# Random Forests and SVM for Handwritten Digits Recognition

Akshara Boppidi, Pooja Jadhav Eshwarlal, Pablo Rivas (Ph.D)

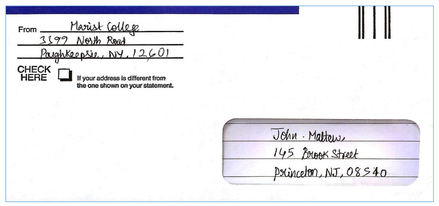
*Department of Computer Science, Marist College, New York*

MARIST

## INTRODUCTION

Digit recognition is a good problem to learn about machine learning.

- Some applications for digit recognition are online handwriting recognition on devices, numeric entries in forms filled by hand, investigation of vehicle license plates and recognition of zip code by postal services.

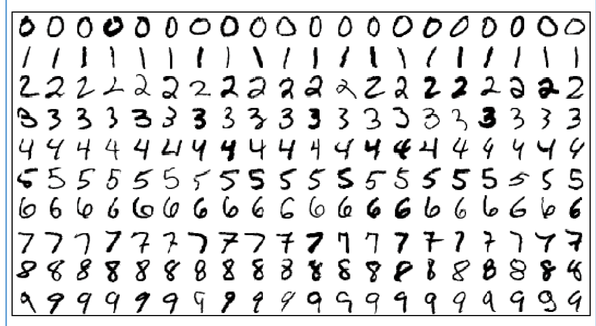



## RECOGNIZING DIGITS

- Our goal is to recognize the handwritten digits (0-9) from the dataset of images.
- While performing handwritten digit recognition we come across many challenges, one being the digits written are not always the same, they may vary in size and thickness.

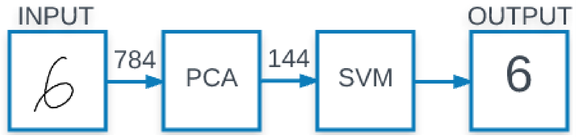
**Dataset:**

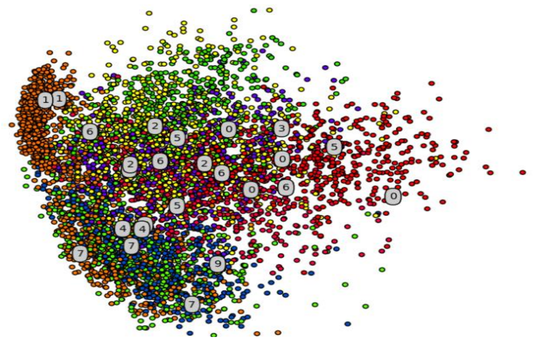
- Digit Recognition is one of the most popular competitions in Kaggle and the data is taken from MNIST dataset.
- Training set has 42000 examples and 784 features. Test set has 28000. The number of classes is 10.



## PRINCIPAL COMPONENT ANALYSIS

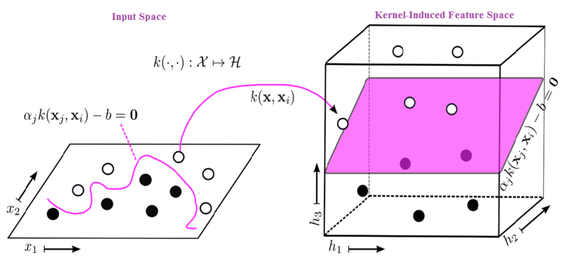
- PCA retains most of the data information while reducing the dimensionality of the dataset.
- Input digit images are projected onto a low dimensional space and based on these images, PCA extracts Eigen-vector based digits.



## SUPPORT VECTOR MACHINES

- SVM is a supervised machine learning algorithm which can be used for both classification and regression.
- It aims to find a hyperplane which best divides the dataset into two classes.
- SVM is very accurate where the number of dimensions is greater than the number of samples.




## RANDOM FOREST ALGORITHM

- Random Forests (RFs) are used for general purpose classification and regression.
- The aggregated result from multiple decision trees would form an ensemble known as a random forest.
- RFs are computationally efficient for real world prediction tasks.

| Input |  |  |  | |
|---|---|---|---|---|
| Tree 1 | 3 | 1 | 6 | 45% |
| Tree 2 | 8 | 1 | 5 | 86% |
| Tree 3 | 8 | 2 | 5 | 64% |
| Majority Vote | 8 | 1 | 5 | |

## EXPERIMENTS/RESULTS

- All the algorithms were implemented using Python scripts.
- Experiments with SVMs using the full 784 dimensions required massive amounts of memory leading to poor efficiency.
- In order to improve efficiency of the SVM training, we used PCA for dimensionality reduction from 784 to 144 features.
- We used 10-fold cross-validation to repeatedly and randomly split the dataset into training and validation set.
- Using RFs we varied the number of trees achieving the following accuracies.
  - n_estimators = 2 , accuracy = 92.7%
  - n_estimators = 10 , accuracy = **99.9**%
- The results are reported on the full 784 features training set. The cross-validated results are shown in the next table.

| Algorithm | Training Set | Vadilation Set | Testing Set |
|---|---|---|---|
| SVM | 97+/-0.4% | 94.9+/-6.3% | 93.9% |
| Random Forests *n. est=2* | 92.6+/-0.1% | 80.2+/-0.7% | 80.7% |
| Random Forests *n. est=10* | 99.9+/-0.1% | 94.1+/-0.4% | 93.7% |

## CONCLUSION

- Random forests have a much simpler algorithm then SVMs and also consume lesser resources than SVM training on highly dimensional datasets.
- SVMs work best for smaller datasets, since the training time taken for larger dataset is typically $O(n^2)$ with respect to the input.
- For both random forest and SVMs, the complexity increases as the number of training samples increase.

## ACKNOWLEDGEMENTS

- This work was partially supported by the Department of Computer Science of Marist College.
- Special thanks to our advisor for supporting us in furthering our research experience.

## REFERENCES

[1] Alaei, Alireza, et.al , "Using modified contour features and SVM based classifier for the recognition of Persian/Arabic handwritten numerals." *Advances in Pattern Recognition, 2009. ICAPR'09. Seventh International Conference on*. IEEE, 2009.

[2] Hearst, Marti A., et al. "Support vector machines." *IEEE Intelligent Systems and their Applications* 13.4 (1998): 18-28.

[3] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer, Berlin: Springer series in statistics, 2001.