SUPPORT VECTOR MACHINE-KERNEL FUNCTIONS

Anna Binoy and Sumegha M. T.
Group 12

CS460: Machine Learning
National Institute of Science Education and Research, Bhubaneswar

February 13, 2023

PART I: QUICK RECAP

PART II: KERNELS

Part I

QUICK RECAP

RECAP: SUPPORT VECTOR MACHINE

SVMs or **Support Vector Machines** are supervised learning algorithm that are used primarily for binary classification of data points which are linearly separable in a given dimension. The objective of SVM is to find the hyperplane that maximizes the margin, that is,

$$max \; \frac{2}{||w||}$$

where

$$w^T x + b = 0$$

is the equation of the hyperplane. In the case of a labelled data set in which there is clear linear separation between the two classes, SVM can be easily applied.

RECAP:SUPPORT VECTOR MACHINE

Suppose that the two classes are well separated, however, there are some outliers that are misclassified as they lie within the maximised margin for which a hyperplane has been decided. This situation introduced us to soft margin SVM and hard margin SVM.

Hard margin SVM do not allow any misclassification. Even if the data is well separated except for negligible outliers, Hard margin SVM does not return the hyperplane that is a good generalisation. Soft margin SVM allows some misclassification such that the hyperplane returned will have maximum margin. This prevents much misclassification in the training data set.

To penalise misclassification, Hinge Loss function is introduced.

$$min(\ max(0, 1 - y_i.f(w^T x_i + b) + \frac{c}{2}||w||))$$

Part II

Primal-Dual Lagrangian

LAGRANGIAN

If we have an optimization problem,
$$min\ f(x) \tag{1}$$
subject to the constraints:
$$g(x_i) \leq 0 \ \text{(where i = 1,2,3,....,k)} \tag{2}$$
$$h(x_i)0 \ \text{(where i = 1,2,3,....,l)} \tag{3}$$
The above function can be converted into **Primal Lagrangian** (Θ_p) using Lagrange multipliers, α and β where $\alpha_i \geq 0$, as
$$\Theta_p = max\ \mathcal{L}(x, \alpha, \beta) = f(x) + \Sigma\alpha_i g(x_i) + \Sigma\beta_i h(x_i) \tag{4}$$
Inference from the above equation

$$\Theta_p = f(x) \ \text{(when constraints are satisfied)}$$

$$\Theta_p = \infty \ \text{(when constraints are not satisfied)}$$

LAGRANGIAN

Hence, when all the constraints are satified, the problem of finding the minimum of f(x) (1) becomes finding the value of α and β such that the Lagrangian $\mathcal{L}$ is maximum. Therefore, the optimization problem is

$$min\ max(L(x, \alpha, \beta)$$

In the dual of the above optimization problem, we have

$$max\ min(L(x, \alpha, \beta)$$

Here we look for the optimal x rather than α, β that minimized the Lagrangian.
In the above case

$$\text{(dual) } max\ min(L(x, \alpha, \beta) = min\ max(L(x, \alpha, \beta) \text{ (primal)}$$

That is, solving the dual solves the primal.
Here, solving the dual is easier than solving the primal as it involves minimization of Lagrangian, that is, Lagragian is minimum for x where

$$\frac{\partial \mathcal{L}}{\partial x} = 0 \tag{5}$$

PRIMAL-DUAL LAGRANGIAN

SVM optimization problem is given by

$$min \ \frac{1}{2}||w||^2 \tag{6}$$

subject to the constraints

$$1 - y_i(wx + b) \leq 0 \text{ (where i=1,2,3,....k)} \tag{7}$$

Here, (1) is analogous to (6) and (2) is analogous to (7). Hence, we can construct primal-dual Lagrangian for SVM.

$$\mathcal{L}(w, \alpha) = \frac{1}{2}||w||^2 + \Sigma\alpha_i(1 - y_i(wx + b)) \tag{8}$$

The primal is

$$\Theta_p = max \ \mathcal{L}(w, \alpha) = \frac{1}{2}||w||^2$$

The optimization problem is then

$$min \ max \ \mathcal{L}(w, \alpha)$$

PRIMAL-DUAL LAGRANGIAN

To convert the primal to dual, we minimize the Lagrangian,

$$\frac{\partial \mathcal{L}}{\partial w} = 0$$

This gives,

$$w = \Sigma \alpha_i y_i x_i$$

Since, b is a constant,

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \tag{9}$$

This gives,

$$\Sigma \alpha_i y_i = 0 \tag{10}$$

PRIMAL-DUAL LAGRANGIAN

Plugging in (9) and (10) to (8), we obtain

$$L(w, \alpha) = \Sigma_i \alpha_i - \frac{1}{2} \Sigma_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j$$
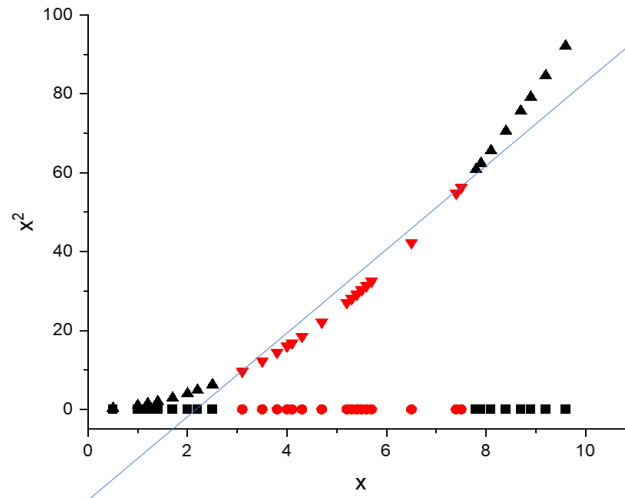
which can be written as a dot product of two vectors

$$L(w, \alpha) = \Sigma_i \alpha_i - \frac{1}{2} \Sigma_{i,j} \alpha_i \alpha_j y_i y_j < x_i, x_j >$$

Part III

KERNEL FUNCTIONS

WHAT IF THE DATA IS LINEARLY NOT SEPARABLE

Suppose we have a set of linearly non separable data points ● and ●. In order to apply SVM to the given data points, we project them to a higher dimension (say, ϕ space) where the data points are linearly separable.

WHAT IF THE DATA IS LINEARLY NOT SEPARABLE?

For example, in this case we have done

$$(x_i) \xrightarrow{\phi} (x_i, x_i^2)$$

where i represents the index of the data points
The respective data points in ϕ space, that is, ▲ and ▲, then undergo binary classification using SVM in that dimension. The straight line in the figure represents the hyperplane in the ϕ space. The classified data points are then projected back to the old lower dimension.
However, projecting each and every data point onto a higher dimension and back is computationally expensive, especially for large data sets. Hence we introduce **Kernel Trick**.

NON-LINEAR SVM

SVM is implemented to classify linearly inseparable data points using a family of functions known as kernel functions. Using kernel functions, relationships between data points in higher dimension space can be calculated, without actually transforming the data points to points in higher dimensions. This method is called Kernel Trick. To carry out the kernel trick, the dot product in dual Lagrangian of SVM is exploited [Ng n.d.]. We have the Lagrangian

$$L(w, \alpha) = \Sigma_i \alpha_i - \frac{1}{2} \Sigma_{i,j} \alpha_i \alpha_j y_i y_j < x_i, x_j >$$

We replace the dot product with

$$K(x_i, x_j) = \phi(x_i)\phi(x_j)$$

Here K is the Kernel Function and ϕ is the feature vector.
Since K returns a real number, it is computationally inexpensive than projecting each data points to a higher dimension, especially for large datasets.
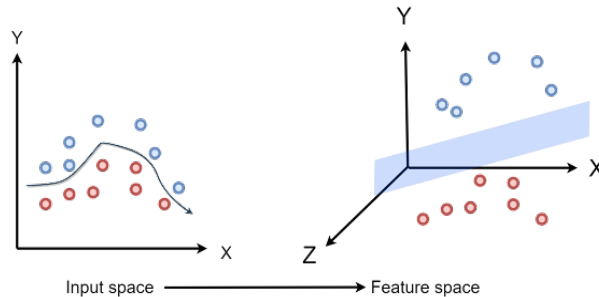
NON-LINEAR SVM

Figure. Here we were able to linearly classify the data using a plane in the ϕ space

The Lagrangian dual for non-linear SVM is hence defined as

$$max_\alpha \ \Sigma_i \alpha_i - \frac{1}{2}\Sigma_{i,j}\alpha_i\alpha_j y_i y_j K(x_i, x_j)$$

subject to

$$\Sigma_i \alpha_i y_i = 0$$
$$C \geq \alpha_i \geq 0$$

where

$$K(x_i, x_j) = \phi(x_i)\phi(x_j)$$

Some of the kernel functions covered in the class are discussed in the following .

POLYNOMIAL KERNEL

Given below is a polynomial kernel of degree up to d,

$$K(x, z) = (xz + c)^d \tag{11}$$

where

  . x,z=two different features of the dataset

  . c=coefficient of the polynomial

  . d=degree of the polynomial

Here c and d are real numbers and are hyperparameters,hence have to be found using cross validation.(give the link to the other slide here) The above equation computes the relationships between pairs of features.

▶ How to find the ϕ for $(xz + c)^d$?

Step 1- Expand $(xz + c)^d$

Step 2- Find which of the elements in the expanded $(xz + c)^d$ give the dot product equal to $\phi(x)\phi(z) = (xz + c)^d$.

Lets take an example of Quadratic Kernel.

QUADRATIC KERNEL

For quadratic kernel d=2.Hence

$$K(x, z) = (xz + c)^2 \tag{12}$$

Taking $c = \frac{1}{2}$ we get

$$(xz + \frac{1}{2})^2 = (xz + \frac{1}{2})(xz + \frac{1}{2}) = xz + x^2z^2 + \frac{1}{4} = (x, x^2, \frac{1}{2}).(z, z^2, \frac{1}{2}) = \phi(x)\phi(z) \tag{13}$$

where

. $\phi(x) = (x, x^2, \frac{1}{2})$
. $\phi(z) = (z, z^2, \frac{1}{2})$

Here the dot product gives us the higher dimensional coordinates for the data (and it behaves like a weighted K nearest neighbour model.)
Taking $c = 1$

$$(xz + 1)^2 = (xz + 1)(xz + 1) = 2xz + x^2z^2 + 1 = (\sqrt{2}x, x^2, 1).(\sqrt{2}z, z^2, 1) = \phi(x)\phi(z) \tag{14}$$

where

. $\phi(x) = (\sqrt{2}x, x^2, 1)$
. $\phi(z) = (\sqrt{2}z, z^2, 1)$

Radial Basis Function(RBF) kernel

It can be used to deal with overlapping data.Given below is the Gaussian Kernel RBF,

$$K(x,z) = e^{-\gamma||x-z||^2} \tag{15}$$

here Gaussian Radial basis function,$\gamma = \frac{1}{\sigma^2}$ and $\gamma > 0$ and is the amount of influence the two points have on each other.Here σ is the spread parameter that plays the same role as standard deviation in normal density function.
Here x and z refers to two different observation in the dataset and $||x-z||^2$ is the squared distance between two observation.
Let

$$s^2 = e^{-\frac{\gamma}{2}(x^2+z^2)} \tag{16}$$

Taylor expansion of $e^{\gamma xz}$ around $\gamma xz = 0$ we get

$$e^{\gamma xz} = \sum_{k=0}^{\infty} \frac{(\gamma xz)^2}{k!} = (1, \sqrt{\frac{\gamma}{1!}}x, \sqrt{\frac{\gamma^2}{2!}}x^2, \sqrt{\frac{\gamma^3}{3!}}x^3, ...).(1, \sqrt{\frac{\gamma}{1!}}z, \sqrt{\frac{\gamma^2}{2!}}z^2, \sqrt{\frac{\gamma^3}{3!}}z^3, ...) \tag{17}$$

Radial Basis Function(RBF) kernel

We can take

$$e^{-\gamma||x-z||^2} = e^{-\gamma(x-z)^2} \tag{18}$$

and using Eqn.16 and Eqn.17 we get

$$e^{-\gamma(x-z)^2} = e^{-\frac{\gamma}{2}(x^2+z^2)}e^{\gamma xz} = s^2 e^{\gamma xz} = s^2(1, \sqrt{\frac{\gamma}{1!}}x, \sqrt{\frac{\gamma^2}{2!}}x^2, \sqrt{\frac{\gamma^3}{3!}}x^3, ...).(1, \sqrt{\frac{\gamma}{1!}}z, \sqrt{\frac{\gamma^2}{2!}}z^2, \sqrt{\frac{\gamma^3}{3!}}z^3, ...)$$

$$\tag{19}$$

$$e^{-\gamma(x-z)^2} = (s, s\sqrt{\frac{\gamma}{1!}}x, s\sqrt{\frac{\gamma^2}{2!}}x^2, s\sqrt{\frac{\gamma^3}{3!}}x^3, ...).(s, s\sqrt{\frac{\gamma}{1!}}z, s\sqrt{\frac{\gamma^2}{2!}}z^2, s\sqrt{\frac{\gamma^3}{3!}}z^3, ...) \tag{20}$$

Hence the above represents a dot product for infinite number of dimensions or we can say that radial basis function maps the input space to the surface of an infinite dimensional hyperspace[Póczos 2014].

REFERENCES I

📄 Ng, Andrew (n.d.). *Support Vector Machines*.

📄 Póczos, Barnabás (2014). *Kernel Methods*.