INDIAN INSTITUTE OF INFORMATION TECHNOLOGY ALLAHABAD

Subject – Data Mining and Warehousing

Topic – Generalization ability of SVM classification based on Markov Sampling

Report By -

IIT2018108 – Ravi Kumar Sharma

Assignment :

You have to understand the algorithm proposed in the paper "Generalization ability of SVM classification based on Markov Sampling ''.

Run the algorithm on the shared pascal and Letter dataset and show the accuracy in terms of the attached image table: (make one more column in the last name MS_SVM with the new algorithm and give the result.

Try to run the algorithm using different kernels as given in the above table.

1) http://host.robots.ox.ac.uk/pascal/VOC/voc2012/ - pascal

2) https://archive.ics.uci.edu/ml/datasets/Letter+Recognition – letter

Kernel	KPCA	SVDD	OCSVM	OCSSVM	OCSSVM with SMO
Linear	0.02	0.09	0.01	0.07	0.04
RBF	0.05	0.07	0.14	0.09	0.04
Intersection	0.18	0.01	0.04	0.26	0.22
Hellinger	0.01	0.02	0.02	0.13	0.10
χ^2	0.18	0.0	0.02	0.18	0.17

Introduction :

SVM – Support Vector Machine is one of the most widely used machine learning algorithm for classification problems in particular for classifying high-dimensional data. Besides their good performance in practical applications they also a good theoretical justification in terms of both universal consistency and learning rates.

Algorithm :

Algorithm 1 Markov Sampling for SVMC

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.

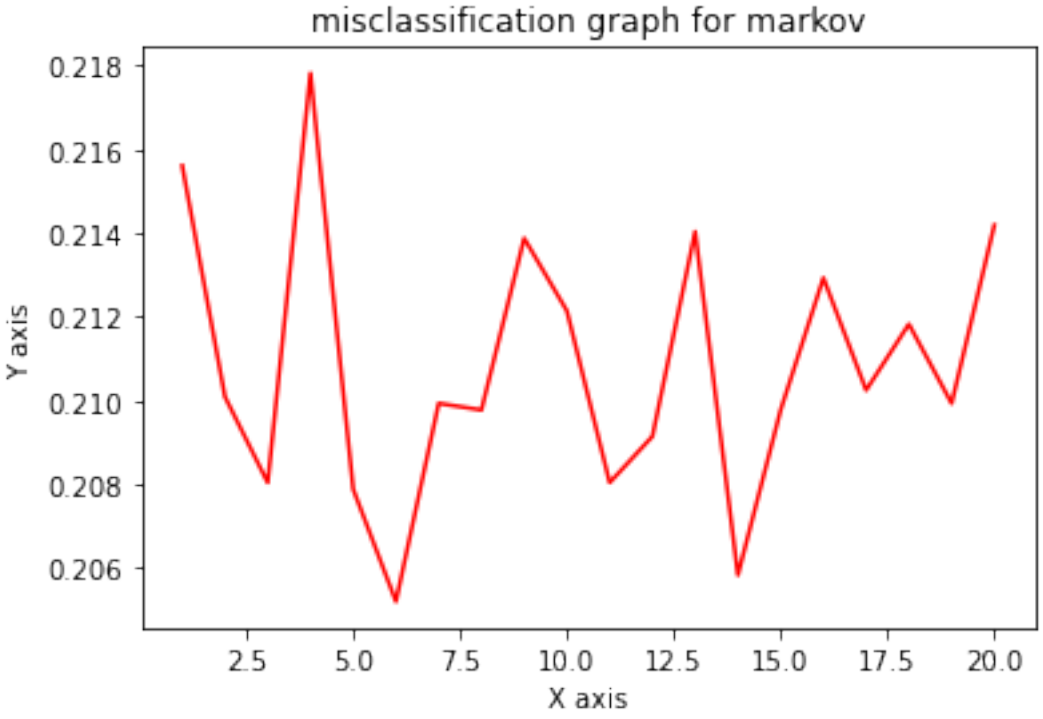
Step 3: Draw randomly another sample from D_{tr} and denote it the candidate sample z_*.

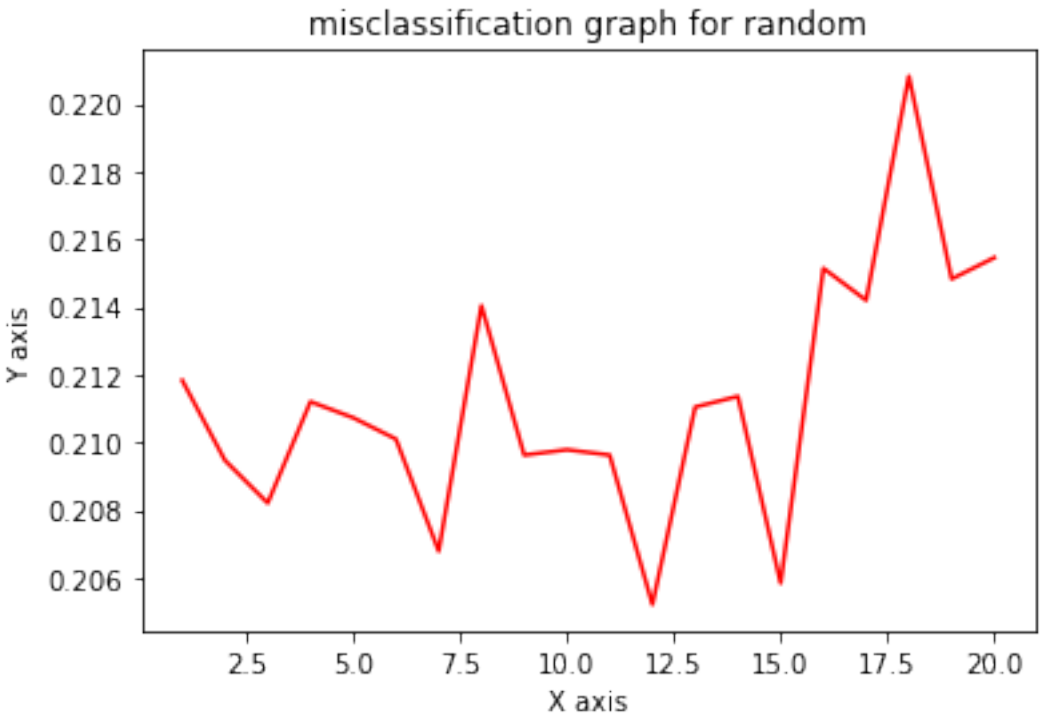
Step 4: Calculate the ratio P of $e^{-\ell(f_0,z)}$ at the sample z_* and the sample z_t, $P = e^{-\ell(f_0,z_*)}/e^{-\ell(f_0,z_t)}$.

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.

Result :


misclassification graph for markov
0.218
0.216
0.214
0.212
0.210
0.208
0.206
Y axis
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
X axis


misclassification graph for random
0.220
0.218
0.216
0.214
0.212
0.210
0.208
0.206
Y axis
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
X axis