# DMW C2 Assignment-3

Shubham Kumar IIT2018115
Raktim Bijoypuri IIT2018125
Aditya Kamble IIT2018126
Aakashdeep IIT2018128
Tejas Mane IIT2018135

*VI Semester, Department of Information Technology,*

*Indian Institute of Information Technology, Allahabad, Prayagraj.*

## Introduction:

SVM is a very widely used supervised learning algorithm that can be used with relatively less data, also it is very fast as compared to other supervised learning algorithms.
However one huge limitation of SVM is that assumes that the data used to train the SVM must be independently and identically distributed that is an IID, however this has huge limitations because many of the real world datasets don't follow this assumption, moreover even if they are identically distributed then also independent is a very strong condition which is not followed most of the times. So this paper tries to focus only on cases where the data is actually a markov chain, this is because many real world examples and situations follow markov chains such as DNA sequences, Natural language etc.

The paper goes on to show that the SVM classifier along with markov sampling achieves good learning performance with real world data sets.

## Algorithm:

**Step 1 :** Take out random samples from the data and construct a base model using support vector machines., let P=0 and N=0(the number of positive and negative samples).

**Step 2 :** Draw 1 sample from the dataset randomly = X1 and if the number of samples is even and if X1 is a positive sample increase P otherwise increase N.

**Step 3 :** Draw another sample from the dataset randomly = X2 if X2 is a positive sample increase P otherwise increase N.

**Step 4 :** Using the base model, calculate the ratio of $e^{-L(x1)} / e^{-L(x2)}$

$-Y_{x1} -Y_{x2}$

**Step 5 :** if ratio = 1 and $Y -$ and accept X2 with probability similarly $_{x1} = 1$ $Y_{x2} = -1$ $e\,/e$ using the rules described in the paper accept X2 with probabilities calculated accordingly increase P and N that is the number of positive and negative samples.
**Step 6 :** If the number of positive samples is less than total/2 Or the number of negative samples is less than total/2 then continue to step 3.

The paper goes to use the algorithm on 10 real world datasets which shows good results and successfully extends the usual assumption of independent and identically distributed data to markov chains.

## Output for Letter Dataset:

| Kernel | Accuracy |
|---|---|
| linear | 89.44056362545876 |
| rbf | 94.08435900153268 |
| poly | 94.78161784444326 |
| chi_square | 96.36111253879433 |
| hellinger | 87.3793509307043 |