EEL-5840/EEL-4930 Elements of Machine Intelligence

Probability Review

Why Bother About Probabilities?

- Accounting for *uncertainty* is a crucial component in decision making (e.g., classification) because of ambiguity in our measurements.
- Probability theory is the proper mechanism for accounting for *uncertainty*.
- Need to take into account reasonable preferences about the state of the world, for example:

"If the fish was caught in the Atlantic ocean, then it is more likely to be salmon than sea-bass

Definitions

- Random experiment
  - An experiment whose result is not certain in advance (e.g., throwing a die)
- Outcome
  - The result of a random experiment
- Sample space
  - The set of all possible outcomes (e.g., {1,2,3,4,5,6})
- Event
  - A subset of the sample space (e.g., obtain an odd number in the experiment of throwing a die = {1,3,5})

Intuitive Formulation

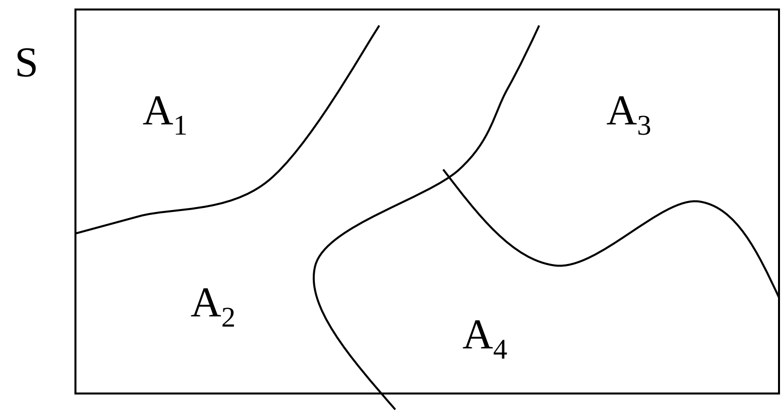
- Intuitively, the probability of an event **a** could be defined as:

$$P(a) = \lim_{n \to \infty} \frac{N(a)}{n}$$

Where ***N(a)*** is the number that event a happens in n trials

- Assumes that all outcomes in the sample space are *equally likely* (Laplacian definition)

Axioms of Probability



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

(2) $P(S) = 1$ (S is the sample space)

(3) If $A_1, A_2, \ldots, A_n$ are mutually exclusive events (i.e., $P(A_i \cap A_j) = 0$), then:

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

Note: we will denote $P(A \cap B)$ as $P(A, B)$)

Prior (Unconditional) Probability

- This is the probability of an event prior to arrival of any evidence.

P(Cavity)=0.1 means that “*in the absence of any other information, there is a 10% chance that the patient is having a cavity*”.

Posterior (Conditional) Probability

- This is the probability of an event given some evidence.

P(Cavity/Toothache)=0.8 means that "*there is an 80% chance that the patient is having a cavity given that he is having a toothache*"

Posterior (Conditional) Probability (cont'd)

- Conditional probabilities can be defined in terms of unconditional probabilities:

$$P(A/B) = \frac{P(A,B)}{P(B)}, \quad P\left(B/A\right) = \frac{P(A,B)}{P(A)}$$

- Conditional probabilities lead to the *chain rule*:

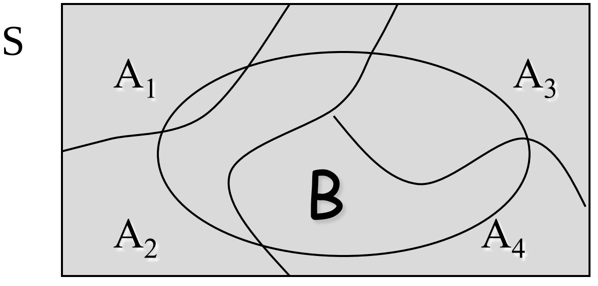
$$P(A,B)=P(A/B)P(B)=P(B/A)P(A)$$

Law of Total Probability

- If $A_1, A_2, \ldots, A_n$ is a partition of mutually exclusive events and B is any event, then

$$P(B) = P(B/A_1)P(A_1) + P(B/A_2)P(A_2) + \ldots + P(B/A_n)P(A_n)$$

$$= \sum_{j=1}^{n} P(B/A_j)P(A_j)$$



- Special case :

$$P(A) = P(A,B) + P(A,\bar{B})$$

- Using the chain rule, we can rewrite the law of total probability using conditional probabilities:

$$P(A) = P(A,B) + P(A,\bar{B}) = P(A/B)P(B) + P(A/\bar{B})P(\bar{B})$$

Law of Total Probability: Example

- My mood can take one of two values
  - Happy, Sad
- The weather can take one of three values
  - Rainy, Sunny, Cloudy
- We can compute *P(Happy)* and *P(Sad)* as follows:

$$P(Happy)=P(Happy/Rainy)+P(Happy/Sunny)+P(Happy/Cloudy)$$

$$P(Sad)=P(Sad/Rainy)+P(Sad/Sunny)+P(Sad/Cloudy)$$

Bayes' Theorem

- Conditional probabilities lead to the *Bayes' rule*:

$$\boxed{P(A/B)=\frac{P(B/A)P(A)}{P(B)}}$$

where $P(B)=P(B,A)+P(B,\overline{A})=P(B/A)P(A)+P(B/\overline{A})P(\overline{A})$

- Example: consider the probability of *Disease* given *Symptom:*

$$P(Disease/Symptom)=\frac{P(Symptom/Disease)P(Disease)}{P(Symptom)}$$

where

$$P(Symptom) = P(Symptom / Disease)P(Disease) + P(Symptom / \overline{Disease})P(\overline{Disease})$$

Bayes' Theorem Example

- Meningitis causes a stiff neck 50% of the time.
- A patient comes in with a stiff neck – what is the probability that he has meningitis?
- Need to know two things:
  - The prior probability of a patient having meningitis (1/50,000)
  - The prior probability of a patient having a stiff neck (1/20)

$$P(M / S) = \frac{P(S / M) P(M)}{P(S)}$$

$P(M/S)=0.0002$

General Form of Bayes' Rule

- If $A_1, A_2, \ldots, A_n$ is a partition of mutually exclusive events and B is any event, then the Bayes' rule is given by:

$$\boxed{P(A_i / B) = \frac{P(B / A_i)P(A_i)}{P(B)}}$$

where

$$P(B) = \sum_{j=1}^{n} P(B / A_j)P(A_j)$$

Independence

- Two events *A* and *B* are independent iff:

$$P(A,B)=P(A)P(B)$$

- From the above formula, we can show:

$$P(A/B)=P(A) \text{ and } P(B/A)=P(B)$$

- *A* and *B* are <u>conditionally independent</u> given *C* iff:

$$P(A/B,C)=P(A/C)$$

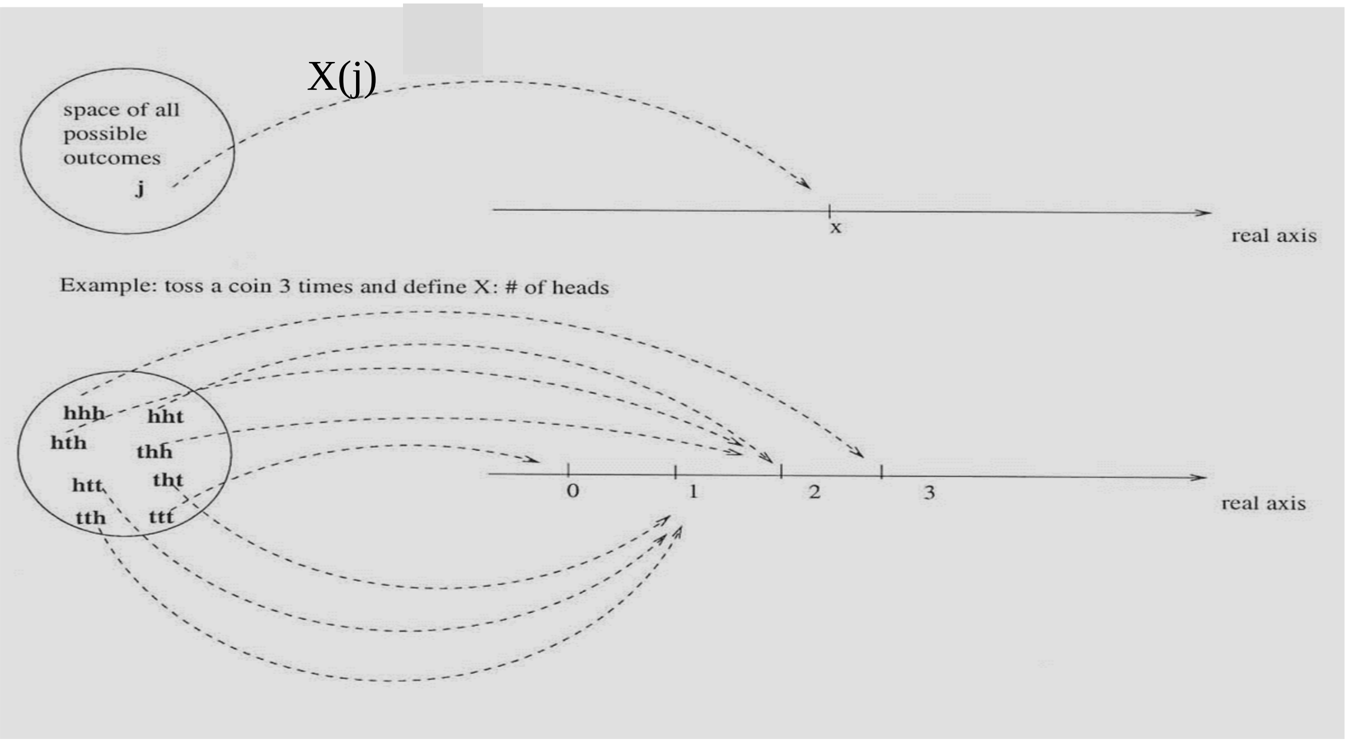
e.g., $P(WetGrass/Season,Rain)=P(WetGrass/Rain)$

Random Variables

- In many experiments, it is easier to deal with a summary variable than with the original probability structure.
- *Example*: in an opinion poll, we ask 50 people whether agree or disagree with a certain issue.
  - Suppose we record a "1" for agree and "0" for disagree.
  - The sample space for this experiment has 2^{50} elements.
  - Suppose we are only interested in the number of people who agree.
  - Define the variable *X=number of "1" 's recorded out of 50*.
  - Easier to deal with this sample space (has only 51 elements).

Random Variables (cont'd)

- A random variable (r.v.) is the value we assign to the outcome of a random experiment (i.e., a function that assigns a real number to each event).



Discrete/Continuous Random Variables

- A *discrete* r.v. can assume only a countable number of values.
- Consider the experiment of throwing a pair of dice

$$X = \text{"sum of dice"}$$

$$\text{e.g.,} \qquad X = 5 \text{ corresponds to } A_5 = \{(1,4),(4,1),(2,3),(3,2)\}$$

$$P(X = x) = P(A_x) = \sum_{s:X(s)=x} P(s) \text{ or}$$

$$P(X = 5) = P((1,4)) + P((4,1)) + P((2,3)) + P((2,3)) = 4/36 = 1/9$$

- A *continuous* r.v. can assume a range of values (e.g., sensor readings).

Probability mass (pmf) and density function (pdf)

- The *pmf* /*pdf* of a r.v. X assigns a probability for each possible value of X.

- **Warning:** given two r.v.'s, X and Y, their *pmf/pdf* are denoted as $p_X(x)$ and $p_Y(y)$; for convenience, we will drop the subscripts and denote them as $p(x)$ and $p(y)$, however, keep in mind that these functions are different !

Probability mass (pmf) and density function (pdf) (cont'd)

- Some properties of the *pmf* and *pdf*:

$$\sum_x p(x) = 1 \text{ (pmf)}$$

$$P(a < X < b) = \sum_{k=a}^{b} p(k) \text{ (pmf)}$$

$$\int_{-\infty}^{\infty} p(x)dx = 1 \text{ (pdf)}$$

$$P(a < X < b) = \int_{a}^{b} p(t)dt \text{ (pdf)}$$

Probability Distribution Function (PDF)

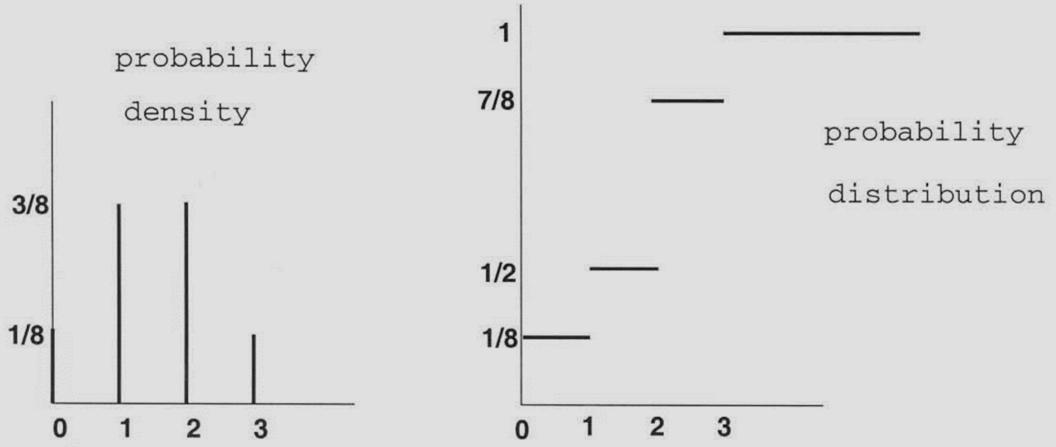
- With every r.v., we associate a function called *probability distribution function* (PDF) which is defined as follows:

$$F(x) = P(X \le x)$$

- Some properties of the PDF are:
  - (1) $0 \le F(x) \le 1$
  - (2) *F(x)* is a non-decreasing function of *x*
- If X is discrete, its PDF can be computed as follows:

$$F(x) = P(X \le x) = \sum_{k=0}^{x} P(X = k) = \sum_{k=0}^{x} p(k)$$

Probability Distribution Function (PDF) (cont'd)



$$F(0) = P(X \le 0) = P(X = 0) = 1/8$$

$$F(1) = P(X \le 1) = P(X = 0) + P(X = 1) = 1/2$$

$$F(2) = P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2) = 7/8$$

$$F(3) = P(X \le 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 1$$

Probability mass (pmf) and density function (pdf) (cont'd)_

- If X is continuous, its PDF can be computed as follows:

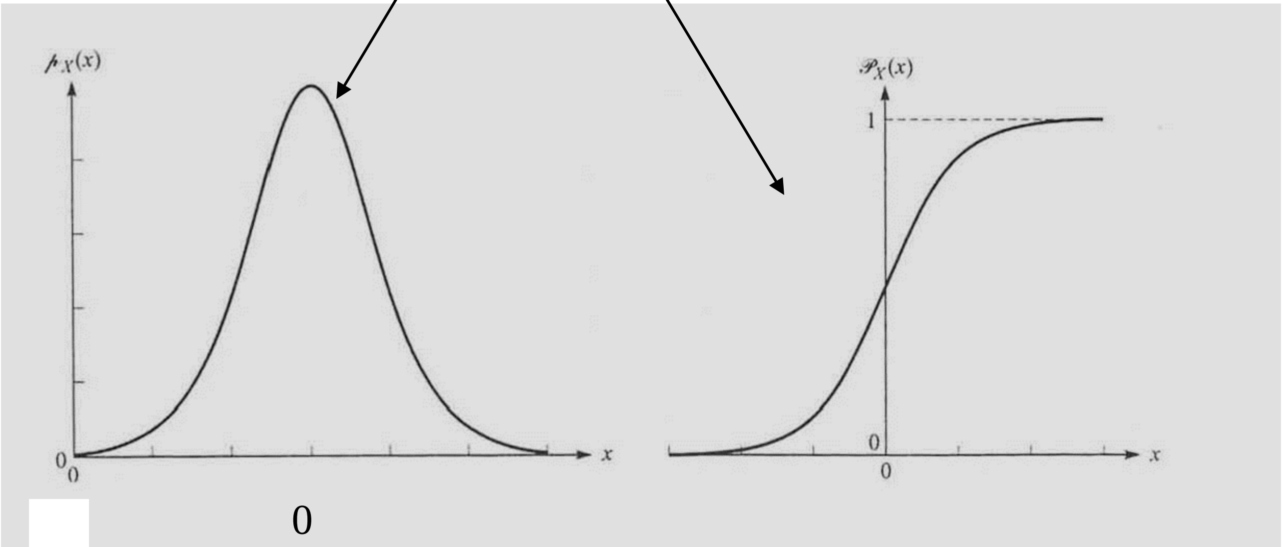
$$F(x) = \int_{-\infty}^{x} p(t)dt \quad for\, all\, x$$

- Using the above formula, it can be shown that:

$$p(x) = \frac{dF}{dx}(x)$$

Probability mass (pmf) and density function (pdf) (cont'd)_

- Example: the Gaussian *pdf* and *PDF*



Joint *pmf* (discrete r.v.)

- For n random variables, the joint *pmf* assigns a probability for each possible combination of values:

$$p(x_1,x_2,\ldots,x_n)=P(X_1=x_1,\ X_2=x_2,\ \ldots,\ X_n=x_n)$$

Warning: the joint *pmf /pdf* of the r.v.'s $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are denoted as $p_{X1X2\ldots Xn}(x_1,x_2,\ldots,x_n)$ and $p_{Y1Y2\ldots Yn}(y_1,y_2,\ldots,y_n)$; for convenience, we will drop the subscripts and denote them $p(x_1,x_2,\ldots,x_n)$ and $p(y_1,y_2,\ldots,y_n)$, keep in mind, however, that these are two different functions.

Joint *pmf* (discrete r.v.) (cont'd)

- Specifying the joint *pmf* requires an enormous number of values
  - k^n assuming n random variables where each one can assume one of k discrete values.
  - things much simpler if we assume independence or conditional independence …

P(Cavity, Toothache) is a 2 x 2 matrix

Joint Probability

	Toothache	Not Toothache
Cavity	0.04	0.06
Not Cavity	0.01	0.89

Sum of probabilities = 1.0

Joint *pdf* (*c*ontinuous r.v.)

For *n* random variables, the joint *pdf* assigns a probability for each possible combination of values:

$$p(x_1, x_2, ..., x_n) \geq 0$$

$$\int_{x_1} ... \int_{x_n} p(x_1, x_2, ..., x_n) dx_1 ... dx_n = 1$$

Discrete/Continuous Probability Distributions

Probability Distributions

	Continuous vars	Discrete vars
P(X)	Function of one variable	M vector
P(X=x)	Scalar*	Scalar
P(X,Y)	Function of two variables	MxN matrix
P(X|Y)	Function of two variables	MxN matrix
P(X|Y=y)	Function of one variable	M vector
P(X=x|Y)	Function of one variable	N vector
P(X=x|Y=y)	Scalar*	Scalar

Interesting Properties

- The conditional *pdf* can be derived from the joint *pdf*:

$$p(y/x) = \frac{p(x,y)}{p(x)} \; or \; p(x,y) = p(y/x)\,p(x)$$

- Conditional *pdf*s lead to the chain rule (*general form*):

$$p(x_1, x_2, ..., x_n) = p(x_1/x_2, ..., x_n)\,p(x_2/x_3, ..., x_n) ... p(x_{n-1}/x_n)\,p(x_n)$$

Interesting Properties (cont'd)

- Knowledge about independence between r.v.'s is *very* powerful since it simplifies things a lot,

  e.g., if *X* and *Y* are independent, then:

$$p(x, y) = p(x)\,p(y)$$

- The law of total probability:

$$p(y) = \sum_{x} p(y / x)\,p(x)$$

Marginalization

- From a joint probability, we can compute the probability of any subset of the variables by marginalization:
  - Example - case of joint *pmf* :

$$p(x) = \sum_{y} p(x, y)$$

  - Examples - case of joint *pdf* :

$$p(x) = \int_{-\infty}^{\infty} p(x, y)dy$$

$$p(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = \int_{-\infty}^{\infty} p(x_1, x_2, \ldots, x_n)dx_i$$

$$p(x_1, x_2) = \int_{x_3} \ldots \int_{x_n} p(x_1, x_2, \ldots, x_n)dx_3 \ldots dx_n$$

Why is the joint pmf (or pdf) useful?

Given $P(X,Y)$, we can always calculate:

$P(X)$	$P(X=x_1)$
$P(Y)$	$P(Y=y_2)$
$P(X\|Y)$	$P(X\|Y=y_1)$
$P(Y\|X)$	$P(Y\|X=x_1)$

By using (1) marginalization and (2) the Chain Rule

Simple example:

		X		
		x_1	x_2	x_3
Y	y_1	0.2	0.1	0.1
	y_2	0.1	0.2	0.3

Why is the joint pmf (or pdf) useful (cont'd)?

P(X,Y)

	x_1	x_2	x_3
y_1	0.2	0.1	0.1
y_2	0.1	0.2	0.3

P(X)

x_1	x_2	x_3
0.3	0.3	0.4

P(Y)

y_1	0.4
y_2	0.6

P(X|Y)

	x_1	x_2	x_3
y_1	0.5	0.25	0.25
y_2	0.167	0.333	0.5

P(Y|X)

	x_1	x_2	x_3
y_1	0.667	0.333	0.25
y_2	0.333	0.667	0.75

$P(X=x_1, Y=y_2) = ?$

$P(X=x_1) = ?$

$P(Y=y_2) = ?$

$P(X|Y=y_1) = ?$

$P(X=x_1|Y) = ?$

Probabilistic Inference

- If we could define all possible values for the probability distribution, then we could read off any probability we were interested in.
- In general, <u>it is not practical to define all possible entries for the joint probability function</u>.
- *Probabilistic inference* consists of computing probabilities that are not explicitly stored by the reasoning system (e.g., marginals, conditionals).

Normal (Gaussian) Distribution

- The Gaussian pdf is defined as follows:

$$p(\mathbf{x}) = \frac{1}{\sigma\sqrt{2\pi}} exp[\frac{(x-\mu)^2}{2\sigma^2}]$$

where μ is the mean and σ the standard deviation.

- The multivariate Gaussian ($\mathbf{x}$ is a vector) is defined as follows:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(\mathbf{x}-\mu)^t\Sigma^{-1}(\mathbf{x}-\mu)]$$

where μ is the mean and Σ the covariance matrix.

Normal (Gaussian) Distribution (cont'd)

- Shape and parameters of Gaussian distribution
  - number of parameters is $d + \frac{d(d+1)}{2}$
  - shape determined by Σ

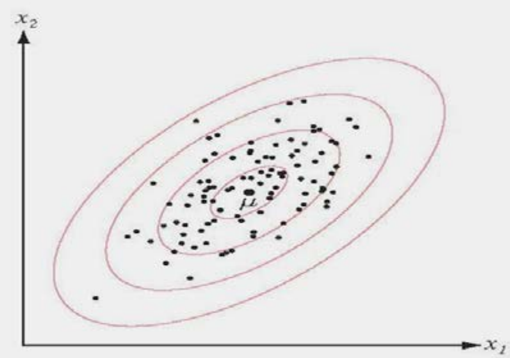


FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}$. The ellipses show lines of equal probability density of the Gaussian. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Normal (Gaussian) Distribution (cont'd)

- Mahalanobis distance:

$$r^2 = (\mathbf{x} - \mu)^t \Sigma^{-1} (\mathbf{x} - \mu)$$

- If variables are <u>independent</u>, the multivariate normal distribution becomes:

$$p(x) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp[\frac{(x_i - \mu_i)^2}{2\sigma_i^2}]$$

Expected Value

- The expected value for a discrete r.v. X is given by

$$E(X) = \sum_x x p(x)$$

Example: Let X denote the outcome of a die roll

$$E(X) = 1\ 1/6 + 2\ 1/6 + 3\ 1/6 + 4\ 1/6 + 5\ 1/6 + 6\ 1/6 = 3.5$$

- The "sample" mean $\bar{x}$ for a r.v. X is given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where x_i denotes the i-th measurement of X.

Expected Value (cont'd)

- The mean and the expected value are related by

$$E(X) = \lim_{n \to \infty} \bar{x}$$

- The expected value for a continuous r.v. is given by

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

Example: $E(X)$ for the Gaussian is μ.

Properties of the Expected Value

- The expected value of a function $g(X)$ is given by:

$$E(g(X)) = \sum_x g(x)p(x) \text{ (discrete case)}$$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)p(x)dx \text{ (continuous case)}$$

- Linearity property

$$E(af(X) + bg(Y)) = aE(f(X)) + bE(g(Y))$$

Variance and Standard Deviation

- The variance $Var(X)$ of a r.v. X is defined by

$$Var(X) = E((X-\mu)^2), \text{ where } \mu = E(X)$$

- The "sample" variance $\overline{Var}$ for a r.v. X is given by

$$\overline{Var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- The standard deviation σ of a r.v. X is defined by

$$\sigma = \sqrt{Var(X)}$$

Example: The variance of the Gaussian is σ^2

Covariance

- The covariance of two r.v. X and Y is defined by:

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

where $\mu_X = E(X)$ and $\mu_Y = E(Y)$

- The correlation coefficient ρ_{XY} between X and Y is given by:

$$\rho_{XY}a = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

- The "sample" covariance matrix is given by:

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n-1}(x_i-\bar{x})(y_i-\bar{y})$$

Covariance Matrix

- The covariance matrix of 2 random variables is given by:

$$C_{XY} = \begin{bmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{bmatrix}$$

where $Cov(X,X) = Var(X), \quad Cov(Y,Y) = Var(Y)$

- The covariance matrix of n random variables is given as:

$$C_X = \begin{bmatrix} Cov(X_1,X_1) & Cov(X_1,X_2) & \dots & Cov(X_1,X_n) \\ Cov(X_2,X_1) & Cov(X_2,X_2) & \dots & Cov(X_2,X_n) \\ \dots & \dots & \dots & \dots \\ Cov(X_n,X_1) & Cov(X_n,X_2) & \dots & Cov(X_n,X_n) \end{bmatrix}$$

where $Cov(X_i,X_j) = Cov(X_j,X_i)$ and $Cov(X_i,X_i) \geq 0$

Example: Σ is the covariance matrix of the multivariate Gaussian.

Uncorrelated r.v.'s

- X and Y are called *uncorrelated*, if:

$$Cov(X, Y) = 0$$

- $X_1, X_2, \ldots, X_n$ are called *uncorrelated*, if:

$$C_X = \Lambda, \qquad \text{where } \Lambda \text{ is a diagonal matrix.}$$

Properties of the covariance matrix

- Since C_X is symmetric, it has *real* eigenvalues ≥ 0

- Any two eigenvectors, with different eigenvalues, are *orthogonal*.

- The eigenvectors corresponding to different eigenvalues define a *basis*.

Moments of r.v.'s

- Definition of moments:

$$m_n = E(x^n)$$

- Definition of central moments:

$$cm_n = E((x-\mu)^n)$$

- Useful moments

m_1: mean
cm_2: variance
cm_3: skewness (measure of asymmetry of a distribution)
cm_4: kurtosis (detects heave and light tails and deformations of a distribution)

Questions?