LAB DAY 3

**Log Reg using Gradient Descent**

Download the two datasets in the folder LogReg\_DT: ex2data1.csv and ex2data2.csv

Datasets are from Andrew Ng’s Machine Learning course.

1. Load the first dataset and plot it by visualizing the two different classes in different colors and different markers.
2. Do you see a linear decision boundary?
3. Split the dataset into train and test
4. Apply Logistic Regression using SciKit learn on train dataset and then predict on the hold out test set.
5. Estimate accuracy of your prediction

**Using Scipy’s Optimize API:**

1. Write a function that calculates the cost function. This function will take parameters of the model and the data as inputs.
2. Write a function that returns the gradient of the model parameters (inputs are again model parameters and the data)
3. Use Scipy’s optimize API to minimize the above cost function. The fprime option will take the gradient returned above as the value.
4. The result of the above optimization is that we now have the optimal model parameters for our data. Write a predict function that will return the right class for a test dataset. The function will take the model parameters and X values of test dataset as inputs.
5. Calculate the accuracy (percent of examples the model correctly predicted)
6. Redo the above exercise by performing cross-validation.

**Second Dataset:**

1. Load the second dataset and plot it. Can you comment on the decision boundary?
2. Create new features that are polynomials of original features, upto degree 3.

For example: Create features like x12, x22,x1 x2, x13, x12x2, x1x22,x23.

1. Split the dataset into train and test and apply Scikit learn’s Logistic regression model and test its accuracy.
2. Redo the prediction using Scipy Optimize like before.
3. Can you add a regularization term to penalize higher order polynomial terms?

**DECISION TREES:**

1. Follow along the example given in scikit learn on how to implement decision trees: <http://scikit-learn.org/stable/modules/tree.html>
2. Now use the titanic dataset in Kaggle : www.kaggle.com/titanic-gettingStarted
3. Keep only the features Pclass, Sex, Age, SibSp, Parch, Fare, Embarked and the label Survived.
4. Use Scikit learn as in the example above to predict on test set.
5. Report on the evaluation metrics given below:

