Short course

A vademecum of statistical pattern recognition and machine learning

Lecture 1
Review of probability and statistics

Massimo Piccardi
University of Technology, Sydney, Australia

Agenda

- Discrete and continuous random variables
- Joint, conditional, marginal probabilities
- Bayes' theorem
- Independence
- Mean, variance, moments
- Expectations
- Covariance matrix, correlation coefficients
- Sample mean, sample covariance
- Gaussian distribution
- Main properties of Gaussian distributions
- Mixture distributions and Gaussian mixture models

Random variables

- A *random variable* can be defined as an instrument to map all the possible outcomes of an event
- Many alternative definitions of random variable are possible, at the same time more comprehensive and requiring more mathematical fluency; the above is enough for us
- A random variable has an associated *probability distribution*
- There are two main types of random variables: *discrete* (countable outcomes) and *continuous* (infinite, continuous outcomes)
- Discrete variables are also called *categorical* or sometimes, with a bit of a stretch, *multinomial*

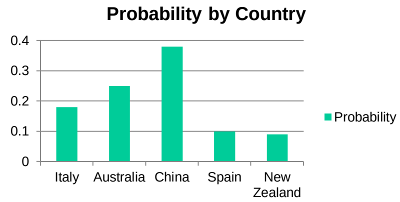
Slight rewording from the Wikipedia entry for random variable, 20 Sept 2010 3pm AEST

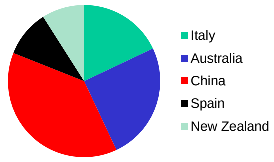
Discrete random variables

- A *discrete* random variable takes values in a finite set of symbols. Examples: variable **Country** can take values in {Italy, Australia, China, ...}; variable **Toss of a Coin** can take values in {Heads, Tails}
- The function assigning a probability value to each of these values is known as the *probability mass function*
- Notation **P(Country)** means the probability of any possible value of Country
- Notation **P(Country = Italy)** means the probability of a specific value; sometimes the shorthand notation **P(Italy)** is used if no ambiguity arises (and sometimes even then...)
- The probability of any value is always ≥ 0!
- The sum of the probabilities of all values is always 1 for the Axiom of Total Probability!

Probability mass function

- Draw it the way you like!






```
p(Country = Italy)        0.18
p(Country = Australia)    0.25
p(Country = China)        0.38
p(Country = Spain)        0.10
p(Country = New Zealand)  0.09
```


Continuous random variables

- A *continuous* random variable takes values in a continuous interval. Examples: variable Height can take values in (0 cm, 280 cm); variable Weight can take values in (0 Kg, much more than you think)
- The function assigning a probability value to each of these values is known as the *probability density function* (pdf)*;* actually, it assigns a *density of probability* to each value
- Notation **p(x)** means the probability density of any value of x
- Beware: many authors – including the yours truly - use the same notation for p and P, assuming you would guess from the context
- The unit of measurement of p(x) is $[x]^{-1}$; to go back to a probability, one must multiply by any interval over x, finite or infinitesimal (dx)
- It is therefore clear that the pdf is defined only up to the chosen unit of measurement and can be made big or small at will
- What do we lose if we measure the length of a finger in terameters? Numerical resolution

Probability density function (pdf)

- The pdf of a continuous random variable, x, defines the *density of probability* for each value of x
- To return to a probability, one must integrate over an interval

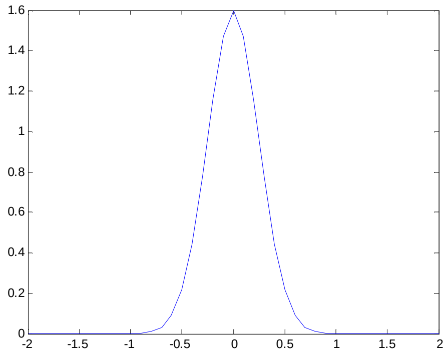
- Some properties:

  - $p(x) \geq 0$

  - $p(x)$ can be > 1!

  - $\int_a^b p(x)\, dx \leq 1$

(1 if over the entire domain of x)



Matlab®, Statistics Toolbox™, command `plot(-2:0.1:2, pdf('norm',-2:0.1:2,0,0.25))`

Joint probability

- Let us now consider two discrete random variables: Weather (W) and Temperature (T)
- Both assumed binary, i.e. only two possible values each:
  - W: rainy (r), sunny (s)
  - T: low (l), high (h)
- Take 100 samples of (W,T) and map the joint frequencies in this table
- Assuming we have enough samples, we call them *joint probabilities*
- We'll use this as a running example for the next few slides

W \ T	l	h
r	0.25	0.10
s	0.05	0.60

Joint probability

- Joint probability of W and T, value by value:

  $p(W = r, T = l) = 25/100 = 0.25$

  $p(W = r, T = h) = 10/100 = 0.10$

  $p(W = s, T = l) = 5/100 = 0.05$

  $p(W = s, T = h) = 60/100 = 0.60$

- The notation with the variables, $p(W,T)$, means any (or all) of these joint probability values
- $p(W,T) = p(T,W)$: the order does not count
- Each of the above values can be noted as $p(r,l)$ for short, instead of $p(W = r, T = l)$, provided there is no ambiguity

Joint probability

- The joint probability values add up to 1, as they cover all possible cases (Axiom of total probability)
- Thus, in the example, only 3 of them can be arbitrarily chosen, as the fourth results from: 1 – the sum of the other 3. There are 3 independent numbers (*degrees of freedom*, dof, or *parameters* of the discrete distribution)
- For two variables with N values each, the joint probability has $N^2 - 1$ dof

Conditional probability

- The concept of *conditional probability* is simple: given two r.v., a conditional probability fixes one of the two and uses the other as **the only random variable**
- The conditional probability reflects the frequencies of the random variable not over all the samples, but on the specific sub-set where the given condition is true
- Example: p(W = r | T = l)
  - reads as: “the probability of Weather being rainy *given that* the Temperature is low”
  - instead of considering all the 100 samples, one just takes those where the temperature is low (30 samples in total)
  - out of the above, compute the frequency of rainy days: 25 out of 30 = 0.83

Conditional probability

- A conditional probability like p(W|T) is still **a function of both variables**, but **a probability in only one**!!!
- No frequency information whatsoever is provided for T!
- Let us fix T = l in the example; then, the only variable is W, with two possible values:

  p(W = r | T = l) = 25/30 = 0.83

  p(W = s | T = l) = 5/30 = 0.17

  they are all the possible cases and as such their sum is 1; we have only one dof
- There are many conditional probabilities! One for each value of the variables in the condition
- For variables with N values, each conditional probability has N – 1 dof

Conditional probability

- In the example:

  $$\left.\begin{array}{l} p(W = r \mid T = l) = 25/30 = .83 \\ p(W = s \mid T = l) = 5/30 = .17 \end{array}\right\} \textit{1 dof}$$

  $$\left.\begin{array}{l} p(W = r \mid T = h) = 10/70 = .14 \\ p(W = s \mid T = h) = 60/70 = .86 \end{array}\right\} \textit{1 dof}$$

  → there are 2 degrees of freedom overall for p(W|T), and N(N-1) for two N-valued r.v.

- NB: p(r, l) < p(r | l) by definition
  (the latter has a smaller denominator!)

Marginal probability

- W and T are jointly called a *random vector*, or, equivalently, a *multivariate random variable*
- One can obtain the marginal probability of either variable by adding up the joint probabilities for all possible values of the other (**marginalisation**):

$$p(W) = \sum_T p(W, T)$$

- The above is informally called the **sum rule**
- For continuous random variables, the sum is replaced by an integral

Marginal probability

With our running example:

p(W = r) = p(W = r, T = l) + p(W = r, T = h) =
25/100 + 10/100 = 35/100

p(W = s) = 65/100 (1 dof)

p(T = l) = 30/100
p(T = h) = 70/100 (1 dof)

Bayes' theorem

$$p(W,T) = p(W\,|\,T)\,p(T)$$

joint probability

conditional probability of W given T

marginal probability of T

- Always holds!
- It is informally called the **product rule**
- It is a powerful tool to break down the complexity of the joint probabilities into the product of simpler probabilities
- Sum rule + product rule: foundations of statistical PR
- Bayes' theorem <u>also applies to pdfs</u>

Bayes' theorem: examples

- Sometimes, you will see it written like this:

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B)}$$

- It applies to any two *sets* of random variables:

$$p(A, B, C, D, E) = p(A, D \mid B, C, E)\, p(B, C, E)$$

- It applies to joint *conditional* probabilities:

$$p(A, B \mid C) = p(A \mid B, C)\, p(B \mid C)$$

Independence

$$p(W, T) = p(W)\, p(T)$$

joint probability

marginal probability of W

marginal probability of T

- If the above holds, the two r.v. are called **independent**
- Often (not always!) a desirable case
- Equivalent to p(W|T) = p(W) and p(T|W) = p(T)
- Does not hold for our running example! For instance:
  – p(r,l) = 0.25
  – p(r) = 0.35; p(l) = 0.30 → p(r) p(l) = 0.105

An example

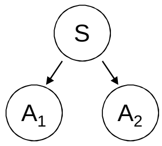
- Given three binary r.v., A_1, A_2 and S, let us assume that

  $p(A_1, A_2 \mid S) = p(A_1 \mid S)\, p(A_2 \mid S)$

  instead of the always true (from Bayes rule):

  $p(A_1, A_2 \mid S) = p(A_1 \mid A_2, S)\, p(A_2 \mid S)$, or

  $p(A_1, A_2 \mid S) = p(A_2 \mid A_1, S)\, p(A_1 \mid S)$
- The above reads as "A_1 and A_2 are independent given S"
- Not equivalent to "A_1 and A_2 are independent"!
- It is a relevant case, with S often called a *state* or *class* and the A_i being *measurements*



assume 90 samples:

#(A1,A2,S=0):

A_1 \ A_2	0	1
0	20	10
1	10	5

#(A1,A2,S=1):

A_1 \ A_2	0	1
0	1	14
1	2	28

#(A1,A2):

A_1 \ A_2	0	1
0	21	24
1	12	33

A note on the argmax of a probability

- Nota Bene: $$A^* = \arg\max_{A,B} p(A,B)$$ $$\neq \arg\max_{A} p(A)$$ $$\neq \arg\max_{A,B} p(A \mid B)$$

A \ B	b_1	b_2	b_3
a_1	30	40	30
a_2	0	60	10
a_3	50	20	20

$$a_2 = \arg\left(\max_{A,B} p(A,B) = 60/260\right)$$
$$a_1 = \arg\left(\max_{A} p(A) = 100/260\right)$$
$$a_3 = \arg\left(\max_{A,B} p(A \mid B) = 50/80\right)$$

- Yet, for any B, $A^* = \arg\max_{A} p(A,B) = \arg\max_{A} p(A \mid B)$

Mean, variance and moments

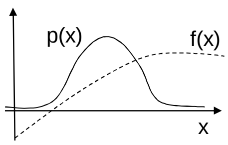
- The pdf of a continuous r.v. describes the probability distribution fully; yet, sometimes we prefer to describe it in a more synthetic way
- Mean, or expected value: $\mu \equiv E[x] = \int_x x\, p(x) dx$
- Variance: $VAR(x) \equiv \sigma^2 \equiv E\left[(x-\mu)^2\right] = \int_x (x-\mu)^2 p(x) dx$
  - The standard deviation, σ, is its square root
  - VAR(x) is also = $E[x^2] - 2\mu E[x] + \mu^2 = E[x^2] - \mu^2$
- N^{th} moment:

$$E\left[x^N\right] = \int_x x^N p(x) dx$$

Expectations

- An expectation is an *averaging operation weighted by p(x)*; it can be extended to any function of x, f(x):

$$E\left[f(x)\right] = \int_x f(x)\, p(x) dx$$



  - E[f(x)] is a scalar value
  - Jensen's inequality: if f(x) convex, $E[f(x)] \geq f(E[x])$; $\leq$ if concave
  - Here x can be discrete!
- The expectation of a function of multiple variables, f(x,y), over x:

$$E\left[f(x,y)\right]_x = \int_x f(x,y)\, p(x) dx$$

"averages out" x and returns a function of the sole y

Expectations

- A marginalisation can be seen as a particular expectation:

$$p(y) = \int_x p(y,x)dx = \int_x p(y \mid x)\, p(x)dx = E[p(y \mid x)]_x$$

- An expectation can also be computed over a conditional probability:

$$E[f(x) \mid y] = \int_x f(x)\, p(x \mid y)dx$$

Sample mean and sample covariance

- At times, either p(x) is not available or the expectation integrals are not easy to compute
- Assuming a set of samples, x_i, i=1…N, is available, it is possible to approximate the mean and the variance as:

$$\mu \equiv E[x] \approx \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \textit{sample mean}$$

$$\sigma^2 \equiv E\left[(x-\mu)^2\right] \approx \frac{1}{N}\sum_{i=1}^{N} (x_i-\mu)^2 \qquad \textit{sample variance}$$

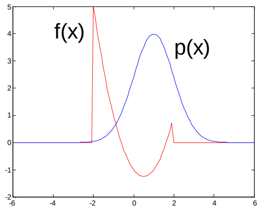
- Other expectations can be approximated in the same way (Monte Carlo methods)

Example

- Let us compute the expected value of function:

$$f(x) = \begin{cases} x^2 - x - 1 & -2 \le x \le 2 \\ 0 & otherwise \end{cases}$$

under Gaussian distribution: $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-1)^2}$



NB: p(x) is magnified 10 times to make it visible on f(x)

 25

Example

- Analytically, the integral is equal to:

$$\int_{-\infty}^{+\infty} f(x)p(x)dx = \int_{-\infty}^{-2} f(x)p(x)dx + \int_{-2}^{+2} f(x)p(x)dx + \int_{+2}^{+\infty} f(x)p(x)dx =$$

$$= 0 + \int_{-2}^{+2} f(x)p(x)dx + 0$$

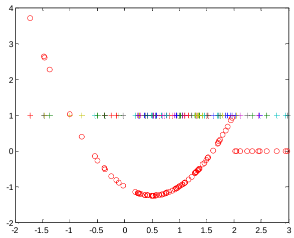
$$\int_{-2}^{+2} f(x)p(x)dx = \int_{-2}^{+2} (x^2 - x - 1) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-1)^2} dx = -\frac{x}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-1)^2} \Big]_{-2}^{+2} =$$

$$= -0.4839 - 0.0089 = -0.4928$$

 26

Example

- Let us approximate this expectation by drawing 100 samples, $\{x_i\}$, i=1…100, from p(x) and computing f(x) at those locations:



- This empirical expectation is equal to (changes at every draw):

$$\frac{1}{100}\sum_{i=1}^{100} f(x_i) = -0.4408$$

Multivariate random variables: mean, covariance and moments

- The same definitions extend to **multivariate r.v., X = [x_1,.. x_D]^T**:
- The **mean** becomes a D x 1 vector:

$$\mu = E[X] = [\mu_1,\ldots,\mu_D]^T = [E[x_1],\ldots,E[x_D]]^T$$

- The variance becomes a D x D **covariance matrix**:

$$COV(X) = \Sigma = E\left[(X-\mu)(X-\mu)^T\right] =$$

$$\begin{bmatrix} E[(x_1-\mu_1)(x_1-\mu_1)] & \ldots & E[(x_1-\mu_1)(x_D-\mu_D)] \\ \ldots & \ldots & \ldots \\ E[(x_D-\mu_D)(x_1-\mu_1)] & \ldots & E[(x_D-\mu_D)(x_D-\mu_D)] \end{bmatrix}$$

Multivariate mean

- Although the multivariate mean is intuitive, it may prove useful to derive it to recap on expectations:

NB: it is a D x 1 quantity

$$E[X] = \int_X X\, p(X) dX = \int_{x_1 \ldots x_D} \begin{bmatrix} x_1 \\ \ldots \\ x_D \end{bmatrix} p(x_1,\ldots,x_D) dx_1 \ldots dx_D =$$

$$= \begin{bmatrix} \int_{x_1 \ldots x_D} x_1\, p(x_1,\ldots,x_D) dx_1 \ldots dx_D \\ \ldots \\ \int_{x_1 \ldots x_D} x_D\, p(x_1,\ldots,x_D) dx_1 \ldots dx_D \end{bmatrix}$$

marginalise $x_2 \ldots x_D$

$$= \int_{x_1} x_1 \left(\underbrace{\int_{x_2 \ldots x_D} p(x_1,\ldots,x_D) dx_2 \ldots dx_D}_{\text{marginalise } x_2 \ldots x_D} \right) dx_1 = \int_{x_1} x_1\, p(x_1) dx_1 = E[x_1]$$

Covariance matrix

- The covariance matrix is a **symmetric matrix** by construction: only D(D + 1)/2 dof

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \ldots & \operatorname{cov}(x_1, x_D) \\ \ldots & \ldots & \ldots \\ \operatorname{cov}(x_D, x_1) = \operatorname{cov}(x_1, x_D) & \ldots & \sigma_D^2 \end{bmatrix}$$

- Terms $\operatorname{cov}(x_i, x_j)$ measure how much x_i and x_j *co-vary*
- A covariance matrix is also (at least) **positive semi-definite**:

  $X^T \Sigma X \geq 0$ for any X

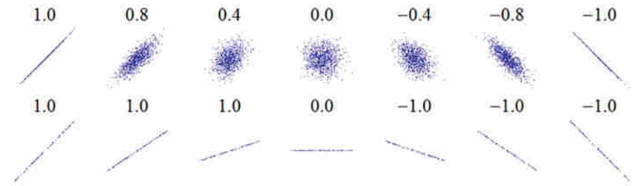
- If it is also non-singular (i.e., full rank, invertible) $X^T \Sigma X$ is strictly > 0 for any $X \neq 0$ and Σ is **positive definite**

Correlation coefficients

- Terms cov(x_i, x_j) are often expressed as *correlation coefficients*, ρ_{ij}:

$$\rho_{ij} = \frac{\mathrm{cov}(x_i, x_j)}{\sigma_i \sigma_j}$$

- NB: $-1 \le \rho_{ij} \le +1$ (corollary of the Cauchy-Schwarz inequality) *Speakers's Notes*



courtesy of Wikipedia

(Un)correlation vs independence

- Two r.v., x_i, x_j, are *uncorrelated* iff $\rho_{ij} = 0$
- Two uncorrelated variables are not independent; they are only in terms of *linear* mutual dependencies
- For two uncorrelated variables, x_i, x_j, it can be easily shown that:

  $E[x_i\, x_j]=E[x_i]\, E[x_j]$

  Proof: $\mathrm{cov}(x_i, x_j) = E[(x_i - \mu_i)(x_j - \mu_j)] = E[x_i\, x_j] - \mu_i\, \mu_j = 0$ by definition
- Independence – $p(x_i,x_j) = p(x_i)p(x_j)$ – is a much stronger property than uncorrelation and guarantees:

  $E[x_i^N\, x_j^M]=E[x_i^N]\, E[x_j^M]$ for any N, M

  and even $E[f(x_i)\, g(x_j)]=E[f(x_i)]\, E[g(x_j)]$ for any f, g

Sample mean and sample covariance

- For multivariate variables, given N X_i samples:

$$\mu \equiv E[X] \approx \frac{1}{N}\sum_{i=1}^{N} X_i \qquad \textit{sample mean}$$

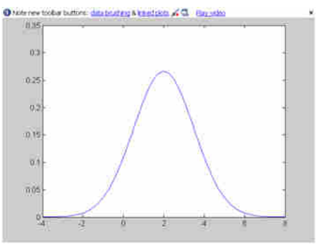
$$\Sigma \equiv E\left[(X-\mu)(X-\mu)^T\right] \approx \frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)(X_i-\mu)^T \qquad \textit{sample covariance}$$

Gaussian distribution

- The Gaussian, or normal, distribution enjoys nice properties making it very popular for pdf modelling
- Gaussian pdf in 1 dimension (univariate):

$$p(x) = N\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

with μ=2, σ=1.5:

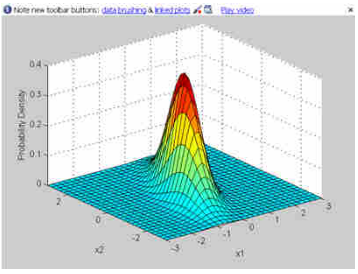


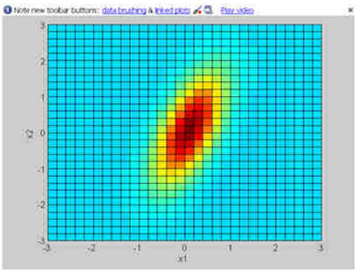
Multivariate Gaussian distribution

- Gaussian pdf in D dimensions ($X=[x_1,.. x_D]^T$):

$$p(X) = N(X \mid \mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)}$$

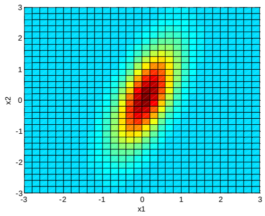
with D=2,
$\mu_1=0$, $\mu_2=0$,
$\Sigma=[.25\ .3;\ .3\ 1]$





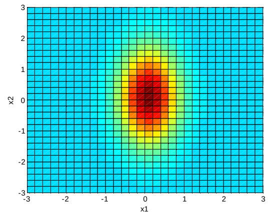
* See the Notes Page *

Full, diagonal, spherical covariance



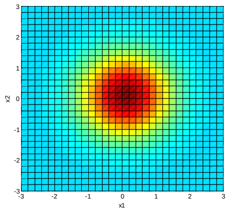
full covariance matrix

$$\Sigma = \begin{bmatrix} 0.25 & 0.3 \\ 0.3 & 1 \end{bmatrix}$$



diagonal covariance matrix

$$\Sigma = \begin{bmatrix} 0.25 & 0 \\ 0 & 1 \end{bmatrix}$$



spherical covariance matrix

$$\Sigma = \begin{bmatrix} 0.8 & 0 \\ 0 & 0.8 \end{bmatrix}$$

Properties of Gaussian distributions

- Mean and variance identify the whole pdf
- Uncorrelation ≡ independence
  - Covariance matrix of joint probability becomes diagonal
- ! Given x_1 and x_2 jointly Gaussian, also their marginal and conditional pdfs are Gaussian, and the mean and covariance are available analytically (see *partitioned Gaussians* in Bishop)
- Linear transformations are Gaussian:

  given $X \sim N(\mu, \Sigma)$

  $Y = A X + K$

  $\rightarrow Y \sim N(A\mu + K, A\Sigma A^T)$

* See the Notes Page *

Properties of Gaussian distributions: example

- Just an example: given two scalar Gaussian r.v., $x_1 \sim N(\mu_1, \sigma_1^2)$ and $x_2 \sim N(\mu_2, \sigma_2^2)$ as marginal probabilities, consider $\mathbf{y = x_1 + x_2}$
- This is equivalent to $X = [x_1,x_2]^T$, $A = [1 \;\; 1]$ and $y = AX$
  $\rightarrow \mu_y = \mu_1 + \mu_2$; $\sigma_y^2 = \sigma_1^2 + 2\,\mathrm{cov}(x_1,x_2) + \sigma_2^2$
- If x_1,x_2 have common variance, σ_x^2:
  $\rightarrow \sigma_y^2 = 2\sigma_x^2 + 2\,\mathrm{cov}(x_1,x_2)$
- If they are also uncorrelated/independent :
  $\rightarrow \sigma_y^2 = 2\sigma_x^2 \quad (\sigma_y = \sqrt{2}\,\sigma_x)$
- If they have maximal, positive correlation (degenerate case: $\det(\Sigma) = 0$):
  $\rightarrow \sigma_y^2 = 4\sigma_x^2 \quad (\sigma_y = 2\,\sigma_x)$
- If they have maximal, negative correlation (degenerate case likewise):
  $\rightarrow \sigma_y^2 = 0$

Sampling the Gaussian

- Assume we have a uniform random number generator in interval (0,1)
- We can use *the Box-Muller method* to generate independent, univariate Gaussian random samples with zero mean and unit standard deviation
- To obtain D-variate samples, X, just concatenate D univariate samples; their distribution has $\mu = 0$ and $\Sigma = I$ (the identity, or unit, matrix)
- Eventually, to obtain D-variate Gaussian samples, Z, with arbitrary μ and Σ, just use the properties of linear combinations of Gaussian distributions:

  $Z = W X + \mu$

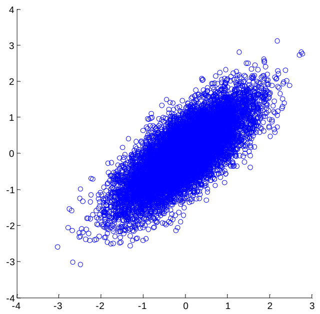
  where W such that $W W^T = \Sigma$ is obtained by the Choleski decomposition (Σ must be full rank)

  Z has therefore mean equal to μ and covariance equal to $W I W^T = \Sigma$

 39

Example

- A scatter plot of 10,000 2D Gaussian samples (μ = [0,0], Σ=[0.61 0.48; 0.48 0.64])

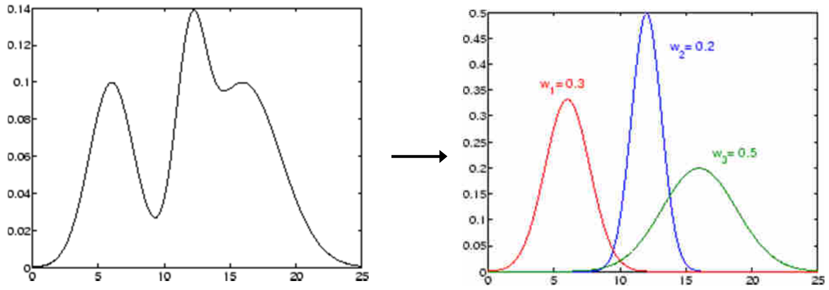


* See the Notes Page *

 40

Mixture distribution

- A *mixture distribution* is a distribution combining a finite number (say, M) of distributions, known as the *components*
- A mixture distribution is often used to represent multi-modal distributions, i.e. distributions with more than one mode:



Mixture distribution

- The principle of a mixture distribution is that **each sample, X, is generated from one of its components**
- A new, discrete random variable is introduced to indicate the component:
  $z \in \{1, \ldots, l, \ldots M\}$
- Each component is described by its pdf, $p(X \mid z = l)$, or $p_l(X)$ if one prefers a shorter notation
- Each component has a prior probability, $p(z = l)$, sometimes noted as α_l or π_l and called the component's "weight"

Mixture distribution: pdf

- The pdf of the mixture distribution, p(X), can be obtained by marginalising the component's index, z:

$$p(X) = \sum_{l=1}^{M} p(X, z = l) =$$
$$\sum_{l=1}^{M} p(X \mid z = l)\, p(z = l) =$$
$$\sum_{l=1}^{M} \alpha_l \; p_l(X)$$

- Given that M is usually small, evaluation (i.e. given X, compute p(X)) is not unreasonably heavy

Mixture distribution: inference

- Variable z is called a **latent** (*hidden*, *unobserved*) **random variable**; instead, X is the value (called *measurement* or *observation*) of an observed random variable
- The process of assigning a probability to z given X, p(z|X), is known as **inference** and plays a major role in statistical pattern recognition
- For the mixture distribution, we have:

$$p(z = l \mid X) = \frac{p(z = l, X)}{p(X)} = \frac{\alpha_l \; p_l(X)}{\sum_{k=1}^{M} \alpha_k \; p_k(X)}$$

Sampling a mixture distribution

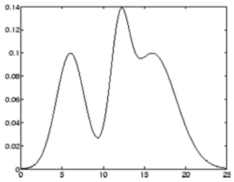
- The mixture distribution can be sampled by *ancestral sampling*:
  - first, draw one value out of M according to discrete distribution p(z); this picks the component
  - second, draw a sample from the selected component
- The so-called *generative model* of the mixture distribution is p(X,z) = p(X|z)p(z). It is represented by the *graphical model* below:

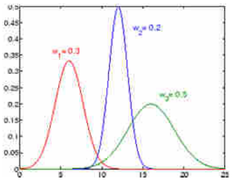
Gaussian mixture model (GMM)

- A Gaussian mixture model (GMM) has components which are Gaussians with their individual mean and covariance
- The pdf of a GMM is given by:

$$p(X) = \sum_{l=1}^{M} \alpha_l \, N(X \mid \mu_l, \Sigma_l)$$

- GMMs are very useful and popular models since they can represent multimodal distributions with Gaussian modes

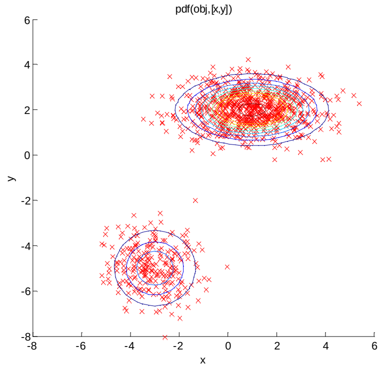




Example

- A scatter plot of 1,000 2D samples generated from a GMM with 2 components

  $\alpha_1 = 0.75$, $\mu_1 = [1,2]$, $\Sigma_1=[2\ 0;\ 0\ 0.5]$, $\alpha_2 = 0.25$, $\mu_2 = [-3,-5]$, $\Sigma_2=[1\ 0;0\ 1]$



* See the Notes Page *

