Kernel Methods for Image Classification

Efforts by: Manish Kumar (201811085), Rishi Dave (201811073).
Supervisor: Prof. Pankaj Kumar.
IT524 – Computer Vision.

ABSTRACT

A support vector machine(SVM) [1] is widely used as a classification task. SVM generates a hyperplane between two classes. But there are many such examples in real life that data is not always linearly separable. So kernel methods (or kernel tricks) are used for making a non-linear classifier from a linear classifier. Kernel methods transform input feature vector to another so that SVM can fit hyperplane more accurately.

1 Introduction:

1.1 Why do we use kernels?

Kernel method transforms one feature vector to another feature vector. There is a mapping function ϕ: X→F, which transforms the input feature vector to another vectors where X is Input Space and F is the transformed feature vector space (Feature Space). SVM uses F for learning and predicting instead of X.
Consider this example:
As shown in figure-1 input vector has two features a and b. There are two classes red and blue for given data points. There is no linear separation between them. So simple SVM cannot generate accurate hyperplane (line in this case). So we transform this 2-dimensional feature vector to 3-dimensional feature vector with this ϕ function: $\phi((a,b)^T) = (p,q,r)^T$ where p = a, q = b and $r = a^2 + b^2$. Figure-2 represents this new transformed data points. Figure-3 represents the front view of figure-2. In figure-3, we can observe that SVM can fit an accurate hyperplane that can separate two classes.

1.2 Dual Representation and Kernel Function.

Figure-4 shows the decision boundary generated by a SVM. SVM learns w and b parameters to generate hyperplane. Objective of SVM is to minimize $< w \bullet w >$ with respect to $y_i(< w \bullet x_i > + b) \geq 1 \;\; (i = 1 \; to \; N)$ constraint. So the primal Lagrangian is [1]

$$L(w,b,\alpha) = \frac{1}{2} < w \bullet w > - \sum_{i=1}^{N} \alpha_i [y_i(< w \bullet x_i > + b) - 1] \quad (\alpha_i \geq 0). \qquad (1)$$

Here N is number of training examples, w and b are the parameters of hyperplane, $x_i \in X$, $y_i \in \{1,-1\}$(class labels), α_i are lagrangian multipliers and $< x \bullet y >$ is the notation of the inner product of vectors(dot product).
If we differentiate and equate it with zero to find extrema, we get $w = \sum_{i=1}^{N} \alpha_i y_i x_i$ (w is a linear combination of input data points) [1].
To find the dual representation replace the w in equation-1 so we get [1]

$$L(w,b,\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j < x_i \bullet x_j >. \qquad (2)$$

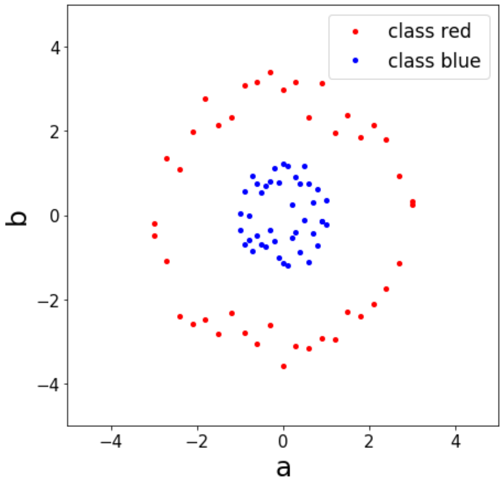


Figure-1: data points of two classes

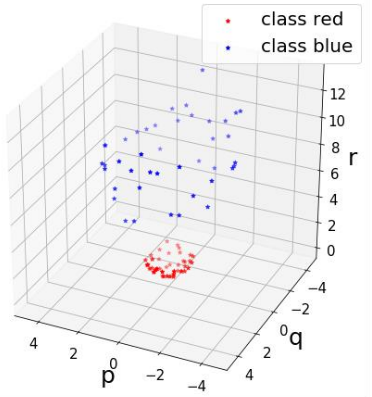


Figure-2: data points after transformation

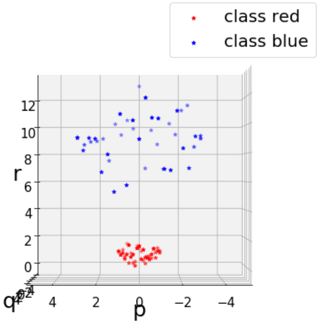


Figure-3: Front View of figure-2

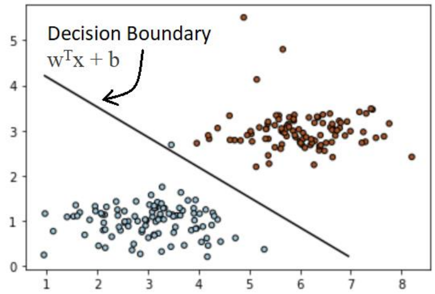


Figure-4: SVM Decision Boundary

From equation-2 we can see that the hypothesis can be represented as a linear combination of the input data points. Dual representation helps to use of kernels. If we transform input feature vector with ϕ(x_i) then new hypothesis becomes [1]

$$L(w, b, \alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j < \phi(x_i) \bullet \phi(x_j) >. \qquad (3)$$

If we directly find the inner product value of $< \phi(x_i) \bullet \phi(x_j) >$ somehow then we can merge two steps(one is transforming input feature vectors second is inner product of $< \phi(x_i) \bullet \phi(x_j) >$) into one. This direct computation method is called **Kernel Method or Trick.** Mapping Function k: X × X ⟶ ℝ is called **Kernel Function.**

$k(x, y) = < \phi(x) \bullet \phi(y) > (x, y \in X)$.

1.3 Gram Matrix and Kernel Matrix.

We can represent the complete training set with one matrix G($G \in \mathbb{R}^{N*N}$) where $G(\, i\, , j\,) = < x_i \bullet x_j >$. From equation-2, we can observe that we need the inner product of the input feature vector. This G is called **Gram Matrix**. So we can train SVM with G.

K($K \in \mathbb{R}^{N*N}$) where $K(\, i\, , j\,) = k(x_i\, , x_j) = < \phi(x_i) \bullet \phi(x_j) >$. Matrix K is called **Kernel Matrix.**

1.4 Mercer Theorem.

Mercer theorem stats validity of a kernel function.
Theorem: Let X be a finite input space with k(x,y) a symmetric function on X. Then k(x,y) is a kernel function if and only if the matrix K is positive semi-definite (has non-negative eigenvalues) [1].

2 Problem Definition:

2.1 Image Classification.

Classification is the technique for assigning objects to a particular class based on some properties (features) of that object. Classification is a vast research topic in Machine Learning. There are many real-life problems where we need classification, such as whether the patient has cancer or not, tweet sentiment analysis (positive or negative or neutral), speaker identification, weather prediction.
Image classification is a technique to classify an image to a particular class. Gender and age(range) classification from an image, detection of facial expression, detection of cancer from MRI images are some examples of image classification. Kernel method helps to increase the accuracy of classification as it can transform features so that the linear separation between classes can be done.

3 Experiments and Results:

3.1 Datasets Used.

MNIST [5].
MNIST dataset is the handwritten English digits. There are 10 classes in the MNIST dataset. 10000 images (nearly 1000 per class) are used for the training SVM and 1000 (nearly 100 per class) images are used for prediction. Accuracy can be improved if we use complete datasets.

Devanagari [6].
Devanagari dataset is the handwritten characters in the Devanagari script. There are 46 classes (2000 images per class) in the Devanagari dataset. We choose only digits for classification. 4000 images (400 per class) are used for the training SVM and 1000 (100 per class) images are used for prediction.

JAFFE [7].
The Japanese Female Facial Expression. Dataset contains 213 images which has 7 different classes (facial expression). SVM is trained on 178 images and tested on 35 images.

COREL [9]
Corel dataset contains 10000 images. There are 100 classes in corel dataset and per class 100 images. We choose 6 different classes PolarBears, Dogs, Butterflies, Arabian Horses, Apes, Rhinos. There are 600 images, SVM is trained with 420 images (70 per class) and tested with 180 images (30 per class).

3.2 Kernels Used.

Polynomial (Poly) [8].
$k(x, y) = < x \bullet y >^{d}\ (x, y \in X, d > 0).$

Radial basis function(RBF)[8].
$k(x, y) = \exp[-\gamma(||x - y||^{2})]\ (x, y \in X, \gamma > 0)$

Cosine [8].
$$k(x, y) = \frac{< x \bullet y >}{||x||||y||}(x, y \in X).$$

Tanh [8].
$k(x,y) = \tanh(a * < x \bullet y > + b)\ (x, y \in X\ \ a, b \in \mathbb{R})$.

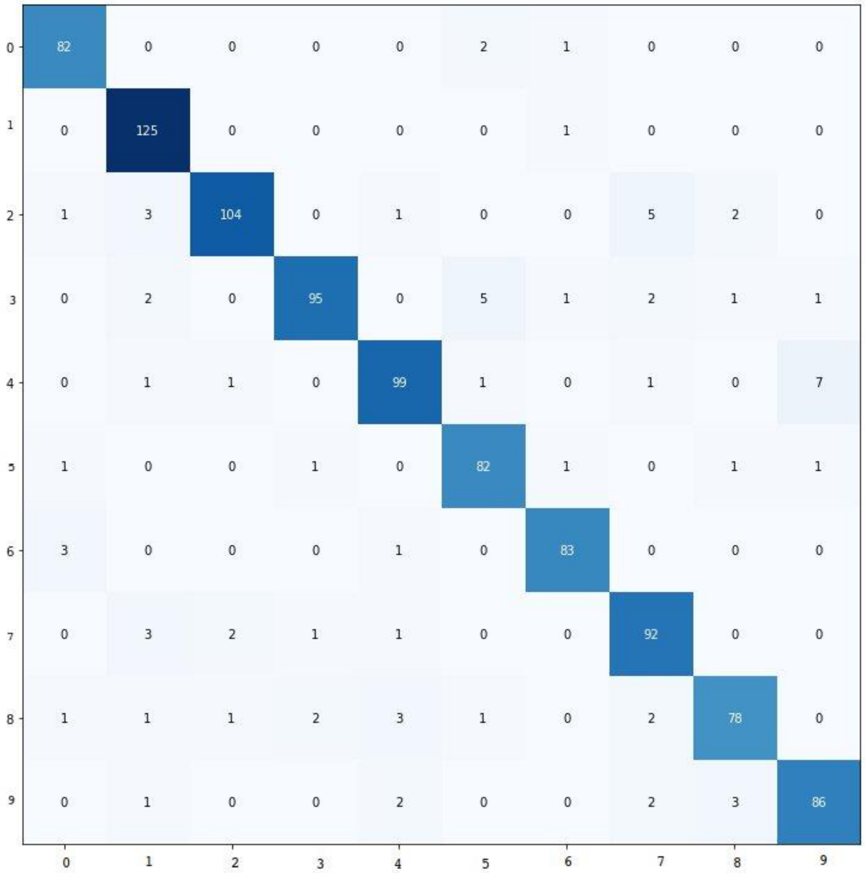
3.3 Results.

One vs Rest is used for the multiclass classification in experiments.

	Linear	Poly	RBF	Cosine	Tanh
MNIST	90.30	84.2	90.60	91.80	92.6
Devaganagari Digits	89.80	91.40	10.00	09.00	09.60
JAFFE	82.85	82.85	54.28	25.71	31.42
Corel	77.22	74.44	32.77	20.55	19.44

Table 1: Accuracy of different types of kernels on different datasets

3.4 Confusion Matrix.



	0	1	2	3	4	5	6	7	8	9
0	82	0	0	0	0	2	1	0	0	0
1	0	125	0	0	0	0	1	0	0	0
2	1	3	104	0	1	0	0	5	2	0
3	0	2	0	95	0	5	1	2	1	1
4	0	1	1	0	99	1	0	1	0	7
5	1	0	0	1	0	82	1	0	1	1
6	3	0	0	0	1	0	83	0	0	0
7	0	3	2	1	1	0	0	92	0	0
8	1	1	1	2	3	1	0	2	78	0
9	0	1	0	0	2	0	0	2	3	86

Figure 5: MNIST Dataset with Tanh kernel

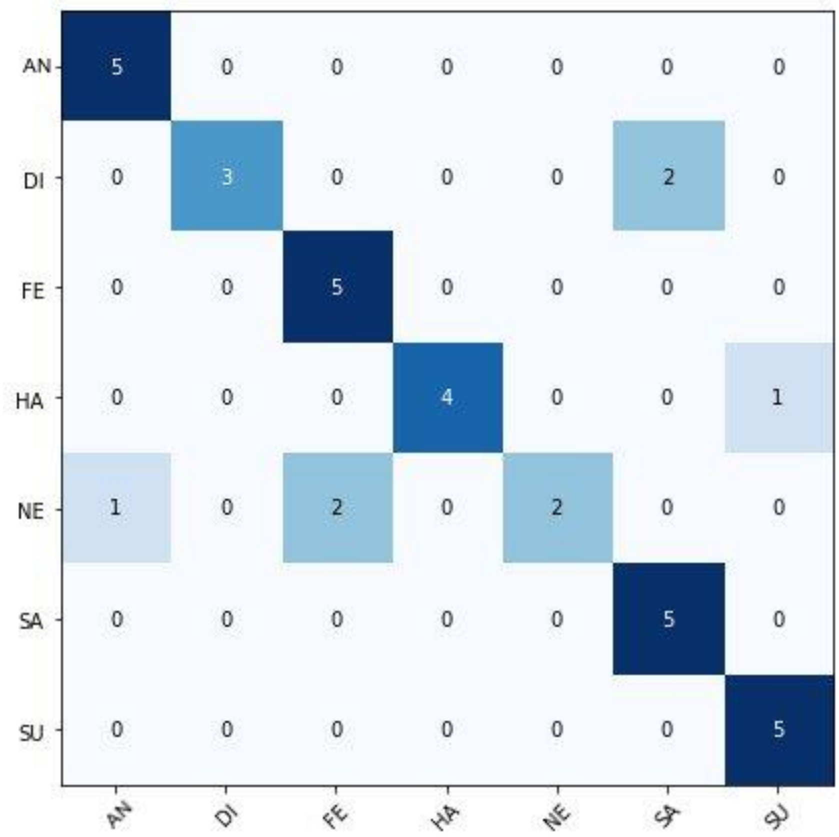


Figure 6: JAFFE Dataset with Poly Kernel

REFERENCES

[1] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, 2000.

[2] Andrew Y. Ng and M. Jordan, On Discriminative vs. Generative Classifiers: A Comparison of logistic regression and naive Bayes, Neural Information Processing System, 2001.

[3] T. Jaakkola and D. Haussler, Exploiting generative models in discriminative classifiers, Advance Neural Information Processing System II, 1998.

[4] P. Moreno, Purdy P. Ho, N. Vasconcelos, Advances in Neural Information Processing System 16, 2003.

[5] Y. LeCun and C. Cortes, “MNIST handwritten digit database,” 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[6] S. Acharya, A.K. Pant and P.K. Gyawali, Deep Learning Based Large Scale Handwritten Devanagari Character Recognition, In Proceedings of the 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), pp. 121-126, 2015.

[7] Michael J. Lyons, Shigeru Akamatsu, Miyuki Kamachi, Jiro Gyoba. Coding Facial Expressions with Gabor Wavelets, 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200-205 (1998)

[8] Sch¨olkopf, B., and Smola, A. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, Dec. 2002. Parts of this book, including an introduction to kernel methods, can be downloaded http://www.learning-with-kernels.org/sections/ here.

[9] Guang-Hai Liu, Jing-Yu Yang, etc,. Content-based image retrieval using computational visual attention model, Pattern Recognition, 48(8) (2015) 2554-2566.