# Data Mining
# Assignment-1

**Group Members :-**

1. Prabal (IIT2018140)
2. Nikhil Kumar (IIT2018152)
3. Sagar Kumar (IIT2018154)
4. Kartik Nema (IIT2018156)
5. Bhupendra (IIT2018163)
6. Prakhar Srivastava (IIT2018172)

## k-Times Markov Sampling for SVMC

**Introduction :-** This paper aims to solve the problem of reducing the training time while using Markov sampling, while keeping the classification rates the same.
To solve this problem a SVMC algorithm is proposed based on the k-times Markov Sampling.
On comparing the performance of SVMC based on k times Markov Sampling and classical SVMC and the SVMC based on Markov sampling, it was found that the SVMC with k-times (k = 1, 2), Markov sampling has the following advantages over the others :-

1. The misclassification rates are smaller;
2. The total time of sampling
and training is less; and
3. The obtained classifiers are more sparse.

**Keywords:-** SVM (support vector machine), SVMC (support vector machine classifiers), Markov Sampling, Balanced samples, unbalanced samples.

SVM :- SVM are supervised machine learning models that use classification algorithms for two-group classification problems.
SVMC :- SVMC has a good theoretical property in universal consistency and learning rates.
Markov Sampling :- It creates samples from a continuous random variable, with probability density proportional to a known function.

Balanced samples :- In data set if positive values are approximately the same as negative values then it is called balanced samples(dataset).
Unbalanced samples :- In a data set if there is a very high difference between the positive values and negative values, then it is called unbalanced samples.

**Algorithm :-**

1. Algorithm - I (for balanced samples)

**Input**: $S_T$, $N$, $k$, $q$, $n_2$
**Output**: $sign(f_k)$

**1:** Draw randomly $N$ samples $S_{iid} := \{z_j\}_{j=1}^{N}$ from $S_T$. Train $S_{iid}$ by SVMC and obtain a preliminary learning model $f_0$. Let $i = 0$.

**2:** Let $N_+ = 0$, $N_- = 0$, $t = 1$.

**3:** Draw randomly a sample $z_t$ from $S_T$, called it the current sample. Let $N_+ = N_+ + 1$ if the label of $z_t$ is $+1$, or let $N_- = N_- + 1$ if the label of $z_t$ is $-1$.

**4:** Draw randomly another sample $z_*$ from $S_T$, called it the candidate sample, and calculate the ratio $\alpha$, $\alpha = e^{-\ell(f_i, z_*)} / e^{-\ell(f_i, z_t)}$.

**5:** If $\alpha \geq 1$, $y_t y_* = 1$ accept $z_*$ with probability $\alpha_1 = e^{-y_* f_i} / e^{-y_t f_i}$. If $\alpha = 1$ and $y_t y_* = -1$ or $\alpha < 1$, accept $z_*$ with probability $\alpha$. If there are $n_2$ candidate samples can not be accepted continually, then set $\alpha_2 = q\alpha$ and accept $z_*$ with probability $\alpha_2$. If $z_*$ is not accepted, go to Step 4, else let $z_{t+1} = z_*$, $N_+ = N_+ + 1$ if the label of $z_{t+1}$ is $+1$ and $N_+ < N/2$, or let $z_{t+1} = z_*$, $N_- = N_- + 1$ if the label of $z_{t+1}$ is $-1$ and $N_- < N/2$ (if the value $\alpha$ (or $\alpha_1$, $\alpha_2$) is bigger than 1, accept the candidate sample $z_*$ with probability 1).

**6:** If $N_+ + N_- < N$, return to Step 4, else we obtain $N$ Markov chain samples $S_{Mar}$. Let $i = i + 1$. Train $S_{Mar}$ by SVMC and obtain a learning model $f_i$.

**7:** If $i < k$, go to Step 2, else output $sign(f_k)$.

2. Algorithm - II (for unbalanced samples)

**Input**: $S_T$, $N$, $k$, $q$, $n_2$
**Output**: $sign(f_k)$

**1:** Draw randomly $N$ samples $S_{iid} := \{z_j\}_{j=1}^{N}$ from $S_T$. Train $S_{iid}$ by SVMC and obtain a preliminary learning model $f_0$. Let $i = 0$.

**2:** Let $N_i = 0$, $t = 1$.

**3:** Draw randomly a sample $z_t$ from $S_T$, called it the current sample. Let $N_i = N_i + 1$.

**4:** Draw randomly another sample $z_*$ from $S_T$, called it the candidate sample. Calculate the ratio $\alpha$, $\alpha = e^{-\ell(f_i, z_*)} / e^{-\ell(f_i, z_t)}$.

**5:** If $\alpha = 1$, $y_t y_* = 1$ accept $z_*$ with probability $\alpha_1 = e^{-y_* f_i} / e^{-y_t f_i}$. If $\alpha = 1$ and $y_t y_* = -1$ or $\alpha < 1$, accept $z_*$ with probability $\alpha$. If there are $n_2$ candidate samples can not be accepted continually, then set $\alpha_2 = q\alpha$ and accept $z_*$ with probability $\alpha_2$. If $z_*$ is not accepted, go to Step 4, else let $z_{t+1} = z_*$, $N_i = N_i + 1$ (if $\alpha$ (or $\alpha_1$, $\alpha_2$) is greater than 1, accept $z_*$ with probability 1).

**6:** If $N_i < N$, return to Step 4, else we obtain $N$ Markov chain samples $S_{Mar}$. Let $i = i + 1$. Train $S_{Mar}$ by SVMC and obtain a learning model $f_i$.

**7:** If $i < k$, go to Step 2, else output $sign(f_k)$.

**Results :-**

The following table presents the accuracy of the k-Times Markov Sampling for SVMC with different Kernels for **letters** dataset.

| Kernel | KPCA | SVDD | OCSVM | OCSSVM | OCSSVM with smo | KT_SVM |
|---|---|---|---|---|---|---|
| Linear | 0.02 | 0.09 | 0.01 | 0.07 | 0.04 | 0.86 |
| RBF | 0.05 | 0.07 | 0.14 | 0.09 | 0.04 | 0.94 |
| chi_square | 0.18 | 0.0 | 0.02 | 0.18 | 0.17 | 0.89 |

**Conclusion :-** This paper proposes SVMC with k times sampling and presents 2 algorithms (one for balanced, one for unbalanced samples). The results show that SVMC with k-times (k = 1, 2) Markov sampling outperforms classical SVMC and the SVMC with Markov sampling in terms of learning efficiency (misclassification speeds, total time of sampling and preparation, and support vector numbers).