Machine Learning

*Analyzing problems, designing a machine learning solution, implementing ML algorithms, and evaluating data sets

1st Saksham Saxena
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

2nd Rushabh Doshi
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

3rd Luke Beaulieu
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

4th Utkarsh Patel
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

***Abstract*—This paper presents the analysis and evaluation on data sets using machine learning techniques. Our method in this assignment consist of working with three data sets and implement ML algorithms on them to understand more about the field of machine learning and its applications in the real world. The three three data sets worked on in this project was Boston housing, breast cancer, and Forest Covertype.**

***Index Terms*—introduction, methodology, result, discussion, conclusion**

I. Introduction

The goal of this final project is to show that we have learned something in the class. It is an opportunity for us to explore ideas that we have see in the lectures and assignments and extend them. This project is a first steps towards research in machine learning or its applications. We will analyze problems, design a machine learning solution, implement ML algorithms, and evaluate them on three data sets (one for classification and one for regression from Small Data Sets and one data set either classification or regression from Large Data Sets).

II. Problem and Data sets

A. Breast Cancer, logistic regression

Breast cancer data set is a classification data set. The output variable of the cancer data was either malignant or benign .

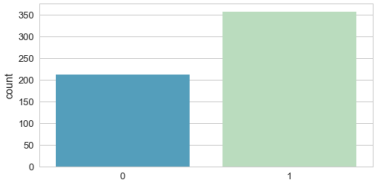

Fig. 1. 0: Malignant 1: Benign

Since this is a classification data set, it was important to go back and look at which methods I could and could not use to work with this data set. I chose logistic regression model in this case. The idea of logistic regression is that there is an optimal decision boundary that separates the two classes of cancer. From class, we learned methods to obtain that optimal decision boundary using cost function.

For the given training data, we want to find parameters Θ that are most likely by maximizing $L(\Theta)$. This is the maximum likelihood estimator.

Cost function

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \text{Cost}(h(\mathbf{x}^{(i)}), y^{(i)})$$

$$\begin{aligned} L(\boldsymbol{\theta}) &= p(y^{(1)}, \ldots, y^{(n)} \mid \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}; \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} p(y^{(i)} \mid \mathbf{x}^{(i)}; \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \sigma(\boldsymbol{\theta}^T \mathbf{x}^{(i)})^{y^{(i)}} (1 - \sigma(\boldsymbol{\theta}^T \mathbf{x}^{(i)}))^{1-y^{(i)}} \end{aligned}$$

Fig. 2. maximum likelihood estimator function

Knowing the likelihood function (above) for this given problem, we look for such Θ that maximizes the probability of obtaining the data we have. To find the Θ, we have an optimization algorithm to find that Θ.

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

Fig. 3. gradient descend where n is the learning rate

Going through different learning rates and iterations, this leaves us with an optimal line that separates the two classification of breast cancer. *n is the learning rate and it determines how big of a step we get to the minimum point in the gradient descend. -Too big of a learning rate causes drastic updates which leads to divergent behavior. -Too small of a learning rate and it may take a lot of iterations to get to the minimum point.

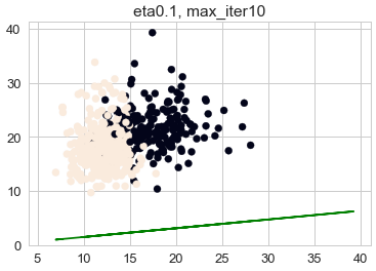

Fig. 4. Small learning rate and iteration

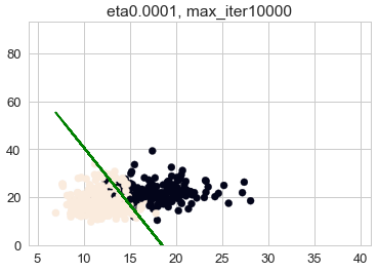

Fig. 5. Good learning rate and iteration

Although I did not have enough time to find the numerical of the errors and accuracy, that is what I would have liked to do if i had more time.(doable using train and test using a 40:60 split model)

The logistic model optimally draws the line between two classes of data, so these results (graphs) do make sense.

B. Breast Cancer, Support Vector Machine

Support Vector Machine or SVM is another method for classification problems such at breast cancer. This works efficiently on datasets that are linearly separable, and so what SVM attempts to do is separate the data by a 'separating hyperplane'. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. [wiki].

The methodology for SVM was to setup the data frame and use a train test 40:60 split. Then using the support vector classifier, or SVC model, I went to predict the accuracy of the model, how well it maps new data or examples for this data.

	precision	recall	f1-score	support
0	0.00	0.00	0.00	84
1	0.63	1.00	0.77	144
avg / total	0.40	0.63	0.49	228

Fig. 6. Prediction with no adjustment

Obviously this was not a good prediction without training, adjustment, and normalizing the data. The model is fairly inaccurate.

For the parameters of the SVC model, we want to use **grid search** to find the best parameter, specifically C and Γ. Grid search runs the same loop with cross-validations to find the best parameters. Once it has found the best combinations, it runs fit on all the data passed to build a single model using the best parameter setting.

```
In [119]: grid.best_params_

Out[119]: {'C': 10, 'gamma': 0.0001, 'kernel': 'rbf'}

In [122]: grid.best_estimator_

Out[122]: SVC(C=10, cache_size=200, class_weight=None, coef0=0.0,
            decision_function_shape='ovr', degree=3, gamma=0.0001, kernel='rbf',
            max_iter=-1, probability=False, random_state=None, shrinking=True,
            tol=0.001, verbose=False)
```

Fig. 7. best parameters & estimator

You can pull out the best parameters and estimators from grid search. With the adjusted parameters and normalization of the train test split data, we get a better accuracy for the model.

```
In [127]: print(classification_report(y_test,grid_predictions))

             precision    recall  f1-score   support

          0       0.99      0.90      0.94        84
          1       0.95      0.99      0.97       144

avg / total       0.96      0.96      0.96       228
```

Fig. 8. Prediction with adjustment + normalization

The SVC model does make sense in terms of mapping an example to a class with an accuracy of 96%. With more time, I would extend this project to set up more classification models, including Neural Networks and K-Mean Clustering. With these different models, an obvious thing to do is compare and contrast the different models. An entire different report on the comparison would include errors margins, efficient margins and more data to describe the different techniques used.

On the side, this project could extend to doing research and problem solving in other areas aside from the data sets we worked on. I would have loved to do a research project on how students past grades affect future classes, but that is for the future and a different time.