Learning a Label-noise Robust Logistic Regression: Analysis and Experiments

Jakramate Bootkrajang and Ata Kabán

School of Computer Science, University of Birmingham,
Edgbaston, Birmingham, B15 2TT, UK
{J.Bootkrajang,A.Kaban}@cs.bham.ac.uk

Abstract. Label-noise robust logistic regression (rLR) is an extension of logistic regression that includes a model of random mislabelling. This paper attempts a theoretical analysis of rLR. By decomposing and interpreting the gradient of the likelihood objective of rLR as employed in gradient ascent optimisation, we get insights into the ability of the rLR learning algorithm to counteract the negative effect of mislabelling as a result of an intrinsic re-weighting mechanism. We also give an upper-bound on the error of rLR using Rademacher complexities.

Keywords: label noise, logistic regression, robust learning, gradient ascent optimisation, generalisation error bounds.

1 Introduction

In the context of supervised learning, a classification rule is to be derived from a set of labelled examples. The training sample is assumed to be drawn i.i.d. from some unknown distribution over tuples of the form $(\mathbf{x}, y)$, where $\mathbf{x}$ is an m-dimensional vector that represents a data point in the m-dimensional space and y is its class label assignment. In this paper we will consider linear classifiers. Geometrically, the classification rule is defined by a hyperplane that separates the classes. Regardless of the learning approach used, in the traditional framework of supervised learning, the induction of the classification rule crucially relies on the given class labels. Unfortunately, often in practice there is no guarantee that the class labels are all correct.

Noise in the labels may originate from the subjective nature of the labelling task, noisy communication channels, or lack of information to determine the true label of a given instance, to name just a few of the most common causes. The presence of class label noise in training samples has been empirically reported to deteriorate the performance of the existing classifiers in a broad range of classification problems including biomedical data analysis [7, 13] and image classification [10, 17]. Although, the problem posed by the presence of class label noise is acknowledged, often it gets naively ignored in practice.

There is an increasing research literature that aims to address the issues related to learning from samples with noisy class labels. [4, 5, 15, 14]. Most previous approaches try to detect mislabelled instances based on various heuristics, and very few take a principled modelling approach such as either a generative [10] or a discriminative model [2, 12]. The latter is our focus in this paper. While there is theoretical analysis on the negative effects of label noise for traditional logistic regression [8, 9, 6, 11], an analysis of the good performance of a label-noise robust logistic regression approach has been lacking, and this is what we address here.

2 Robust Logistic Regression

Consider a set of training data $S = \{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^m$ represent an observation and $\tilde{y}_i \in \{0, 1\}$, denotes the given label of $\mathbf{x}_i$. In the classical scenario for binary classification, the log likelihood for logistic regression model is defined as:

$$L(\mathbf{w}) = \sum_{i=1}^{n} \tilde{y}_i \log(\tilde{P}_i^1) + (1 - \tilde{y}_i) \log(\tilde{P}_i^0), \tag{1}$$

where $\tilde{P}_i^k := p(\tilde{y}_i = k|\mathbf{x}_i, \mathbf{w})$. If the labels were presumed to be correct, we would have

$$p(\tilde{y} = 1|\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + e^{(-\mathbf{w}^T\mathbf{x})}}, \tag{2}$$

and whenever this is above 0.5 we would decide that $\mathbf{x}$ belongs to class 1. Here, $\mathbf{w}$ is the weight vector orthogonal to the decision boundary and it determines the orientation of the separating plane. However, when there is label noise present, then it is no longer valid to make predictions in this way. Instead we introduce a latent variable y, to represent the true label, and we rewrite $\tilde{P}_i^k$ as the following:

$$\tilde{P}_i^k = \sum_{j=0}^{1} p(\tilde{y}_i = k|y_i = j)p(y_i = j|\mathbf{x}_i, \mathbf{w}) \tag{3}$$

where, $\gamma_{jk} := p(\tilde{y}_i = k|y_i = j)$ represents the probability that the label has flipped from the true label j to the observed label k. It is worth noting that γ_{jk} is a global value for all observations. These additional parameters form a transition table Γ of class label flippings and these parameters need to be estimated.

The flipping probabilities for a two-class problem are summarised in Table 2. The table will be referred to as the 'gamma matrix' from now on. Hence, instead of predicting the label using the posterior probability of $\tilde{y}$ as in eq. (2), we are now able to calculate the posterior probability of the true label y, and use this to decide that $\mathbf{x}$ belongs to class 1 whenever $p(y = 1|\mathbf{x}, \mathbf{w}) \geq 0.5$.

		$\tilde{y}$	
		0	1
y	0	γ_{00}	γ_{01}
	1	γ_{10}	γ_{11}

Table 1. Probabilistic relationships between given label and true label.

3 Results

3.1 Understanding rLR learning via an interpretation of gradient ascent optimisation

We have seen in the last section that rLR makes use of a latent variable y to model the true label. However, it is not obvious at all why the introduction of this extra variable will lead to a more robust model. In this section we give an intuitive explanation for why this is the case indeed, by looking at the gradient ascent optimisation process for learning the rLR model. The gradient update rule for the j-th iteration is the following:

$$\mathbf{w}^j = \mathbf{w}^{j-1} - \eta \times \mathbf{g} \tag{4}$$

where η is usually referred to as the 'learning rate' and $\mathbf{g} = \nabla_{\mathbf{w}} L(\mathbf{w})$ is the gradient of the likelihood objective function.

Assume a fixed training set S of size n, and assume also that the gamma matrix is known. Let n_1 be the number of points that have been assigned the label $\tilde{y} = 1$ and n_0 the number of points with $\tilde{y} = 0$. Denote by $\mathbf{g}^+$ the terms of the gradient that correspond to points with $\tilde{y} = 1$, and $\mathbf{g}^-$ those of points with $\tilde{y} = 0$, so that $\mathbf{g} = \mathbf{g}^+ + \mathbf{g}^-$. For classical logistic regression without label noise modelling, these terms are:

$$\begin{aligned}
\mathbf{g}^+ &= \frac{\partial \sum_{i=1}^n \tilde{y}_i \log(\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}_i}})}{\partial \mathbf{w}} &&= \sum_{i=1}^{n_1} [p(y_i = 0|\mathbf{x}_i, \mathbf{w}) \cdot \mathbf{x}_i] \\
\mathbf{g}^- &= \frac{\partial \sum_{i=1}^n (1-\tilde{y}_i) \log(\frac{e^{-\mathbf{w}^T\mathbf{x}_i}}{1+e^{-\mathbf{w}^T\mathbf{x}_i}})}{\partial \mathbf{w}} &&= \sum_{i=1}^{n_0} [-p(y_i = 1|\mathbf{x}_i, \mathbf{w}) \cdot \mathbf{x}_i]
\end{aligned} \tag{5}$$

Let us now look at the corresponding two terms of the gradient for our robust logistic regression. From the definition of data likelihood using eqs. (1)-(3) we get:

$$\mathbf{g}^+_{rLR} = \sum_{i=1}^{n_1} \left[\frac{\tilde{y}_i(\gamma_{11} - \gamma_{01})}{\tilde{P}_i^1} \sigma(\mathbf{w}^T\mathbf{x}_i)(1 - \sigma(w^T\mathbf{x}_i)) \cdot \mathbf{x}_i \right]$$

Now, define $\alpha_i = \frac{(\gamma_{11}-\gamma_{01})p(y_i=1|\mathbf{x}_i,\mathbf{w})}{\tilde{P}_i^1}$, and we can rewrite this as:

$$\mathbf{g}^+_{rLR} = \sum_{i=1}^{n_1} [\alpha_i \cdot p(y_i = 0|\mathbf{x}_i, \mathbf{w}) \cdot \mathbf{x}_i] \tag{6}$$

Likewise, define $\beta_i = \frac{(\gamma_{00}-\gamma_{10})p(y_i=0|\mathbf{x}_i,\mathbf{w})}{\tilde{P}_i^0}$, and find:

$$\mathbf{g}^-_{rLR} = \sum_{i=1}^{n_0} [-\beta_i \cdot p(y_i = 1|\mathbf{x}_i, \mathbf{w}) \cdot \mathbf{x}_i] \tag{7}$$

We see that both α_i and β_i act as weighting coefficients. This weighting mechanism adjusts the contribution that particular data points make to the gradient. The weight itself, for example α_i is composed of two components, the first is a global weight: $(\gamma_{11} - \gamma_{01})$; the second is a data point specific weight: $p(y_i = 1|\mathbf{x}_i, \mathbf{w})/\tilde{P}_i^1$. That is, each data point within a class will be weighted equally by the global weight, in accordance with the extent of label noise in that class. In addition, the individual points that have been potentially mislabelled are multiplied by the data-point-specific weighting factor.

Thus, $\mathbf{g}^+_{rLR}$ will take into account the wrong information to a lesser extent. Similar reasoning applies to $\mathbf{g}^-_{rLR}$in a symmetric manner. Through this weighting mechanism, rLR is then able to distinguish between correctly labelled points and mislabelled points. As a consequence, it is expected that rLR will perform better in terms of generalisation error in label noise conditions, and will be able to detect mislabelled instances with high accuracy.

Finally, we find it instructive to look at the behaviour of robust logistic regression in a label-noise free scenario, i.e. when the training labels are actually all correct. In this case $\tilde{y} = y$, $\gamma_{01} = 0$ and $\gamma_{11} = 1$. In consequence $\alpha_i = 1$, and likewise $\beta_i = 1$. Hence is this case we recover $\mathbf{g}^+_{rLR} = \mathbf{g}^+$ and $\mathbf{g}^-_{rLR} = \mathbf{g}^-$ as in classical logistic regression.

3.2 Error analysis

Although there is no framework to analyse the error of predicting the true labels since those are hidden even in the training phase, in this section we derive a bound on generalisation error in predicting $\tilde{y}$ for a new input $\mathbf{x}$. Such bound is informative because it gives a guarantee that the rLR model (that includes the hidden variable y) is able to explain the *observed* data pairs.

We will use Rademacher complexities. Let $l(h, \mathbf{x}, \tilde{y})$ denote the loss of a classifier (or hypothesis) $h \in \mathcal{H}$ on the input point $\mathbf{x}$, where $\mathcal{H}$ is the hypothesis class considered. Define $L_D(h) = \mathbb{E}_{(\mathbf{x},\tilde{y})\sim D} l(h, \mathbf{x}, \tilde{y})$ to be the generalisation error, and $L_S(h) = (1/n) \cdot \sum_{i=1}^n l(h, \mathbf{x}_i, \tilde{y}_i)$ to be its empirical estimate. The Rademacher complexity of the composite function of the hypothesis class $\mathcal{H}$ and a training set S is defined as:

$$R(\mathcal{H} \circ S) = \frac{1}{n} \mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^n} \left[\sup_{h\in\mathcal{H}} \sum_{i=1}^n \sigma_i h(\mathbf{x}_i) \right] \tag{8}$$

If we can obtain the Rademacher complexity for our classifier then we can use the following lemma, which gives an upper bound on the generalisation error of an Empirical Risk Minimisation (ERM) classifier in terms of its Rademacher complexity.

Lemma 1 (Barlett and Mendelson [1]). *For a training set S of size n, let h_s be an ERM hypothesis. Assume that for all $\mathbf{x} \in S$ and $h \in \mathcal{H}$ we have the ρ-Lipschitz loss function l satisfying $|l(h, \mathbf{x}, \tilde{y})| \leq c$ for some positive constant c. Then, $\forall\delta \in (0, 1)$, with probability at least $1 - \delta$*

$$L_D(h_S) - L_S(h_S) \leq 2\rho R(\mathcal{H} \circ S) + c\sqrt{\frac{\log(1/\delta)}{2n}} \tag{9}$$

To apply this lemma, we first define the loss function $l(h, \mathbf{x}, \tilde{y})$ associated with rLR that we will show to be Lipschitz. We then calculate the Rademacher complexity of the hypothesis class of rLR. Finally, the generalisation error bound is obtained by plugging the Rademacher complexity into Lemma 1.

Let us begin with the loss function associated with robust logistic regression. Recall the data log-likelihood in eq. (1). Because of monotonicity of logarithm, maximising the log-likelihood is equivalent to minimising the negative log-likelihood. It then follows that we can define the loss function to be the negative of the data log-likelihood:

$$l(h, \mathbf{x}_i, \tilde{y}_i) = -\tilde{y}_i \log(\tilde{P}_i^1) - (1 - \tilde{y}_i)\log(\tilde{P}_i^0) \tag{10}$$

We are now ready to consider the Rademacher complexity of rLR. It has been shown that the complexity of the hypothesis class of a linear halfspace classifier is bounded as:

Lemma 2 (Adapted from Shalev-Shwartz [16]). *Let $\mathcal{H} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle, \|\mathbf{w}\|_2 \leq \lambda\}$ defines the hypothesis class of linear classifiers. Let $S = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ be a set of vectors in $\mathbb{R}^m$. Then,*

$$R(\mathcal{H} \circ S) \leq \frac{\lambda \max\limits_i \|\mathbf{x}_i\|_2}{\sqrt{n}} \tag{11}$$

Note that rLR can be regarded as a linear classifier because the loss is a function of a linear function of $\mathbf{w}$. Hence the complexity of its hypothesis class is also defined by Lemma 2. Using these, we state and prove our result as the following:

Theorem 1. *Let $h_s \in \mathcal{H}_{rLR}$ be an ERM hypothesis from the class of rLR classifiers, and let l be the loss function defined in eq. (10). Let S be a training set of n examples drawn i.i.d. from an unknown distribution over $\mathbb{R}^m$. We assume that $\|\mathbf{x}_i\|_2 < \infty, \forall i = 1 : n$, and also that γ_{00} and γ_{11} are bounded away from zero. Then, $\forall \delta \in (0, 1)$, with probability at least $1 - \delta$ the following bound holds:*

$$L_D(h_s) \leq L_S(h_s) + 2\frac{\lambda \max\limits_i \|\mathbf{x}_i\|_2}{\sqrt{n}} + c\sqrt{\frac{\log(1/\delta)}{2n}} \tag{12}$$

Proof. We first show that the rLR loss function is a Lipschitz function with Lipschitz constant $\rho = 1$. Note that $\tilde{y}$ can either take value 0 or 1 but not both at the same time, so the loss function decouples into two terms. In order to obtain the Lipschitz constant, we shall show that the absolute value of the derivative of each term is bounded by 1. Without loss of generality, let us consider the first term, $\log(\tilde{P}_i^1)$. Also for convenience, we define $\mathbf{a} = \mathbf{w}^T\mathbf{x}$. Then,

$$\begin{aligned} \tilde{P}_i^1 &= \frac{\gamma_{11}}{1 + e^{-\mathbf{a}}} + \frac{\gamma_{01}e^{-\mathbf{a}}}{1 + e^{-\mathbf{a}}} \quad = \frac{\gamma_{11} + \gamma_{01}e^{-\mathbf{a}}}{1 + e^{-\mathbf{a}}} \\ \frac{\partial \log(\tilde{P}_n^1)}{\partial \mathbf{a}} &= \frac{\partial \log(\gamma_{11} + \gamma_{01}e^{-\mathbf{a}}) - log(1 + e^{-\mathbf{a}})}{\partial \mathbf{a}} \quad = \left| \frac{-\gamma_{01}e^{-\mathbf{a}}}{\gamma_{11} + \gamma_{01}e^{-\mathbf{a}}} + \frac{e^{-\mathbf{a}}}{1 + e^{-\mathbf{a}}} \right| \\ &= \left| \frac{1}{1 + e^{\mathbf{a}}} - \frac{1}{1 + \frac{\gamma_{11}}{\gamma_{01}}e^{\mathbf{a}}} \right| \quad = |f(\alpha)|, \end{aligned} \tag{13}$$

where in the last step we divided through by $e^{-\mathbf{a}}$, and we defined $\alpha := \gamma_{11}/\gamma_{01}$. Since γ_{11} and γ_{01} are probabilities, their values are between 0 and 1. It follows that the domain of α is $[0, \infty)$. Observe that eq. (13) attains its maximum either 1) at its end points $f(0)$ and at the limit $f(\infty)$ or 2) at a point where $f'(\alpha) = 0$. The first case can be easily checked by plugging the extreme values into eq. (13), and we see that $f(0) \in [-1, 0]$ and $f(\infty) \in [0, 1]$. The remaining case can be verified by calculating $f'(\alpha), \alpha \in (0, \infty)$ which turns out to be,

$$f'(\alpha) = \frac{e^{\mathbf{a}}}{(1 + \alpha e^{\mathbf{a}})^2} \tag{14}$$

It can be seen that eq. (14) is non-negative and can only be zero when $\alpha = \infty$, which is the case we have already considered. Hence, we learn that the absolute value of eq. (13) is bounded by 1. It follows that $\log(\tilde{P}_i^1)$ is a 1-Lipschitz function. $\log(\tilde{P}_i^0)$ can be shown to be 1-Lipschitz using similar technique.

Next we need to show that the loss function is bounded. Without loss of generality, let us consider eq. (10) where $\tilde{y}_i = 1$. It can be shown that the term is bounded by some finite number c.

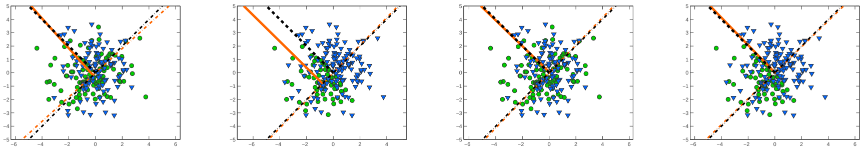
$$\begin{aligned}\Big| -\log(\tilde{P}_i^1)\Big| &= \Big| -\log(\frac{\gamma_{01}e^{-\mathbf{a}}}{1+e^{-\mathbf{a}}} + \frac{\gamma_{11}}{1+e^{-\mathbf{a}}})\Big| \\ &= \big|\log(1+e^{-\mathbf{a}}) - \log(\gamma_{01} + \gamma_{11}e^{-\mathbf{a}})\big| \\ &\leq \big|\log(1+e^{-\mathbf{a}})\big| + \big|\log(\gamma_{01} + \gamma_{11}e^{-\mathbf{a}})\big| \\ &\leq 1 + ||\mathbf{w}||_2||\mathbf{x}||_2 + \big|\log(\gamma_{01} + \gamma_{11}e^{-\mathbf{a}})\big| \end{aligned} \tag{15}$$

where we have used the triangle inequality to get the third line. The last line makes use of a bound on the loss function of standard logistic regression which takes the form $|\log(1 + e^{-\mathbf{a}})| \leq 1 + ||\mathbf{w}||_2||\mathbf{x}||_2$. For the second term we will show that its value is finite, that is we show that an argument to logarithm is never be 0 or infinity. Since $\mathbf{a} = \mathbf{w}^T\mathbf{x}$ is bounded because of the assumptions and, γ_{11} is bounded away from zero, it can be shown that 1) the maximum value of $\gamma_{01} + \gamma_{11}e^{-\mathbf{a}}$ is finite, and 2) its minimum is bounded away from 0. As a consequence, we have that the loss function of rLR is bounded by some positive constant c, as required.

Finally, since we have that the loss function of robust logistic regression is 1-Lipschitz, and it is bounded by a finite constant c, we apply lemma 1 to conclude the statement of our theorem. □

4 Experiments

We first illustrate the effect of two types of label noise, symmetric and asymmetric using synthetic data. The decision boundary obtained from both models are plotted in Fig. (1). It is obvious that symmetric label noise does not alter the decision boundary of traditional logistic regression as expected. By contrast, LR suffers from asymmetric mislabelling. Robust logistic regression, however, was able to perform well in both situations. Next we demonstrate the benefit of having label noise modelling using four UCI data sets: *Adult*, *Boston*, *Liver*, and



(a) LR-symm. (b) LR-asym. (c) rLR-symm. (d) rLR-asym.

Fig. 1. Decision boundary obtained by LR and rLR at 40% symmetric and 40% asymmetric label noise respectively. The orange lines are the estimates and the black lines are the true values.

Pima. More applications of rLR and L1-regularised rLR may be found in our previous works [2] and [3]. We also compare rLR with its generative counterpart: robust Normal Discriminant Analysis (rNDA) [10]. We first show the result on clean datasets where the labels are all perfect. This is to demonstrate that label noise formulation does not deteriorate the performance when there is no wrong label. The mean error and standard deviation from 50 independent runs are presented in Table 2 (top). It is clear from the result that the rLR reduces to classical LR when there is no wrong label.

Further, we artificially inject 10% and 30% asymmetric label noise into the training sets. We hold out 20% of the data for testing and train on the remaining points. Table (2) (middle and bottom) summarises the empirical results. As expected, and in agreement with our analysis, rLR improves upon traditional LR in label noise conditions. Observe also that rNDA is competitive only when the Gaussianity assumption holds true, and is inferior to rLR, otherwise.

Table 2. Empirical error rate of LR vs rLR vs rNDA in predicting the true labels from partially mislabelled training set. The mislabelling rate was 10% and 30%, of asymmetric type. P-values from Wilcoxon test at 5% level are presented. Boldface highlights statistically best method(s) while italics indicates the second best method.

Dataset	Error at 0 % Noise			p-values		
	rNDA	LR	rLR	rNDA v. LR	rNDA v. rLR	LR v. rLR
Adult	22.70 ± 0.03	**19.87 ± 0.02**	**20.13 ± 0.02**	$6.26e^{-6}$	$4.81e^{-5}$	0.62
Boston	15.53 ± 0.03	**13.86 ± 0.03**	**13.08 ± 0.03**	0.022	0.001	0.23
Liver	33.74 ± 0.05	32.03 ± 0.05	31.33 ± 0.04	0.075	0.02	0.39
Pima	24.84 ± 0.03	23.47 ± 0.03	24.36 ± 0.04	0.073	0.50	0.26
	10 % Noise					
Adult	22.06 ± 0.04	**19.94 ± 0.03**	**19.70 ± 0.02**	0.001	$6.52e^{-4}$	0.84
Boston	16.62 ± 0.04	*14.55 ± 0.03*	**13.06 ± 0.03**	0.015	$1.26e^{-5}$	0.021
Liver	33.80 ± 0.06	32.14 ± 0.06	32.20 ± 0.06	0.15	0.092	0.81
Pima	23.53 ± 0.03	22.77 ± 0.03	23.66 ± 0.03	0.34	0.72	0.14
	30 % Noise					
Adult	*24.88 ± 0.07*	26.87 ± 0.05	**21.66 ± 0.05**	0.023	0.008	$4.17e^{-7}$
Boston	*20.27 ± 0.04*	22.29 ± 0.05	**14.94 ± 0.04**	0.044	$3.31e^{-8}$	$1.19e^{-10}$
Liver	37.68 ± 0.09	39.36 ± 0.07	**33.62 ± 0.07**	0.11	0.032	$2.11e^{-4}$
Pima	**24.35 ± 0.04**	28.05 ± 0.04	**25.09 ± 0.04**	$6.28e^{-6}$	0.30	$4.22e^{-4}$

5 Conclusion and Future Work

We have presented a theoretical analysis of the label-noise robust logistic regression (rLR) classifier, showing that its error is bounded, and that rLR behaves the same as LR when there is no label noise or when the label flipping is symmetric. We have also demonstrated that rLR improves upon LR when there is asymmetric label flipping, and this is due to a weighting mechanism that our analysis has revealed. This was achieved by decomposing and interpreting the gradient as employed in gradient ascent optimisation. Future work is required to modify the formalism in a way to have a loss function defined on the latent true label rather than the observed noisy label.

References

1. P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *JMLR*, 3:463–482, 2003.
2. J. Bootkrajang and A. Kabán. Label-noise robust logistic regression and its applications. In *Proceedings of ECML-PKDD'12*, pp. 143-158, 2012.
3. J. Bootkrajang and A. Kabán. Classification of mislabelled microarrays using robust sparse logistic regression. *Bioinformatics*, 29(7):870-877, 2013.
4. C. E. Brodley and M. A. Friedl. Identifying and eliminating mislabeled training instances. In *Proceedings of AAAI'96*, pp. 799–805, 1996.
5. C. E. Brodley and M. A. Friedl. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11:131–167, 1999.
6. R. S. Chhikara and J. McKeon. Linear discriminant analysis with misallocation in training samples. *Journal of the American Stat. Assoc.*, 79(388):899–906, 1984.
7. T. Krishnan and S. C. Nandy. Efficiency of discriminant analysis when initial samples are classified stochastically. *Pattern Recognition*, 23(5):529–537, 1990.
8. P. A. Lachenbruch. Discriminant analysis when the initial samples are misclassified. *Technometrics*, 8(4):657–662, 1966.
9. P. A. Lachenbruch. Discriminant analysis when the initial samples are misclassified ii: Non-random misclassification models. *Technometrics*, 16(3):419–424, 1974.
10. N. D. Lawrence and B. Schölkopf. Estimating a kernel fisher discriminant in the presence of label noise. In *Proceedings of ICML'01*, pp. 306–313, 2001.
11. G. Lugosi. Learning with an unreliable teacher. *Pattern Recogn.*, 25:79–87, 1992.
12. L. S. Magder and J. P. Hughes. Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*, 146(2):195–203, 1997.
13. A. Malossini, E. Blanzieri, and R. T. Ng. Detecting potential labeling errors in microarrays by data perturbation. *Bioinformatics*, 22(17):2114–2121, 2006.
14. F. Muhlenbach, S. Lallich, and D. A. Zighed. Identifying and handling mislabelled instances. *Journal of Intelligent Information Systems*, 22(1):89–109, 2004.
15. J. S. Sánchez, R. Barandela, A. I. Marqués, R. Alejo, and J. Badenas. Analysis of new techniques to obtain quality training sets. *Pattern Recognition Letters*, 24(7):1015–1022, 2003.
16. S. Shalev-Shwartz. Introduction to machine learning. The Hebrew University of Jerusalem , `http://www.cs.huji.ac.il/~shais/Handouts.pdf`, 2009.
17. Y. Yasui, M. Pepe, L. Hsu, B.-L. Adam, and Z. Feng. Partially supervised learning using an EM-boosting algorithm. *Biometrics*, 60(1):199–206, 2004.