**MMD**

Maximum mean discrepancy (MMD) is a statistic to quantify the mean discrepancy of two data distributions in kernel space in order to determine if two samples are drawn from different distributions [1]. Let and be two independent probability distributions, and (shorthand notation for ) denotes the mathematical expectation of with under the probability density . The statistic definition between and is:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | (1) | (1) |

where the function class is a unit ball in the reproducing Hilbert space (RKHS) and is the mean embedding of and respectively i.e., the mean of the feature mapping in the kernel space. The function class is universal meaning that if and only if . Therefore, MMD is the largest difference in expectations over functions in and can only be zero if the two samples were drawn from the same distribution.

In practise, we use the square MMD in order to be able to use kernel functions. Let and denote the independent and identically distributed (i.i.d.) samples from distribution and respectively. An unbiased estimation of can be obtained using U-statistic:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | (2) | (1) |

Under the null hypothesis, the test statistic is expected to be close to 0 and the value of the estimate should be smaller than the test bounds given by the kernel two sample test.

**OCSVM**

One-class support vector machines (OCSVMs) is a one-class classification technique, which aims to classify instances into one of two classes, the inlier and outlier class. The method, first presented by Schölkopf et. al [2], utilizes a training data set with normal data to learn the boundaries of the normal data points so that data points which lie outside of the region to be classified as outliers. OCSVMs utilize an implicit transformation function defined by the kernel to project data to a higher dimensional space. The algorithm learns the decision boundary (a hyperplane) which achieves the maximum separation of the majority of the data points from their origin. Only a small fraction of data points are allowed to lie on the other side of the decision boundary and those data are considered outliers.

The OCSVM algorithm returns a function that takes the value +1 for the normal region and -1 elsewhere. Hence, function is called a decision function and is defined as:

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

where is the vector perpendicular to the decision boundary and is the bias term. The decision boundary is defined as:

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

Given that the distance of any arbitrary data point to the decision boundary can be calculated with the following:

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

and the fact that the origin's value when plugged to is , the distance of the origin to the decision boundary is . The OCSVM algorithm essentially attempts to maximise the distance by solving the minimisation problem of .

Formally the primary objective of OCSVM is defined by the following equation:

|  |  |  |
| --- | --- | --- |
|  |  | (6) |
|  | subject to |  |

where is the slack variable for a point which allows it to lie on the other side of the decision boundary, is the size of the training dataset and is the regularization parameter. As shown in Equation 6 the objective is not only to minimise the distance of the origin to the decision boundary but also minimise the slack variables for all points. The regularization parameter represents the upper bound limit of the fraction of outliers and a lower bound on the number of support vectors. In other words, specifies the number of training points which are guaranteed to be misclassified and the number of training examples being support vectors. As mentioned above and therefore a percentage, where a high value may lead to over-fitting and a low value to under-fitting. The value controls the trade-off between and .

In order to reduce the number of variables to a single vector and utilise the kernel trick, the primary objective is transformed into a dual objective:

|  |  |  |
| --- | --- | --- |
|  |  | (7) |
|  | subject to: |  |

where is the kernel matrix and the Lagrange multipliers. In our research we utilised the Gaussian radial basis kernel function (RBF kernel), a special case of RBF when data are centred around the origin. The Gaussian RBF kernel is defined by:

|  |  |  |
| --- | --- | --- |
|  |  | (7) |

where is a kernel parameter and is a dissimilarity measure such as the square Euclidean distance.

Using the above, the decision function now becomes:

|  |  |  |
| --- | --- | --- |
|  |  | (7) |

[1] A. Gretton, “A Kernel Two-Sample Test,” vol. 13, pp. 723–773, 2012.

[2] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Piatt, “Support vector method for novelty detection,” *Adv. Neural Inf. Process. Syst.*, no. May 2014, pp. 582–588, 2000.