Data Mining
Assignment-3

Group Members :-

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

The Generalization Ability of SVM Classification Based on Markov Sampling

Introduction :- The authors study the generalization ability of SVMC based on uniformly ergodic Markov chain (u.e.M.c.) samples. They analyze the excess misclassification error of SVMC based on u.e.M.c. samples, and obtain the optimal learning rate of SVMC for u.e.M.c. samples. They have presented a new Markow Sampling Algorithm for SVMC to generate u.e.M.C samples and also have presented its numerical studies.

Keywords :- SVM (support vector machine), SVMC (support vector machine classifiers), Markov Sampling, Markov Chain.

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.

Algorithm :-

The algorithm proposed by the authors is mentioned below :-

Step 1: Let m be the size of training samples and $m\%2$ be the remainder of m divided by 2. m_+ and m_- denote the size of training samples which label are $+1$ and -1, respectively. Draw randomly $N_1 (N_1 \leq m)$ training samples $\{z_i\}_{i=1}^{N_1}$ from the dataset D_{tr}. Then we can obtain a preliminary learning model f_0 by SVMC and these samples. Set $m_+ = 0$ and $m_- = 0$.

Step 2: Draw randomly a sample from D_{tr} and denote it the current sample z_t. If $m\%2 = 0$, set $m_+ = m_+ + 1$ if the label of z_t is $+1$, or set $m_- = m_- + 1$ if the label of z_t is -1.

Step 3: Draw randomly another sample from D_{tr} and denote it the candidate sample z_*.

Step 4: Calculate the ratio P of $e^{-\ell(f_0,z)}$ at the sample z_* and the sample z_t, $P = e^{-\ell(f_0,z_*)}/e^{-\ell(f_0,z_t)}$.

Step 5: If $P = 1$, $y_t = -1$ and $y_* = -1$ accept z_* with probability $P' = e^{-y_* f_0}/e^{-y_t f_0}$. If $P = 1$, $y_t = 1$ and $y_* = 1$ accept z_* with probability $P' = e^{-y_* f_0}/e^{-y_t f_0}$. If $P = 1$ and $y_t y_* = -1$ or $P < 1$, accept z_* with probability P. If there are k candidate samples z_* can not be accepted continuously, then set $P'' = qP$ and with probability P'' accept z_*. Set $z_{t+1} = z_*$, $m_+ = m_+ + 1$ if the label of z_t is $+1$, or set $m_- = m_- + 1$ if the label of z_t is -1 [if the accepted probability P'(or P'', P) is larger than 1, accept z_* with probability 1].

Step 6: If $m_+ < m/2$ or $m_- < m/2$ then return to Step 3, else stop it.

Results :-

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

Kernel	KPCA	SVDD	OCSVM	OCSSVM	OCSSVM with smo	FS_SVM
Linear	0.02	0.09	0.01	0.07	0.04	0.85
RBF	0.05	0.07	0.14	0.09	0.04	0.95
chi_square	0.18	0.0	0.02	0.18	0.17	0.88

Conclusion :- To study the generalization performance of SVMC based on u.e.M.c. samples, the authors first establish two new concentration inequalities for u.e.M.c. samples, then we analyze the excess misclassification error of SVMC with u.e.M.c. samples, and obtain the optimal learning rate for SVMC with u.e.M.c. samples.