Programming Project 4: Adaline and Logistic Regression

Ricca D. Callis RCALLIS1@JH.EDU
Whiting School of Engineering
Engineering for Professionals, Data Science
Johns Hopkins University

Abstract

This project provided students enrolled in an Introduction to Machine Learning course (605.649.83.SU20), at Johns Hopkins University, the opportunity to implement two different linear classifier algorithms: Adaline and Logistic Regression. Adaline predicts a class value (0 or 1) using a linear transformation, whereas Logistic Regression applies a log-odds transformation to the target class in order to predict the likelihood an instance belongs to a class. Here, both algorithms are trained using gradient descent and are applied to 5 data sets obtained from the UCI Machine Learning Repository.

Keywords: Adaline, Logistic Regression, UCI Machine Learning Repository

1 Problem Statement & Hypothesis

A discriminant is a function that takes an input vector, x, and assigns it to one of K classes, denoted C_k. Fisher (1936) originally formulated the Linear Discriminant to solve a 2-class problem, for which the decision surface was a hyperplane. A simple representation of a linear discriminant function is obtained by taking a linear function of the input vector so that

$$y(x) = w^T x + w_0$$

where,

w: $weight\ vector$
w_0: $bias; determines\ location\ of\ the\ decision\ surface$
x: $input\ vector$
$y(x)$: $signed\ measure\ of\ perpendicular\ distance\ r\ of\ point\ x\ from\ the\ decision\ surface$

An input vector, x, is assigned to class C_1 if $y(x) \geq 0$ and to class C_2 otherwise. The corresponding decision boundary is, therefore, defined by the relation $y(x) = 0$, which corresponds to a (D-1)-dimensional hyperplane within the D-dimensional input space.

Multiple Classes. Rao (1948) later generalized linear discriminant to multiple classes. Rather than building a K-class discriminant by combining a number of two-class discriminant functions, we instead consider the use of K-1 classifiers. Each K-1 classifier solves a two-class problem of separating points in a particular class C_k from points not in that class. This is known as one-versus-the-rest classifier.

1.1 This Project

This paper introduces two techniques to find a linear discriminant between classes, based on a set of class attributes. The parameters of the linear discriminant are learned via

gradient descent. The two algorithms utilized for this project were Adaline and Logistic Regression.

1.2 Expectations

Both of these techniques are linear and are expected to perform comparably on any given linear problem. However, both techniques are expected to be unable to learn classifications on non-linearly separable problems unless features are transformed prior to learning. Thus, classification accuracy is expected to be proportional to the degree to which instances are linear separable. Noisy class labels will yield proportionally lower accuracy scores.

It is expected that both Adaline and Logistic Regression will not innately classify discrete-multivalued attributes. Thus, for this assignment, all datasets with discrete-multivalued attributes were transformed using one-hot encoding and then rescaled such that all values ranged between 0 and 1. This rescale allowed for better convergence of the training algorithm.

Logistic Regression assumes variables are independent. Thus, it does not have the ability to identify interactions or correlations between input features. Similar expectations are placed upon Adaline.

2 Description of Algorithm

2.1 Logistic Regression

Discriminant Development. The logistic regression algorithm implemented here creates a linear discriminant between two classes. Given k classes $(C_1, \dots, C_k)$, the logistic regression classifier calculates the probability that a given instance, x, belongs to class C_k, based on it's attributes, according to:

$$logit\ P(C_k|x) = log\frac{P(C_1|x)}{1 - P(C_1|x)}$$

We can write using Bayes' Rule as:

$$log\frac{P(C_k|x)}{1 - P(C_k|x)} = log\frac{P(x|C_k)}{1 - P(x|C_k)} + log\frac{P(C_k)}{1 - P(C_k)}$$

Which becomes a linear discriminant:

$$og\frac{P(x|C_k)}{1 - P(x|C_k)} + log\frac{P(C_k)}{1 - P(C_k)} = w_k^T x + w_{k_0}$$

To normalize the output for each class relative to the likelihoods predicted by the other classes:

$$P(C_k|x) = \frac{exp\left(w_k^T x + w_{k_0}\right)}{\sum_{j=1}^{k} exp\left(w_k^T x + w_{k_0}\right)}$$

Thus,

$$\lim_{w_k^T x + w_{k_0} \to +\infty} P(C_k|x) = 1$$
$$\lim_{w_k^T x + w_{k_0} \to -\infty} P(C_k|x) = 0$$

Learning Parameters. Here, Logistic Regression relies on the minimization of the binary cross-entropy:

$$E(w_k^T x|x) = -\sum_t r^t \log y^t + (1 - r^t) log(1 - y^t)$$

where,

$y^t = w_k^T x^t$ $(1st\ column\ of\ X\ is\ a\ vector\ of\ ones\ for\ the\ intercept)$
r^t: $actual\ class\ of\ the\ instance$

Thus, the gradient of the loss function is:

$$\frac{dE}{dw_k} = -\sum_t (r^t - y^t) x^t$$

And the update rule:

$$\Delta w_k = -\eta \frac{dE}{dw_k} = \eta \sum_t (r^t - y^t) x^t$$

$$w_k^{n+1} = w_k^n + \Delta w_k^n$$

Which continues until the convergence criteria is met or the maximum number of iterations is met.

Prediction. To make a prediction, the outputs of each of the *k* discriminants are calculated and the largest is chosen.

2.2 Adaline

Training. Adaline is also a linear discriminant estimated with gradient descent. Unlike Logistic Regression, however, Adaline attempts to minimize the squared distance between it's outputs and the correct value:

$$E(w, x^t) = \sum_t (w^T x^t - r^t)^2$$

where,

$w_k = weights$
$r^t \in \{0,1\}\ is\ the\ correct\ label$

Then, the gradient of the loss function is:

$$\frac{dE}{dw} = \sum_t (w^T x^t - r^t) x^t$$

And, the update rule is:

$$\Delta W = \eta (w^T x^t - r^t) x^t$$
$$W^{n+1} = W^n + \Delta W^n$$

Note that Adaline updates the weights by comparing the real-value output to the actual class label. After we train the model, we use a threshold function to turn the real-valued output into a class label (see below).

Prediction. To make a prediction, the output is discretized to 1 (if > 0.5) or 0 (if < 0.5).

3 Experimental Approach

This analysis was conducted on 5 data sets, each obtained from the UCI Machine Learning Repository:

(1) Breast Cancer Data Set

(2) Glass Data Set

(3) Iris Data Set

(4) Soybean Data Set

(5) House Votes Data Set

For each data set, pre-processing included the elimination of null-values and transformation of all multi-valued discrete features into one-hot encoded variables. Each of the data sets were scaled to [-1, 1], such that all values were divided by the maximum absolute value in the column.

Descriptive statistics were calculated for all features and for each feature grouped by class label.

Five-fold stratified cross-validation was used to estimate the out-of-sample performance on various subsets of the data. And, for each data set, the accuracy of each algorithm's classification was measured.

4 Experiment Results

4.1 Breast Cancer Dataset

Data Description. Classifies tumors as either malignant or benign, based on 10 feature attributes: id number, clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitosis (Wolberg, 1992; Wolberg & Mangasarian, 1990). This is a multivariate data set with 699 instances, where each attribute's instance is represented as a discrete value integer, ranging from 1 to 10.

Data Cleaning & Transformation: All instances with missing data were dropped from the data set (16 rows out of 699). The attribute id number was also dropped from the data set, as it represented a unique identifier that would not serve to teach class attributes. As a simple two-class problem (benign or malignant), with continuous-valued inputs, no data transformation was needed. The features were scaled using the maxscaler described above.

Exploratory Data Analysis & Results: Roughly 65.47% of the actual data instances were classified as benign, where as roughly 34.53% of the actual data instances were classified as malignant (see Figure 1). Each feature was also described by class (descriptive statistics included mean, standard deviation, minimum, maximum, range, 1st quartile, median, and 3rd quartile) and were plotted using a box-plot. The results of the experiment are shown in Table 1.

Algorithm	Accuracy
Baseline	65.01%
Logistic Regression	95.75%
Adaline	96.20%

Table 1. Breast cancer dataset experiment results.

4.2 Glass Dataset

Data Description: Classifies origin of broken glass, based on 10 feature attributes of the broken shards: id number, refractive index, sodium, magnesium, aluminum, silicon, potassium, calcium, barium, and iron (German, 1987). The class attribute has 6 classifications: building windows float processed, building windows nonfloat processed, vehicle windows float processed, containers, tableware, or headlamp. This is a multivariate data set with 214 instances, where each attribute's instance is represented as a continuous value float.

Data Cleaning & Transformation: All instances with missing data were dropped from the data set (16 rows out of 699). The attribute id number was also dropped from the data set, as it represented a unique identifier that would not serve to teach class attributes. Due to the fact that inputs were all continuous values, a one-vs-all approach was used for the Adaline algorithm. The features were scaled using the maxscaler approach described above.

Exploratory Data Analysis: There were 76 instances classified as 2 (building windows nonfloat processed), 70 instances classified as 1 (building windows float processed), 29 instances classified as 7 (headlamps), 17 instances classified as 3 (vehicle windows float processed), 13 instances classified as 5 (containers), and 9 instances classified as 6 (tableware; see Figure 3). Each feature was also described by class (descriptive statistics included mean, standard deviation, minimum, maximum, range, 1st quartile, median, and 3rd quartile) and were plotted using a box-plot. The results can be found in Table 2.

Algorithm	Accuracy
Baseline	35.55%
Logistic Regression	49.90%
Adaline	52.27%

Table 2. Glass dataset experiment results.

4.3 Iris Dataset

Data Description: Classifies Iris species (Iris Setosa, Iris Versicolour, or Iris Virginica) based on 4 feature attributes from leaf measurements: sepal length, sepal width, petal length, and petal width (Fisher, 1988). As mentioned, the class attribute has 3 classifications: Iris Setosa, Iris Versicolour, or Iris Virginica. This is a multivariate data set with 150 instances, where each attribute's instance is represented as a continuous value float.

Data Cleaning & Transformation: Due to the fact that inputs were all continuous values, a one-vs-all approach was used for the Adaline algorithm. The features were scaled using the maxscaler approach described above.

Exploratory Data Analysis: There were 50 instances classified as Iris Versicolor, 50 instances classified as Iris Virginica, and 49 instances classified as Iris Setosa (see Figure 5). Each feature was also described by class (descriptive statistics included mean, standard deviation, minimum, maximum, range, 1st quartile, median, and 3rd quartile) and were plotted using a box-plot. The results can be found in Table 3.

Algorithm	Accuracy
Baseline	32.87%
Logistic Regression	96.64%
Adaline	83.84%

Table 3. Iris dataset experiment results.

4.4 Soybean Dataset

Data Description: Classifies soybean disease based on 36 feature attributes of the crop: date, plant stand, precipitation, temperature, hail, crop history, area damaged, severity, seed tmt, germination, plant growth, leaves, leafspots-halo, leafspots-marg, leafspot size, leaf shred, leaf malf, leaf mild, stem, lodging, stem cankers, canker-lesion, fruiting bodies, external decay, mucelium, int-discolor, sclerotia, fruit pods, fruit spots, seed, mold growth, seed discolor, seed size, shriveling, and roots (Michalski, 1987). The class attribute has 4 classifications: D0, D1, D2, or D3. This is a multivariate data set with 47 instances, where each attribute's instance is represented as a discrete value integer.

Data Cleaning & Transformation: Some attributes only had a single value and were dropped due to the fact that these algorithms would be unable to learn classifications that way. All remaining attributes were mapped into one-hot encodings.

Exploratory Data Analysis: There were 17 instances classified as D3, 10 instances classified as D2, 10 instances classified as D1, and 9 instances classified as D0 (see Figure 7). Each feature was also described by class (descriptive statistics included mean, standard deviation, minimum, maximum, range, 1st quartile, median, and 3rd quartile) and were plotted using a box-plot. The results can be found in Table 4.

Algorithm	Accuracy
Baseline	36.83%
Logistic Regression	100%
Adaline	100%

Table 4. Soybean dataset experiment results.

4.5 House Votes Dataset

Data Description: Classifies party (republican or democrat) based on 16 feature attributes of legislation votes: handicapped infants, water project cost sharing, adoption of the budget resolution, physician fee freeze, el Salvador aid, religious groups in schools, anti-satellite test ban, aid to Nicaraguan contras, mx missile, immigration, synfuels corporation cutback, education spending, superfund right to sue, crime, duty free exports, and export administration act south Africa (Congress Quarterly Almanac, 1984). This is a multivariate data set with 435 instances, where each attribute's instance is represented as a discrete Boolean value, where 1 indicates that the congressperson voted for a measure and 0 indicates that the congressperson voted against a measure.

Data Cleaning & Transformation: All of the features were one-hot encoded. Null values were encoded as their own Boolean values.

Exploratory Data Analysis: There were 124 instances classified as Republican and 108 instances classified as Democrat. The results of the experiment are shown in Table 5.

Algorithm	Accuracy
Baseline	53.45%
Logistic Regression	93.99%
Adaline	96.97%

Table 5. House votes dataset experiment results.

5 Behavior of Algorithms

In all experiments conducted for this assignment, both algorithms outperformed baseline. This indicates that each data set contained classes which could be linearly separable.

Training algorithms required some trial and error in order to determine the appropriate learning rate and number of iterations.

The largest observed difference between Logistic Regression and Adaline was found in the Iris data set. It is unclear why. Interestingly, both Logistic Regression and Adaline performed with 100% accuracy in the Soybean data set.

6 Summary

This project provided students the opportunity to implement two linear discriminant algorithms, both of which were trained using gradient descent. The two algorithms, Adaline and Logistic Regression were used on 5 data sets obtained from the UCI Machine Learning Repository. Adaline and Logistic Regression both out-performed base-line class predictions and yielded similar accuracy for each of the 5 data sets.

References

Alpaydın, E. (2020). *Introduction to Machine Learning*. Cambridge, MA: MIT Press.

Congressional Quarterly Almanac (1984). 98th Congress, 2nd Session, 1984, Volume XL: Congressional Quarterly Inc., Washington, D.C. Retrieved June 8, 2020 from https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

German, B. (1987, September 1). Glass Identification Data Set. Retrieved June 8, 2020, from https://archive.ics.uci.edu/ml/datasets/Glass+Identification

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 2: 179-188.

Fisher, R.A. (1988). Iris Data Set. Retrieved June 8, 2020, from https://archive.ics.uci.edu/ml/datasets/Iris

Michalski, R. S. (1987, January 1). Soybean (Small) Data Set. Retrieved June 8, 2020, from https://archive.ics.uci.edu/ml/datasets/Soybean+%28Small%29

Rao, C.R. (1948). The utilization of multiple measurements in problems of biological classification. Journal of the Royal Statistical Society, B, 10, 2: 159-203.

Wolberg, W. (1992, July 15). Breast Cancer Wisconsin (Original) Data Set. Retrieved June 8, 2020, from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceedings of the National Academy of Sciences, 87, 9193-9196.