18-794 - Pattern Recognition Theory, Fall 2013
Notes On Probability, Statistics & MATLAB

September 16th, 2013

Abstract This is a set of notes that summarizes the concepts outlined in the recitation. These notes are meant to be a quick refresher of fundamental formulae and mathematical principles related to probability and statistics that will come in handy later on in the course. For a more complete understanding of these concepts, please read up on material mentioned in the references section. More notes and links related to these topics will be posted on blackboard.

Terminology: P always represents probability of an event, E represents expectation and A, B, C etc are used to denote events. Please do not get confused between scalars, vectors, matrices and Random Variables (RVs). The convention followed is that RVs are always capitalized (X) and are capitalized, underlined if they are random vectors ($\underline{X}$). Matrices are capitalized, in bold and have a bar below them ($\underline{\mathbf{\Sigma}}$). Vectors are in lower case, not in bold and have a bar below them ($\underline{\mu}$), while scalars are simply represented in lower case, are not in bold and have no bar below them (s).

1 Basic Probability Theory

The probability of an event A can be determined using observations and the relative frequency approach so that we can finally assign the value $P(A) = \lim_{n\to\infty} \frac{n_A}{n}$ if we make n observations and the event A occurs n_A times. This is one of the oldest definitions of probability but is still worth keeping in mind.

The axioms of probability (which seem quite obvious) are:

1. $0 \leq P(A) \leq 1$
2. The probability of a certain event equals 1 and that of an impossible event equals 0
3. If the events A and B are mutually exclusive i.e. $P[A \cap B] = 0$ then $P[A \cup B] = P[A] + P[B]$

Many other theorems of probability follow from axioms of set theory. For example if A is an event and $\bar{A}$ is its complement event, then using set theory it becomes easy to see that $P(\bar{A}) = 1 - P(A)$. Similarly using a Venn diagram it can be observed that $A \cup B = A + B - A \cap B$ and hence $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ and if A and B are mutually exclusive, then $A \cap B = 0$ and hence we end up with axiom 3.

Conditional Probability and Bayes Theorem The conditional probability of an event A, given that an event B has occurred is given by

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

assuming of course that $P(A)$ and $P(B)$ are non zero. This rule, known as Bayes rule is widely used and almost any course at CMU that involve the use of probability start with a description of it. At this point it becomes necessary to go into the details of two more terms.

Two events A and B are said to be mutually exclusive if they cannot occur together at the same so that $P(AB) = 0$ and hence $P(A \cup B) = P(A) + P(B)$. Two events A and B are statistically independent if the

occurrence of A does not in any way affect the occurrence of B. This is a very powerful concept and makes life very simple when computing complex probabilistic events composed of the occurrence of several small events because it leads to the fact that $P(AB) = P(A)P(B)$. Using this in Bayes rule leads to some more results. If A and B are independent then we can observe that

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(B)P(A)}{P(B)} = P(A) \tag{2}$$

Total Probability Law The total probability law states that if $U = [A_1...A_n]$ is a partition of S into n mutually exclusive events and B is an arbitrary event, then

$$P(B) = P(B|A_1)P(A_1) + ...P(B|A_n)P(A_n) \tag{3}$$

and hence using Bayes rule we can also determine that

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + ...P(B|A_n)P(A_n)} \tag{4}$$

The simplest way of determining the probability of an event is to enumerate all possible ways the event can occur and then determine the total possible outcomes and simply take the ratio of the two. Here are a couple of examples to illustrate this and the use of Bayes rule.

Example 1 The birthday problem: Assume there are n people at a party (including you), what is the probability that (a) Two or more people share the same birthday (b) Someone in the group has the same birthday as you?

Answer 1 (a) To approach this problem it would be easier to determine the probability of the complementary event (A) and determine the probability of no two people sharing a birthday. Let us also use $N = 365$. Among n people, each person could have a birthday on any of the N days, hence the total outcomes possible are N^n. If no two people can share the same birthday, then the first person could have a birthday on any of N days, the second one on $N-1$ days, the third on $N-2$ and so so that the last one can have a birthday on only $N-(n-1)$ days. So the P(No two people have same birthday)= $P(A) = \frac{N(N-1)...(N-n+1)}{N^n} = \prod_{i=1}^{n-1}(1-\frac{i}{N})$. The probability of the complementary event is hence $P(\bar{A} = 1 - P(A)$ and is the solution to (a). It can be simplified further using the property that $1 - x \simeq e^{-x}$.

(b) Now to obtain the probability of the event B (the event that someone in the group has the same birthday as you), it is again easier to compute $P(\bar{B})$ and look at the ways that no one in the group has the same birthday as you. For this to happen, each of the n people can have a birthday from among $N-1$ days (all except yours), so $P(\bar{B}) = \frac{(N-1)(N-1)...(N-1)}{N^n} = \frac{(N-1)^n}{N^n} = (1-\frac{1}{N})^n$, which can be approximated as $e^{-n/N}$. Hence, $P(B) = 1 - e^{-n/N}$.

Example 2 - Courtesy 10-701 Fall 2009 Course - John has to answer a True or False question in an exam. The probability that he knows the correct answer is p, and thus the probability that he does not know the correct answer is $1-p$. Since there is no negative marking, he decides to guess in case he does not know the answer. With probability q he chooses the true option and $1-q$ the false option. However, he is ignorant that the professor is biased when he decides which of the two options is the correct one. To be precise, the professor with probability δ assigns the correct answer to the true option and with $1-\delta$ to the false one. John gives an answer to the question and he answers correctly! What is the chance that he actually knew the answer?

Answer 2 P(John knew answer|He answers correctly) = P(John knew answer and answers correctly)/P(John answers correctly). Let the event John knew answer = K, the event John gets answer correct = C and the event John guessed = $\bar{K}$. Hence what we want is to find $P(K|C) = P(KC)/P(C)$. Now the numerator $P(KC) = p$. The denominator is more complex. $P(C) = P(C|K)P(K) + P(C|\bar{K})P(\bar{K})$. Now again,

$P(C|\bar{K}) = P(T_J, T_P|\bar{K}) + P(\bar{T_J}, \bar{T_P}|\bar{K})$ where T_J is the event that John chooses to guess True and T_P is the event that the professor also chooses to set the correct answer as True.

So finally we obtain $P(K|C) = P(KC)/P(C) = \frac{p}{p+(1-p)[q\delta+(1-q)(1-\delta)]}$.

2 Random Variables, PDFs, CDFs etc

A random variable (RV) takes on different values with probabilities that sum up to 1. More formally, to an experiment specified by S, to each outcome i of this experiment we assign a probability given by $X(i)$ where X is the random variable (always represented by a capital letter) and x represents a scalar value that X can take on. For a vector of RVs we use $\underline{X}$. An RV can be continuous, discrete or mixed. Two fundamental functions associated with an RV are the CDF and PDF.

The cumulative distribution function (CDF) is defined as

$$F_X(x) = P(X \leq x) \tag{5}$$

Properties of the CDF are:

$$F(+\infty) = 1, F(-\infty) = 0 \tag{6}$$

and

$$F(x_2) - F(x_1) = P(x_2 < x \leq x_1) \tag{7}$$

It must also be noted that the CDF of an RV is a non decreasing function.

The probability density (or mass) function (PDF - for continuous RVs or PMF - for discrete RVs) is the derivative of the CDF and is defined as

$$f_X(x) = \frac{d}{dx}(F_X(x)) \tag{8}$$

Thus, for a continuous RV we have that

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx \tag{9}$$

and for a discrete RV

$$F_X(x) = \sum_{x \leq x_i} f_X(x_i) \tag{10}$$

Properties of the PDF are:

$$f_X(x) \geq 0 \tag{11}$$

and

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1 \tag{12}$$

The PDF or PMF must sum up to 1, usually some constant is placed outside the integral or summation sign to ensure this as will be seen later when we go over commonly used PDFs.

Mean and Variance The mean of a PDF is called expectation $E(X)$. It is simply obtained as

$$\mu = E(X) = \int_{-\infty}^{+\infty} x f_X(x)dx \tag{13}$$

for a continuous RV and as

$$\sum_{-\infty}^{+\infty} x f_X(x) \tag{14}$$

for a discrete RV.

The variance σ^2 of an RV is the second moment about the mean and is defined as

$$\sigma^2 = E[(X - E(X))^2] = E(X^2) - E^2(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f_X(x) dx \tag{15}$$

for a continuous RV and as

$$\sum_{-\infty}^{+\infty} (x-\mu)^2 f_X(x) \tag{16}$$

for a discrete RV. The square root of variance is called std. deviation (σ).

Properties of Mean and Variance Some of the most important properties of the Expectation operator and Variance are:

1. $E(X + c) = E(X) + E(c) = E(X) + c$
2. $E(cX) = cE(X)$
3. $E(X + Y) = E(X) + E(Y)$
4. $E(XY) \neq E(X)E(Y)$ unless X and Y are uncorrelated
5. $var(aX) = a^2 var(X)$

Commonly Used PDFs The following PDFs are the ones you will most frequently encounter and it will be beneficial to remember their forms.

Please note that the mean of an RV X is not the value of X that maximizes its PDF ($f_X(x)$). The value of X which does this is known as the mode and is the most probable value of X.

1. Bernoulli

$$f_X(x) = x^p (1-x)^{(1-p)} \quad X = 0, 1 \tag{17}$$

which is simply $P(X = 0) = p$ and $P(X = 1) = q = (1 - p)$. $E(X) = p$ and $var(X) = p(1 - p)$. This distribution simply assigns probability of success ($X = 1$) as p and probability of failure ($X = 0$) as $q = (1-p)$.

2. Binomial - Repeated Bernoulli trials result in a binomial distribution who's PDF gives a way of calculating the probability of k successes in n trials.

$$P(X = k) = \binom{n}{k} p^k q^{n-k} \quad p + q = 1 \quad k = 0, 1, 2...n \tag{18}$$

$E(X) = np$ and $var(X) = npq = np(1 - p)$.

3. Geometric - X is said to be a geometric RV if

$$P(X = k) = pq^{k-1} \quad k = 1, 2, 3...\infty \tag{19}$$

This measures the probability of getting first success on the k-th trial. $E(X) = \frac{1}{p}$ and $var(X) = \frac{1-p}{p^2}$.

4. Poisson - X is a Poisson RV with parameter λ if X takes on the values 0, 1, 2...∞ with

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad k = 0, 1, 2...\infty \tag{20}$$

$E(X) = \lambda$ and $var(X) = \lambda$.

5. Uniform - X is said to be uniformly distributed in (a, b) $-\infty < a < b < +\infty$ if

$$f_X(x) = \frac{1}{b-a} \quad a \le x \le b$$
$$= 0 \quad otherwise \tag{21}$$

$E(X) = \frac{a+b}{2}$ and $var(X) = \frac{(b-a)^2}{12}$.

6. Normal or Gaussian - The most common distribution you will use. A normal RV X is normally distributed with parameters μ and σ^2 and is represented as $X \sim N(\mu, \sigma^2)$. The PDF of a univariate normal distribution is given as.

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty \le x \le +\infty \tag{22}$$

The CDF of such a distribution is given by

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2} dy \triangleq G(\frac{x-\mu}{\sigma}) \tag{23}$$

where the function

$$G(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy \tag{24}$$

is available in tabulated form and represents the area under a standard normal curve $X \sim N(0,1)$. To determine the area under a normal variable (Z) with different mean and variance, transform the variable into standard normal form by subtracting its mean and dividing by its std devn. and then use the normal table on it. There are other representations that are used to determine the area under a normal curve such as the error function ($erf(x)$) and the Q function ($Q(x)$).

The error function is defined as follows

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-y^2} dy \tag{25}$$

Similarly there is a complementary error function $erfc(x)$ defined as

$$erfc(x) = 1 - erf(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-y^2} dy \tag{26}$$

These are the same functions used in MATLAB. It is clearly different from $G(x)$ and the relationship between the two is that

$$G(x) = \frac{1}{2}[1 + erf(x/\sqrt{2})] \tag{27}$$

The Q function is defined as

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-y^2/2} dy = 1 - G(x) \tag{28}$$

Hence the relationship between $Q(x)$ and $erf(x)$ is

$$Q(x) = \frac{1}{2}[1 - erf(x/\sqrt{2})] \tag{29}$$

7. Exponential - An RV X is exponentially distributed with parameter λ if its density function is given by

$$f_X(x) = \lambda e^{-\lambda x} \quad x \ge 0$$
$$= 0 \quad otherwise \tag{30}$$

$E(X) = \frac{1}{\lambda}$ and $var(X) = \frac{1}{\lambda^2}$.

The first 4 distributions on the list are discrete while the next 3 are continuous.

3 Joint Distributions

Now we need to address the joint PDF of more than 1 RV. We start with the bivariate case (joint PDF of 2 RVs) and can extend this to deal with the multivariate case (many RVs). The joint CDF of 2 RVs X and Y is given by

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) \tag{31}$$

The joint PDF of the two RVs is given by

$$f_{X,Y}(x,y) = \frac{\delta^2 F_{X,Y}(x,y)}{\delta x \delta y} \tag{32}$$

The marginal PDFs of X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \tag{33}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx \tag{34}$$

These equations were simply obtained by marginalizing out one of the variables by integrating it over its full range of values. The same trick can be applied to understand how to obtain $F_X(x) = P(X \leq x)$ or $F_Y(y) = P(Y \leq y)$ from the joint distribution.

X and Y are SI (Statistically Independent) if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \tag{35}$$

This makes life easy when applying Bayes rule (as shown earlier). There is another closely related concept called uncorrelated variables. X and Y are uncorrelated if

$$E(X,Y) = E(X)E(Y) \tag{36}$$

It must be noted that SI does imply uncorrelated variables, but not vice versa i.e. two uncorrelated variables may not be SI. SI is the more powerful concept and is harder to prove. However, if X and Y are both normal RVs, and are uncorrelated, then they will be SI too. This is a property of only normal RVs and can not be applied in general.

Now it becomes important to understand the terms correlation, covariance etc. The covariance between 2 RVs X and Y is defined as

$$cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y) \tag{37}$$

For uncorrelated variables, $cov(X,Y) = 0$. Can you see why? Now the term correlation between 2 RVs X and Y is defined by the correlation coefficient ($\rho_{X,Y}$)

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{38}$$

$-1 \leq \rho_{X,Y} \leq 1$. A value of 1 implies a strong positive correlation, a value of 0 means no correlation (uncorrelated and maybe SI) and a value of -1 implies strong negative correlation between the RVs X and Y.

If we now have n RVs $X_1, X_2, X_3 ... X_n$ we represent this multivariate random vector as a column vector $\underline{X} = [X_1 \ X_2 \ X_3 \ ... \ X_n]^T$. Now the following become clear:

$$F_{\underline{X}}(\underline{x}) = F_{X_1,X_2...X_n}(x_1, x_2...x_n) = P(X_1 \leq x_1 ... X_n \leq x_n) = \int_{-\infty}^{x_n} ... \int_{-\infty}^{x_1} f_{X_1,X_2...X_n}(x_1, x_2...x_n)dx_1...dx_n \tag{39}$$

$$F_{\underline{X}}(-\infty \ \ldots \ -\infty) = 0 \quad F_{\underline{X}}(\infty \ \ldots \ \infty) = 1 \tag{40}$$

$$f_{\underline{X}}(\underline{x}) = P(X_1 = x_1 ... X_n = x_n) = f_{X_1,X_2...X_n}(x_1, x_2 ... x_n) = \frac{dF_{\underline{X}}(\underline{x})}{d\underline{X}} = \frac{\delta^n F_{X_1,X_2...X_n}(x_1, x_2 ... x_n)}{\delta x_1 ... \delta x_n} \tag{41}$$

$$f_{\underline{X}}(\underline{x}) \geq 0 \tag{42}$$

$$\int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} f_{X_1,X_2...X_n}(x_1, x_2 ... x_n) dx_1 ... dx_n = 1 \tag{43}$$

$$\underline{\mu} = E[\underline{X}] = E[X_1 \ \ldots \ X_n]^T = [E[X_1] \ \ldots \ E[X_n]]^T \tag{44}$$

Thus, the bivariate case is a simple case of the multivariate case with $\underline{X} = [X_1 \ X_2]^T$. In the univariate case the mean is a scalar, now it is a vector ($\underline{\mu}$), similarly we now need to define a covariance matrix which is analogous to the variance term in the univariate case. This covariance matrix ($\underline{\mathbf{\Sigma}}$) stores the covariance between each of the n RVs and is defined as follows.

$$\underline{\mathbf{\Sigma}} = \begin{bmatrix} cov(X_1,X_1) & cov(X_1,X_2) & ... & cov(X_1,X_n) \\ cov(X_2,X_1) & cov(X_2,X_2) & ... & cov(X_2,X_n) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ cov(X_n,X_1) & cov(X_n,X_2) & ... & cov(X_n,X_n) \end{bmatrix} = \begin{bmatrix} \sigma^2_{X_1} & cov(X_1,X_2) & ... & cov(X_1,X_n) \\ cov(X_2,X_1) & \sigma^2_{X_2} & ... & cov(X_2,X_n) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ cov(X_n,X_1) & cov(X_n,X_2) & ... & \sigma^2_{X_n} \end{bmatrix} \tag{45}$$

which can be expressed in matrix form as

$$\underline{\mathbf{\Sigma}} = E[(\underline{X} - E(\underline{X}))(\underline{X} - E(\underline{X}))^T] = E[(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})^T] = E[\underline{X}\underline{X}^T] - \underline{\mu}\underline{\mu}^T \tag{46}$$

This covariance matrix is going to be a recurring topic in this course. It is an $n \times n$ square matrix and is positive-semidefinite. Its diagonal entries are simply the variances of the variables X_1, X_2, ... X_n respectively while its non-diagonal entries are the covariance terms. It is clear that it is also a symmetric matrix as $cov(X_1, X_2) = cov(X_2, X_1)$.

As with the bivariate case it can be shown that n RVs are SI if

$$f_{X_1...X_n}(x_1 ... x_n) = \prod_i^n (f_{X_i}(x_i)) \tag{47}$$

It becomes important to see what the multivariate normal distribution looks like. If $\underline{X}$ is a multivariate normal RV it is represented as $\underline{X} \sim N(\underline{\mu}, \underline{\mathbf{\Sigma}})$ and its PDF is now given by the equation below.

$$f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^{n/2}|\underline{\mathbf{\Sigma}}|^{1/2}} exp[\frac{-(\underline{x} - \underline{\mu})^T \underline{\mathbf{\Sigma}}^{-1} (\underline{x} - \underline{\mu})}{2}] \tag{48}$$

Another property is worth mentioning at this stage. If $\underline{X}$ is $N(\underline{\mu}, \underline{\mathbf{\Sigma}})$ then $\underline{Y} = \underline{\mathbf{A}}\underline{X}$ is $N(\underline{\mathbf{A}}\underline{\mu}, \underline{\mathbf{A}}\underline{\mathbf{\Sigma}}\underline{\mathbf{A}}^T)$.

This property becomes useful when trying to diagonalize a covariance matrix. The process is called whitening and will be covered again in class. Let us assume we have an random vector $\underline{X}$ that is $N(\underline{\mu}, \underline{\mathbf{\Sigma}})$ i.e. its covariance matrix is $\underline{\mathbf{\Sigma}}$. We obtain the eigenvectors of the covariance matrix and stack them columnwise in matrix $\underline{\mathbf{V}}$ and store the eigenvalues along the diagonal of matrix $\underline{\mathbf{\Lambda}}$ so $\underline{\mathbf{\Sigma}} = \underline{\mathbf{V}}\underline{\mathbf{\Lambda}}\underline{\mathbf{V}}^T$, $\underline{\mathbf{\Lambda}}^T = \underline{\mathbf{\Lambda}}$ and $\underline{\mathbf{V}}^T\underline{\mathbf{V}} = \underline{\mathbf{I}}$.

Now we generate a new random vector $\underline{Y} = \underline{\mathbf{A}}_w^T\underline{X}$ where $\underline{\mathbf{A}}_w = \underline{\mathbf{V}}\underline{\mathbf{\Lambda}}^{-1/2}$. This will lead to $\underline{Y}$ being $N(\underline{\mathbf{A}}_w^T\underline{\mu}, \underline{\mathbf{A}}_w^T\underline{\mathbf{\Sigma}}\underline{\mathbf{A}}_w) = N(\underline{\mathbf{A}}_w^T\underline{\mu}, \underline{\mathbf{\Lambda}}^{-1/2}\underline{\mathbf{V}}^T\underline{\mathbf{V}}\underline{\mathbf{\Lambda}}\underline{\mathbf{V}}^T\underline{\mathbf{V}}\underline{\mathbf{\Lambda}}^{-1/2}) = N(\underline{\mathbf{A}}_w^T\underline{\mu}, \underline{\mathbf{\Lambda}}^{-1/2}\underline{\mathbf{\Lambda}}\underline{\mathbf{\Lambda}}^{-1/2}) = N(\underline{\mathbf{A}}_w^T\underline{\mu}, \underline{\mathbf{I}})$.

4 Conditional Probability Revisited

We now revisit Bayes theorem and apply it to PDFs and CDFs. Let X and Y be two continuous RVs, then the conditional PDF $f_{X|Y}(x|y)$ is given by

$$f_{X|Y}(x|y) = P(X = x|Y = y) = f_{X,Y}(x, y)/f_Y(y) = P(X = x, Y = y)/P(Y = y) \tag{49}$$

Suppressing the subscripts for convenience, we can express this above rule in many ways

$$f(x|y) = f(x, y)/f(y) \tag{50}$$

$$f(y|x) = f(x, y)/f(x) \tag{51}$$

$$f(x, y) = f(x|y)f(y) = f(y|x)f(x) \tag{52}$$

and hence

$$f(x|y) = \frac{f(x, y)}{f(y)} = \frac{f(x|y)f(y)}{f(y)} = \frac{f(y|x)f(x)}{f(y)} = \frac{f(y|x)f(x)}{\int_{-\infty}^{\infty} f(y|x)f(x)dx} \tag{53}$$

The last part of the above formula was obtained by marginalizing the joint PDF.

If RV Y is being determined by observing RV X then we use the notation:

$f_Y(y)$ - a priori probability (or prior)

$f_{X|Y}(x|y)$ - Likelihood

$f_{Y|X}(y|x)$ - a posteriori probability (or posterior)

5 Useful MATLAB Functions

The following are useful MATLAB functions when dealing with PDFs, CDFs etc:

1. rand/randn - The rand/randn function generates arrays of random numbers whose elements are uniformly/normally distributed in the interval (0,1)/N(0,1).

2. randperm - p = randperm(n) returns a random permutation of the integers 1:n.

3. pdf - pdf('name',X,A1,A2,A3) returns a matrix of densities, where 'name' is a string containing the name of the distribution. X is a matrix of values, and A1, A2, and A3 are matrices of distribution parameters. Depending on the distribution, some of the parameters may not be necessary. Look up all the different PDFs that can be generated using MATLAB help. One such example is p = pdf('Normal',-2:2,0,1).

4. normpdf - normpdf(X,MU,SIGMA) computes the normal pdf at each of the values in X using the corresponding parameters in MU and SIGMA. X, MU, and SIGMA can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array with the same dimensions as the other inputs. The parameters in SIGMA must be positive.

5.mvnpdf - y= mvnpdf(X) returns the n-by-1 vector y, containing the probability density of the multivariate normal distribution with zero mean and identity covariance matrix, evaluated at each row of the n-by-d matrix X. Rows of X correspond to observations and columns correspond to variables or coordinates.

y = mvnpdf(X,mu) returns the density of the multivariate normal distribution with mean mu and identity covariance matrix, evaluated at each row of X. mu is a 1-by-d vector, or an n-by-d matrix. If mu is a matrix,

the density is evaluated for each row of X with the corresponding row of mu. mu can also be a scalar value, which mvnpdf replicates to match the size of X.

y = mvnpdf(X,mu,SIGMA) returns the density of the multivariate normal distribution with mean mu and covariance SIGMA, evaluated at each row of X. SIGMA is a d-by-d matrix, or an d-by-d-by-n array, in which case the density is evaluated for each row of X with the corresponding page of SIGMA, i.e., mvnpdf computes y(i) using X(i,:) and SIGMA(:,:,i). Specify [] for mu to use its default value when you want to specify only SIGMA. If X is a 1-by-d vector, mvnpdf replicates it to match the leading dimension of mu or the trailing dimension of SIGMA.

6. cdf - P = cdf('name',X,A1,A2,A3) returns a matrix of probabilities, where name is a string containing the name of the distribution, X is a matrix of values, and A, A2, and A3 are matrices of distribution parameters. Depending on the distribution, some of these parameters may not be necessary. Vector or matrix inputs for X, A1, A2, and A3 must have the same size, which is also the size of P. A scalar input for X, A1, A2, or A3 is expanded to a constant matrix with the same dimensions as the other inputs. Look up all the different PDFs that can be generated using MATLAB help. One such example is p = cdf('Normal',-2:2,0,1).

7. erf, erfc, erfinv, erfcinv - All related to computing area under the normal curve. See MATLAB help for syntax and how to use these functions.

Hopefully these notes served as a quick revision of concepts related to probability. More details can be found in the references mentioned.

References

[1] Useful Definitions and Results in Probability Theory - Notes By Prof. Vijaykumar Bhagavatula for Pattern Recognition

[2] Athanasios Papoulis, S. Unnikrishna Pillai, "Probability, Random Variables and Stochastic Processes," *TMH* 4^{th} *edition*, 2002

[3] Richard O. Duda, Peter E. Hart, David G. Stork, "Pattern Classification," *Wiley* 2^{nd} *edition*, 2007

[4] MATLAB Help