REPORT -3

# Multiclass Classification: One-vs-all

Now we will approach the classification of data when we have more than two categories. Instead of y = {0,1} we will expand our definition so that y = {0,1...n}.

Since y = {0,1...n}, we divide our problem into n+1 (+1 because the index starts at 0) binary classification problems; in each one, we predict the probability that 'y' is a member of one of our classes.

Train a logistic regression classifier h\_\theta(x)*hθ*​(*x*) for each class i to predict the probability that Py = i￼ ￼.

To make a prediction on a new x, pick the class ￼that maximizes h\_\theta (x) *hθ*​(*x*)

SVM

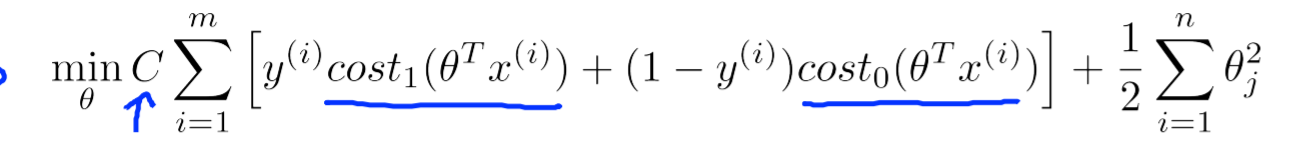
The SVM (Support Vector Machine) is a supervised machine learning algorithm typically used for binary classification problems. It’s trained by feeding a dataset with labeled examples (xᵢ, yᵢ).

In this algorithm, we try to find a decision boundary to classify the elements which is done by finding the optimal values of theta in its hypothesis.

We try to modify the logistic regression cost function equation based on some tricks and intuition to try and create a new expression, that is, the cost function for SVM and using this create the decision boundary for the dataset.

In SVM we use kernels to help in reducing the computational cost and get better results. Kernels can also be used for logistic regression but it will be quite slow and wont be able to utilize the advantage of kernels which we do in SVM using modifications done to the hypothesis and cost function.

Cost function for SVM is –



Where C=1/T , T is lambda( regularization term )

And now for y=1 we need to have the h(x) term to be greater than 1

And for y=0 , h(x)<-1.

The SVM produces the decision boundary that has the largest margin so as to get the maximum value of projection of a feature and there by reduce the value of theta as we have :

Icon

Description automatically generated

(Theta)^t dot product with x which can be written as – magnitude of both of them multiplied by cos(angle between them )

Which can be written as projection of X on theta multiplied by magnitude of theta .

So by increasing the value of projection we are decreasing the value of theta .

Now considering the use of kernels in SVM, we have different kernels but here we are using the gaussian kernel in which we find the term similarity which is defined by

Text

Description automatically generated

So instead of using the x term we are using the f term whose values are between 0 and 1 which helps in saving computational time.

New cost function of SVM is using gaussian kernel –

Text

Description automatically generated

Now, considering the parameters of the SVM we have C and sigma^2.

If the data is overfitting we try to decrease the C term (that is we are increasing the regularization term) and increase the sigma^2 term (that is we are trying to make the curve a bit more smooth) , and do vice versa for underfitting .

For a multi-class classification we can use the one vs all method ,i.e, define k SVM and find the probabilities for them and pick the largest one .

How to choose between SVM and logistic regression :

Graphical user interface, text, application

Description automatically generated

K-means Clustering

K-means clustering is a unsupervised machine learning algorithm , i.e, no labels are given in the dataset .

In this algorithm we try to classify the dataset into k clusters , which we do by finding out the distance between a particular element and k-centroids and then assigning it to the cluster centroid which is nearest to it . After this we update the centroid as the mean of the elements in the cluster and carry out the process until we have low value of the distortion function .

Text

Description automatically generated

Text

Description automatically generated

We randomly assign the values of the cluster centroids in the start to the values of the elements in the dataset and carry on the algorithm , but this can lead to different outcomes based on what was taken as the values so we run this algo multiple times and take the one which has the least value of distortion function as the appropriate model .