right) \log(1-p)\end{aligned}$$

$$\hat{p} = \frac{\sum_t x^t}{N}$$

Gaussian (Normal) Density

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], -\infty < x < \infty$$

$$\mathcal{X} = \{x^t\}_{t=1}^N \text{ with } x^t \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mathcal{L}(\mu, \sigma|\mathcal{X}) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_t (x^t-\mu)^2}{2\sigma^2}$$

$$\begin{aligned} m &= \frac{\sum_t x^t}{N} \\ s^2 &= \frac{\sum_t (x^t-m)^2}{N} \end{aligned}$$

Information Theory

Shanon Entropy

Information $\propto \frac{1}{Possibility}$

$$\text{Information} \propto \frac{1}{P(x)}$$

Adaditivity: For independent events $Z_1$and Z_2

$$\mathbf{Inf}(Z_1 + Z_2) \propto \mathbf{Inf}(Z_1) + \mathbf{Inf}(Z_2)$$

$$I(X) = -\log P(x)$$

Covers two mwntioed properties

Shannon entropy $H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)]$.

$inf(x) \propto \frac{1}{p(x)}$, $inf(x,y) = inf(x) + inf(y)$ ← $inf(x) = \log\left(\frac{1}{p(x)}\right) = -\log p(x)$

x, y indep

$x \to p(x) \to -\log p(x)$

The **entropy** of a random variable X with distribution p, denoted by $\mathbb{H}(X)$ or sometimes $\mathbb{H}(p)$, is a measure of its uncertainty. In particular, for a discrete variable with K states, it is defined by

$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k)$$

$0 \log 0 = 0$

In order to co ssage of length 3 bits.

$E[f(x)] = \sum f(x)p(x)$

$X = \{0, 1\}$

$$\mathrm{H}[x] = -8 \times \frac{1}{8}\log_2\frac{1}{8} = 3 \text{ bits}.$$

$H(x) = E[-\log p(x)] = -\sum p(x)\log p(x)$ Entropy

Event $\{a, b, c, d, e, f, g, h\}$

Probability $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}$

$$\mathrm{H}[x] = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64}$$

Coding

0, 10, 110, 1110, 111100, 111101, 111110, 111111

$$\text{average code length} = \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + 4\times\frac{1}{64}$$

$X = \{0, 1\}$

$p(x) = \frac{1}{2} \quad \frac{1}{2}$

$-\left(\log 0 \times \frac{1}{2} - \log\left(1 \times \frac{1}{2}\right)\right) = -\frac{1}{2} \times \log 0 = \infty$

$\log_2 \infty$

$x = \log_2$

$y = \ln x$

$x = e^y$

$X = \{0, 1\}$

$p(x) = 1 \quad 0$

$-0 \times \log 1 + 1 \log 0 = -\log 0 = \infty$

jel

$-\sum \log p(x) \cdot p(x) = -\left(1 \times \log 1 + 0 \times \log 0\right) = 0$

jel

$-\sum p(x) \log p(x)$

$= -\left[\frac{1}{2} \times \log \frac{1}{2} + \frac{1}{2} \times \log\left(\frac{1}{2}\right)\right]$

$= -\log \frac{1}{2} = \log 2$

0

↓

1

$\log_1 \times 1$

Kullback-Leibler (KL)divergence

Coding by inexact dist

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{x}) \ln q(\mathbf{x})\, d\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\, d\mathbf{x}\right)$$

$$= -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} d\mathbf{x}.$$

Coding by exact dist

p(x): True unknown distribution

q(x): An approximation of p(x)

Extra information for decoding

$$\mathrm{KL}(p\|q) \not\equiv \mathrm{KL}(q\|p).$$

$$\mathrm{KL}(p\|q) \geqslant 0$$

P

x_1	x_2	x_3	x_4
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

Q

x_1	x_2	x_3	x_4
$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$

x_1	x_2	x_3	x_4
$\frac{1}{4}$	$\frac{1}{2}$	0	$\frac{1}{4}$

$$E(f_{(x)}) = \sum f_{(x)} P(x) \cdot \log p(x)$$

Kullback-Leibler (KL) divergence: The similarity of two distributions

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{x\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{x\sim P}[\log P(x) - \log Q(x)]$$

$-H(P)$

$$D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$$

cross-entropy

$$H(P,Q) = H(P) + D_{\mathrm{KL}}(P\|Q) = -\mathbb{E}_{x\sim P}\log Q(x)$$

$f_{(x)}$

$E_{x\sim P}(-\log p(x))$

$$argmin_Q H(P,Q) = argmin D_{KL}(P||Q)$$

$$-\int p(x)\log p(x) \quad \left(-\int p(x)\log q(x)\right)$$

ML and Cross Entropy Equivalence

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log p_{\mathrm{model}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}).$$

$$\Rightarrow \quad \boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x} \sim \hat{p}_{\mathrm{data}}} \log p_{\mathrm{model}}(\boldsymbol{x}; \boldsymbol{\theta}) = \mathit{argmin}\ -\mathbb{E}_{\boldsymbol{x} \sim \hat{p}_{\mathrm{data}}} [\log p_{\mathrm{model}}(\boldsymbol{x})]$$

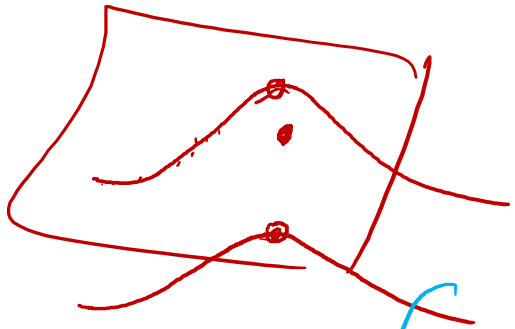
$$= \mathit{argmin}\ H(P, Q)$$

$$D_{\mathrm{KL}}(\hat{p}_{\mathrm{data}} \| p_{\mathrm{model}}) = \mathbb{E}_{\boldsymbol{x} \sim \hat{p}} [\log \hat{p}_{\mathrm{data}}(\boldsymbol{x}) - \log p_{\mathrm{model}}(\boldsymbol{x})]$$

One way to interpret maximum likelihood estimation is to view it as minimizing the dissimilarity between the empirical distribution $\hat{p}_{\mathrm{data}}$ defined by the training set and the model distribution, with the degree of dissimilarity between the two measured by the KL divergence.

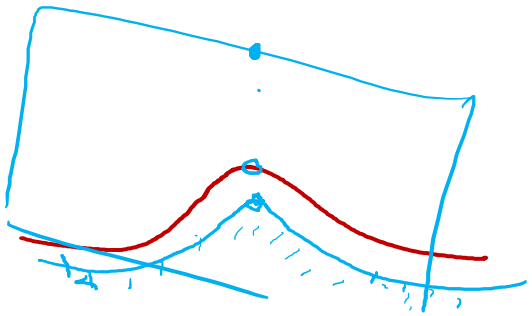
P, Q

P, Q



P, Q

$$-\int p(x) \log p(x)$$



$$\int p(x) \log q(x)$$

Mutual Information

$$\begin{aligned} \mathrm{I}[\mathbf{x},\mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x},\mathbf{y})\|p(\mathbf{x})p(\mathbf{y})) \\ &= -\iint p(\mathbf{x},\mathbf{y}) \ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}\right) \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \end{aligned}$$

$$\mathrm{I}[\mathbf{x},\mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]$$

Conditional log-Likelihood

If $\boldsymbol{X}$ represents all our inputs and $\boldsymbol{Y}$ all our observed targets, then the conditional maximum likelihood estimator is

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{Y} \mid \boldsymbol{X}; \boldsymbol{\theta}).$$

If the examples are assumed to be i.i.d., then this can be decomposed into

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log P(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta}).$$