PA1 - Linear Models for Supervised Learning

CSE 574 - Introduction to Machine Learning

Group No: 14

ANOOSHA SEELAM	(lseelam@buffalo.edu)	50388620
VISWAPUJITHA SURESH	(viswapuj@buffalo.edu)	50388544
BHAVISHA KONDHOL	(bhavisha@buffalo.edu)	50388359

Problem 1 - Linear Regression with Direct Minimization

	Without Intercept	*With Intercept*
Train data	**138.20**	**46.77**
Test data	**326.76**	**60.89**

RMSE is significantly less for both training data and testing data with intercept. This is because without the intercept, for any weight vector 'w', the line will always pass through the origin and the dataset is not always aligned like this.

Problem 2 - Linear Regression with Gradient Descent

	RMSE
Train data	**48.08**
Test data	**54.76**

Although Linear Regression with Direct minimization has a better training error, Linear Regression with Gradient Descent has a better testing error (generalization error).

Problem 3 - Perceptron using Gradient Descent

	Accuracy
Train data	**0.84**
Test data	**0.84**

Problem 4 - Logistic Regression Using Newton's Method

	Accuracy
Train data	**0.84**
Test data	**0.86**

Problem 5 - Support Vector Machines Using Gradient Descent

	Accuracy
Train data	**0.87**
Test data	**0.88**

Problem 6 - Comparing Linear Classifiers

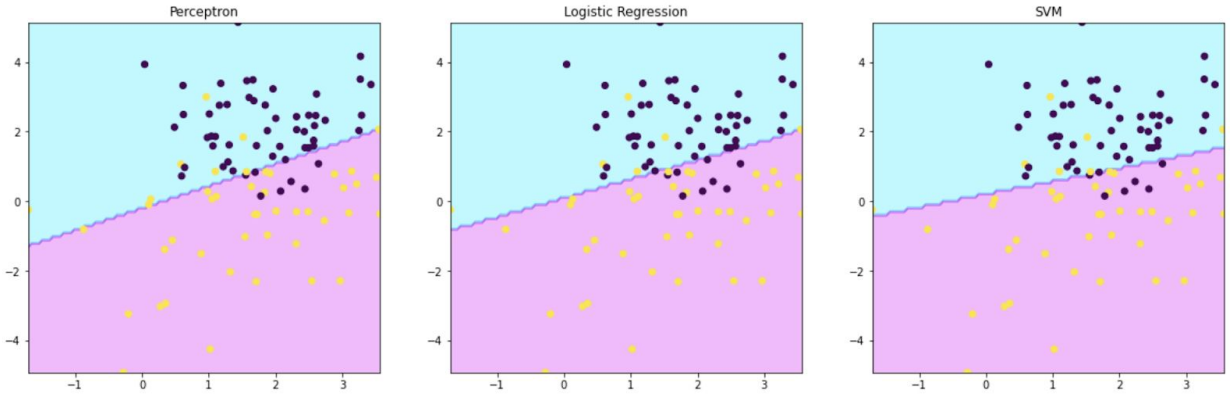
We get different decision boundaries for each model because each model uses different loss functions to minimize. Perceptron uses squared loss function, Logistic regression uses logistic-loss function and SGD based SVM uses a hinge-loss function to get the decision boundaries. By comparing the accuracy values of the test data, it seems SVM has better accuracy. But, depending on the order in which samples were taken in SVM, sometimes Logistic regression has better accuracy and sometimes SVM has a better one. But, all three models have similar accuracy for this particular dataset.

- Perceptron is trying to minimize the squared loss, i.e, it is vulnerable to the outliers in the dataset.

- Logistic regression is trying to maximize the likelihood of the data. The farther the data lies from the decision boundary on the correct side, the better the Logistic regression is.

- SVM tries to maximize the margin, that is, it is trying to maximize the distance of the closest point to the margin.

Boundaries on test data



Of all models, SVM is more robust as it is trained only on a set of samples and it tends to perform better when it sees a new dataset (generalization).