Journal of Physics: Conference Series **473** (2013) 012001 doi:10.1088/1742-6596/473/1/012001

Statistical-mechanics analysis of Gaussian labeled-unlabeled classification problems

Toshiyuki Tanaka

Graduate School of Informatics, Kyoto University,
36-1 Yoshida Hon-machi, Sakyo-ku, Kyoto-shi, Kyoto 606-8501, Japan

E-mail: tt@i.kyoto-u.ac.jp

Abstract. The labeled-unlabeled classification problem in semi-supervised learning is studied via statistical-mechanics approach. We analytically investigate performance of a learner with an equal-weight mixture of two symmetrically-located Gaussians, performing posterior mean estimation of the parameter vector on the basis of a dataset consisting of labeled and unlabeled data generated from the same probability model as that assumed by the learner. Under the assumption of replica symmetry, we have analytically obtained a set of saddle-point equations, which allows us to numerically evaluate performance of the learner. On the basis of the analytical result we have observed interesting phenomena, in particular the coexistence of good and bad solutions, which may happen when the number of unlabeled data is relatively large compared with that of labeled data.

1. Introduction

In the basic framework of classification problems, one is given a dataset of labeled data $\mathcal{D}_l = \{(\boldsymbol{x}^\mu, y^\mu) : \mu = 1, \ldots, L\}$, where $\boldsymbol{x}^\mu$ and y^μ are the feature vector and the class label of datum μ, respectively, and infers the rule of the classification underlying the dataset $\mathcal{D}_l$. This framework is categorized as supervised learning in learning theory, because one can regard that a supervisor has provided the class labels y^μ on the basis of the features $\boldsymbol{x}^\mu$. Although classification problems have mainly been studied as supervised learning, it might not be reasonable to expect in real-world classification problems that a dataset containing a sufficient number of labeled data is available. In some applications only human experts can serve as the supervisor providing the class labels in the dataset. In some other cases it is very expensive and/or time-consuming to obtain the class labels. Hence, a typical situation in this respect is that only a limited number of labeled data are available, whereas a relatively larger number of unlabeled data are easily obtained. In the era of "big data," one should face with such a situation more than before, in a wide variety of applications. Classification problems in such situations are generically called labeled-unlabeled classification problems [1].

One may consider two extreme strategies for a labeled-unlabeled classification problem. One extreme is to just ignore unlabeled data and to make use of available labeled data only in learning. This strategy is of course not optimal in typical problem settings, since unlabeled data may provide some additional information about the underlying classification rule. The other extreme is first to perform learning on labeled data, to estimate labels for unlabeled data on the basis of the result of the learning, and then to perform learning using the labeled data as well as the unlabeled data accompanied with the estimated labels. The process may be iterated until

convergence is achieved. This strategy can be problematic if only a few number of labeled data are available so that the result of the initial learning, on the basis of which one estimate labels of unlabeled data, is unreliable.

Several methods have been proposed to deal with labeled-unlabeled classification problems. Readers are referred to surveys [1, 2] and a book [3] for reviews of these methods. They have been tested empirically, often with significant performance gains compared with the extreme strategy of ignoring unlabeled data, and in some cases performing almost as good as the fully-supervised cases. To the author's knowledge, however, there are only a few analytical studies in the literature on the theoretical upper limit of how well one can utilize unlabeled data in labeled-unlabeled classification problems.

In this paper we report some results of our preliminary analysis on the labeled-unlabeled problem. We focus on the problem with two classes, and assume that the classes are represented with multivariate Gaussian distributions. In contrast to the standard asymptotic theory in statistics, where one fixes the dimension of the feature space and takes the limit of large numbers of data [4, 5], we study in this paper the case where the dimension of the feature space tends to infinity proportionally to the numbers of labeled and unlabeled data, which is a typical problem setting in the framework of the statistical-mechanics approach to information processing [6, 7].

2. Formulation

We consider the two-class labeled-unlabeled classification problem defined as follows. We assume that feature vectors of positively- and negatively-labeled data are generated from N-dimensional Gaussian distributions centered at $N^{-1/2}\boldsymbol{w}_0$ and $-N^{-1/2}\boldsymbol{w}_0$, respectively, where $\boldsymbol{w}_0 \in \mathbb{R}^N$. The covariance matrices of these Gaussian distributions are assumed to be equal to $\lambda_0^{-1}I$, where I denotes the identity matrix. The dataset of labeled data is given by

$$\mathcal{D}_l = \{(\boldsymbol{x}^\mu, y^\mu) \in \mathbb{R}^N \times \{-1, 1\} : \mu = 1, \ldots, L\}, \tag{1}$$

where $\boldsymbol{x}^\mu \in \mathbb{R}^N$ and $y^\mu \in \{-1, 1\}$ denote the feature vector and the class label of datum μ, respectively. The feature vector $\boldsymbol{x}^\mu$ is assumed to have been generated according to the conditional distribution $p(\boldsymbol{x}^\mu|y^\mu) = \mathcal{N}(y^\mu N^{-1/2}\boldsymbol{w}_0, \lambda_0^{-1}I)$. The dataset of unlabeled data is

$$\mathcal{D}_u = \{\boldsymbol{x}^\mu \in \mathbb{R}^N : \mu = 1, \ldots, U\}, \tag{2}$$

where $\boldsymbol{x}^\mu$ is assumed to have been generated according to the distribution $p(\boldsymbol{x}^\mu) = \sum_{y=\pm 1}(1/2)\mathcal{N}(yN^{-1/2}\boldsymbol{w}_0, \lambda_0^{-1}I)$. All the labeled and unlabeled data are assumed independent.

The learner in our labeled-unlabeled classification problem assumes an equal-weight mixture of symmetrically-located Gaussian distributions:

$$\begin{aligned} p(\boldsymbol{x}, y|\boldsymbol{w}) &= p(\boldsymbol{x}|y, \boldsymbol{w})p(y), \\ p(\boldsymbol{x}|y, \boldsymbol{w}) &= \left(\frac{\lambda}{2\pi}\right)^{N/2} e^{-\lambda\|\boldsymbol{x} - yN^{-1/2}\boldsymbol{w}\|^2/2}, \\ p(y) &= \frac{1}{2}, \quad y \in \{-1, 1\}, \end{aligned} \tag{3}$$

where $\boldsymbol{w} \in \mathbb{R}^N$ is the parameter vector of the model assumed by the learner. The learner is supposed to estimate $\boldsymbol{w}$ on the basis of the dataset of labeled and unlabeled data $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$. This is the formulation of the labeled-unlabeled classification problem we discuss in this paper.

3. Replica analysis

The likelihood function of the parameter $\boldsymbol{w}$ given the dataset of labeled and unlabeled data $\mathcal{D}$ is

$$p(\mathcal{D} \mid \boldsymbol{w}) = \left(\frac{\lambda}{2\pi}\right)^{N(L+U)/2} \prod_{\mu\in\mathcal{D}_l} e^{-\lambda\|\boldsymbol{x}^\mu - y^\mu N^{-1/2}\boldsymbol{w}\|^2/2} \times \prod_{\mu\in\mathcal{D}_u} \left[\frac{1}{2}\left(e^{-\lambda\|\boldsymbol{x}^\mu - N^{-1/2}\boldsymbol{w}\|^2/2} + e^{-\lambda\|\boldsymbol{x}^\mu + N^{-1/2}\boldsymbol{w}\|^2/2}\right)\right]. \tag{4}$$

We also assume that the learner has as the prior distribution of the parameter vector $\boldsymbol{w}$ the Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \kappa^{-1}I)$, where $\kappa^{-1} = \|\boldsymbol{w}_0\|^2/N$. The posterior distribution of $\boldsymbol{w}$ given the dataset $\mathcal{D}$ is thus given via the Bayes formula as

$$p(\boldsymbol{w} \mid \mathcal{D}) = \frac{p(\boldsymbol{w})p(\mathcal{D} \mid \boldsymbol{w})}{\int p(\boldsymbol{w}')p(\mathcal{D} \mid \boldsymbol{w}')\, d\boldsymbol{w}'}. \tag{5}$$

Quantities of our interest are those represented as posterior means

$$\mathbb{E}_{\boldsymbol{w}|\mathcal{D}}[f(\boldsymbol{w})] = \int f(\boldsymbol{w})\, p(\boldsymbol{w} \mid \mathcal{D})\, d\boldsymbol{w}. \tag{6}$$

Such quantities include the squared error of the posterior mean estimate of $\boldsymbol{w}_0$. Those posterior means also depend on the dataset $\mathcal{D}$ used by the learner. Since the dataset $\mathcal{D}$ is assumed to be generated randomly in our problem setting, the posterior means are also random quantities. We are thus interested in their averages over randomness of the dataset $\mathcal{D}$, that is,

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\boldsymbol{w}|\mathcal{D}}[f(\boldsymbol{w})]] &= \iint f(\boldsymbol{w})\, p(\boldsymbol{w} \mid \mathcal{D})p(\mathcal{D} \mid \boldsymbol{w}_0)\, d\boldsymbol{w}\, d\mathcal{D} \\ &= \int \frac{\int f(\boldsymbol{w})\, p(\boldsymbol{w})p(\mathcal{D} \mid \boldsymbol{w})\, d\boldsymbol{w}}{\int p(\boldsymbol{w}')p(\mathcal{D} \mid \boldsymbol{w}')\, d\boldsymbol{w}'} p(\mathcal{D} \mid \boldsymbol{w}_0)\, d\mathcal{D}.\end{aligned} \tag{7}$$

Analytically evaluating the averages with respect to the randomness of the dataset $\mathcal{D}$ poses the major challenge. We adopt the statistical-mechanics approach and apply the replica method, in which we evaluate

$$\Xi_n = \mathbb{E}_{\mathcal{D}}\left[\left(\int p(\boldsymbol{w})p(\mathcal{D} \mid \boldsymbol{w})\, d\boldsymbol{w}\right)^n\right]. \tag{8}$$

The dataset $\mathcal{D}$ serves as the quenched randomness in our problem. The standard prescription of the replica method is that one evaluates the free energy[1] averaged over the randomness of the dataset $\mathcal{D}$ as

$$\begin{aligned}\mathcal{F} = \lim_{N\to\infty} \mathbb{E}_{\mathcal{D}}\left[\frac{1}{N}\log\int p(\boldsymbol{w})p(\mathcal{D} \mid \boldsymbol{w})\, d\boldsymbol{w}\right] &= \lim_{N\to\infty}\frac{1}{N}\lim_{n\to 0}\frac{\partial \log \Xi_n}{\partial n} \\ &= \lim_{n\to 0}\frac{\partial}{\partial n}\lim_{N\to\infty}\frac{1}{N}\log\Xi_n,\end{aligned} \tag{9}$$

where in the last equality we have interchanged the order of the operations on n and the thermodynamic limit $N \to \infty$, assuming that this interchange does not affect the end result.

[1] Note that in the definition above the overall sign is reversed from the conventional physics definition of "free energy". This does not affect the following analysis.

We then evaluate $\lim_{N\to\infty}(1/N)\log\Xi_n$ via the saddle-point method. In evaluating this quantity we temporarily assume that n is a natural number and rewrite Ξ_n as

$$\begin{aligned}\Xi_n &= \int\cdots\int\prod_{a=1}^{n}[p(\boldsymbol{w}_a)p(\mathcal{D}\mid\boldsymbol{w}_a)]p(\mathcal{D}\mid\boldsymbol{w}_0)\,d\mathcal{D}\prod_{a=1}^{n}d\boldsymbol{w}_a\\ &= \int\cdots\int\left(\int\prod_{a=0}^{n}p(\mathcal{D}\mid\boldsymbol{w}_a)\,d\mathcal{D}\right)\prod_{a=0}^{n}p(\boldsymbol{w}_a)\,d\boldsymbol{w}_a,\end{aligned}\tag{10}$$

where we have formally introduced $p(\boldsymbol{w}_0)$ to make the resulting formula symmetric in the replica index a. One can use as $p(\boldsymbol{w}_0)$ the "true prior" if one is interested in quantities which are further averaged over randomness of $\boldsymbol{w}_0$.

Let $W=(\boldsymbol{w}_0,\boldsymbol{w}_1,\ldots,\boldsymbol{w}_n)$. A key observation is that the quantity

$$\int\prod_{a=0}^{n}p(\mathcal{D}\mid\boldsymbol{w}_a)\,d\mathcal{D},\tag{11}$$

appearing in the integrand of (10), depends on W only through its normalized Gram matrix $Q=N^{-1}W^TW$. Omitting details of derivations, we have

$$\lim_{N\to\infty}\frac{1}{N}\log\Xi_n=\sup_{Q}[\mathcal{G}(Q)-\mathcal{I}(Q)]\tag{12}$$

as the result of applying the saddle-point method, where

$$\mathcal{G}(Q)=-\frac{\alpha+\alpha'}{2}\operatorname{tr}Q\Lambda+\alpha\frac{\mathbf{1}^T\Lambda Q\Lambda\mathbf{1}}{2(\lambda_0+n\lambda)}+\alpha'\log\left(\frac{1}{2^{n+1}}\sum_{\boldsymbol{s}\in\{-1,1\}^{n+1}}e^{\boldsymbol{s}^T\Lambda Q\Lambda\boldsymbol{s}/2(\lambda_0+n\lambda)}\right),\tag{13}$$

$$\mathcal{I}(Q)=\sup_{\tilde{Q}}\left(\sum_{a\le b}\tilde{Q}_{ab}Q_{ab}+\sum_{a=1}^{n}\tilde{Q}_{0a}Q_{0a}-\log\mathcal{M}(\tilde{Q})\right),\tag{14}$$

$$\mathcal{M}(\tilde{Q})=\mathbb{E}\left(e^{\sum_{a\le b}\tilde{Q}_{ab}w_aw_b+\sum_{a=1}^{n}\tilde{Q}_{0a}w_0w_a}\right),\tag{15}$$

and where we let $\alpha=L/N$, $\alpha'=U/N$, and $\Lambda=\operatorname{diag}(\lambda_0,\lambda,\ldots,\lambda)$. The expectation in the last equation is taken over $(w_1,\ldots,w_n)\sim\mathcal{N}(\mathbf{0},\lambda^{-1}I)$. We have also introduced $\tilde{Q}$ as the $(n+1)\times(n+1)$ symmetric order-parameter matrix conjugate to the Gram matrix Q.

The function $\mathcal{G}(Q)$ represents the exponent of the integrand (11) in terms of the Gram matrix Q. Large-deviations theory [8] tells us that the rate function $\mathcal{I}(Q)$ is obtained via the Legendre transform of the cumulant generating function $\log\mathcal{M}(\tilde{Q})$ for the multivariate Gaussian random variable $(w_1,\ldots,w_n)\sim\mathcal{N}(\mathbf{0},\lambda^{-1}I)$.

The saddle-point conditions are

$$\tilde{Q}_{aa}=-(\alpha+\alpha')\frac{\lambda(\lambda_0+(n-1)\lambda)}{2(\lambda_0+n\lambda)},\tag{16}$$

$$\tilde{Q}_{0a}=\frac{\lambda_0\lambda}{\lambda_0+n\lambda}\left(\alpha+\alpha'\frac{\sum_{\boldsymbol{s}\in\{1,-1\}^{n+1}}s_0s_ae^{\boldsymbol{s}^T\Lambda Q\Lambda\boldsymbol{s}/2(\lambda_0'+n\lambda)}}{\sum_{\boldsymbol{s}\in\{1,-1\}^{n+1}}e^{\boldsymbol{s}^T\Lambda Q\Lambda\boldsymbol{s}/2(\lambda_0'+n\lambda)}}\right),\tag{17}$$

$$\tilde{Q}_{ab}=\frac{\lambda^2}{\lambda_0+n\lambda}\left(\alpha+\alpha'\frac{\sum_{\boldsymbol{s}\in\{1,-1\}^{n+1}}s_as_be^{\boldsymbol{s}^T\Lambda Q\Lambda\boldsymbol{s}/2(\lambda_0'+n\lambda)}}{\sum_{\boldsymbol{s}\in\{1,-1\}^{n+1}}e^{\boldsymbol{s}^T\Lambda Q\Lambda\boldsymbol{s}/2(\lambda_0'+n\lambda)}}\right),\tag{18}$$

$$Q_{ab}=\langle w_aw_b\rangle,\tag{19}$$

where $\langle \cdots \rangle$ is defined as

$$\langle \cdots \rangle := \frac{\mathbb{E}\left[(\cdots) e^{\sum_{a \leq b} \tilde{Q}_{ab} w_a w_b + \sum_{a=1}^n \tilde{Q}_{0a} w_0 w_a} \right]}{\mathbb{E}\left[e^{\sum_{a \leq b} \tilde{Q}_{ab} w_a w_b + \sum_{a=1}^n \tilde{Q}_{0a} w_0 w_a} \right]}. \tag{20}$$

In order to proceed further we make the assumption of replica symmetry (RS), where we assume that the order parameters are invariant under permutations of replica indices $a = 1, \ldots, n$. We introduce RS order parameters by letting

$$\begin{aligned} &Q_{aa} = p, && Q_{ab} = q, && Q_{0a} = m, \\ &\tilde{Q}_{aa} = \tilde{p}, && \tilde{Q}_{ab} = \tilde{q}, && \tilde{Q}_{0a} = \tilde{m}. \end{aligned} \tag{21}$$

Under the RS assumption one has

$$\begin{aligned} \mathcal{F} = &-\frac{(\alpha + \alpha')\lambda}{2} \left(\frac{\lambda_0 - \lambda}{\lambda_0} p + \frac{\lambda}{\lambda_0} q \right) + \alpha \lambda m + \alpha' \sqrt{\frac{\lambda_0}{2\pi q}} \int e^{-\frac{\lambda_0}{2q}(z-m)^2} \log \cosh \lambda z \, dz \\ &- m\tilde{m} - p\tilde{p} + \frac{1}{2} q\tilde{q} + \frac{1}{2} \frac{\tilde{m}^2 + \kappa \tilde{q}}{\kappa(\kappa - 2\tilde{p} + \tilde{q})} + \frac{1}{2} \log \left(\frac{\kappa}{\kappa - 2\tilde{p} + \tilde{q}} \right). \end{aligned} \tag{22}$$

The saddle-point equations for the RS order parameters are

$$\partial_m \mathcal{F} = \alpha\lambda + \alpha'\lambda \sqrt{\frac{\lambda_0}{2\pi q}} \int e^{-\frac{\lambda_0}{2q}(z-m)^2} \tanh \lambda z \, dz - \tilde{m} = 0, \tag{23}$$

$$\partial_p \mathcal{F} = -\frac{(\alpha + \alpha')\lambda(\lambda_0 - \lambda)}{2\lambda_0} - \tilde{p} = 0, \tag{24}$$

$$\partial_q \mathcal{F} = -\frac{(\alpha + \alpha')\lambda^2}{2\lambda_0} + \frac{\alpha'\lambda^2}{2\lambda_0} \sqrt{\frac{\lambda_0}{2\pi q}} \int e^{-\frac{\lambda_0}{2q}(z-m)^2} (1 - \tanh^2 \lambda z) \, dz + \frac{\tilde{q}}{2} = 0, \tag{25}$$

$$\partial_{\tilde{m}} \mathcal{F} = -m + \frac{\tilde{m}}{\kappa(\kappa - 2\tilde{p} + \tilde{q})} = 0, \tag{26}$$

$$\partial_{\tilde{p}} \mathcal{F} = -p + \frac{\tilde{m}^2 + \kappa\tilde{q}}{\kappa(\kappa - 2\tilde{p} + \tilde{q})^2} + \frac{1}{\kappa - 2\tilde{p} + \tilde{q}} = 0, \tag{27}$$

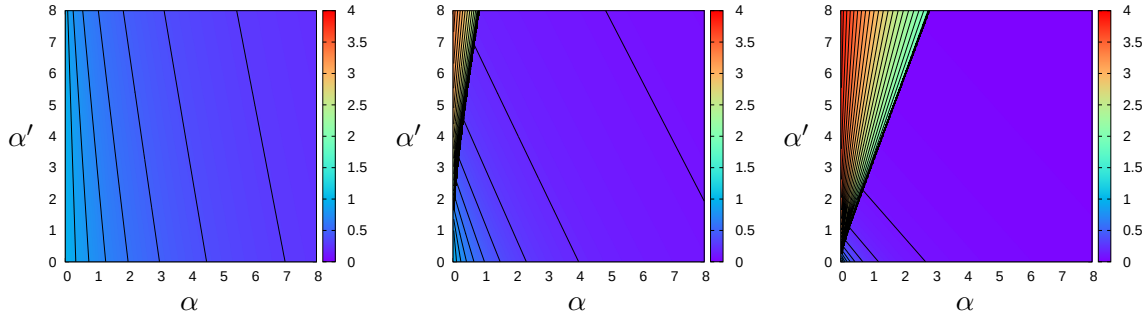
$$\partial_{\tilde{q}} \mathcal{F} = \frac{q}{2} - \frac{1}{2} \frac{\tilde{m}^2 + \kappa\tilde{q}}{\kappa(\kappa - 2\tilde{p} + \tilde{q})^2} = 0. \tag{28}$$

In the following of this paper we focus on the case where the learner assumes the true noise variance, that is, where $\lambda_0 = \lambda$ holds. One has $\tilde{p} = 0$ under this condition. Rescaling the RS order parameters as $\lambda m \to m$, $\lambda q \to q$, $\tilde{m}/\lambda \to \tilde{m}$, and $\tilde{q}/\lambda \to \tilde{q}$, in order to simplify the argument, the saddle-point equations take the following simple form:

$$\tilde{m} = \alpha + \alpha' \sqrt{\frac{1}{2\pi q}} \int e^{-(z-m)^2/2q} \tanh z \, dz, \tag{29}$$

$$\tilde{q} = \alpha + \alpha' \sqrt{\frac{1}{2\pi q}} \int e^{-(z-m)^2/2q} \tanh^2 z \, dz, \tag{30}$$

$$m = \frac{a^2 \tilde{m}}{1 + a\tilde{q}}, \quad q = \frac{a^3 \tilde{m}^2 + a^2 \tilde{q}}{(1 + a\tilde{q})^2}, \tag{31}$$



(a) $a = 0.33$ (b) $a = 1$ (c) $a = 3.33$

Figure 1. Mean-squared errors of posterior-mean estimates.

where we have let $a = \lambda/\kappa$, which represents the signal-to-noise ratio of the problem. The normalized mean-squared error of the posterior mean estimate $\boldsymbol{w}$ against $\boldsymbol{w}_0$ is given in terms of a saddle-point solution as

$$\mathcal{E} = \frac{q - 2m + a}{a}. \tag{32}$$

Specifying the signal-to-noise ratio a, as well as the normalized numbers of labeled and unlabeled data α and α', one can numerically solve the above saddle-point equations to evaluate the theoretical performance of the labeled-unlabeled classification problem in terms of the normalized mean-squared error. This is the main result of this paper.

4. Numerical evaluations

We have numerically evaluated the normalized mean-squared error of the posterior mean estimate $\boldsymbol{w}$ using the analytical result presented in the previous section. The result is summarized in figure 1. One can observe in these figures that $\boldsymbol{w}_0$ can be estimated accurately when α' is small. As α' increases, mean-squared error becomes even smaller, implying that unlabeled data are utilized in learning. Let us define utility r of unlabeled data as a function of (α, α') as

$$r = \frac{\partial \mathcal{E}}{\partial \alpha'} \Big/ \frac{\partial \mathcal{E}}{\partial \alpha}, \tag{33}$$

that is, the utility of unlabeled data represents how effective an unlabeled datum is in reducing the mean-squared error in comparison with a labeled data. The utility of unlabeled data in the small-α' regime depends on the signal-to-noise ratio a, in such a way that the utility approaches 0 and 1 as a becomes smaller and larger, respectively. One can show that when $a\alpha \gg 1$ and $\alpha' \ll 1$ the utility of unlabeled data is given by

$$r = f(a) = \int \sqrt{\frac{a}{2\pi}} e^{-a(z-1)^2/2} \tanh az \, dz. \tag{34}$$

The shape of the function $f(a)$ is depicted in figure 2. It shows that the utility is close to 1 when the signal-to-noise ratio a is 10 or larger, and that it drops below 0.1 when a is less than 0.1.

One can also observe in figure 1 that coexistence of good and bad solutions occurs when α is small and α' is large enough. Indeed, in the α-α' plane, there is a "spinodal" line $\alpha' = \alpha'_c(\alpha)$, below which only a solution with good performance exists, and above which the coexistence of good and bad solutions takes place. It should be noted that in figure 1 only the bad solution is

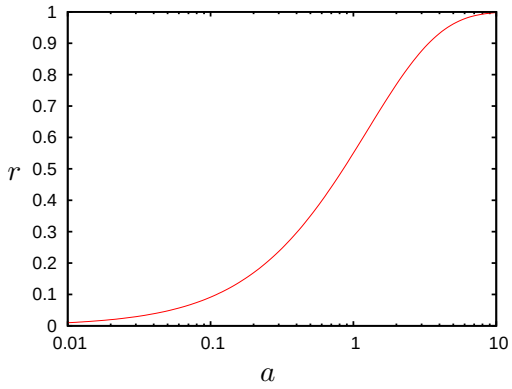


Figure 2. Utility r of unlabeled data as a function of the signal-to-noise ratio a.

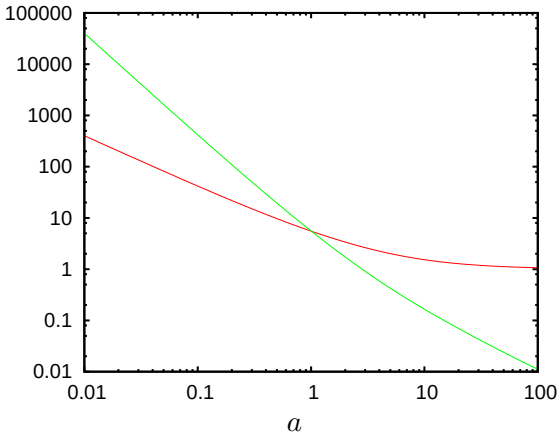


Figure 3. Slope and intercept of asymptotic of spinodal line when α and α' are large enough, versus the signal-to-noise ratio a. Red and green curves represent the slope and the intercept, respectively.

shown where the coexistence occurs. With perturbation analysis, one can show that the spinodal line has the following asymptotic form for small enough α:

$$\alpha'_c(\alpha) = \frac{1}{a^2}\left(1 + \sqrt{8a(a+1)\alpha}\right) \tag{35}$$

This asymptotic formula for the spinodal line tells us that, even when $L = o(N)$, that is, when the number of labeled data is vanishingly small compared with the dimension of the feature space, the coexistence of multiple solutions does not occur when $\alpha' < 1/a^2$. It also shows that a small increase in α from 0 causes a significant increase in the critical α' value, above which the coexistence of multiple solutions occurs. Indeed, the increase is superlinear, being proportional to $\alpha^{1/2}$.

When α and α' are large enough, on the other hand, one can show, again via perturbation analysis, that the spinodal line is asymptotically linear. In figure 3 we show the slope and the intercept of the asymptotic spinodal line as a function of a. Details of the perturbation analysis will be presented elsewhere.

5. Summary and outlook

We have reported some results of our preliminary statistical-mechanics-based analysis on the labeled-unlabeled classification problem. In our analysis we have dealt with the simplest setting, in which a learner with the equal-weight mixture of two symmetrically-located Gaussians is assumed. Even in this simplest setting we have observed interesting phenomena, in particular the coexistence of good and bad solutions, which may happen when the number of unlabeled data is relatively large compared with that of labeled data. We have obtained explicit expressions describing asymptotic behaviors of the spinodal line that marks the boundary between the region with the coexistence and the region without it.

On the basis of the basic results presented so far, one can pose several problems to be explored. First of all, the unequal-weight case, in which a learner is to estimate the class weights as well as the class centers, should be explored to make the setting a bit more realistic. Another interesting problem would be to study the case where unlabeled data are generated from a probability model which is similar to but not the same as that for labeled data. The utility of unlabeled data would decrease as the probability model of unlabeled data becomes less similar to that of labeled data. It would also be important to consider the multi-class problem as well. From the viewpoint of statistical-mechanics analysis, it is necessary to study stability of the RS assumption in order to see if the RS assumption is valid. Finally, algorithms to efficiently solve the labeled-unlabeled classification problem should be studied as well, in view of practical applications.

Acknowledgments

The author would like to thank Professor Seiji Miyoshi with Kansai University, Osaka, Japan, for the discussion they had in March, 2013, which has inspired me to do the study.

References

[1] Seeger M 2002 Learning with labeled and unlabeled data Tech. Rep. EPFL-REPORT-161327 École Polytechnique Fédérale de Lausanne URL `http://infoscience.epfl.ch/record/161327/files/review.pdf`
[2] Zhu X 2008 Semi-supervised learning literature survey Technical Report 1530 Department of Computer Sciences, University of Wisconsin Madison
[3] Chapelle O, Schölkopf B and Zien A (eds) 2006 *Semi-Supervised Learning* (MIT Press)
[4] O'Neill T J 1978 *Journal of the American Statistical Association* **73** 821–6
[5] Ganesalingam S and McLachlan G J 1978 *Biometrika* **65** 658–62
[6] Nishimori H 2001 *Statistical Physics of Spin Glasses and Information Processing: An Introduction* (Oxford University Press)
[7] Mézard M and Montanari A 2009 *Information, Physics, and Computation* (Oxford University Press)
[8] Dembo A and Zeitouni O 1998 *Large Deviations Techniques and Applications* 2nd ed (Springer)