Implementing Logistic Regression and Adaline on Various Datasets

Mohammed Momin

1. Introduction

Logistic Regression is one of the most popular algorithms used for classification. This is due to its easy implementation and good performance on linearly separable data classes. There are two types of logistic regression problems: binary and multiclass. In this project I have explored both types on different datasets. The Adaptive Linear Neuron or Adaline is a binary classification algorithm and a single layer neural network. The Adaline algorithm works very similar to the logistic regression algorithm but has some distinct differences which include a different activation function and a discrete output rather than a continuous one.

Hypothesis:

I hypothesize that each algorithm will work very well for the binary case due to the nature of the activation function. The multiclass cases may have a lower accuracy due to the added complexity of the model. I believe the more classes the dataset may have the lower accuracy the algorithms will have. Another thing to consider is the size of the dataset. I think the larger datasets would perform better since there would be more data to train on and hence a more learned model.

1. Experimental Approach

Logistic Regression:

In the training phase, the main goal is to find the optimal weight vector using the gradient descent algorithm. We first initialize an empty weight vector with the size equal to the number of features. It is initialized to random values between -.01 and .01. Another array of the same size is initialized with all zeros to represent the weight change vector. The algorithm then loops through each datapoint in the training set and makes a probability prediction by calculating the weighted sum of the feature values and inputting that value into the activation function which is the sigmoid function:

$$f_w(x) = \frac{1}{1 + e^{-w \cdot x}}$$

If the function is greater than or equal to .5 then the prediction is class 1 (positive class) otherwise its class 0 (negative class).

After that, the gradient of the cost function is evaluated which in this case is the cross-entropy equation. The actual class value and the predicted class value is plugged in to find the gradient value. This value is added to the weight change vector. After looping through each datapoint in the training set, we multiply the weight change vector by the learning rate which then gets added to the weight vector itself. The algorithm is then run again with this new weight vector and it continues to run until the weights stop changing or we reached the max iterations threshold. In the testing phase, we use the final weight vector from the training phase to make predictions on the given testing dataset via the sigmoid function.

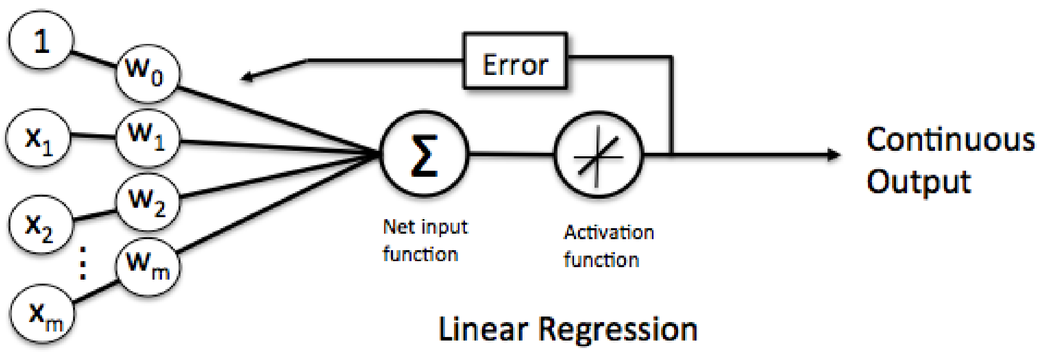
The above experimental approach is the algorithm for the binary class logistic regression problem; however, it would not work for a multiclass dataset since the predictions are only for a positive and negative class. In order to handle a multiclass problem, I use the one-vs-all strategy in which the algorithm makes separate probability predictions for each class and the prediction is the class with the highest probability.

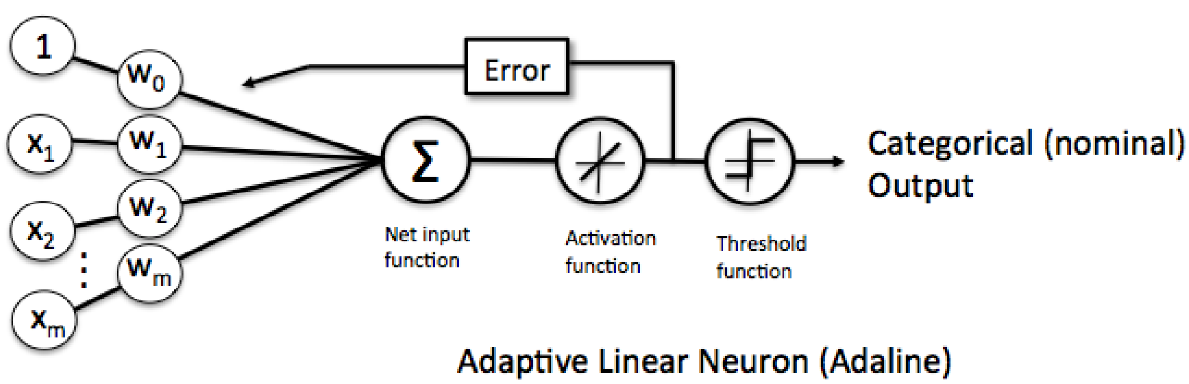
Adaline:

The Adaline algorithm works almost the same as the logistic regression algorithm but with a different activation function and the output being categorical rather than continuous. The activation function for adaline is :

$$f_w(x) = w^T x$$

Visually, this is how the key differences in calculation of output between adaline and logistic regression would look:





[Picture from "Understanding Logistic Regression"]

Like logistic regression, I also used the one-vs-all strategy to handle the multiclass problem for Adaline since its basic algorithm is only for the binary case.

Datasets and preprocessing:

1. Breast Cancer
   This breast cancer databases was obtained from the University of Wisconsin. The feature values were changed to binary values for the algorithms to handle the data better. If the feature value was greater than its median, then the value was changed to 1. If the feature value was less than its median, the value was changed to 0. For the 16 missing attribute values, the "?" was changed to a random number (0 or 1)
2. Glass

The study of classification of types of glass was motivated by criminological investigation. The feature values were changed to binary values for the algorithms to handle the data better. If the feature value was greater than its median, then the value was changed to 1. If the feature value was less than its median, the value was changed to 0.

3. Iris
   The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. The feature values were changed to binary values for the algorithms to handle the data better. If the feature value was greater than its median, then the value was changed to 1. If the feature value was less than its median, the value was changed to 0.
4. Soybean (small)
   A small subset of the original soybean database. The feature values were changed to binary values for the algorithms to handle the data better. If the feature value was greater than its median, then the value was changed to 1. If the feature value was less than its median, the value was changed to 0.
5. Vote
   This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the Congressional Quarterly Almanac. All of the "y" was changed to 1 and all of the "n" was changed to 0. The "?" was changed to a random number (0 or 1).

2. Results

1. Breast Cancer Dataset:
   a. Logistic Regression
      Accuracy for all 5 folds: [0.95035461, 0.95, 0.95, 0.96402878, 0.96402878]
      Average Accuracy: 95.56824327771827%

```
Learned Model:
                      Bias Clump Thickness Uniformity of Cell Size  \
Benign weights     3.67035        -2.38503                -1.22494
Malignant weights -3.67035         2.38503                 1.22494

                  Uniformity of Cell Shape Marginal Adhesion  \
Benign weights                    -1.57873          -2.01721
Malignant weights                  1.57873           2.01721

                  Single Epithelial Cell Size Bare Nuclei Bland Chromatin  \
Benign weights                       -1.19117    -3.08803        -1.27065
Malignant weights                     1.19117     3.08803         1.27065

                  Normal Nucleoli Mitoses
Benign weights          -0.977592  -2.5233
Malignant weights        0.977592   2.5233
```

   b. Adaline
      Accuracy for all 5 folds: [0.90780142, 0.95, 0.89285714, 0.9352518, 0.96402878]
      Average Accuracy: 92.99878273672854%

Learned Model:

	Bias	Clump Thickness	Uniformity of Cell Size
Benign weights	0.949492	-0.276092	-0.125888
Malignant weights	0.0505074	0.276125	0.125598

	Uniformity of Cell Shape	Marginal Adhesion
Benign weights	-0.117151	-0.149026
Malignant weights	0.117406	0.149085

	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin
Benign weights	-0.117739	-0.344981	-0.106922
Malignant weights	0.117804	0.34489	0.106903

	Normal Nucleoli	Mitoses
Benign weights	-0.106475	0.0551341
Malignant weights	0.106504	-0.0551634

2. Glass Dataset
    a. Logistic Regression
       Accuracy for all 5 folds: [0.6, 0.56818182, 0.55813953, 0.64285714, 0.675]
       Average Accuracy: 60.88356991845364%

Learned Model:

	Bias	Refractive Index
building_windows_float_processed weights	-0.453714	-0.263651
building_windows_non_float_processed weights	-0.977078	0.583227
containers weights	-7.99696	-0.434558
headlamps weights	-14.9581	-3.52481
tableware weights	-7.14205	1.74145
vehicle_windows_float_processed weights	-1.62573	-1.14272

	Sodium	Magnesium	Aluminum
building_windows_float_processed weights	-0.359652	1.1402	-2.30407
building_windows_non_float_processed weights	-0.833929	0.339368	1.07383
containers weights	-9.76518	-5.84202	11.8024
headlamps weights	9.75027	-7.94542	-0.620679
tableware weights	8.6504	-9.08615	2.20358
vehicle_windows_float_processed weights	1.06277	-0.0102368	-1.19048

	Silicon	Potassium	Calcium
building_windows_float_processed weights	0.448162	0.145033	0.175751
building_windows_non_float_processed weights	-0.304939	0.0994436	-0.676297
containers weights	-4.86451	-3.72651	5.39815
headlamps weights	3.64732	4.81128	3.50508
tableware weights	0.706055	-7.89915	-4.60435
vehicle_windows_float_processed weights	-1.32551	0.43537	0.0375538

	Barium	Iron
building_windows_float_processed weights	-0.86819	0.108631
building_windows_non_float_processed weights	-0.957885	0.765459
containers weights	-1.68703	-3.94665
headlamps weights	7.02332	-3.68074
tableware weights	-10.4858	-6.18164
vehicle_windows_float_processed weights	-0.624931	-0.123107

    b. Adaline
       Accuracy for all 5 folds: [0.4888889, 0.63636364, 0.62790698, 0.5952381, 0.525]
       Average Accuracy: 57.467951944696125%

Learned Model:

	Bias	Refractive Index	Sodium	Magnesium	Aluminum	Silicon	Potassium	Calcium	Barium	Iron
building_windows_float_processed weights	0.418026	-0.0487949	-0.0841239	0.174099	-0.348089	0.0598545	0.00937144	0.0894426	-0.112927	-0.000491284
building_windows_non_float_processed weights	0.312361	0.180001	-0.106212	0.0777681	0.185157	-0.0147126	-0.0189084	-0.247871	-0.201808	0.148264
containers weights	0.0340907	0.0549383	-0.101323	-0.0528258	0.163908	-0.0287055	-0.00533466	0.0890371	-0.0272858	-0.0893952
headlamps weights	0.0078983	-0.0854053	0.138351	-0.0849243	0.00918147	0.0872543	0.0100895	0.0389817	0.547419	-0.0499512
tableware weights	0.0753776	0.00411689	0.0847273	-0.0842137	0.0307399	0.0224976	-0.0547188	-0.00422969	-0.152022	-0.0272023
vehicle_windows_float_processed weights	0.0831656	-0.0821856	0.0927832	-0.0139594	-0.0265043	-0.0991079	0.0790754	0.0421464	-0.0610567	0.0274198

3. Iris Dataset
    a. Logistic Regression
    Accuracy for all 5 folds: [0.86666667, 0.9, 0.73333333, 0.76666667, 0.8]
    Average Accuracy: 81.33333333333333%

Learned Model:

	Bias	sepal length in cm	sepal width in cm	petal length in cm	petal width in cm
Iris-setosa weights	-0.559869	-5.54843	8.29735	-5.67022	-5.16808
Iris-versicolor weights	0.0653798	1.16841	-1.87935	0.254228	-1.54781
Iris-virginica weights	-10.5239	-0.132648	-0.341594	5.35079	6.26627

    b. Adaline
    Accuracy for all 5 folds: [0.76666667, 0.8, 0.86666667, 0.83333333, 0.73333333]
    Average Accuracy: 80%

```
Learned Model:
                               Bias  sepal length in cm  sepal width in cm  \
Iris-setosa weights        0.480614           -0.200326           0.382175
Iris-versicolor weights    0.543968            0.110597          -0.404582
Iris-virginica weights   -0.0263451           0.0859404          0.0241862

                         petal length in cm  petal width in cm
Iris-setosa weights               -0.267579          -0.168228
Iris-versicolor weights            0.165117          -0.345271
Iris-virginica weights             0.133615           0.486329
```

4. Soybean (small) Dataset
    a. Logistic Regression
       Accuracy for all 5 folds: [0.9, 0.9, 0.88888889, 0.88888889, 1.0]
       Average Accuracy: 91.55555555555554%

```
Learned Model:
               Bias      date plant-stand      precip       temp       hail  \
D1 weights -3.76528   1.15765  0.00701485  0.00941969   -0.55505   -1.59167
D2 weights -4.42298   1.04581 -0.00470579 -0.00104453    1.50523 -0.0166234
D3 weights  7.32559  -1.64475 -0.00909448 -0.00020988  -0.997582   -3.58669
D4 weights  -9.0277  -2.13072 -0.00232337 -0.00660641  -0.600599     3.9087

           crop-hist area-damaged  severity  seed-tmt  ... int-discolor  \
D1 weights  -1.59867     -1.23163 -0.743133  -1.98084  ...     -1.05034
D2 weights -0.543476      2.73345  -2.09769  -1.23437  ...      3.01441
D3 weights  -2.31375     -3.20665  -2.22747  0.637527  ...     -3.12365
D4 weights   2.82277    -0.201081   2.83174  0.478432  ...    -0.963741

           sclerotia  fruit-pods       fruit         seed  mold-growth  \
D1 weights  -1.03973  0.00767834  0.00952958   -0.0080872    0.0059018
D2 weights   3.02295 -0.00264798  0.00358416   0.00267143   -0.0027266
D3 weights  -3.11346  0.00751764  0.00165538  0.000166737   0.00624337
D4 weights -0.966937 -0.00408704 -0.00664828   0.00248675 -0.000821415

           seed-discolor   seed-size  shriveling    roots
D1 weights   -0.00347581  0.00797091 -0.00535703 -4.59704
D2 weights    0.00672712 -0.00170688  0.00953383 -2.77005
D3 weights   -0.00593553 -0.00506926  0.00649813 -7.62052
D4 weights    0.00185276   0.0015294 -0.00785796  9.22458
```

    b. Adaline
       Accuracy for all 5 folds: [1.0, 1.0, 0.88888889, 1.0, 1.0]
       Average Accuracy: 97.77777777777779%

```
Learned Model:
                 Bias       date plant-stand       precip        temp  \
D1 weights   0.148726   0.086103  0.00405332  0.000838724  -0.0307614
D2 weights  0.0288639  0.0349214 -0.00656794  -0.00626186   0.0217641
D3 weights   0.646612 -0.0796257 -0.00857368   0.00713734 -0.00728879
D4 weights  0.0431591 -0.0263343   -0.000288   0.00485627  -0.0116093

                  hail   crop-hist area-damaged    severity    seed-tmt  ...  \
D1 weights -0.00870158 -0.00815136   -0.0502705  -0.0131745  -0.0735418  ...
D2 weights  -0.0153555 -0.00117249     0.137152   0.0161663 -0.00174452  ...
D3 weights   -0.044404  -0.0356256   -0.0596476 -0.00892799    0.124318  ...
D4 weights   0.0464014   0.0421473    0.0187642   0.0329004  -0.0241576  ...

           int-discolor  sclerotia  fruit-pods        fruit        seed  \
D1 weights   -0.0311872 -0.0294643   0.0069293  -0.00632993  0.00737605
D2 weights      0.26059   0.267958  0.00280568  0.000523899  0.00721785
D3 weights    -0.189256  -0.187964  0.00191663   0.00330122 0.000610038
D4 weights   -0.0239206 -0.0155619 -0.00853825  -0.00107726  -0.00416969

           mold-growth seed-discolor   seed-size  shriveling      roots
D1 weights  0.00572097     0.0092764  0.00660567 -0.00799377  -0.166694
D2 weights -0.00678365   -0.00851943  0.00242301 -0.00706406 -0.0298783
D3 weights -0.00574538    0.00921072  0.00119603 -0.00277684  -0.450046
D4 weights  -0.0067297   -0.00276615 -0.00535934  0.00771627   0.775593
```

5. Vote Dataset
   a. Logistic Regression
      Accuracy for all 5 folds: [0.9545455, 0.977273, 0.90804598, 0.976744, 0.906977]
      Average Accuracy: 94.47170178124468%

```
Learned Model:
                      Bias handicapped-infants water-project-cost-sharing  \
democrat weights    11.3555            -1.63709                    1.34192
republican weights -11.3555             1.63709                   -1.34192

                   adoption-of-the-budget-resolution physician-fee-freeze  \
democrat weights                             5.59574             -12.2073
republican weights                          -5.59573              12.2072

                   el-salvador-aid religious-groups-in-schools  \
democrat weights          -4.32935                      1.9773
republican weights         4.32934                     -1.9773

                   anti-satellite-test-ban aid-to-nicaraguan-contras  \
democrat weights                  -5.27046                  -3.33033
republican weights                 5.27045                   3.33032

                   mx-missile immigration synfuels-corporation-cutback  \
democrat weights      4.69408    -5.24537                      5.06974
republican weights   -4.69407     5.24536                     -5.06973

                   education-spending superfund-right-to-sue     crime  \
democrat weights             -1.44114               -1.69667 -0.410851
republican weights            1.44114                1.69667   0.41085

                   duty-free-exports export-administration-act-south-africa
democrat weights             1.20671                               0.0942851
republican weights          -1.20671                              -0.0942841
```

   b. Adaline
      Accuracy for all 5 folds: [0.977273, 0.95455, 0.954023, 0.9418605, 0.9534884]
      Average Accuracy: 95.62380015066464%

```
Learned Model:
                         Bias handicapped-infants water-project-cost-sharing  \
democrat weights      0.79872          0.00320143                  0.0281335
republican weights   0.185394         -0.00229484                 -0.0273305

                   adoption-of-the-budget-resolution physician-fee-freeze  \
democrat weights                            0.208173            -0.652849
republican weights                         -0.206858             0.653504

                   el-salvador-aid religious-groups-in-schools  \
democrat weights        -0.0689333                   0.0313436
republican weights       0.0775433                  -0.0309787

                   anti-satellite-test-ban aid-to-nicaraguan-contras  \
democrat weights                -0.0832377                 -0.027173
republican weights               0.0841007                 0.0331276

                   mx-missile immigration synfuels-corporation-cutback  \
democrat weights    0.0818377  -0.0582488                     0.109054
republican weights -0.0786083   0.0582604                    -0.108065

                   education-spending superfuned-right-to-sue        crime  \
democrat weights           -0.0438209             -0.00710112 -0.000365067
republican weights          0.0454895              0.00791634   0.00161023

                   duty-free-exports export-administration-act-south-africa
democrat weights           0.0058523                             -0.0107708
republican weights       -0.00446549                              0.0119653
```

3. Discussion

The logistic regression algorithm outperformed the adaline algorithm for 3/5 datasets, but the results were not too drastically different when comparing algorithms. When comparing datasets however, the glass dataset performed very poorly. The soybean dataset had the highest-level accuracy across the board with the Adaline algorithm (97.777%) only misclassifying one datapoint in all of the 5 folds.

4. Conclusion

My hypothesis was correct in that both the logistic regression and adaline algorithms performed very well with the binary class datasets, but the multiclass datasets such as iris and glass did not perform as well. The hypothesis that the higher unique classes resulting in lower classification accuracy held true especially for the glass dataset which performed the worst and had a much higher number of unique classes than the rest. The soybean dataset was an anomaly however in that it had a very high accuracy for a multiclass case and it had the smallest amount of data to learn from. I hypothesized that the larger amount of data would result in a more trained and accurate model. This may still be the case despite the results from the soybean dataset because I believe the inflated accuracy scores for that dataset is a result of overfitting. We would need to test the algorithms on much more soybean data to test that however.

5. References

Kathuria, A. (2019, February 10). Adaptive linear neuron (ADALINE). Retrieved March 18, 2021, from https://arjunkathuria.com/ml/adaline/

Kathuria, A. (2019, February 20). Understanding logistic regression. Retrieved March 18, 2021, from https://arjunkathuria.com/ml/logistic_regression/