**Machine Learning**

**Session 2**

1. **Probability Theory**: A branch of mathematics concerned with the analysis of random phenomena.  
     
   Probability of an event occurring = (No. of ways it can occur) divided by (Total number of outcomes)

A probability space consists of three elements:

* 1. Sample space (Omega): set of all possible outcomes.
  2. Event space (A): an event being a set of outcomes in the sample space.
  3. Probability function (P): assigns each event in the event space a probability, which is a number between 0 and 1.

1. **Axioms of probability**:
   1. P(A) more than or equal to 0: The probability of an event happening is always more than or equal to zero.
   2. P(Omega) = 1: The probability that at least one of all the possible outcomes of a process (such as rolling a die) will occur is 1.
   3. P(A union B) = P(A) + P(B): If two events A and B are **mutually exclusive**, then the probability of either A or B occurring is the probability of A occurring plus the probability of B occurring. Mutually exclusive means that events A and B cannot both occur at once.
2. **Probability mathematics**:
   1. P(A union B) = P(A) + P(B) - P(A intersection B): *The sum rule*. The probability that an event in A or B will happen is the sum of the probability of an event in A and the probability of an event in B, minus the probability of an event that is in both A and B.
   2. P(A intersection B) = P(A) multiplied by P(B given A): The probability that B occurs given that A has occurred.

This formula says that we can multiply the probabilities of two events, but we need to take the first event into account when considering the probability of the second event.

If the events are independent, one happening doesn't impact the probability of the other, and in that case, P(B given A)= P(B).

1. **Conditional probability**: A measure of the probability of an event occurring, given that another event has already occurred. The mathematical formula is given as the probability an event that is in both A and B divided by the probability of an event in B.  
   P(A given B) = P(A intersection B) / (P(B))
   1. **Total probability law**: The rule states that if the probability of an event is unknown, it can be calculated using the known probabilities of several distinct events.  
        
      For example, there are three events: A, B, and C. Events B and C are distinct from each other while event A intersects with both events. We do not know the probability of event A. However, we know the probability of event A under condition B and the probability of event A under condition C.

The total probability rule states that by using the two conditional probabilities, we can find the probability of event A.

P(A) = summation of P(A union B) = summation of P(A given B) multiplied with P(B)

* 1. **Bayes Rule:** Bayes rule provides us with a way to update our beliefs based on the arrival of new, relevant pieces of evidence. For example, if we were trying to provide the probability that a given person has cancer, we would initially just say it is whatever percent of the population has cancer. However, given additional evidence such as the fact that the person is a smoker, we can update our probability, since the probability of having cancer is higher given that the person is a smoker. This allows us to utilize prior knowledge to improve our probability estimations.  
       
     P(A given B) = (P(B given A) multiplied with P(A)) / P(B)   
       
     In this formula, **A** is the event we want the probability of, and **B** is the new evidence that is related to A in some way.

**P(A given B)**is called the **posterior**; this is what we are trying to estimate. In the above example, this would be the “probability of having cancer given that the person is a smoker”.

**P(B given A)** is called the**likelihood**; this is the probability of observing the new evidence, given our initial hypothesis. In the above example, this would be the “probability of being a smoker given that the person has cancer”.

**P(A)** is called the **prior**; this is the probability of our hypothesis without any additional prior information. In the above example, this would be the “probability of having cancer”.

**P(B)** is called the **marginal likelihood**; this is the total probability of observing the evidence. In the above example, this would be the “probability of being a smoker”. In many applications of Bayes Rule, this is ignored, as it mainly serves as normalization.  
<https://www.investopedia.com/terms/b/bayes-theorem.asp>

1. **Independence**: When one event has no influence on the probability of another.  
     
   Event *A* and *B* are independent if P(*A* given *B*) = P(*A*) or, alternatively, P(*A* intersect *B*) = P(*A*) times P(*B*).  
     
   If A1, A­2, A3 are mutually independent, then:  
     
   P(A1 intersect A­2 intersect A3) = P(A1) times P(A­2) times P(A3)  
     
   Pair-wise independence is necessary but not sufficient for mutual independence.  
     
   *Properties*:
   1. If A and B are independent:
      1. A complement C and Bcomplement C are mutually independent
      2. A complement C and B or A and A complement C are independent
   2. If A, B and C are mutually independent then:
      1. A and B intersect C are independent
      2. A and B union C are independent

*Example*:

Tossing two fair coins (elements of sample space are ordered pairs):

State space (omega) = {(Head, Head), (Head, Tail), (Tail, Head), (Tail, Tail)}

By assigning ‘classical’ probabilities:

P({(Head, Head)}) = 1/4 & P({(Head, Tail)}) = 1/4 , etc.

With these probabilities, coin 1 / coin 2 events are independent:

P(coin 1 = Head given coin 2 = Head) = P({(Head, Head)}) divided by P(coin 2 = Head) = 1/2 = P(coin 1 = Head)

(The same holds for all other combinations of Head and Tail.)

Hence the experiments are independent.

In general, relative frequency view of repeated experiments only holds if the experiments are

independent.

1. **Cumulative Distribution Function (cdf)**:   
     
   A random variable (RV), X, is a measurable function from a set of possible outcomes to a measurable space.  
     
   Discrete random variables take on a countable number of distinct values. Consider an experiment where a coin is tossed three times. If X represents the number of times that the coin comes up heads, then X is a discrete random variable that can only have the values 0, 1, 2, 3 (from no heads in three successive coin tosses to all heads). No other value is possible for X.  
     
   P(x) = P(X = x)  
     
   P(X = x) refers to **the probability that the random variable X is equal to a particular value**, denoted by x. As an example, P(X = 1) refers to the probability that the random variable X is equal to 1.  
     
   Continuous random variables can represent any value within a specified range or interval and can take on an infinite number of possible values (called cdf). An example of a continuous random variable would be an experiment that involves measuring the amount of rainfall in a city over a year or the average height of a random group of 25 people.  
     
   P(x) = the derivative of P(X less than or equal to x)
2. **Sample mean**: Measure of the average of all samples. Calculated by summing up the samples and dividing by no. of total samples.
3. **Expected value** (EV): The expected value is calculated by multiplying each of the possible outcomes by the likelihood each outcome will occur and then summing all those values.   
     
   EV for discrete RV’s = the summation of P(X\_i) multipled with X\_i, where i represents the iteration number.  
     
   EV for continuous RV’s = the integral of P(X\_i ) multiplied with (X\_i ) ranging from positive to negative infinity and with respect to x.  
     
   The EV of a random variable gives a measure of the centre of the distribution of the variable. Essentially, the EV is the long-term average value of the variable.
4. **Law of Large Numbers**: States that an observed sample average from a large sample will be close to the true population average and that it will get closer the larger the sample.

The law of large numbers does not guarantee that a given sample, especially a small sample, will reflect the true population characteristics or that a sample which does not reflect the true population will be balanced by a subsequent sample.

1. **Variance**: Variance is a measure of dispersion of data points from the mean. Low variance indicates that data points are generally similar and do not vary widely from the mean. High variance indicates that data values have greater variability and are more widely dispersed from the mean.  
     
   To calculate the variance follow these steps:
   1. Work out the Mean (the simple average of the numbers)
   2. Then for each number: subtract the Mean and square the result (the *squared difference*).
   3. Then work out the average of those squared differences.
2. **Standard deviation**: A measure of how spread-out numbers are.  
     
   It is the square root of variance.
3. **Skewness**: Measures the lopsidedness of a distribution and based upon polynomial expectations, (E{X3}, E{X4}, . . .), e.g.:

Skewness = E{( (sample X – mean) / variance )}

1. **Kurtosis**: Measures how much of the probability mass lies within the tails. Positive kurtosis is also termed ‘heavy-tailed’. ‘Heavy’ with respect to a Gaussian: hence the −3.  
     
   Kurtosis = E{( (sample X – mean)4 / variance )} – 3
2. **Gaussian Distribution**: Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.
3. **Multiple Random Variables (RVs)**: Extend the idea of RVs to two or more dimensions,   
     
   e.g., RVs X (age) and Y (wealth)   
     
   RVs X and Y have respective distributions pX(x) and pY(y) but this does not tell us about the relation between X and Y.
4. **Joint probability distribution**:   
     
   Px,y(x, y) = the probability of any joint area / delta X times delta Y  
     
   Also see PDF in same folder for further explanation.
5. **Marginal density**:   
     
   Defined as the integral from + infinity to – infinity of Pxy(x, y) w.r.t dy  
     
   So: Area of slice through pXY (x, y) at X = x is equal to pX(x).  
     
   Also see PDF in same folder for further explanation.
6. **Conditional densities**:  
     
   PY given X (y given x) = PXY(x, y) / f­X(x)
7. **Independent RVs**:   
     
   Events A and B are independent if P(A, B) = P(A) times P(B).   
     
   We say the RVs X and Y are independent RVs if:   
     
   fXY (x, y) = fX(x) times fY(y)   
     
   So the joint density has a factorial (or product) form.
8. **Covariance:** When two random variables, X and Y, are defined on a probability space, it is useful to describe how they vary together. A common measure of the relationship between the two random variables is the covariance.  
     
   The covariance of two RVs X and Y is:  
     
   Cov(X, Y ) ≡ sigma\_{XY} (variance of XY) = E{(X – mean\_X )(Y – mean\_Y )} = E{XY} − m\_X times m\_Y
9. **Correlations**: refers to the degree to which a pair of variables are linearly related.  
     
   PXY = Cov(X, Y) / varianceX times varianceY  
     
   NOTE: Covariance and correlation will have the same sign (positive or negative).
10. **Independent vs Uncorrelated RVs**:   
      
    If X and Y are independent scalar RVs, we have:   
      
    PXY (x, y) = PX(x) times PY(y)  
      
    Hence, E{XY} = E{X} times E{Y } = mX times mY.  
      
    Hence, Cov(X, Y) = E{XY} − mXtimesmY = 0 and so also PXY = 0  
      
    Therefore: Independent implies Uncorrelated.  
      
    But: Uncorrelated does not imply Independent.  
      
    Covariance Cov(X, Y) only measures 2nd order statistics. Independence is much stronger: all nonlinear statistics must be unrelated.
11. **Covariance Matrix**: A multivariate generalization of the definition of variance for a scalar random variable. Variance and covariance are often displayed together in a variance-covariance matrix, (aka, a covariance matrix). The variances appear along the diagonal and covariances appear in the off-diagonal elements.  
      
    The equivalent of multi-dimensional variance is more complex.   
      
    There are several ways of generalising E(X2) to N-dimensions.   
      
    To capture the full generality we define the covariance matrix:   
      
    Cov(vector X) = E{(vector X – mean of vector X )(vector X – mean of vector X) all raised to the power of T}   
      
    Note that Cov(vector X) is an N × N matrix.   
      
    For two N-dimensional RVs vector X and vector Y, we can define the cross-covariance:   
      
    Cov(vector X , vector Y) = E{( vector X – mean of vector X )( vector Y – mean of vector Y ) all raised to the power of T}
12. **N-dimensional Gaussian RVs**:   
      
    Let,   
      
    a = determinant of covariance of vector X raised to the power of -0.5  
    b = 2 times pi raised to the power N/2  
    c = -0.5 times x – mean of vector X, all raised to the power T  
    d = inverse of the covariance of vector X times x – mean of vector X  
      
    Therefore, the Gaussian distribution has a natural N-dimensional form:  
      
    pX(x) = (a/b) times exponent of c times d.