S21CSE471: Statistical Methods in AI **Spring 2021**

KMPP, SVM Kernels and GMM Fundamentals

Prepared by: Devershi, Balavarun & Jagaran

This scribe is primarily aimed at discussing three things.

- KMPP clustering
- SVM Kernels
- Fundamental Statistics required to understand the Guassian Mixture Model.

1 K Means ++

1.1 Introduction

The K means ++ algorithm was developed in 2007 by David Arthur and Sergei Vassilvitskii, as an approximation algorithm for the NP-hard k-means problem—a way of avoiding the sometimes poor clusterings found by the standard k-means algorithm.

One disadvantage of the K-means algorithm is that it is sensitive to the initialization of the centroids or the mean points. So, if a centroid is initialized to be a "far-off" point, it might just end up with no points associated with it, and at the same time, more than one cluster might end up linked with a single centroid. Similarly, more than one centroids might be initialized into the same cluster resulting in poor clustering

1.2 Algorithm

The intuition behind this approach is that spreading out the k initial cluster centers is a good thing: the first cluster center is chosen uniformly at random from the data points that are being clustered.

After which each subsequent cluster center is chosen from the remaining data points with probability proportional to its squared distance from the point's closest existing cluster center.

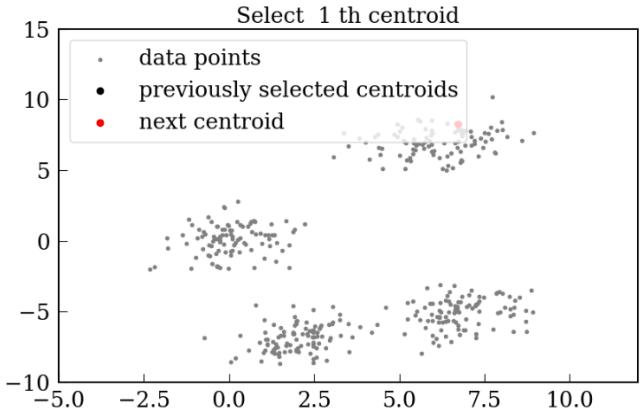
The steps to initialize the centroids using K-Means++ are:

- The first cluster is chosen uniformly at random from the data points that we want to cluster. This is similar to what we do in K-Means, but instead of randomly picking all the centroids, we just pick one centroid here
- Next, we compute the distance $D(x)$ of each data point x from the cluster center that has already been chosen
- Then, choose the new cluster center from the data points with the probability of x being proportional to $D(x)^2$
- We then repeat steps 2 and 3 until k clusters have been chosen

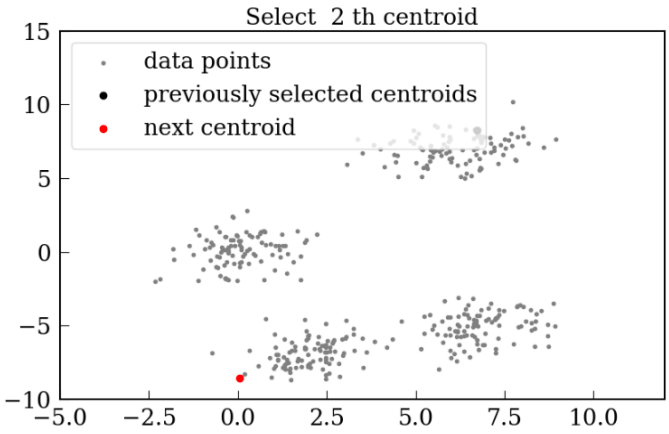
1.3 Implementation

An implementation of K means ++ can be found here

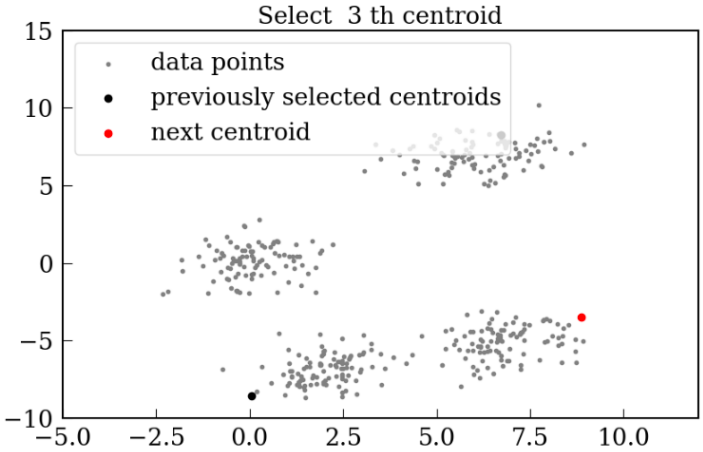
The first step is to take a random point as a cluster centroid



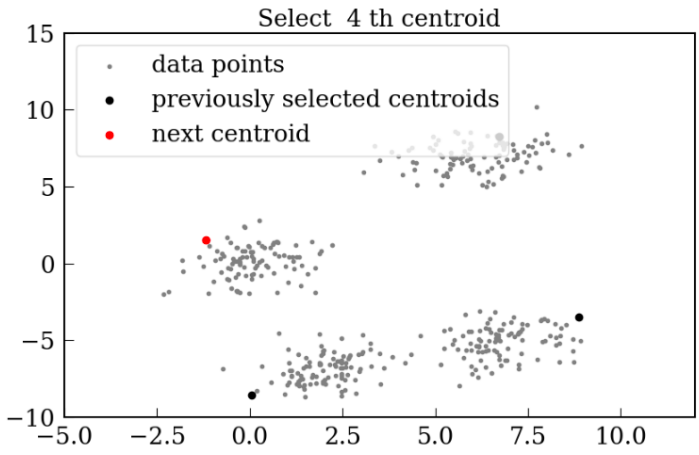
The next centroid will be the one whose squared distance $D(x)^2$ is the farthest from the current centroid:



Continuing the process,



The next step would be,



2 SVM Kernels

SVM algorithms utilize a bunch of numerical functions that are characterized as the kernel. The function of kernel is to accept information as input and change it into the necessary structure. Diverse SVM algorithms utilize various kinds of kernel functions. These functions can be various sorts. For instance linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid.

Basically, It returns the inner product between two points in a standard feature dimension.

2.1 Types of SVM Kernels

- Standard Kernel Function

$$K(\bar{x}) = 1, if||\bar{x}|| <= 1$$
$$K(\bar{x}) = 0, Otherwise$$

Figure 1:

- Guassian Kernel Function: It is used to perform transformation, when there is no prior knowledge about data.

$$K(x, y) = e^{-}(\frac{||x-y||^2}{2\sigma^2})$$

Figure 2:

- Gaussian Kernel Radial Basis Function (RBF) : Same as above kernel function, adding radial basis method to improve the transformation. (Shown in figure 3)
- Sigmoid Kernel: this function is equivalent to a two-layer, perceptron model of neural network, which is used as activation function for artificial neurons. (Shown in figure 4)
- Polynomial Kernel: It represents the similarity of vectors in training set of data in a feature space over polynomials of the original variables used in kernel.

$$K(x,y) = e^{-}(\gamma||x-y||^2)$$
$$K(x,x1) + K(x,x2)(Simplified - Formula)$$
$$K(x,x1) + K(x,x2) > 0(Green)$$
$$K(x,x1) + K(x,x2) = 0(Red)$$

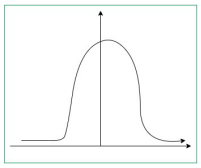


Figure 3: RBF formula and graph

$$K(x,y) = tanh(\gamma.x^T y + r)$$

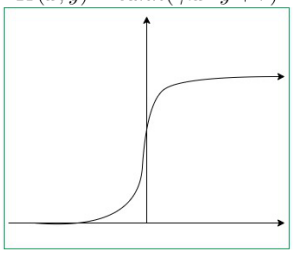


Figure 4: Sigmoid formula and graph

$$K(x,y) = tanh(\gamma.x^T y + r)^d, \gamma > 0$$

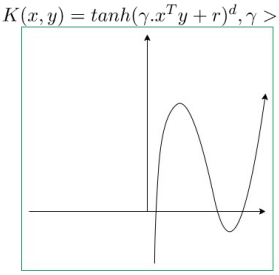


Figure 5: Polynomial formula and graph

2.2 Choosing the correct Kernel

Given a dataset, if you don't know what kernel to choose, I recommend to start with simpler kernels and go to more complex models.So, the linear kernel works fine if your dataset if linearly separable; however, if your dataset isn't linearly separable, a linear kernel isn't going to cut it.

For simplicity (and visualization purposes), let's assume our dataset consists of 2 dimensions only. Below, I plotted the decision regions of a linear SVM on 2 features of the iris dataset:

Now compare it with IRIS dataset on RBF kernel

Now, it looks like both linear and RBF kernel SVM would work equally well on this dataset. So, why prefer the simpler, linear hypothesis?

Linear SVM is a parametric model, an RBF kernel SVM isn't, and the complexity of the latter grows with the size of the training set. Not only is it more expensive to train an RBF kernel SVM, but you also have to keep the kernel matrix around, and the projection into this "infinite" higher dimensional space where the data becomes linearly separable is more expensive as well during prediction. Furthermore, you have more hyperparameters to tune, so model selection is more expensive as well! And finally, it's much easier to overfit a complex model.

In this case, a RBF kernel would make so much more sense.

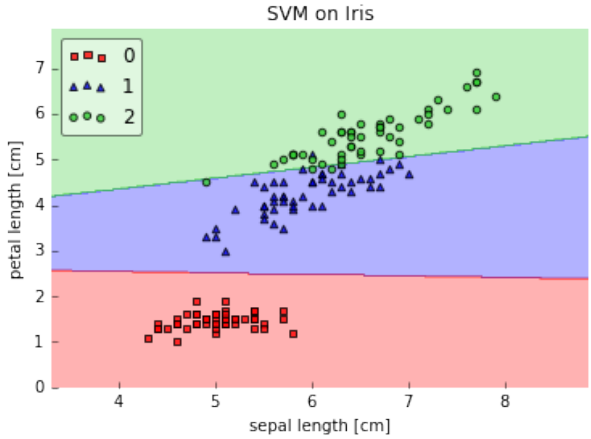


Figure 6: Iris dataset under Linear SVM

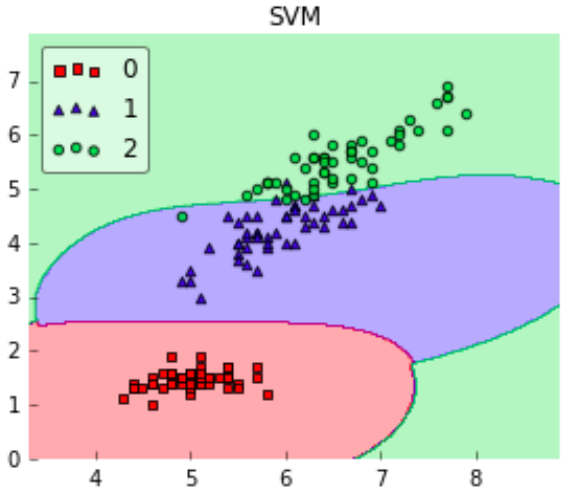


Figure 7: Iris dataset under RBF SVM

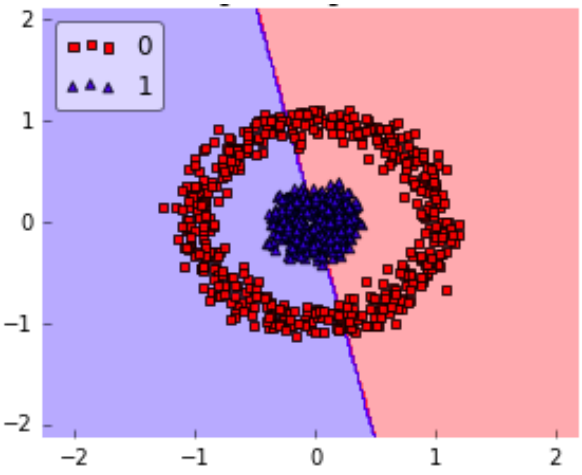


Figure 8: Non Linear dataset under Linear SVM

In any case, I wouldn't bother too much about the polynomial kernel. In practice, it is less useful for efficiency (computational as well as predictive) performance reasons. So, the rule of thumb is: use linear SVMs (or logistic regression) for linear problems, and nonlinear kernels such as the Radial Basis Function

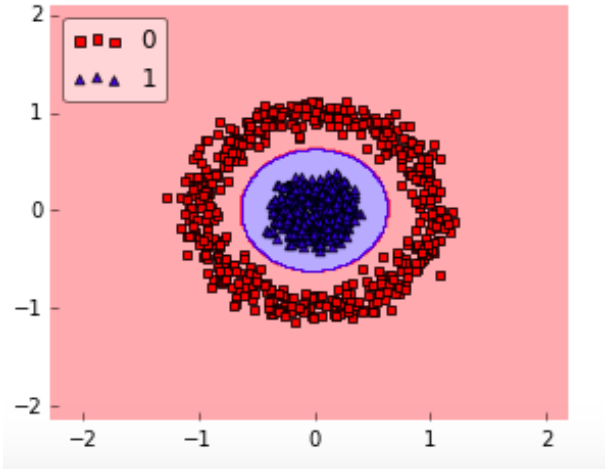


Figure 9: Non Linear dataset under RBF SVM

kernel for non-linear problems.

The RBF kernel SVM decision region is actually also a linear decision region. What RBF kernel SVM actually does is to create non-linear combinations of your features to uplift your samples onto a higher-dimensional feature space where you can use a linear decision boundary to separate your classes.

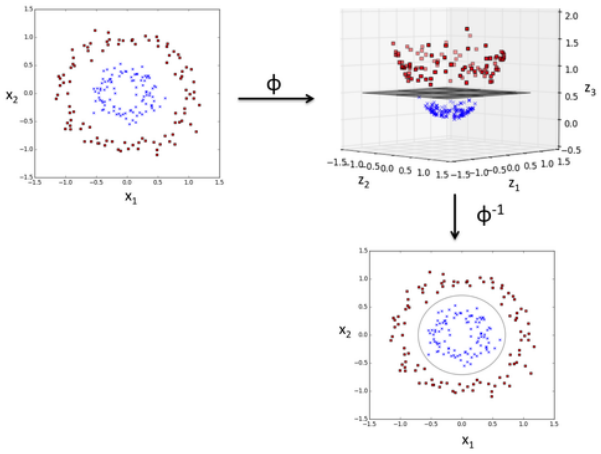


Figure 10: RBF kernel compressing non linear lower dimensional data to linear higher dimensional data.

3 Mathematics in Machine Learning

The purpose of this scribe is to discuss the various mathematical concepts introduced during our lecture. The topics that we would be discussing are as follows:-

1. Mean
2. Variance
3. Co variance
4. Expectation
5. Gaussian Distribution
6. Sampling distribution

The remaining portion of the document will discuss these topics in detail.

3.1 Mean

Mean is one of the measures of central tendency. It along with the other two measures namely median and mode tries to describe a set of data based on central position.

Mean gives us the average value of the dataset. Mathematically, it is defined as :

$$\bar{X} = \frac{1}{n}\Sigma x_i \tag{1}$$

Here, $\bar{X}$ represents the mean of a dataset whose values are signified by x_i. i ranges from 1 to n.

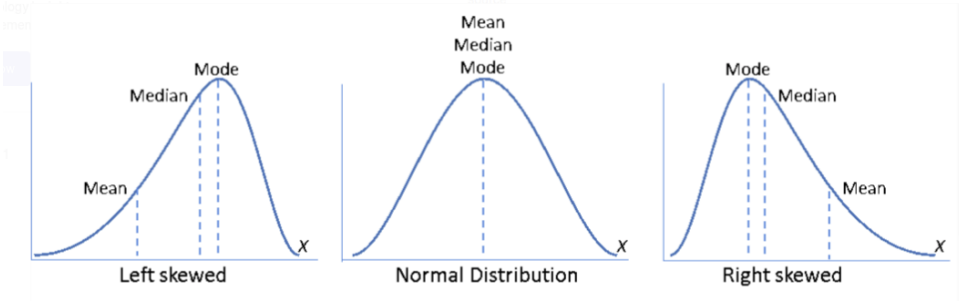


Figure 11: The three measures of central tendency Image source [1]

3.2 Variance

Variance tells us about the spread of the dataset with respect to its mean. To calculate the variance of a dataset we first need to find the derivation(difference) of each data point with respect to its mean.

There are two formulae to calculate variance.

3.2.1 Variance from entire population

When we calculate the variance by considering each and every data in the population we use the following formula : -

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N} \tag{2}$$

Here, σ^2 is the variance , μ is the mean. We consider the population to have N datapoints.

3.2.2 Variance from sample

When we consider only a sample of the dataset present, the variance should be less than the actual variance. Likewise, we update the formula to be as follows :-

$$S^2 = \frac{\Sigma(X - \bar{x})^2}{n - 1} \tag{3}$$

Unlike calculating the variance from entire population, when we are calculating the variance from a sample we choose n-1 , where n is the same size. $\bar{x}$ is the mean of the sample.

The unit of variance is much larger than that of the dataset, hence usually standard deviation which is the square root of variance is used.

3.3 Co variance

In the topics until now we have discussed the statistical methods used on data from a single distribution/dataset. Covariance tells us how the data from two different distribution vary with respect to each other.

Mathematically,

$$Cov(X, Y) = \frac{\Sigma((X - \mu)(Y - \nu))}{n - 1} \tag{4}$$

Here, μ and ν are the mean/expected values of the two distribution X and Y respectively.

Intuitively, X and Y can vary in the same direction together i.e. an increase in X sees a likewise increase in Y, can vary in opposite direction i.e. an increase in X sees a decrease in the value of corresponding Y value or may not vary in any such patterns. Hence, co variance can be positive (suggesting they move in same direction) , negative or zero.

In Figure 12 we can see the positive and negative covariance relationship between two distribution. In case of zero covariance, no such distinctive pattern would be noticed.

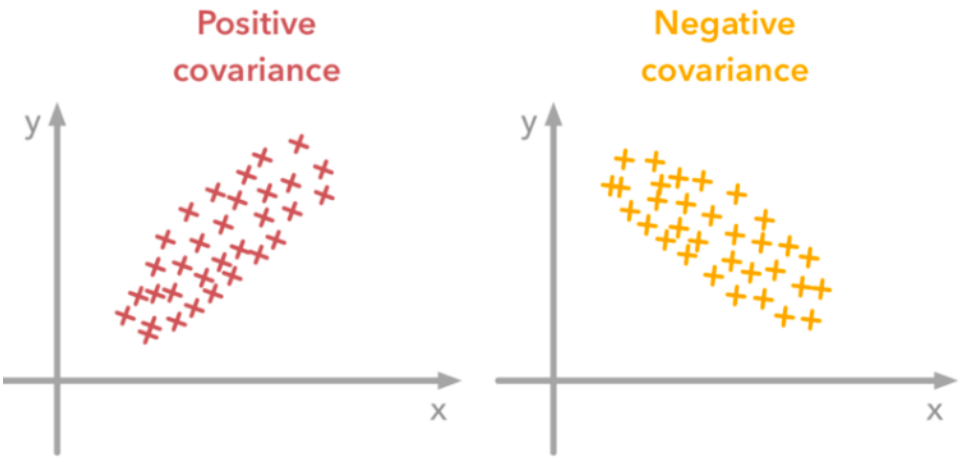


Figure 12: Positive and Negative covarianceImage source KDNuggets Tutorial

3.4 Expectation

The expectation of a random variable X is defined as the weighted average of a large number of independent realisations of X. It is denoted by $E(X)$.

For discrete random variables, it is defined as :

$$E(X) = \Sigma x_i.p(x_i) \tag{5}$$

For continuous random variables, it is defined as :

$$E(X) = \int_a^b xf(x)\ dx \tag{6}$$

The expected value of a random variable where the probabilities of all the outcomes are equal is given by the mean of the outcomes.

3.5 Gaussian Distribution

Gaussian distribution, also known as the Normal distribution, is a continuous 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.

The probability density function for the normal distribution is defined as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{7}$$

where μ is the mean or expectation of the distribution and σ is the standard deviation.

3.6 Sampling distribution

In statistics, a sampling distribution or finite-sample distribution is the probability distribution of a given random-sample-based statistic. If an arbitrarily large number of samples, each involving multiple observations (data points), were separately used in order to compute one value of a statistic (such as, for example, the sample mean or sample variance) for each sample, then the sampling distribution is the probability distribution of the values that the statistic takes on.

References

[1] https://blog.minitab.com/en/quality-business/common-assumptions-about-data-part-2-normality-and-equal-variance

[2] https://towardsdatascience.com/

[3] https://www.kdnuggets.com/

[4] https://en.wikipedia.org/

[5] https://www.geeksforgeeks.org