Re-Implementation of Learning with Noisy Labels: ECE50024 Final Project

Ian Greene [1]

Abstract

Noisy labels in data sets can hinder the performance of machine learning models, making it crucial to develop effective methods to handle such imperfections. In this paper, Learning with Noisy Labels, which contains two robust methods for dealing with noise, is re-implemented. The method of unbiased estimators was implemented using logistic loss and the method of label-dependent costs was implemented using C-SVM. The aim is to demonstrate the successful re-implementation of the methods through a series of data sets from the original paper and on new data sets from the UCI benchmarks. Specific focus is placed on non-uniform noise rates as that is a harder situation for most classifiers and was not thoroughly explored in the original paper. These methods had very high accuracies (greater than 95%) even with non-uniform noise rates of 40% or higher on benchmark data sets.

1. Introduction

A common issue for supervised learning algorithms is being able to learn from data sets with noisy labels. This is a critical issue because noisy labels can result in extremely poor performance for classification algorithms that would otherwise perform perfectly well on the clean data set. Additionally, these algorithms perform better on large data sets, yet these larger data sets are likely to have more noise. Since this is such a common issue, many papers have been written on overcoming this issue. The paper, Learning with Noisy Labels (abbreviated LNL), provides two robust supervised learning methods for when your data set has noisy labels (Natarajan et al., 2013). Specifically, LNL works in the presence of class-conditional random label noise where the noise rates depend on the class label. Additionally, LNL can perform without any assumptions on the true distribution.

This paper re-implements both methods from Learning with Noisy Labels and seeks to confirm the claims and results achieved in the paper. I accomplish this by first reaching the same lemmas and theorems as LNL in the method section. Then both methods are implemented in Python and analyzed on the same data set that LNL utilizes. However, special focus is placed on cases with non-uniform noise rates as these are where the baseline methods struggle the most. These results are all compared in the experiment section along with applying the methods to a new data set on banknotes.

2. Related Work

Of critical importance is what is meant by noisy labels. Learning with Noisy Labels and this paper focus on class-conditional random label noise. Class-conditional random label noise is defined as follows:

$$\rho_{+1} = P(\bar{Y} = -1 \mid Y = +1)$$
$$\rho_{-1} = P(\bar{Y} = -1 \mid Y = +1)$$
$$\rho_{+1} + p_{-1} < 1$$

where $\bar{Y}$ represents a noisy label.

Thus, for n clean samples in a data set represented as $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $Y \in \{\pm 1\}$, the application of the above model for class-conditional noise results in the clean samples being changed to corrupted samples represented as $\left(X_1, \bar{Y}_1\right), \ldots, \left(X_n, \bar{Y}_n\right)$ pairs. The model injects random classification noise independently for each sample. The corrupted samples are what the classifiers are trained on and the classifiers have no access to the clean samples.

Additionally, the data set is built from samples of a distribution. Thus, there is a clean distribution D from which the clean samples are taken and a dirty distribution D_ρ from which the dirty samples are taken.

In Learning with Noisy Labels, they implement two robust supervised learning methods: the method of unbiased estimators and the method of label-dependent costs (Natarajan et al., 2013). For the method of unbiased estimators, in LNL and in this re-implementation, logistic loss ($\ell_{\log}$) was chosen as a representative algorithm. For the method of label-

[1]Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA. Correspondence to: Ian Greene <greenei@purdue.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

dependents costs, in LNL and in this re-implementation, C-SVM was chosen as a representative algorithm.

Logistic loss is defined as:

$$\ell_{\log}(t, y) = \log(1 + \exp(-ty))$$

There are also some important quantities associated with loss functions. One important quantity is the risk which is defined as the expectation of the loss function. For the clean distribution we have:

$$R_{\ell,D}(f) := \mathbb{E}_{(X,Y)\sim D}[\ell(f(X), Y)]$$

The main goal of a learning algorithm is to minimize the risk. However, since the true distribution is not known the risk cannot be calculated. Thus, we use an approximation called empirical risk which is the average of the loss of the training set. Hence, the empirical $\tilde{\ell}$-risk on the noisy samples is:

$$\hat{R}_{\tilde{\ell}}(f) := \frac{1}{n}\sum_{i=1}^{n} \tilde{\ell}(f(X_i), \tilde{Y}_i)$$

Furthermore, as n grows, we expect the empirical risk to be close to the actual risk (expected loss) for the noisy distribution. In other words:

$$\hat{R}_{\tilde{\ell}}(f) := \mathbb{E}_{(X,\tilde{Y})}[\tilde{\ell}(f(X), \tilde{Y})]$$

C-SVM is support vector machine (SVM) that also has penalty parameter (C). Normal SVMs find a hyperplane that maximizes the margin (the distance between the hyperplane and the nearest data points of the class). The parameter 'C' adds a penalty for each misclassified data-point and thus determines the degree to which the algorithm tolerates misclassification errors (Yildrim, 2020). This allows for a "soft-margin" SVM that can often generalize better. Another important aspect of SVMs is they employ the kernel trick which allows them to classify data that is not linearly separable in the original feature but is linearly separable in a higher-dimensional space.

3. Method

As mentioned earlier, in Learning with Noisy Labels, they implement two robust supervised learning methods: the method of unbiased estimators and the method of label-dependent costs (Natarajan et al., 2013). For the method of unbiased estimators, an unbiased estimate is found for the loss function allowing for empirical risk minimization in the presence of class-conditional random noise. For the method of label-dependents costs, a weighted loss function is created where the weights are label-dependent. Some key assumptions have been made for these methods and will be repeated here. First, the label depends on the class label and the label noise is random rather than adversarial. Second, $\rho_{+1} + p_{-1} < 1$. Third, while the noise rates can actually be found through tuning them using cross-validation, it is assumed they are known to the learner in the method section.

3.1. Method of Unbiased Estimators

The method of unbiased estimator primarily revolves around Lemma 1 which defines the unbiased estimator of the loss for noisy labels. We have:

Lemma 1. Let $\ell(t, y)$ be any bounded loss function. Then, if we define,

$$\tilde{\ell}(t, y) := \frac{(1 - \rho_{-y})\,\ell(t, y) - \rho_y \ell(t, -y)}{1 - \rho_{+1} - \rho_{-1}}$$

we have, for any t, y, $\mathbb{E}_{\tilde{y}}[\tilde{\ell}(t, \tilde{y})] = \ell(t, y)$.

This lemma is found using the definition of unbiasedness for $\hat{\ell}$:
For every t,

$$\mathbb{E}_{\tilde{y}\sim y}[\tilde{\ell}(t, \tilde{y})] = \ell(t, y)$$

Thus, for the cases $y = +1$ and $y = -1$, we have:

$$(1 - \rho_{+1})\,\tilde{\ell}(t, +1) + \rho_{+1}\tilde{\ell}(t, -1) = \ell(t, +1),$$
$$(1 - \rho_{-1})\,\tilde{\ell}(t, -1) + \rho_{-1}\tilde{\ell}(t, +1) = \ell(t, -1).$$

Solving these equations for $\tilde{\ell}(t, +1)$ and $\tilde{\ell}(t, -1)$ results in:

$$\tilde{\ell}\,(t, +1) = \frac{(1 - \rho_{-1})\,\ell(t, +1) - \rho_{+1}\ell\,(t, -1)}{1 - \rho_{+1} - \rho_{-1}},$$
$$\tilde{\ell}(t, -1) = \frac{(1 - \rho_{+1})\,\ell\,(t, -1) - \rho_{-1}\ell\,(t, +1)}{1 - \rho_{+1} - \rho_{-1}},$$

Replacing +1 and -1 with y:

$$\tilde{\ell}\,(t, y = +1) = \frac{(1 - \rho_{-y})\,\ell(t, y) - \rho_y \ell\,(t, -y)}{1 - \rho_{+1} - \rho_{-1}},$$
$$\tilde{\ell}(t, y = -1) = \frac{(1 - \rho_{-y})\,\ell\,(t, y) - \rho_y \ell\,(t, -y)}{1 - \rho_{+1} - \rho_{-1}},$$

Combining the two equations together to get the overall loss for y:

$$\tilde{\ell}(t, y) := \frac{(1 - \rho_{-y})\,\ell(t, y) - \rho_y \ell(t, -y)}{1 - \rho_{+1} - \rho_{-1}} \checkmark$$

The equation found matches Lemma 1.

Since the loss function is now unbiased:

$$\ell(t, y) = \mathbb{E}_{\tilde{y}}[\tilde{\ell}(t, \tilde{y})]$$

Thus, the empirical $\tilde{\ell}$-risk can be found by finding the average of the loss:

$$\hat{R}_{\tilde{\ell}}(f) := \frac{1}{n}\sum_{i=1}^{n} \tilde{\ell}(f(X_i), \tilde{Y}_i)$$

Finally, by minimizing the empirical risk a good classifier on noisy labels can be found:

$$\hat{f} \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \widehat{R}_{\tilde{\ell}}(f).$$

3.2. Method of Label-Dependent Costs

This method is based around two lemmas:

Lemma 2. Denote $P(Y = 1 \mid X)$ by $\eta(X)$ and $P(\bar{Y} = 1 \mid X)$ by $\tilde{\eta}(X)$. The Bayes classifier under the noisy distribution, $\tilde{f}^* = \operatorname{argmin}_f E_{(X,Y)} \sim D_\rho \left[1_{\{\operatorname{sign}(f(X)) \neq \bar{Y}\}}\right]$ is given by:

$$\hat{f}^*(x) = \operatorname{sign}(\tilde{\eta}(x) - 1/2) = \operatorname{sign}\left(\eta(x) - \frac{1/2 - \rho_{-1}}{1 - \rho_{+1} - \rho_{-1}}\right)$$

This lemma is found is using the definition of the optimal bayes classifier. We know the optimal bayes classifier under D_ρ thresholds $\tilde{\eta}(X) = P(\tilde{Y} = 1 \mid X)$ at $1/2$. Expanding the probabilities for y = 1 and y = -1:

$$\begin{aligned}
\tilde{\eta}(X) &= P(\tilde{Y} = 1, Y = 1 \mid X) + P(\tilde{Y} = 1, Y = -1 \mid X) \\
&= P(\tilde{Y} = 1 \mid Y = 1) P(Y = 1 \mid X) \\
&\quad + P(\tilde{Y} = 1 \mid Y = -1) P(Y = -1 \mid X) \\
&= (1 - \rho_{+1})\, \eta(X) + \rho_{-1}(1 - \eta(X)) \\
&= (1 - \rho_{+1} - \rho_{-1})\, \eta(X) + \rho_{-1}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\operatorname{sign}(\tilde{\eta}(x) - 1/2) &= \operatorname{sign}\left((1 - \rho_{+1} - \rho_{-1})\, \eta(x) + \rho_{-1} - 1/2\right. \\
&= \operatorname{sign}\left(\eta(x) - \frac{1/2 - \rho_{-1}}{1 - \rho_{+1} - \rho_{-1}}\right).
\end{aligned}$$

Lemma 3. The α^*-weighted Bayes optimal classifier under noisy distribution coincides with that of 0-1 loss under clean distribution. In other words:

$$\underset{f}{\operatorname{argmin}} R_{\alpha^*, D_\rho}(f) = \underset{f}{\operatorname{argmin}} R_D(f) = \operatorname{sign}(\eta(x) - 1/2).$$

Furthermore, the "noisy" Bayes classifier is the minimizer of a weighted $0 - 1$ loss. First, the 0-1 loss can be written as 'label-dependent' costs for binary classification as:

$$1_{\{\operatorname{sign}(f(X)) \neq Y\}} = 1_{\{Y=1\}} 1_{\{f(X) \leq 0\}} + 1_{\{Y=-1\}} 1_{\{f(X) > 0\}}$$

Then the α-weighted version of the $0 - 1$ loss where $\alpha \in (0, 1)$ is:

$$U_a(t, y) = (1 - \alpha) 1_{\{y=1\}} 1_{\{t \leq 0\}} + \alpha 1_{\{y=-1\}} 1_{\{t > 0\}}$$

Combining the two lemmas results in the following empirical risk minimization problem for noisy labels:

$$\hat{f}_\alpha = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell_\alpha \left(f\left(X_i\right), \tilde{Y}_i\right).$$

Where ℓ_α is an α-weighted margin loss function of the form:

$$\ell_\alpha(t, y) = (1 - \alpha) 1_{\{y=1\}} \ell(t) + \alpha 1_{\{y=-1\}} \ell(-t)$$

4. Experiment

4.1. Evaluation Metric

Accuracy was used to evaluate and compare the different classifiers. Accuracy was defined as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Accuracy was calculated with respect to the clean distribution.

4.2. Linearly Separable Data Set from LNL

First, a 2D synthetic linearly separable data set was recreated to match the one in Figure 1 of Learning with Noisy Labels (Natarajan et al., 2013). The clean data set is shown below in Figure 1.

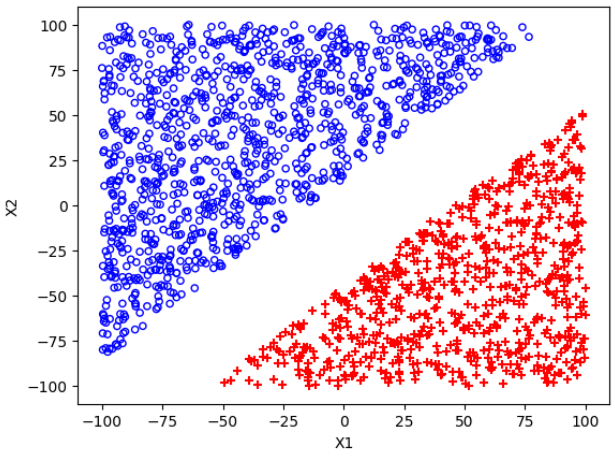

Figure 1. Noise-free linearly separable data set from LNL with 2500 points

Next, random classification noise is added according to noise rates of $\rho_{+1} = 0.2$ and $\rho_{-1} = 0.4$. The data set with noise is shown in Figure 2.

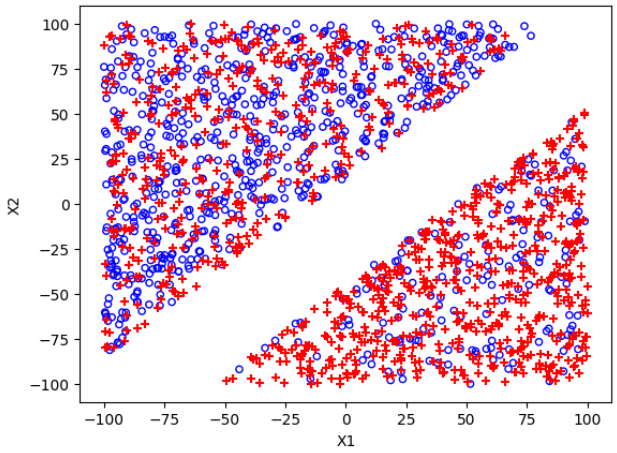

Figure 2. Corrupted linearly separable data set from LNL with $\rho_{+1} = 0.2$ and $\rho_{-1} = 0.4$

Then the corrupted data is classified using the baseline classifier for the method of unbiased estimators which is simply normal logistic regression.

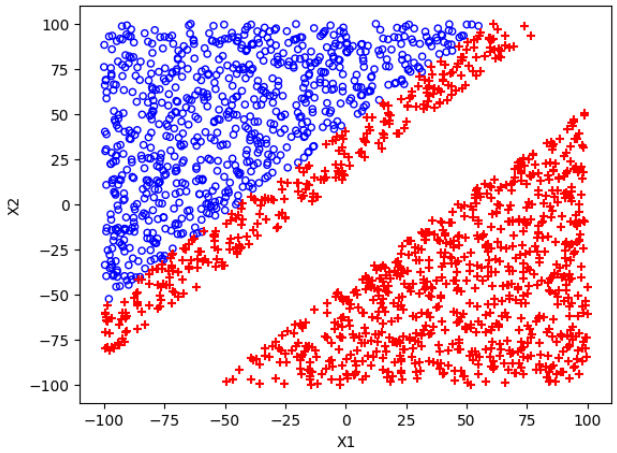

Figure 3. Classification results for the baseline (normal logistic regression) on the linearly separable data set from LNL with $\rho_{+1} = 0.2$ and $\rho_{-1} = 0.4$

After that the corrupted data is classified using the method of unbiased estimators with the logistic loss function where the noise rates are known.

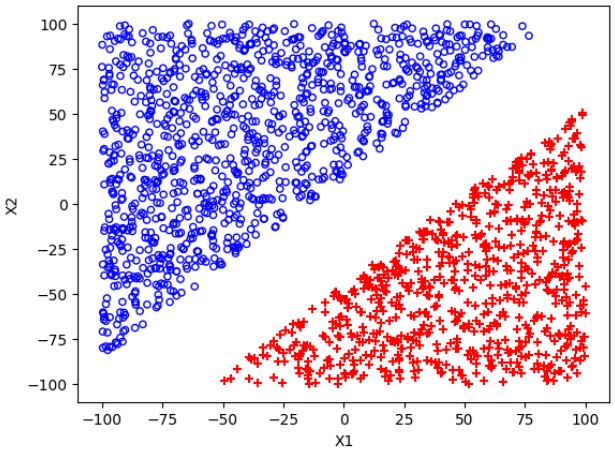

Figure 4. Classification results for the method of unbiased estimators with known noise rates on the linearly separable data set from LNL with $\rho_{+1} = 0.2$ and $\rho_{-1} = 0.4$

Finally, the corrupted data is classified using the method of unbiased estimators when the noise rates are unknown. This is accomplished through k-fold cross validation to tune the possible values for the noise rates to find a classifier without knowing the noise rates.

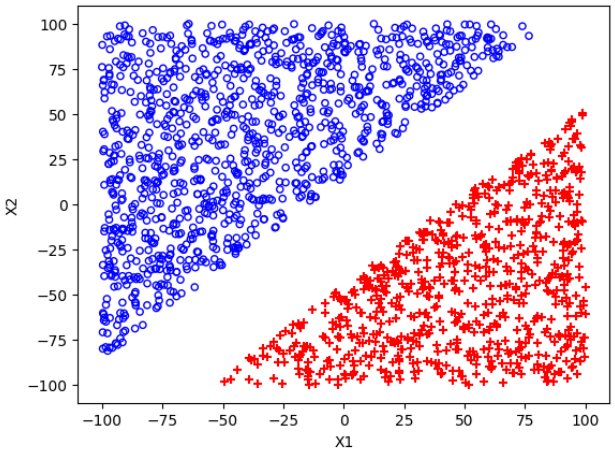

Figure 5. Classification results for the method of unbiased estimators with unknown noise rates on the linearly separable data set from LNL with $\rho_{+1} = 0.2$ and $\rho_{-1} = 0.4$

From Figures 3, 4, and 5 it is clear the method of unbiased estimators successfully classifies noisy data while the baseline struggles. These visual results are backed up by the accuracy for each classifier.

Table 1. Classification accuracy in percent for baseline, method of unbiased estimators with noise rates, and method of unbiased estimators (abbreviated MUE), without noise rates

CLASSIFIER	ACCURACY
BASELINE	83.08
MUE WITH NOISE RATES	100.00
MUE WITHOUT NOISE RATES	100.00

The method of unbiased estimators works very well for the very simple case of a 2D linearly separable data with noise. In fact, MUE without noise rates actually finds the best noise rates to be $\rho_{+1} = 0.2$ and $\rho_{-1} = 0.4$ just as expected. Furthermore, both versions of the method of unbiased estimators have extremely high accuracies of 100.0 especially compared to the baseline's accuracy of 83.08. Overall, the method achieves extremely high accuracies even when faced with high, non-uniform noise rates.

4.3. Banana Data Set from LNL

Next, a 2D synthetic "banana" data set was used to test the method of label-dependent costs. The banana data set is often used for classification as a non-separable data set and was provided from an article by Saravanan Jaichandaran (Jaichandaran, 2017). The clean data set is shown below in Figure 6.

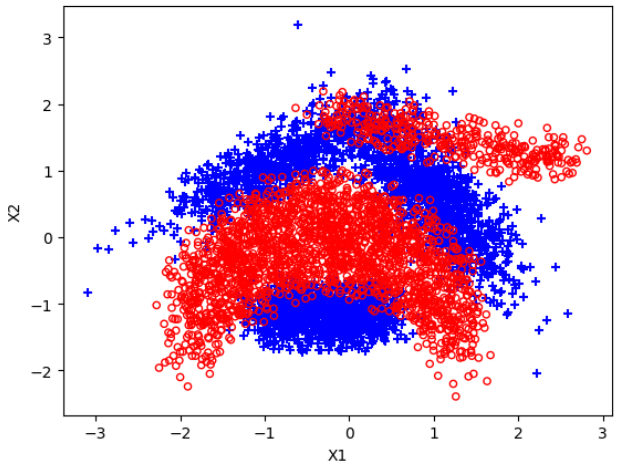

Figure 6. Noise-free non-separable "banana" data set

Next, random classification noise is added according to noise rates of $\rho_{+1} = 0.4$ and $\rho_{-1} = 0.1$. The data set with noise is shown in Figure 7.

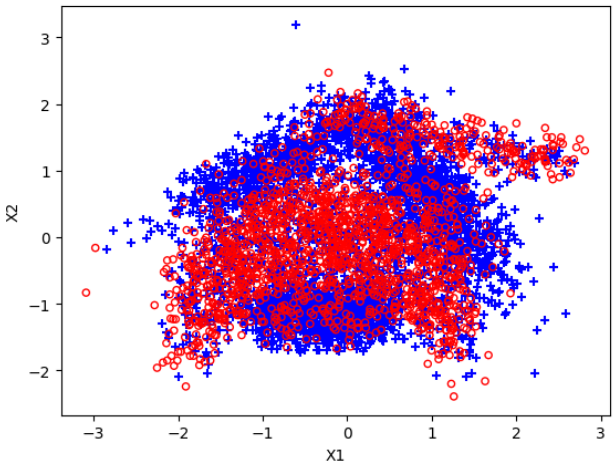

Figure 7. Corrupted non-separable "banana" data set with $\rho_{+1} = 0.4$ and $\rho_{-1} = 0.1$

Then the corrupted data is classified using the baseline classifier for the method of label-dependent costs which is simply a normal SVM.

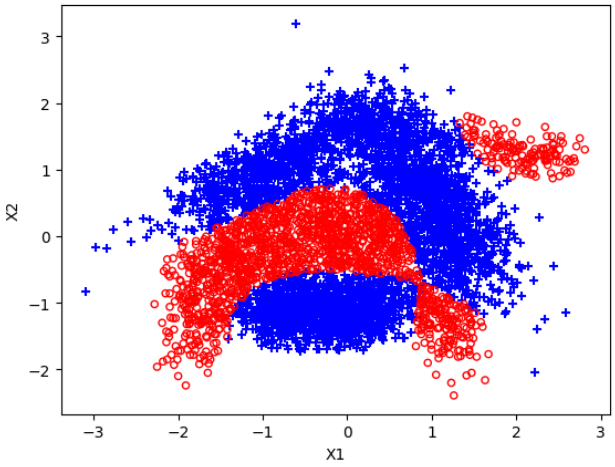

Figure 8. Classification results for the baseline (normal SVM) on the non-separable "banana" data set with $\rho_{+1} = 0.4$ and $\rho_{-1} = 0.1$

After that the corrupted data is classified using the method of label-dependent costs with C-SVM where the noise rates are known.

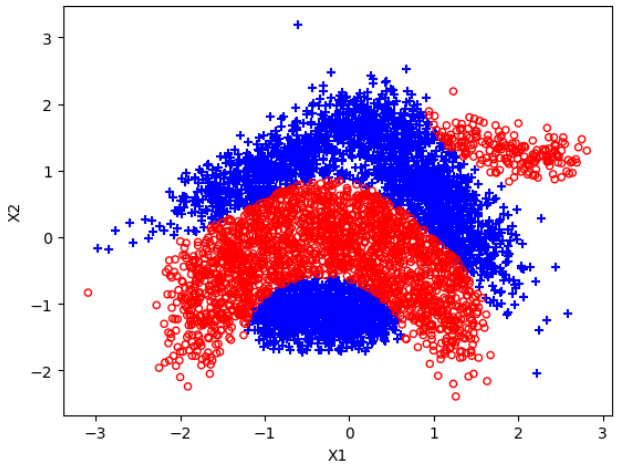

Figure 9. Classification results for the method of label-dependent costs with known noise rates on the non-separable "banana" data set with $\rho_{+1} = 0.4$ and $\rho_{-1} = 0.1$

Finally, the corrupted data is classified using the method of label-dependent costs when the noise rates are unknown. This is accomplished through k-fold cross validation to tune the possible values for α^* to find a classifier without knowing the noise rates.

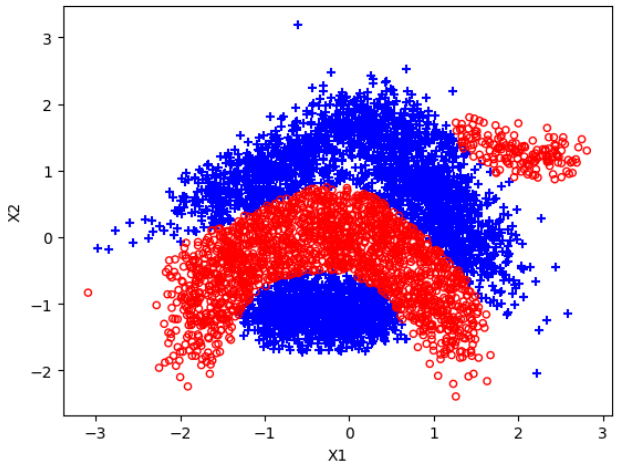

Figure 10. Classification results for the method of label-dependent costs with unknown noise rates on the non-separable "banana" data with $\rho_{+1} = 0.4$ and $\rho_{-1} = 0.1$

From Figures 8, 9, and 10 it is once again clear that the method of label-dependent costs successfully classifies noisy data where the baseline struggles. These visual results are backed up by the accuracy for each classifier.

Table 2. Classification accuracy in percent for baseline, method of label-dependent costs with noise rates, and method of label-dependent costs (abbreviated MLDC), without noise rates

CLASSIFIER	ACCURACY
BASELINE	84.91
MLDC WITH NOISE RATES	89.87
MLDC WITHOUT NOISE RATES	88.81

The method of label-dependent costs works well even for a non-separable "banana" data set. MLDC with noise rates works slightly better than MLDC without noise rates which make sense because the MLDC with noise rates has more information about the problem. Both arrived at a similar α^* values of $\sim$0.4. However, both versions of the method of label-dependent costs have high accuracies of $\sim$89 and perform better than the baseline by $\sim$5 percent. Overall, the method achieves high accuracies even when faced with high, non-uniform noise rates.

4.4. Breast Cancer Data Set from UCI

Both methods were then tested on the UCI benchmark data set Breast Cancer with noise rates of $\rho_{+1} = 0.1$ and $\rho_{-1} = 0.4$ (Patrício et al., 2018).

Table 3. Classification accuracy in percent on the UCI benchmark data set Breast Cancer for baselines, method of unbiased estimators (abbreviated MUE) with and without noise rates, and method of label-dependent costs (abbreviated MLDC) with and without noise rates (abbreviated NR)

CLASSIFIER	NOISE RATES	ACCURACY
MUE BASELINE	$\rho_{+1} = 0.1, \rho_{-1} = 0.4$	52.59
MUE W/ NR	$\rho_{+1} = 0.1, \rho_{-1} = 0.4$	68.97
MUE W/O NR	$\rho_{+1} = 0.1, \rho_{-1} = 0.4$	71.55
MLDC BASELINE	$\rho_{+1} = 0.1, \rho_{-1} = 0.4$	56.90
MLDC W/ NR	$\rho_{+1} = 0.1, \rho_{-1} = 0.4$	79.31
MLDC W/O NR	$\rho_{+1} = 0.1, \rho_{-1} = 0.4$	79.31

Once again it is clear that the methods work well even for a non-separable, real-life, complex data set. Both methods have high accuracy and perform significantly better than the baselines by $\sim$20 percent. Overall, the methods achieve high accuracies even when faced with high, non-uniform noise rates.

4.5. New Banknote Data Set from UCI

Finally, both methods were then tested on the UCI benchmark data set Banknote Authentication with noise rates of $\rho_{+1} = 0.5$ and $\rho_{-1} = 0.1$ (Lohweg, 2012).

Table 4. Classification accuracy in percent on the UCI benchmark data set Banknote Authentication for baselines, method of unbiased estimators (abbreviated MUE) with and without noise rates, and method of label-dependent costs (abbreviated MLDC) with and without noise rates (abbreviated NR)

CLASSIFIER	NOISE RATES	ACCURACY
MUE BASELINE	$\rho_{+1} = 0.5, \rho_{-1} = 0.1$	71.64
MUE W/ NR	$\rho_{+1} = 0.5, \rho_{-1} = 0.1$	95.34
MUE W/O NR	$\rho_{+1} = 0.5, \rho_{-1} = 0.1$	95.19
MLDC BASELINE	$\rho_{+1} = 0.5, \rho_{-1} = 0.1$	81.78
MLDC W/ NR	$\rho_{+1} = 0.5, \rho_{-1} = 0.1$	99.27
MLDC W/O NR	$\rho_{+1} = 0.5, \rho_{-1} = 0.1$	99.93

Table 4 serves to enforce that the methods work well even for a non-separable, real-life, complex data set. Both methods have extremely high accuracy and perform significantly better than the baselines by $\sim$15 percent. Overall, the methods achieve high accuracies even when faced with high, non-uniform noise rates. In this case, the second method seems to perform better than the first method. However, this is likely because the first method is based off logistic loss and thus can only linearly classify data. However, the second method uses SVM which can have a kernel allowing it to classify more than just a linear line.

4.6. Access to Implementation

The whole implementation can be found at:

https://github.com/iangreene/ECE50024_FinalProject

5. Conclusion

The paper Learning with Noisy Labels was successfully re-implemented. Additionally, the two methods specified in the original paper were successfully applied to both the benchmark data sets in the original paper as well as to a new UCI benchmark data set. Both methods resulted in high accuracy during classification and were competitive with the results obtained in the original paper. They also still had high accuracy even when high levels of non-uniform noise was applied. In the future, models for more diverse noise could be considered. For example, methods for dealing with adversarial noise where the noise is not random or where the noise is not class-based. Finally, the original paper provided performance guarantees and convexity guarantees for their methods which this paper did not check but in the future could be examined more closely.

Acknowledgements

This paper was based off topics taught by Prof. Guo in his lectures for ECE50024. His expertise and guidance were instrumental in navigating the challenges associated with implementing this paper.

References

Jaichandaran, S. Standard classification with banana dataset. Kaggle, December 22 2017. URL https://www.kaggle.com/code/saranchandar/standard-classification-with-banana-dataset/input. Retrieved May 4, 2023.

Lohweg, V. UCI machine learning repository, 2012. URL http://archive.ics.uci.edu/ml.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/3871bd64012152bfb53fdf04b401193f-Paper.pdf.

Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., and Caramelo, F. Using resistin, glucose, age and bmi to predict the presence of breast cancer. *BMC Cancer*, 18(1), 2018.

Yildrim, S. Hyperparameter tuning for support vector machines - c and gamma parameters. Towards Data Science, June 2020. URL https://towardsdatascience.com/hyperparameter-tuning-for-support-vector-machines