IE 506: Machine Learning - Principles and Techniques **Jan-May 2023**

End-term Project Report:Class-Prior Estimation From Positive Labelled data

Team Name: Deep Minds-1 *Team Members: Aayushi Patil, Aakash Roy*

Abstract

In this project, we have considered the problem of learning a classifier using only positive and unlabeled samples. We have estimated the class-prior on an unlabeled dataset with additional samples coming only from positive class. We have used penalized L_1- divergence for model fitting to cancel the error caused by the absence of negative samples. We experimentally demonstrate the usefulness of penalized f-divergences method on MNIST hand-written digit dataset.

1 Introduction

In typical classification problems, it is assumed that all training examples have labels. However, in some practical scenarios, only a portion of positive samples may be labeled. This situation arises in tasks such as detecting outliers or novelties, where the labels of some inlier samples are available, or in one-class land-cover identification, where it is necessary to identify land-cover areas belonging to the same class as the labeled instances. This project demonstrates that by incorporating extra samples that exclusively belong to the positive class, it is possible to accurately estimate the class prior of the unlabeled dataset. For this task, we hacve used properly penalized divergences for model fitting to cancel the error caused by the absence of negative samples.

2 Literature Survey

To estimate the class-prior various works has been done earlier. In the study which was done by du Plessis and Sugiyama in 2012, they estimated the class-prior using f-divergences. Their work is as follows: Suppose we have two datasets X and X' which are i.i.d samples from probability distributions with density $p(x|y=1)$ and $p(x)$ respectively:

$$X = \{x_i\}_{i=1}^{n} - \; p(x|y=1), \quad X' = \{x'_j\}_{j=1}^{n'} - \; p(x)$$

where X is a set of samples with a positive class and X' is a set of unlabeled samples, consisting of a mixture of positive and negative samples. The unlabeled dataset is distributed as

$$p(x) = \pi p(x|y=1) + (1-\pi)p(x|y=-1)$$

where $\pi = p(y=1)$ is the class prior. We have to estimate this class-prior π . So, if a mixture of class-input densities

$$q'(x;\theta) = \theta p(x|y=1) + (1-\theta)p(x|y=-1)$$

is fitted on the unlabeled input density $p(x)$ then the class prior can be obtained as :

$$\theta := \arg\min_{0\leq\theta\leq1} \int \left[f\left(\frac{q'(x,\theta)}{p(x)} \right) \right] p(x)dx$$

where $f(t)$ is a convex function with $f(1) = 0$Ḃut this this of estimation requires labeled samples from both the classes.

So, their next study which was done by du Plesssis and Sugiyama in 2014, they have considered that they have labelled samples only from positive class. In this work, they have considered a partial model ,

$$q(x;\theta) = \theta p(x|y=1)$$

which is fitted on the unlabeled input density $p(x)$Ḣere the class prior will be estimated as:

$$\theta^{*} = \arg\min_{\theta} \ PE(\theta)$$

where $PE(\theta)$ denotes the PE divergence from $\theta p(x|y=1)$ to $p(x)$:

$$PE(\theta) = \frac{1}{2} \int f \left(\theta \frac{p(x|y=1)}{p(x)} - 1 \right)^{2} p(x)\, dx$$

This methods works well when only positive labeled samples are available along with unlabeled samples. But the issue with this method of estimation is that , it over estimates the class-prior in the absence f negative labeled samples. To cope up with this problem of over estimation, some modifications are required to be done with this approach of class-prior π estimation., which we will see in the next section.

3 Methods and Approaches

3.1 Work done before mid-term project review

We have seen in the previous section that the methods proposed earlier overestimates the class-prior π in the absence of negative labeled samples. So, to solve this problem of overestimation, we have used (du Plessis and Sugiyama in 2016) properly penalized $f-$ divergences for model fitting to cancel the error caused by the absence of negative labeled samples.

Suppose we have two datasets X and X' which are i.i.d samples from probability distributions with density $p(x|y=1)$ and $p(x)$ respectively:

$$X = \{x_i\}_{i=1}^{n} - \ p(x|y=1), \quad X' = \{x'_j\}_{j=1}^{n'} - \ p(x)$$

where X is a set of samples with a positive class and X' is a set of unlabeled samples, consisting of a mixture of positive and negative samples. The unlabeled dataset is distributed as

$$p(x) = \pi p(x|y=1) + (1-\pi)p(x|y=-1)$$

where $\pi = p(y = 1)$ is the class prior. We have to estimate this class-prior π .They have considered a partial model ,

$$q(x;\theta) = \theta p(x|y=1)$$

which is fitted on the unlabeled input density $p(x)\dot{\text{H}}$ere the class prior will be estimated using penalized $f-$ divergences : Penalized f is given by :

$$\tilde{f}(t) = \begin{cases} f(t) & 0 \leq t \leq 1 \\ \infty & \text{otherwise} \end{cases}$$

Using this penalized f class-prior can be estimated as follows:

$$\theta^* = \arg\min_{0 \leq \theta \leq 1} Div_{\tilde{\mathrm{f}}}(\theta)$$

where

$$Div_{\tilde{f}}(\theta) = \tfrac{\int}{f}\left(\theta \tfrac{p(x|y=1)}{p(x)}\right) p(x)\, dx$$

So, now we have to evaluate this penalized $f-$ divergences, since they are required for estimation of class-prior π.

Direct evaluation of penalized $f-$ divergences

To evaluate penalized $f-$ divergence, they have used the $Fenchel\ duality\ bounding\ technique\ for\ f-divergences$, which is based on $Fenchels' s inequality$:

$$f(t) \geq z - f(z^*)$$

where $f^*(z)$ is the Fenchel dual or convex conjugate defined as

$$f^*(z) = sup_{t'}\ t'z - f(t')$$

Now applying the Fenchel bound in a point wise manner, we obtain,

$$f(\tfrac{\theta p(x|y=1)}{p(x)}) \geq r(x)(\tfrac{\theta p(x|y=1)}{p(x)}) - f^*(r(x))$$

where $r(x) = \frac{p(x|y=1)}{p(x)}$ fulfills the role of z . Now multiplying both sides by $p(x)$ gives,

$$f(\tfrac{\theta p(x|y=1)}{p(x)})p(x) \geq \theta r(x) p(x|y=1) - f^*(r(x))p(x)$$

Now integrating the above equation and selecting the tightest bound and replacing all the integrals with sample averages gives,

$$Div_{\tilde{f}}(\theta) \geq \sup_r \left\{ \tfrac{\theta}{n} \textstyle\sum_{i=1}^{n} r(x_i) - \tfrac{1}{n'} \textstyle\sum_{i=1}^{n'} f^*(r(x'_j)) \right\} \quad ---------(i)$$

Now, we have to estimate the right-hand side of above expression to estimate $Div_{\tilde{f}}$.

Let us consider ,

$$\tilde{f}(t) = \begin{cases} -(t-1) & t \leq 1 \\ c(t-1) & t \geq 1 \end{cases}$$

Consider the linear model for the sake of convenience,

$$r(x) = \sum_{l=1}^{b} \alpha_l \phi_l(x) - 1$$

where $\phi_l(x) = exp(\frac{-||x-c_l||^2}{2\sigma^2})$ are the Gaussian kernel centered at all the sample points c_l. Now, substituting $r(x)$ and $c =$ in the right-hand side of equation (i) and after simplifying, we get

$$\hat{\alpha}_1, \hat{\alpha}2, \ldots, \hat{\alpha}b = \arg\min_{\alpha_1,\ldots,\alpha_b} \sum_{l=1}^{b} \frac{\lambda}{2}\alpha_l^2 - \sum_{l=1}^{b} \alpha_l \beta_l$$

where,

$$\beta_l = \frac{\theta}{n} \sum_{i=1}^{n} \phi_l(x_i) - \frac{1}{n'} \sum_{j=1}^{n'} \phi_l(x'_j)$$

which can be solved as,

$$\hat{\alpha}_l = \frac{1}{\lambda} \max(0, \beta_l)$$

So, finally our penalized L_1- distance, that is the maximizer of the right-hand side of equation (i) is obtained as:

$$\hat{pen}L_1(\theta) = \frac{1}{\lambda} \sum_{l=1}^{b} \max\left(0, \beta_1\right) \beta_l - \theta + 1$$

The class-prior is then selected so as to minimize the above estimator,

$$\theta^* = \hat{\pi} = \arg\min_\theta \hat{pen}L_1(\theta)$$

So ,now the clas label can be assigned as follows:

$$\hat{y} = \begin{cases} 1 & \text{if } \pi\hat{r}(x) \geq 0.5 \\ -1 & \text{otherwise} \end{cases}$$

4 Dataset Details

We have used MNIST handwritten digit dataset for performing experiment of the proposed method. We have selected one digit(1) as the positive class and the remaining digits as negative class(2,3,4,5). The dataset was reduced to 4 dimensions using principal component analysis.

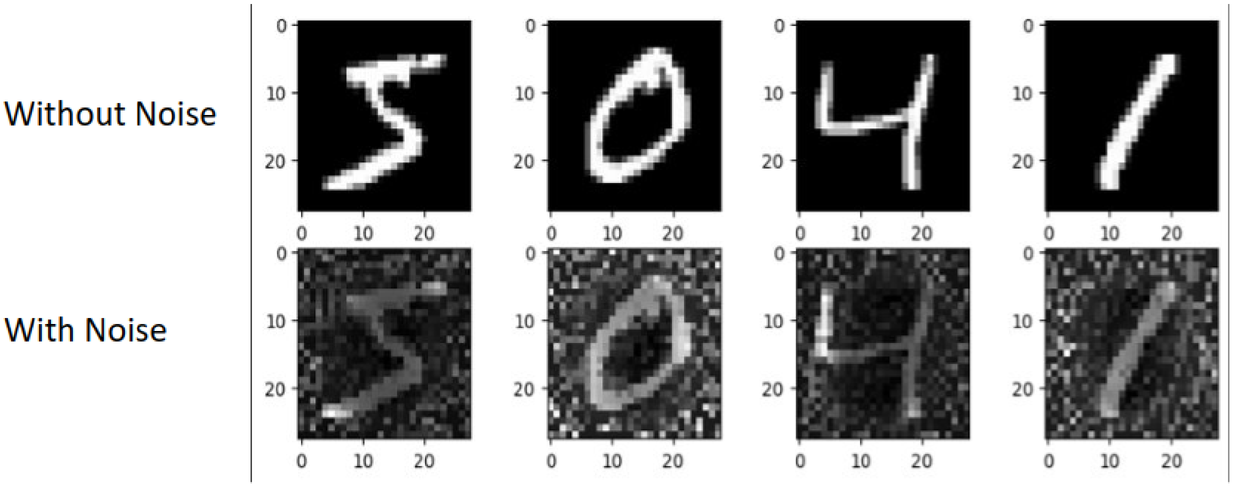

Figure 1: This is an image from a text that uses color to teach music.

5 Experiments

5.1 Work done before mid-term project review

In the initial experiment setup, we utilized a dataset of 400 samples for training our model. We divided the data into 80% for training and 20% for testing. The proportion of positively labeled data was 0.375, and with this configuration, we were able to achieve an accuracy of around 40%.

Now, we have scaled up our experiment and are using a larger dataset of 2000 samples for training. We are maintaining the same training-testing split of 80% and 20%, respectively. With this expanded dataset, we hope to gain deeper insights into the performance of our model and its ability to accurately classify positive and negative labeled data.

Proportion of positive Class Data	λ	σ	Accuracy
0.2	0.001	8.433	76.501
0.4	0.001	8.433	60.001
0.6	0.001	8.433	39.991
0.8	0.001	8.433	19.975

5.2 Work done after mid-term project review

We have a dataset with 2000 training samples. The dataset was tested with and without noise, resulting in different values for the sigma parameter: 1.20476081, 10.84284733, 20.48093385, and 30.11902037 without noise, and 1.36887136, 12.31984221, 23.27081307, and 34.22178393 with noise.

To assess the performance of our model, we used 5-fold cross-validation, splitting the data into 5 equal parts and training the model on 4 parts while testing on the remaining part. We used different values for the lambda parameter, which controls the regularization strength, specifically 0.1, 1, 10, and 100.

In addition, we considered different prior values, namely 0.2, 0.4, 0.6, and 0.8. The prior represents our beliefs about the distribution of the weights before observing the data, and can influence the final results of the model.

Overall, this information provides a glimpse into the process of training and evaluating a machine learning model on a dataset with different parameters and hyperparameters.

6 Results

6.0.1 Without Noise

True Class Prior	Estimated Class prior	λ	σ	Accuracy
0.2	0.2	0.1	10.843	81.25
0.4	0.4	0.1	1.205	84.25
0.6	0.6	0.1	1.205	62.50
0.8	0.6	0.1	1.205	50.00

6.0.2 Without Noise

True Class Prior	Estimated Class prior	λ	σ	Accuracy
0.2	0.2	0.1	11.299	81.25
0.4	0.2	0.1	1.368	83.33
0.6	0.2	0.1	12.319	75.00
0.8	0.6	0.1	1.205	50.00

7 Future Work

- Increase the amount of samples used in the analysis to improve the accuracy of the model.
- Modify the code to run on GPU to reduce the processing time for large data sets.
- Expand the range of prior values used to further optimize the model.
- Evaluate the impact of different hyperparameters on model performance and adjust them accordingly.
- Explore the use of more advanced techniques, such as transfer learning or ensemble models, to improve the accuracy of the model.

8 Conclusion

In the absence of noise, we found that:

- The accuracy was highest for the true class prior value of 0.4.

- All of the estimated class priors were the same as the true class prior for all values except 0.8.
- More accurate class priors resulted in better accuracy.

In the presence of noise, we found that:

- The accuracy was highest for the true class prior value of 0.4.
- All of the estimated class priors were different from the true class prior for all values except 0.2.
- The accuracy was slightly lower compared to the results without noise.

References

[1]M. C. du Plessis, G. Niu, and M. Sugiyama. Analysis of learning from positive and unlabeled data. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 27, pages 703–711, 2014.
[2]M. C. du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. In ICML 2012, pages 823– 830, Jun. 26–Jul. 1 2012.
[3]Christoffel, Marthinus, Gang Niu, and Masashi Sugiyama. "Class-prior estimation for learning from positive and unlabeled data." Asian Conference on Machine Learning. PMLR, 2016.