Project Report: Compressed Learning

Yash Gupta(180050121) Pratyush Agarwal(180050078)

June 20, 2020

1 SVM: Theory and Results

1.1 Soft SVM: Introduction

Whenever the training examples are not linearly separable soft margin SVM's are used. The idea is to simultaneously maximize the margin and minimize the empirical hinge loss. Let $[(x_1, y_1), ..., (x_M, y_M)]$ be a set of M labeled instances sampled i.i.d from some distribution D. For any linear classifier $w \in R^n$ we define its true hinge loss as,

$$H_D(w) = E_{(x,y)\in D}[1 - yw^T x] \tag{1}$$

and its empirical hinge loss

$$\hat{H}_S(w) = E_{(x_i,y_i)\in S}[1 - y_i w^T x_i] \tag{2}$$

We also define the true regularization loss of a classifier w as

$$L(w) = H_D(w) + \frac{1}{2C}\|w\|^2 \tag{3}$$

and the empirical regularization loss

$$\hat{L}(w) = \hat{H}_S(w) + \frac{1}{2C}\|w\|^2 \tag{4}$$

Soft margin SVM's minimize the empirical regularization loss which is a convex optimization program.

1.2 Theorem 2.1

Let $S = [(x_1, y_1), ..., (x_M, y_M)]$ be a set of M examples chosen i.i.d from some distribution D, and let w be the SVM classifier obtained by minimizing Equation (4). Then

$$w = \sum_{i=1}^{M} \alpha_i y_i x_i, \tag{5}$$

where $i : 0 \leq \alpha_i \leq \frac{C}{M}$ and $\|w\|^2 \leq C$.

1.3 Terminology

Classifier in data domain - w Classifier in measurement domain - z

$$w^* = argminL(w) \tag{6}$$

$$z^* = aargminL(z) \tag{7}$$

$$\hat{w}_S = argmin\hat{L}_S(w) \tag{8}$$

$$\hat{z}_{AS} = argmin\hat{L}_{AS}(w) \tag{9}$$

2 Compressed Learning

2.1 Theorem 4.1

If the entries of $\sqrt{m}A$ are sampled i.i.d from either Standard Gaussian Distribution $N(0,1)$ or Bernoulli Distribution $U(-1,1)$ and $m = \Omega(klog(\frac{n}{k}))$, then except with probability $e^{-c(\epsilon)m}$, A satisfies the restricted isometry property with parameters (k, ϵ)

2.1.1 Proof Outline

Concentration Condition:
For random matrices $\Phi(w)$ (generated by independent random sampling):

$$\Pr(|\ \|\Phi(\omega)x\|_{\ell_2^n}^2 - \|x\|_{\ell_2^N}^2| \geq \epsilon\|x\|_{\ell_2^N}^2) \leq 2e^{-nc_0(\epsilon)}, \quad 0 < \epsilon < 1,$$

where the probability is taken over all n x N matrices $\Phi(w)$ and $c_0(\epsilon)$ is a constant depending only on ϵ and such that for all $\epsilon \in (0,1)$, $c_0(\epsilon) > 0$.
Concentration condition is valid for Bernoulli and Gaussian random matrices with $c_0(\epsilon) = \frac{\epsilon^2}{4} - \frac{\epsilon^3}{6}$
Proof : Using Tail Bounds
Now, using the concentration condition, we prove the following:
Let $\Phi(w)$ be a random matrix of size n x N drawn according to any distribution that satisfies the concentration inequality. Then, for any set T with $\|T\| = k < n$ and any $0 < \delta < 1$, we have

$$(1-\delta)\|x\|_2 \leq \|\Phi(w)x\|_2 \leq (1+\delta)\|x\|_2 \tag{10}$$

for all $x \in X_T$ with probability $\geq 1 - 2(12/\delta)^k e^{-c_0(\delta/2)n}$
Using this, we prove:
Suppose that n, N, and $0 < \delta < 1$ are given. If the probability distribution generating the n x N matrices $\Phi(w)$, satisfies the concentration inequality, then there exist constants c1, c2 > 0 depending only on δ such that the RIP holds for $\Phi(w)$ with the prescribed δ and any $k \leq \frac{c_1 n}{log(N/k)}$ with probability $\geq 1 - e^{-c_2 n}$.

2.2 Theorem 4.2 and 4.3

Let A_{mxn} be the measurement matrix satisfying $(2k, \epsilon)$-RIP , and x, x' be two k-sparse vectors in R^n, such that $\|x\|^2 \leq R$, $\|x'\|^2 \leq R$. Then

$$(1+\epsilon)x^T.x' - 2R^2\epsilon \leq (Ax)^T(Ax') \leq (1-\epsilon)x^T.x' + 2R^2\epsilon \tag{11}$$

2.3 Theorem 4.4

This is essentially theorem 4.2 for linear combination of k-sparse vectors.
Let A_{mxn} be the measurement matrix satisfying $(2k, \epsilon)$-RIP. Let M,N be two integers, and $[(x_1, y_1), ..., (x_M, y_M)]$, $[(x'_1, y'_1), ..., (x'_N, y'_N)] \in X$. Let $\alpha_1, ..., \alpha_M, \beta_1, ..., \beta_N$ be non-negative numbers, such that $\sum_{i=1}^M \alpha_i \leq C$ and $\sum_{i=1}^N \beta_i \leq C$ for some $C, D \geq 0$. Let $\alpha = \sum_{i=1}^M \alpha_i y_i x_i$ and $\beta = \sum_{i=1}^N \beta_i y'_i x'_i$. Then:

$$|\alpha^T\beta - (A\alpha)^T A\beta| \leq 3CDR^2\epsilon \tag{12}$$

2.4 Theorem 5.1

Let A_{mxn} satisfy $(2k, \epsilon)$-RIP. Also let $S = [(x_1, y_1), ..., (x_M, y_M)]$ be the the training set of size M, where each example is sampled i.i.d from some distribution D in data domain. Let $\hat{w}_S$ be the soft-margin SVM trained on S, and $A\hat{w}_S$ be the vector in the measurement domain, obtained by projecting down $\hat{w}_S$. Then

$$L_D(A\hat{w}_S) \leq L_D(\hat{w}_S) + O(CDR^2\epsilon) \tag{13}$$

2.5 Theorem 5.2 corollary (proof not covered)

Let $\hat{w}_S$ be the SVM's classifier. Then with probability 1 - δ,

$$L_D(\hat{w}_S) \leq L_D(w^*) + O(\frac{Clog(1/\delta)}{M}) \tag{14}$$

2.6 Theorem 2.1

Combining all of the above and the fact that $L_D(z^*) \leq L_D(\hat{A}w_S)$ (as z^* is the best classifier in the data domain),

$$H_D(\hat{z}_{AS}) \leq H_D(w_0) + \frac{\|w_0\|^2}{2C} + O(\frac{Clog(1/\delta)}{M} + CDR^2\epsilon) \tag{15}$$

Let w_0 be a good linear classifier in the data domain, with low hinge loss, and large margin (hence small $\|w_0\|_2$). Then, for a suitable C and a $(2k, \epsilon)$-RIP matrix A, with probability $1 - 2\delta$ over AS :

$$H_D(\hat{z}_{AS}) \leq H_D(w_0) + O(\sqrt{\|w_0\|^2(\frac{log(1/\delta)}{M} + DR^2\epsilon)}) \tag{16}$$

i.e the hinge loss of the SVMs classifier $\hat{z}_{AS}$ in measurement domain is close to the hinge loss of the oracle best classifier w_0 in data domain.

2.7 Summary

We have a total of 6 classifiers in the proofs. We have the oracle best classifier w^* and its data-driven approximation $\hat{w}_S$. Similarly, for the measurement domain we have z^* and $\hat{z}_{AS}$ respectively. Lastly we have the projected classifier $A\hat{w}_S$. Throughout the paper, the following chain of results leads us to the final result.

$$L_D(\hat{z}_{AS}) \leq L_D(z^*) \leq L_D(\hat{A}w_S) \leq L_D(w_S) + O(f(C,\epsilon)) \leq L_D(w^*) + O(g(C,\epsilon,\delta)) \leq L_D(w_0) + O(h(C,\epsilon,\delta)) \tag{17}$$

where f,g,h are some functions.

3 The case with Unknown Basis

If data is not sparse in the observed domain but there exists a possibly unknown basis such that data can be represented sparsely in that domain. More precisely, now there exists an orthonormal basis such that each instance can be represented as:

$$x = \Psi s \tag{18}$$

where s is k-sparse. Consequently, the measurement domain is

$$M = (Ax, y) = (A\Psi s, y) = (\Phi s, y) \tag{19}$$

Since Ψ is orthonormal, and entries of A are sampled iid form the Gaussian distribution, A and AΨ have the same distribution. Thus, AΨ also follows RIP with a high probability and we note from (16) that SVM performance between sparse data-domain and measurement domain will be close. Now, since Ψ is orthonormal, it preserves exact norms and thus follows 'RIP' with an $\epsilon = 0$. We get a very string bound by putting it in (16). Thus data domain and sparse-data domain are close. We conclude that data-domain and measurement domain are close. Therefore, compressed learning is universal with respect to bases, the SVM's classifier in the measurement domain works almost as well as the best classifier in the data domain provided that there exists some basis in which the instances have a sparse representation.

4 Experimental Results

Datsets used

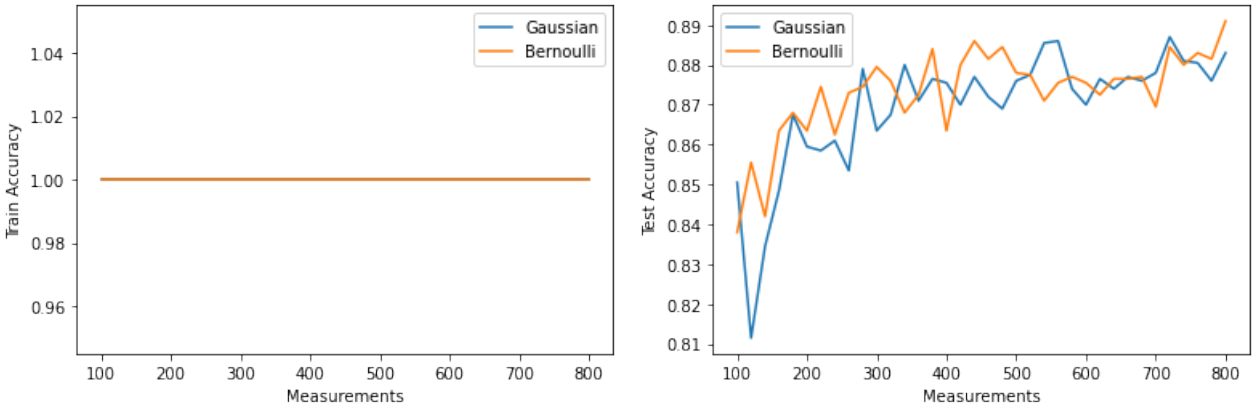
- MNIST : Used this for binary classification using SVM's. To classify whether the digit is 3 or not.
- Brodatz : Used to classify whether the images are horizontal or not

Matrices A Used :

- Gaussian : m*n matrix with entries of $\sqrt{m}A$ sampled from N(0,1)
- Bernoulli : m*n matrix with entries of $\sqrt{m}A$ +1 or -1 with equal probability

4.1 Results for MNIST

Stats for MNIST dataset
No. of Training Samples = 500
No. of Testing Samples = 2000
Dimensions of images 28*28
Number of features in data domain (originally) = 784
Train accuracy in original domain = 1.0
Test accuracy in original domain = 0.882
Train and Test accuracy in compressed domain wrt number of measurements



4.2 Brodatz

Done in accordance with the paper (measurements=2048 as given in paper)
Stats for Brodatz dataset

No. of Training Samples = 70
No. of Testing Samples = 42
Dimensions of image 128 * 128
Number of features in data domain (originally) = 16384
Number of measurements taken (No. of features in measurement domain) = 2048
Train accuracy in original domain = 1.0
Test accuracy in original domain = 0.76
Train accuracy when random Gaussian matrix used 1.0
Test accuracy when random Gaussian matrix used 0.71
Train accuracy when random Bernoulli matrix used 1.0
Test accuracy when random Bernoulli matrix used 0.71

5 Observations

- For MNIST the algorithm works quite well as expected, with test accuracy in measurement domain being almost equal to that in data domain (0.88) with sufficient measurements .
  We start achieving test accuracy of 0.87 around 200 measurments whereas the original domain had 784 features. This shows that the technique works quite well.
  As expected , the test accuracy increases with the number of measurements taken.
  Since the data is linearly seperable , the train accuracy is almost 1, since SVM is able to find a separating plane.
  Also the nice performance is due to the fact that there is huge amount of data available and we take positive samples = negative sample (ie digits 3 and not 3 are taken in equal amount)

- For Brodatz , the data available is less (112 images) , out of which 70 is taken for training and 42 for testing. Hence the SVM test accuracy is less.
  But still , the test accuracy is comparable for both the measurement domain(0.71) and data domain(0.76) , verifying the claims of the paper .
  Possible reasons of relatively poor performance of SVM is less data and the uneven distribution of positive and negative samples (25/112 images are horizontal and rest aren't).

6 Extension

The Theorem 4.4 shows that the given technique can be used for any classification/ regression algorithm which depend linearly on the data , since the inner product is preserved using such matrices.// Hence , we implemented classification of MNIST dataset based on SVM on the top 100 prinicipal components of original data(X) as well as top 100 principal components of the compressed data (AX) and compared the performance.
The test accuracy is comparable. The intuition behind this is that if 100 components of original data cover, say 95% variance of the data, then 100 components of the compressed data will also cover around 95% or more variance. Hence if 100 components of X is sufficient to classify digits, then 100 components of AX will also be sufficient.

6.1 Mathematical intuition

Suppose v_i for $i <= k$ be the top k principal components of the centred data X=$\{x_i$ for $i <= n\}$ and they cover variance T, ie

$$\frac{\sum_{i=1}^{i=n}\sum_{j=1}^{j=k}|x_i^T v_j|^2}{\sum_{i=1}^{i=n}|x_i^T x_i|^2} = T$$

then Av_i for $i <= k$ will also cover around T variance of the data AX=$\{Ax_i$ for $i <= n\}$, since by theorem 4.4
$x_i^T x_i \approx (Ax_i)^T Ax_i$ and $x_i^T v_j \approx (Ax_i)^T Av_j$ along with $(Av_i)^T(Av_j) \approx v_i^T v_j = 0$ for $i! = j$ and Av_i will also be approximately unit vector since $(Av_i)^T(Av_i) \approx v_i^T v_i = 1$
Hence

$$\frac{\sum_{i=1}^{i=n}\sum_{j=1}^{j=k}|(Ax_i)^T Av_j|^2}{\sum_{i=1}^{i=n}|(Ax_i)^T Ax_i|^2} \approx T$$

Hence since the orthonormal set Av_i for $i <= k$ covers around T variance of the data AX, the actual top k principal components of AX will cover $>= T$ variance of AX.

Claim: The principal components v_i of X=$\{x_i$ for $i <= n\}$ also satisfy theorem 4.4, since they are linear combination of $x_i's$ only.
Proof of the claim
The principal components are eigenvectors of the matrix $C = X^T X$.
The columns of C are linear combination of $x_i's$ as the ith column $C_i = \sum_{j=1}^{j=n} x_{ji}\mathbf{x_j}$.
The eigenvectors of C are linear combination of columns of C. This can be shown similarly as above using $C = US^2U^T$, where U's columns are the eigenvectors.
$US^2 = CU$, the ith column of US^2 is $s_i\mathbf{v_i}$ where s_i is the ith eigenvalue and v_i is the ith eigenvector.
Hence ith column of US^2 is linear combination of columns of C, so

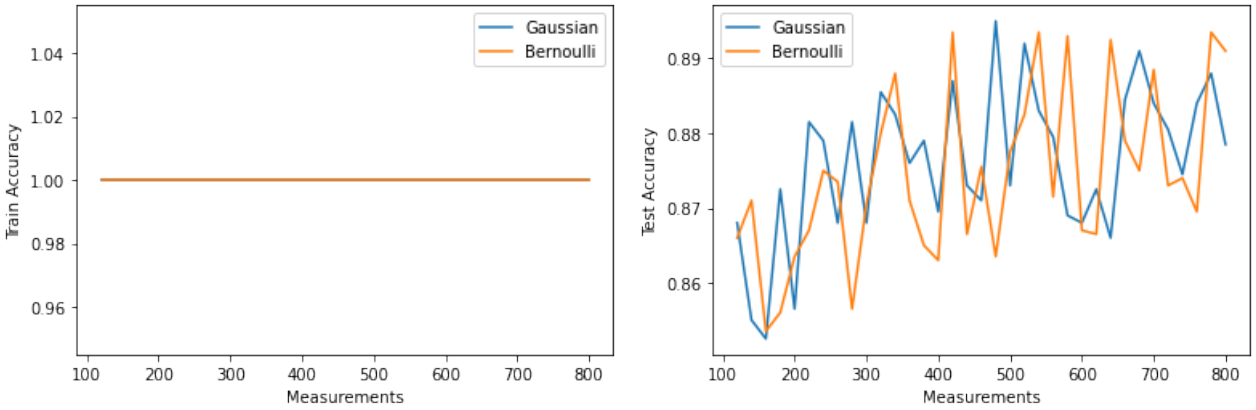
$$v_i = \sum_{j=1}^{j=p}\frac{U_{ji}}{s_i}\mathbf{C_j}$$

where the data has p feaures. The columns of C are linear combination of the data samples.
Hence the principal compoents are linear combination of data samples.

6.2 Experimental results for classification using PCA

Stats for MNIST dataset SVM using PCA with 100 components
No. of Training Samples = 500
No. of Testing Samples = 2000
Dimensions of images 28*28
Number of features in data domain (originally) = 784
Number of PCA components = 100 for both data domain and measurement domain
Train accuracy in original domain = 1.0
Test accuracy in original domain = 0.875

Train and Test accuracy in compressed domain wrt number of measurements



6.3 Observation

As we use sufficient amount of measurements, we get train accuracy(0.86) comparable to that in data domain (0.875), which shows that the technique of using PCA with same number of components in both the domains works quite well.
Here also, generally the test accuracy increases with increasing number of measurements, with few dips in between, but still the test accuracy remains close to that in original data domain.