Theoretical Analysis of Label Distribution Learning

Jing Wang, Xin Geng[*]
MOE Key Laboratory of Computer Network and Information Integration
School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
{wangjing91, xgeng}@seu.edu.cn

Abstract

As a novel learning paradigm, *label distribution learning* (LDL) explicitly models label ambiguity with the definition of label description degree. Although lots of work has been done to deal with real-world applications, theoretical results on LDL remain unexplored. In this paper, we rethink LDL from theoretical aspects, towards analyzing learnability of LDL. Firstly, risk bounds for three representative LDL algorithms (AA-kNN, AA-BP and SA-ME) are provided. For AA-kNN, Lipschitzness of the label distribution function is assumed to bound the risk, and for AA-BP and SA-ME, rademacher complexity is utilized to give data-dependent risk bounds. Secondly, a generalized plug-in decision theorem is proposed to understand the relation between LDL and classification, uncovering that approximation to the conditional probability distribution function in absolute loss guarantees approaching to the optimal classifier, and also data-dependent error probability bounds are presented for the corresponding LDL algorithms to perform classification. As far as we know, this is perhaps the first research on theory of LDL.

Introduction

Traditional learning paradigms include single-label learning (SLL) and multi-label learning (MLL) (Zhang and Zhou 2014). SLL assumes that an instance is associated with one label, while MLL assumes that multiple labels are assigned to an instance. Essentially both SLL and MLL aim at finding related label/labels to describe the instance, while neither SLL nor MLL support relative importance of labels assigned to the instance. However in some real-word applications, labels are related to the instance with different relative importance degree. Thus it seems reasonable to label an instance with a soft label rather than a hard one (e.g., single label or a set of labels). Inspired by this, (Geng 2016) proposes a novel learning paradigm, Label Distribution Learning (LDL), which labels an instance with a distribution of description degree over label space, called label distribution, and learns a mapping from instance to label distribution directly. Compared with MLL, where positive labels are commonly treated equally (with description degree equals $1/c$ implicitly for c positive labels), LDL allows explicitly modeling of different relative importance of labels assigned to the instance, which is more suitable for many real scenarios.

Recently, LDL has been extensively applied in many real-world applications, which can be classified into three classes according to the source of label distribution. The first one features that label distribution is from data itself, which includes pre-release rating prediction on movies (Geng and Hou 2015), emotion recognition (Zhou, Xue, and Geng 2015), et al. The second one is characterized by that label distribution is originated from pre-knowledge, among which applications include age estimation (Geng, Yin, and Zhou 2013), head pose estimation (Geng and Xia 2014), et al. The third one is attributed to that label distribution is learned from data automatically. Applications of such class include label-importance-aware multi-label learning (Li, Zhang, and Geng 2015), beauty sensing (Ren and Geng 2017), video parsing (Geng and Ling 2017), et al. The secret of successes of LDL being applied in a variety of fields is that explicit introduction of label ambiguity with label distribution boosts performance of real-world applications.

In this paper, we re-examine LDL from theoretical aspects, towards analyzing generalization of LDL algorithms. Precisely, there are at least two arguments related to the generalization of LDL. The first one is on generalization of LDL itself, and the second one is on generalization of LDL to perform classification. On one hand, LDL can be regarded as one kind of multi-output regression (Borchani et al. 2015), and generalization of multi-output regression algorithms (Kakade, Sridharan, and Tewari 2009) can be transfered to LDL somewhat. However, specializations of LDL should not be neglected. The first thing to note is *probability simplex constraint* of label distribution. As suggested by (Geng 2016), the target label distribution is real-valued and satisfies distribution constraints, i.e., $\eta_{\mathbf{x}}^{y} \in [0, 1]$, and $\sum_{y \in \mathcal{Y}} \eta_{\mathbf{x}}^{y} = 1$ (formally defined in *Preliminary*). This constraint is often satisfied by applying a softmax function onto each output of a multi-output regression model, which complicates complexity of the corresponding multi-output regression model. Furthermore, the second thing to note is *specialization of measures* for LDL. Although Lipschitzness of loss function is a general assumption when bounding the risk, Kullback-Leibler (KL) divergence as loss function for LDL (e.g., SA-ME) does not satisfy Lipschitzness instead. On another hand, we find that LDL is usually adopted to per-

[*]Corresponding author.

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

form classification implicitly. For examples, in applications of age estimation (Geng, Yin, and Zhou 2013), head-pose estimation (Geng and Xia 2014), pre-release rating prediction on movies (Geng and Hou 2015), et al., label corresponding to the maximum predicted description degree is treated as the predicted label. As far as we know, generalization of such framework has never been touched. The main contributions of this paper are summarized as followings,

1) We establish *risk bounds* for three representative LDL algorithms, i.e., AA-BP, SA-ME, and AA-kNN, where bounds for the first two are data-dependent risk bounds, and bound for AA-kNN is consistency bound with Lipschitz assumption.
2) We generalize the binary *plug-in decision theorem* into multi-class classification, discovering that LDL dominates classification.
3) We provide, to the best of our knowledge, the first *data-dependent error probability bounds* for three representative LDL algorithms to perform classification.

The rest of this paper is organized as follows. Firstly, related works are briefly reviewed. Secondly, notations are introduced in preliminary. Then risk bounds for three representative LDL algorithms are provided. Next generalization of LDL to perform classification is examined. Finally, we conclude the paper.

Related Work

LDL (Geng 2016) is a novel learning paradigm, which labels an instance with a label distribution and learns a mapping from instance to label distribution straightly. Existing studies mainly focus on algorithm design and improvement.

Three strategies are embraced to design algorithms for LDL (Geng 2016). The first one is *Problem Transformation* (PT). Algorithms of this class firstly transform dataset equipped with label distribution into single-label dataset via re-sampling, and then SLL algorithms are employed to learn the transformed single-label dataset. Two representative algorithms are PT-SVM and PT-Bayes, which apply SVM and Bayes classifier respectively. The second one is *Algorithm Adaptation* (AA), which extends certain existing learning algorithms to handle label distribution seamlessly. Concretely, k-NN is adapted in such that, for an instance, mean of label distribution of its k nearest neighbors is calculated as the predicted label distribution, which is denoted by AA-kNN. Besides, three-layer back-propagation neural network with multi-output is also adapted to minimize the sum-squared loss of output of neural network compared with real label distribution, which is denoted by AA-BP. The third one is *Specialized Algorithms* (SA), which matches characteristics of LDL. Two specialized algorithms, i.e., SA-IIS and SA-BFGS, are proposed by applying maximum entropy model (Berger, Pietra, and Pietra 1996) with KL divergence as loss function to learn the label distribution. Notice that all above algorithms are designed without learnability guarantee.

Recently, two improvements on LDL are noteworthy. The first one tackles shortage of label distribution dataset. Compared with SLL and MLL, LDL dataset is more difficult to acquire. (Xu, Tao, and Geng 2018) proposes to boost logical label (from SLL/MLL dataset) into real-valued label distribution by graph laplacian label enhancement. (Hou et al. 2017) utilizes semi-supervised learning, with abundant unlabeled data, to adaptively learn label distribution. (Xu and Zhou 2017) deals with the situation when label distribution is partially observed, and jointly recovers and learns it by matrix completion with trace-norm regularizer. The second one exploits label correlations of LDL. (Zhou, Xue, and Geng 2015) represents label correlations by Pearson's correlation coefficient. (Jia et al. 2018) addresses label correlations by adding a regularizer, which encodes label correlations into distance between labels. (Zheng, Jia, and Li 2018) considers label correlation is local, which is encoded into a local correlation vector, and learns optimal encoding vector and LDL simultaneously. Algorithm improvements are mainly according to practice guide, without taking theoretical feasibility into consideration.

There are few work on theoretical research of LDL. One recent paper (Zhao and Zhou 2018) studies generalization of LDL partially. However, the major motivation of the paper is to take structures of label space into consideration by applying optimal transport distance (Villani 2008) instead of traditional LDL measures. And the proposed risk bound is too abroad, without providing a bound for rademacher complexity of the hypothesis space. Besides, characteristics of LDL measures are not taken into consideration at all. Risk bounds we developed in this paper match specializations of LDL (i.e., probability simplex constraint and loss function characteristics), which are broadly applicable.

For generalization of LDL to perform classification, one possible related topic is the plug-in decision theorem (Devroye, Györfi, and Lugosi 1996) (formally described in Theorem 8). It states that error probability difference between a decision function and the optimal one is bounded by expectation of absolute difference between the plug-in function and the conditional probability distribution function. Plug-in decision theorem has been widely used to prove consistency of many decision rules, such as Stone's Theorem (Stone 1977), strong consistency of k-NN rule (Biau and Devroye 2015), consistency of kernel rule (Devroye and Krzyźak 1989), et al. Note that the classic plug-in decision theorem is only established in the binary setting, which limits its applications. In this paper, we generalize the plug-in decision theorem and develop data-dependent error probability bounds for the corresponding LDL algorithms.

Preliminary

Denote input space by $\mathcal{X} \in \mathbb{R}^d$, and label space by $\mathcal{Y} = \{y_1, y_2, \ldots, y_m\}$. Define label distribution function $\eta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which satisfies $\eta(\mathbf{x}, y) \geq 0$ and $\sum_y \eta(\mathbf{x}, y) = 1$. Let the training set be $S = \{(\mathbf{x}_1, \eta(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, \eta(\mathbf{x}_n))\}$, where $\mathbf{x}_i$ is sampled according to an underlying probability distribution $\mathcal{D}$, and $\eta(\mathbf{x}_i) = \{\eta_{\mathbf{x}_i}^{y_1}, \ldots, \eta_{\mathbf{x}_i}^{y_m}\}$ is determined by an unknown label distribution function η with $\eta_{\mathbf{x}_i}^{y_j} = \eta(\mathbf{x}_i, y_j)$ for convenience of notation. The goal of LDL is to learn the unknown function η.

Given a function class $\mathcal{H}$ and a loss function $\ell : \mathbb{R}^m \times$

$\mathbb{R}^m \to \mathbb{R}_+$, for function $h \in \mathcal{H}$, its corresponding risk and empirical risk are defined as $L_{\mathcal{D}}(h) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\ell(h(\mathbf{x}), \eta(\mathbf{x}))]$ and $L_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), \eta(\mathbf{x}_i))$, respectively. Recall the definition of rademacher complexity w.r.t. S and ℓ,

$$\hat{\mathcal{R}}_n(\ell \circ \mathcal{H} \circ S) = \mathbb{E}_{\epsilon_1,\dots,\epsilon_n}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), \eta(\mathbf{x}_i))\epsilon_i\right],$$

where $\epsilon_1, \dots, \epsilon_n$ are n independent rademacher random variables with $\mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = 1/2$. Then

Lemma 1. (Bartlett and Mendelson 2003; Mohri, Rostamizadeh, and Talwalkar 2012) *Let $\mathcal{H}$ be a family of functions. For a loss function ℓ bounded by μ, then for any $\delta > 0$, with probability at least $1 - \delta$, for all $h \in \mathcal{H}$ such that*

$$L_{\mathcal{D}}(h) \le L_S(h) + 2\hat{\mathcal{R}}_n(\ell \circ \mathcal{H} \circ S) + 3\mu\sqrt{\frac{\log 2/\delta}{2n}}. \quad (1)$$

Risk Bounds for LDL Algorithms

We now provide risk bounds for three representative algorithms from two classes, i.e., AA-kNN and AA-BP from algorithm adaptation, SA-ME (maximum entropy) from specialized algorithms. Notice that (Geng 2016) proposes two specialized algorithms, i.e., SA-IIS and SA-BFGS, which differ only in the underlying optimization methods. In this paper we focus on the generalization ability of algorithms, and SA-ME represents SA-IIS and SA-BFGS collectively from the perspective of the underlying model, i.e., maximum entropy model. Also algorithms from problem transformation are not covered, for the reason that this type of algorithms circumvent learning label distribution with resampling and single-label learning algorithms instead of learning it directly.

AA-kNN

AA-kNN extends kNN to deal with label distribution. Given a new instance $\mathbf{x}$, its k nearest neighbors are firstly selected in the training set. Then mean of label distribution of k nearest neighbors is calculated as the label distribution of $\mathbf{x}$, i.e.,

$$\tilde{\eta}^{y_i}_{\mathbf{x}} = \frac{1}{k}\sum_{j\in N_k(\mathbf{x})} \eta^{y_i}_{\mathbf{x}_j}, (i = 1, 2, \dots, m),$$

where $\tilde{\eta} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^m$ is the output function of AA-kNN with $\tilde{\eta}^{y_i}_{\mathbf{x}} = \tilde{\eta}(\mathbf{x}, y_i)$ for simplicity of notation, and $N_k(\mathbf{x})$ is the index set of k nearest neighbors of $\mathbf{x}$. For convenience of analysis, we assume $\mathcal{X} = [0, 1]^d$.

Theorem 2. *Assume $\eta(\cdot, y_i)$ be c_i-Lipschitz w.r.t. $\mathcal{X}$, and $c = \sum_{i=1}^{m} c_i$, then for any $\delta > 0$, with probability at least $1 - \delta$ such that*

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\sum_{i=1}^{m} |\eta^{y_i}_{\mathbf{x}} - \tilde{\eta}^{y_i}_{\mathbf{x}}|\right] \le \frac{4c\sqrt{d}}{\delta}\left(\frac{2k}{n}\right)^{1/(d+1)}. \quad (2)$$

Proof. With Lipschitz assumption of $\eta(\cdot, y_i)$, we have

$$\begin{aligned}\mathbb{E}_{S,\mathbf{x}}\left[\sum_{i=1}^{m} |\eta^{y_i}_{\mathbf{x}} - \tilde{\eta}^{y_i}_{\mathbf{x}}|\right] &\le \mathbb{E}_{S,\mathbf{x}}\left[\sum_{i=1}^{m}\frac{1}{k}\sum_{j\in N_k(\mathbf{x})} |\eta^{y_i}_{\mathbf{x}} - \eta^{y_i}_{\mathbf{x}_j}|\right] \\ &\le \frac{c}{k}\mathbb{E}_{S,\mathbf{x}}\left[\sum_{j\in N_k(\mathbf{x})} ||\mathbf{x}_j - \mathbf{x}||_2\right].\end{aligned}$$

Thus to prove Theorem (2), it suffices to prove

$$\mathbb{E}_{S,\mathbf{x}}\left[\sum_{j\in N_k(\mathbf{x})} ||\mathbf{x}_j - \mathbf{x}||_2\right] \le 4\sqrt{d}k\left(\frac{2k}{n}\right)^{1/(d+1)}, \quad (3)$$

which is tricky and left to appendices. Then apply Markov's inequality to Eq. (3), and Theorem (2) follows directly. □

AA-BP

AA-BP adapts back-propagation neural network to perform label distribution learning. The three-layer neural network has m output units, each of which outputs the description degree of a label. To make sure the output of neural network satisfies probability simplex constraint, softmax activation function is applied in each output unit. Similar to multi-output regression, AA-BP minimizes sum-squared loss of the output of neural network with the real label distribution. Observing that AA-BP can be regarded as a combination of softmax function and function family of the three-layer neural network, rademacher complexity of which w.r.t. S for loss function ℓ is bounded as following,

Theorem 3. *Denote softmax function by* SF. *Let $\mathcal{H}$ be a family of functions for three-layer neural network with m outputs (identity activation on the output layer), and $\mathcal{H}_j$ be a family of functions for j-th output. For a loss function ℓ with Lipschitz constant L_ℓ, we have*

$$\hat{\mathcal{R}}_n(\ell \circ \text{SF} \circ \mathcal{H} \circ S) \le 2\sqrt{2}mL_\ell \sum_{j=1}^{m} \hat{\mathcal{R}}_n(\mathcal{H}_j \circ S), \quad (4)$$

Proof. Define function $\phi(\cdot, \cdot)$ as $\ell(\text{SF}(\cdot), \cdot)$. Next, we show that ϕ is Lipschitz. For probability distribution $\mathbf{p}, \mathbf{q} \in \mathbb{R}^m$,

$$|\phi(\mathbf{p}, \cdot) - \phi(\mathbf{q}, \cdot)| \le L_\ell ||\text{SF}(\mathbf{p}) - \text{SF}(\mathbf{q})||_1,$$

where the inequality is according to Lipschitzness of ℓ. Next, right-hand side of the proceeding equation equals

$$L_\ell \sum_{i=1}^{m} \left|\frac{1}{1 + \sum_{j\ne i} \mathrm{e}^{p_j - p_i}} - \frac{1}{1 + \sum_{j\ne i} \mathrm{e}^{q_j - q_i}}\right|,$$

where p_i, q_i is i-th element of $\mathbf{p}$ and $\mathbf{q}$, respectively. Observing that for $\mathbf{v} \in \mathbb{R}^m$, function $\frac{1}{1+\sum_i \exp(v_i)}$ is actually 1-Lipschitz, thus the preceding equation is bounded by

$$\begin{aligned}& L_\ell \sum_{i=1}^{m} ||\mathbf{p} - \mathbf{1}\cdot p_i - \mathbf{q} + \mathbf{1}\cdot q_i||_2 \\ \le\ & L_\ell \sum_{i=1}^{m} (||\mathbf{p} - \mathbf{q}||_2 + \sqrt{m}|p_i - q_i|) \\ \le\ & L_\ell (m||\mathbf{p} - \mathbf{q}||_2 + \sqrt{m}\sum_{i=1}^{m} |p_i - q_i|) \\ \le\ & 2mL_\ell ||\mathbf{p} - \mathbf{q}||_2,\end{aligned}$$

where the last inequality is according to Cauchy-Schwarz's inequality. Recall the definition of rademacher complexity

$$\hat{\mathcal{R}}_n(\ell \circ \text{SF} \circ \mathcal{H} \circ S) = \mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \phi(h(\mathbf{x}_i), \eta(\mathbf{x}_i))\epsilon_i\right],$$

and according to (Maurer 2016), with ϕ being $2mL_\ell$-Lipschitz, right-hand side of above equation is bounded by

$$\sqrt{2}(2mL_\ell)\mathbb{E}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\epsilon_{i,j}h_j(\mathbf{x}_i)\right], \qquad (5)$$

where $h_j(\mathbf{x}_i)$ is j-th component of $h(\mathbf{x}_i)$, and $\epsilon_{i,j}$ are $n\times m$ i.i.d. random variables. Notice that function class of a multi-output neural network can be regarded as a direct sum of function classes of multiple scalar-output neural networks. Suppose $\mathcal{H}_1,\dots,\mathcal{H}_m$ be classes of functions for three-layer neural network with scalar output, then $\mathcal{H} = \oplus_{j\in[m]}\mathcal{H}_j = \left\{\mathbf{x}\to[h_1(\mathbf{x})\dots h_m(\mathbf{x})]^T : h_j\in\mathcal{H}_j\right\}$, and Eq. (5) equals

$$\begin{aligned}&2\sqrt{2}mL_\ell\mathbb{E}\left[\sup_{h\in\bigoplus_{j\in[m]}\mathcal{H}_j}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\epsilon_{i,j}h_j(\mathbf{x}_i)\right]\\ &\le 2\sqrt{2}mL_\ell\sum_{j=1}^{m}\mathbb{E}\left[\sup_{h_j\in\mathcal{H}_j}\frac{1}{n}\sum_{i=1}^{n}\epsilon_{i,j}h_j(\mathbf{x}_i)\right]\\ &\le 2\sqrt{2}mL_\ell\sum_{j=1}^{m}\hat{\mathcal{R}}_n(\mathcal{H}_j\circ S).\end{aligned}$$

which finishes proof of Theorem 3. □

Rademacher complexity of class of functions for three-layer neural network with scalar output satisfies

Lemma 4. (Bartlett and Mendelson 2003; Gao and Zhou 2016). *Let σ be Lipschitz with constant L_σ. Define class of functions $\mathcal{H}_j = \{\mathbf{x}\mapsto\sum_i w_{j,i}\sigma(\mathbf{v}_i\cdot\mathbf{x}) : ||\mathbf{w}_j||_2 \le B_1, ||\mathbf{v}_i||_2\le B_0\}$, rademacher complexity of which satisfies*

$$\hat{\mathcal{R}}_n(\mathcal{H}_j\circ S)\le\frac{L_\sigma B_0 B_1}{\sqrt{n}}\max_{i\in[n]}||\mathbf{x}_i||_2.$$

Accordingly, right-hand side of Eq. (4) is bounded as

$$\hat{\mathcal{R}}_n(\ell\circ\mathrm{SF}\circ\mathcal{H}\circ S)\le\frac{2\sqrt{2}m^2L_\ell L_\sigma B_0B_1}{\sqrt{n}}\max_{i\in[n]}||\mathbf{x}_i||_2.$$

For AA-BP, sum-squared loss is 2-Lipschitz, and bounded by 2. Finally data-dependent risk bound for AA-BP is

Theorem 5. *Let $\mathcal{F}$ be the family of functions for AA-BP defined above, with sum-squared loss and weight constraints as Theorem 4, for any $\delta>0$, with probability at least $1-\delta$, for all $f\in\mathcal{F}$ such that*

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\sum_{i=1}^{m}(f_{\mathbf{x}}^{y_i}-\eta_{\mathbf{x}}^{y_i})^2\right]&\le\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}(\eta_{\mathbf{x}_i}^{y_j}-f_{\mathbf{x}_i}^{y_j})^2\\ &+\frac{8\sqrt{2}m^2L_\sigma B_0B_1}{\sqrt{n}}\max_{i\in[n]}||\mathbf{x}_i||_2+6\sqrt{\frac{\log 2/\delta}{2n}}.\end{aligned}\qquad (6)$$

SA-ME

SA-ME applies maximum entropy model to learn label distribution, i.e.,

$$\tilde{\eta}_{\mathbf{x}}^{y_j}=\frac{1}{Z}\exp(\mathbf{w}_j\cdot\mathbf{x}), \qquad (7)$$

where $Z=\sum_j\exp(\mathbf{w}_j\cdot\mathbf{x})$ is the normalization factor. Actually Eq. (7) can be regarded as a combination of softmax function and multi-output linear regression, namely $\mathrm{SF}\circ\mathcal{H}$, where $\mathcal{H}$ represents a class of functions of multi-output linear regression. SA-ME uses KL divergence as loss function, which is denoted by $\mathrm{KL}:\mathbb{R}^m\times\mathbb{R}^m\to\mathbb{R}_+$. Rademacher complexity of SA-ME w.r.t. S for loss function KL satisfies

Theorem 6. *Let $\mathcal{H}$ be a family of functions for multi-output linear regression, and $\mathcal{H}_j$ be a family of functions for the j-th output. Rademacher complexity of SA-ME with* KL *loss satisfies*

$$\hat{\mathcal{R}}_n(\mathrm{KL}\circ\mathrm{SF}\circ\mathcal{H}\circ S)\le(\sqrt{2m}+\sqrt{2})\sum_{j=1}^{m}\hat{\mathcal{R}}_n(\mathcal{H}_j\circ S), \qquad (8)$$

Proof. Note that $\mathrm{KL}(\mathbf{u},\cdot)$ is not ρ-Lipschitz over $\mathbb{R}^m$ for any $\rho\in\mathbb{R}$ and $\mathbf{u}\in\mathbb{R}^m$, thus Theorem 3 cannot be applied directly. Define function $\phi(\cdot,\cdot)$ as $\mathrm{KL}(\cdot,\mathrm{SF}(\cdot))$. Next we show that $\phi(\mathbf{u},\cdot)$ satisfy Lipschitzness. For $\mathbf{p},\mathbf{q}\in\mathbb{R}^m$,

$$|\phi(\mathbf{u},\mathbf{p})-\phi(\mathbf{u},\mathbf{q})|=|\mathrm{KL}(\mathbf{u},\mathrm{SF}(\mathbf{p}))-\mathrm{KL}(\mathbf{u},\mathrm{SF}(\mathbf{q}))|,$$

which equals

$$\begin{aligned}&\left|\sum_{i=1}^{m}u_i\left(\ln\frac{\exp(p_i)}{\sum_{j=1}^{m}\exp(p_j)}-\ln\frac{\exp(q_i)}{\sum_{j=1}^{m}\exp(q_j)}\right)\right|\\ &\le\sum_{i=1}^{m}\left|\ln(1+\sum_{j\ne i}\mathrm{e}^{p_j-p_i})-\ln(1+\sum_{j\ne i}\mathrm{e}^{q_j-q_i})\right|u_i.\end{aligned}$$

Observing that $\ln(1+\sum_j\exp v_i)$ is 1-Lipschitz for $\mathbf{v}\in\mathbb{R}^m$, thus right-hand side of preceding equation is bounded by

$$\begin{aligned}&\sum_{i=1}^{m}u_i||\mathbf{p}-\mathbf{1}\cdot p_i-\mathbf{q}+\mathbf{1}\cdot q_i||_2\\ &\le||\mathbf{p}-\mathbf{q}||_2+\sqrt{m}\sum_{i=1}^{m}u_i|p_i-q_i|\\ &\le(\sqrt{m}+1)||\mathbf{p}-\mathbf{q}||_2,\end{aligned}$$

namely, ϕ is $(\sqrt{m}+1)$-Lipschitz. Similar to the discussion of bounding rademacher complexity for AA-BP, we have

$$\hat{\mathcal{R}}_n(\mathrm{KL}\circ\mathrm{SF}\circ\mathcal{H}\circ S)\le(\sqrt{2m}+\sqrt{2})\sum_{j=1}^{m}\hat{\mathcal{R}}_n(\mathcal{H}_j\circ S),$$

which concludes the proof. □

As discussed above, although KL alone does not satisfy Lipschitzness, KL∘SF, however, is $(\sqrt{m}+1)$-Lipschitz. Define class of functions of j-th output with weight constraints as $\mathcal{H}_j=\{\mathbf{x}\to\mathbf{w}_j\cdot\mathbf{x}:||\mathbf{w}_j||_2\le 1\}$. According to (Kakade, Sridharan, and Tewari 2009), rademacher complexity of $\mathcal{H}_j$ satisfies

$$\hat{\mathcal{R}}_n(\mathcal{H}_j\circ S)\le\frac{\max_{i\in[n]}||\mathbf{x}_i||_2}{\sqrt{n}}.$$

Then right-hand side of Eq. (8) is bounded as

$$\hat{\mathcal{R}}_n(\mathrm{KL}\circ\mathrm{SF}\circ\mathcal{H}\circ S)\le\frac{(\sqrt{2m}+\sqrt{2})m}{\sqrt{n}}\max_{i\in[n]}||\mathbf{x}_i||_2.$$

Accordingly, data-dependent risk bound for SA-ME is

Theorem 7. *Let $\mathcal{F}$ be the family of functions for SA-ME defined above with KL divergence as loss function bounded by a constant b, for any $\delta > 0$, with probability at least $1-\delta$, for all $f \in \mathcal{F}$ such that*

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\sum_{j=1}^{m} \eta_{\mathbf{x}}^{y_j} \ln \frac{\eta_{\mathbf{x}}^{y_j}}{f_{\mathbf{x}}^{y_j}}\right] \leq \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} \eta_{\mathbf{x}_i}^{y_j} \ln \frac{\eta_{\mathbf{x}_i}^{y_j}}{f_{\mathbf{x}_i}^{y_j}} + 3b\sqrt{\frac{\log 2/\delta}{2n}} + \frac{2(\sqrt{2m}+\sqrt{2})m}{\sqrt{n}} \max_{i\in[n]} ||\mathbf{x}_i||_2. \tag{9}$$

As (Cha 2007) suggests that 0 is replaced by a very small value, say $\gamma > 0$, for division by 0 when implementing KL divergence, then for probability distribution $\mathbf{p}, \mathbf{q} \in \mathbb{R}^m$ with $p_i \geq \gamma, q_i \geq \gamma$

$$\mathrm{KL}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{m} p_i \ln \frac{p_i}{q_i} \leq \sum_{i=1}^{m} p_i \ln \frac{1}{\gamma} \leq -\ln \gamma,$$

thus there exists a constant $b \geq -\ln \gamma$ such that $\mathrm{KL}(\cdot,\cdot) \leq b$ (e.g., $b = 35$ for $\gamma = 1 \times 10^{-15}$).

Relation between LDL and Classification

LDL aims at learning the unknown label distribution function η by minimizing distance (or maximize similarity) between the given distribution and the output distribution. However, in practice, LDL is usually applied to perform classification. Firstly, a label distribution function $\tilde{\eta}$ is learned according to training sample S with label distribution. Secondly, a given instance $\mathbf{x}$ is classified by $\tilde{\eta}$, with label corresponding to the maximum predicted label description degree as the predicted label, i.e., $\tilde{y} = \arg\max_{y\in\mathcal{Y}} \tilde{\eta}(\mathbf{x}, y)$. In this part we tackle feasibility of performing classification of LDL algorithms. To make the analysis possible, we stay in probabilistic setting, where the underlying label distribution function η is the conditional probability distribution function, i.e., $\eta(\mathbf{x}, y) = \mathbb{P}(y|\mathbf{x})$.

Unlike LDL, where the unknown label distribution function η is deterministic though unknown, label variable for classification is stochastic, and sampled according to the conditional probability distribution. Thus we start with the plug-in decision theorem to bound error probability with risk of LDL (with absolute loss), then get the upper bound for error probability using union bound.

Generalized Plug-in Decision Theorem

As a preliminary, we firstly introduce the well-known *plug-in decision* theorem. The classic plug-in decision theorem applies to binary decision. For a given $\mathbf{x}$ with label y, then the Bayes decision function h^* is

$$h^*(\mathbf{x}) = \begin{cases} y_0 & \text{if } \eta_1(\mathbf{x}) \leq 1/2, \\ y_1 & \text{otherwise,} \end{cases}$$

where $\eta_1(\mathbf{x}) = \mathbb{P}(y = y_1|\mathbf{x})$. For another decision h with plug-in decision function

$$h(\mathbf{x}) = \begin{cases} y_0 & \text{if } \tilde{\eta}_1(\mathbf{x}) \leq 1/2, \\ y_1 & \text{otherwise,} \end{cases}$$

where $\tilde{\eta}_1(\mathbf{x})$ $(0 \leq \tilde{\eta}_1(\mathbf{x}) \leq 1)$ is an approximation of $\eta_1(\mathbf{x})$. Then we have

Theorem 8. (Devroye, Györfi, and Lugosi 1996) *For the plug-in decision function h defined above, difference of error probability between h and h^* satisfies*

$$\mathbb{P}(h(\mathbf{x}) \neq y) - \mathbb{P}(h^*(\mathbf{x}) \neq y) \leq 2\mathbb{E}_{\mathbf{x}}\left[|\tilde{\eta}_1(\mathbf{x}) - \eta_1(\mathbf{x})|\right].$$

The theorem states that if $\tilde{\eta}_1(\mathbf{x})$ is close to $\eta_1(\mathbf{x})$ in absolute value, then 0/1 risk of decision h is near that of the optimal decision function h^*. The preceding plug-in decision theorem only applies to binary classification, and we generalize it to multi-class classification. Formally, given an instance $\mathbf{x}$ with the conditional probability distribution function η (multi-class), the corresponding Bayes decision is

$$h^*(\mathbf{x}) = \arg\max_{y\in\mathcal{Y}} \eta(\mathbf{x}, y).$$

Similarly, assume we have access to $\tilde{\eta}$ that approximates η and the plug-in decision is defined as

$$h(\mathbf{x}) = \arg\max_{y\in\mathcal{Y}} \tilde{\eta}(\mathbf{x}, y).$$

Theorem 9. *For the plug-in decision function h defined above, we have*

$$\mathbb{P}(h(\mathbf{x}) \neq y) - \mathbb{P}(h^*(\mathbf{x}) \neq y) \leq \mathbb{E}_{\mathbf{x}}\left[\sum_{i=1}^{m} |\eta_i(\mathbf{x}) - \tilde{\eta}_i(\mathbf{x})|\right],$$

where $\eta_i(\mathbf{x})$, $\tilde{\eta}_i(\mathbf{x})$ represent $\eta(\mathbf{x}, y_i)$, $\tilde{\eta}(\mathbf{x}, y_i)$ respectively.

Proof. Given an instance $\mathbf{x}$, then the conditional error probability of h

$$\begin{aligned} \mathbb{P}(h(\mathbf{x}) \neq y|\mathbf{x}) &= 1 - \mathbb{P}(h(\mathbf{x}) = y|\mathbf{x}) \\ &= 1 - \sum_{i=1}^{m} \mathbb{P}(h(\mathbf{x}) = y_i, y = y_i|\mathbf{x}) \\ &= 1 - \sum_{i=1}^{m} \mathrm{Kr}(h(\mathbf{x}), y_i)\eta_i(\mathbf{x}), \end{aligned}$$

where $\mathrm{Kr}(\cdot,\cdot)$ is the Kronecker delta function. Then difference of conditional error probability between h and h^*

$$\begin{aligned} &\mathbb{P}(h(\mathbf{x}) \neq y|\mathbf{x}) - \mathbb{P}(h^*(\mathbf{x}) \neq y|\mathbf{x}) \\ &= \sum_{i=1}^{m} \left(\mathrm{Kr}(h^*(\mathbf{x}), y_i) - \mathrm{Kr}(h(\mathbf{x}), y_i)\right) \eta_i(\mathbf{x}). \end{aligned}$$

Without loss of generality, let k be the prediction of h and j be the prediction for h^*, then

$$\mathbb{P}(h(\mathbf{x}) \neq y|\mathbf{x}) - \mathbb{P}(h^*(\mathbf{x}) \neq y|\mathbf{x}) = \eta_j(\mathbf{x}) - \eta_k(\mathbf{x}). \tag{10}$$

Lemma 10. *Let $a \geq b$ and $d \geq c$, then $|a-b| \leq |a-c| + |b-d|$.*

Proof. If $c \leq b$ or $d \geq a$, the inequality is obvious. If $c \geq b$ and $a \geq d$, then $|a-b| = |a-c| + |c-b| \leq |a-c| + |b-d|$. □

For $k \neq j$, then $\eta_j(\mathbf{x}) \geq \eta_k(\mathbf{x})$ and $\tilde{\eta}_k(\mathbf{x}) \geq \tilde{\eta}_j(\mathbf{x})$, and according to Lemma 10, $|\eta_j(\mathbf{x}) - \eta_k(\mathbf{x})| \leq |\eta_j(\mathbf{x}) - \tilde{\eta}_j(\mathbf{x})| + |\eta_k(\mathbf{x}) - \tilde{\eta}_k(\mathbf{x})|$. For $k = j$, left-hand side of (10) reduces to 0. Thus we have

$$\mathbb{P}(h(\mathbf{x}) \neq y|\mathbf{x}) - \mathbb{P}(h^*(\mathbf{x}) \neq y|\mathbf{x}) \leq \sum_{i=1}^{m} |\eta_i(\mathbf{x}) - \tilde{\eta}_i(\mathbf{x})|. \tag{11}$$

Take expectation of both sides of Eq. (11) and we finish proof of Theorem 9. □

Table 1: Measures for LDL

Measure	Formula
Chebyshev ↓	$\mathrm{Dis}_1(\mathbf{p},\mathbf{q}) = \max_i \lvert p_i - q_i \rvert$
Clark ↓	$\mathrm{Dis}_2(\mathbf{p},\mathbf{q}) = \sqrt{\sum_i \frac{p_i - q_i)^2}{(p_i+q_i)^2}}$
Canberra ↓	$\mathrm{Dis}_3(\mathbf{p},\mathbf{q}) = \sum_i \frac{\lvert p_i - q_i \rvert}{p_i+q_i}$
Kullback-Leibler ↓	$\mathrm{Dis}_4(\mathbf{p},\mathbf{q}) = \sum_i p_i \ln \frac{p_i}{q_i}$
Cosine ↑	$\mathrm{Sim}_1(\mathbf{p},\mathbf{q}) = \frac{\sum_i p_i q_i}{\sqrt{\lvert\lvert\mathbf{p}\rvert\rvert_2}\sqrt{\lvert\lvert\mathbf{q}\rvert\rvert_2}}$
Intersection ↑	$\mathrm{Sim}_2(\mathbf{p},\mathbf{q}) = \sum_i \min(p_i, q_i)$

Measures for LDL

As we already have discussed in Theorem 9 that in probabilistic setting, *LDL with absolute loss is sufficient for classification*, because approximation to the conditional probability distribution function with absolute loss guarantees approximation to the optimal classifier. (Geng 2016) suggests six measures for LDL, i.e., Chebyshev distance, Clark distance, Canberra metric, KL divergence, cosine coefficient, and intersection similarity, which belong to Minkowski family, the χ^2 family, the L_1 family, the Shannon's entropy family, the inner product family, and the intersection family, respectively (Cha 2007). For probability distribution $\mathbf{p},\mathbf{q} \in \mathbb{R}^m$, formulation of the six measures are summarized in Table 1, where ↓ after the distance measures indicates "the smaller the better", and ↑ after the similarity measures indicates "the larger the better".

Theorem 11. *All measures in Table 1 are sufficient for LDL to perform classification.*

Proof. Denote absolute loss as $\mathrm{Dis}(\mathbf{p},\mathbf{q}) = \sum_i \lvert p_i - q_i \rvert$. For Chebyshev, Clark and Canberra distances, it's trivial to validate that $\mathrm{Dis}_1 \geq \frac{1}{m}\mathrm{Dis}$, $\mathrm{Dis}_3 \geq \frac{1}{2}\mathrm{Dis}$ and $\mathrm{Dis}_2 \geq \frac{1}{2\sqrt{m}}\mathrm{Dis}$. For KL distance, we have $\sqrt{\mathrm{Dis}_4} \geq \frac{1}{2}\mathrm{Dis}$. For cosine similarity, we have $1 - m\mathrm{Sim}_1 \geq \frac{1}{2}\mathrm{Dis}^2$, and for intersection similarity, it satisfies that $1 - \mathrm{Sim}_2 = \frac{1}{2}\mathrm{Dis}$. In conclusion, all measures in Table 1 bound absolute loss somehow. Details of the proof is left to appendix. □

Data-dependent Error Probability Bounds for LDL

As discussed above, error probability is directly correlated with absolute loss, and Theorem 11 states that measures in Table 1 bound absolute loss somehow. Accordingly, risk bounds for AA-kNN (absolute loss), AA-BP (sum-squared loss), and SA-ME (KL loss) can be extended to error probability bounds. Formally, let f be a learned label distribution function, define corresponding *decision function* as $g(f(\mathbf{x})) = \arg\max_{y \in \mathcal{Y}} f(\mathbf{x}, y)$.

Notice that the optimal error probability L^* exists in the left-hand side of Theorem 9, which can be bounded with Hoeffding's inequality. Moreover, the optimal classifier h^* outputs the label corresponding to the maximum conditional probability, and empirical error probability of the optimal classifier $h^*$is $L_S(h^*) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \max_{y \in \mathcal{Y}} \eta_{\mathbf{x}_i}^{y}\right)$. By Hoeffding's inequality, for any $\epsilon > 0$ such that

$$\mathbb{P}\left\{\lvert L_S(h^*) - L^* \rvert \geq \epsilon\right\} \leq 2\exp(-2n\epsilon^2),$$

namely for any $\delta > 0$, with probability at least $1 - \delta$,

$$\lvert L_S(h^*) - L^* \rvert \leq \sqrt{\frac{\ln 2/\delta}{2n}}. \tag{12}$$

Finally combine Theorem 9 and Equation 12 with the results of Theorem 2, Theorem 5 and Theorem 7 respectively, and we conclude with following theorems.

Theorem 12. *Let η_i be c_i-Lipschitz. Let $\tilde{\eta}$ be the output function of AA-kNN. Then for any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$\mathbb{P}(g(\tilde{\eta}(\mathbf{x})) \neq y) \leq \frac{8c\sqrt{d}}{\delta}\left(\frac{2k}{n}\right)^{1/(d+1)} + \sqrt{\frac{\ln 4/\delta}{2n}} + 1 - \frac{1}{n}\sum_{i=1}^{n}\max_{y \in \mathcal{Y}} \eta_{\mathbf{x}_i}^{y}.$$

Different with AA-kNN, AA-BP minimizes sum-squared loss, and note that for random variable $\mathbf{z} \in \mathbb{R}^m$

$$\mathbb{E}_{\mathbf{z}}\lvert\lvert\mathbf{z}\rvert\rvert_1 \leq \sqrt{m}\mathbb{E}_{\mathbf{z}}\lvert\lvert\mathbf{z}\rvert\rvert_2 \leq \sqrt{m}\sqrt{\mathbb{E}\lvert\lvert\mathbf{z}\rvert\rvert_2^2},$$

where the first inequality is according to Cauchy-Schwarz's inequality and the second one is according to Jensen's inequality. Thus error probability for AA-BP satisfies

Theorem 13. *Let $\mathcal{F}$ be a family of functions for AA-BP defined above. For any $\delta > 0$, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$, such that*

$$\mathbb{P}(g(f(\mathbf{x})) \neq y) \leq \sqrt{m}\surd\left(\frac{1}{n}\sum_{i,j}(\eta_{\mathbf{x}_i}^{y_j} - f_{\mathbf{x}}^{y_j})^2 + 7\sqrt{\frac{\log\frac{4}{\delta}}{2n}} + \frac{8\sqrt{2}m^2 L_\sigma B_0 B_1}{\sqrt{n}}\max_{i \in [n]}\lvert\lvert\mathbf{x}_i\rvert\rvert_2 + 1 - \frac{1}{n}\sum_{i=1}^{n}\max_{y \in \mathcal{Y}}\eta_{\mathbf{x}_i}^{y}\right),$$

where $\surd$ is the root operator.

Observing that for probability distribution $\mathbf{p},\mathbf{q} \in \mathbb{R}^m$, $\lvert\lvert\mathbf{p} - \mathbf{q}\rvert\rvert_1 \leq 2\sqrt{\mathrm{KL}(\mathbf{p},\mathbf{q})}$, which implies that

Theorem 14. *Let $\mathcal{F}$ be family of functions for SA-ME defined above, and $b' = 3(b+1)$. For any $\delta > 0$, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$ such that*

$$\mathbb{P}(g(f(\mathbf{x})) \neq y) \leq 2\surd\left(\frac{1}{n}\sum_{i,j}\eta_{\mathbf{x}_i}^{y_j}\ln\frac{\eta_{\mathbf{x}_i}^{y_j}}{f_{\mathbf{x}_i}^{y_j}} + b'\sqrt{\frac{\log\frac{4}{\delta}}{2n}} + \frac{2(\sqrt{2m}+\sqrt{2})m}{\sqrt{n}}\max_{i \in [n]}\lvert\lvert\mathbf{x}_i\rvert\rvert_2 + 1 - \frac{1}{n}\sum_{i=1}^{n}\max_{y \in \mathcal{Y}}\eta_{\mathbf{x}_i}^{y}\right),$$

Conclusion

This paper studies learnability of LDL from two aspects, i.e., generalization of LDL itself, and generalization of LDL to perform classification. On one hand, for generalization of LDL, risk bounds for three representative LDL algorithms, i.e., AA-kNN, AA-BP and SA-ME are provided, with convergence rate $\mathcal{O}((\frac{k}{n})^{1/d+1})$, $\mathcal{O}(\frac{1}{\sqrt{n}})$, $\mathcal{O}(\frac{1}{\sqrt{n}})$ respectively, which indicates learnability of LDL. On the other hand, for generalization of LDL to perform classification, a generalized plug-in decision theorem is proposed, discovering that minimizing absolute loss is sufficient for a corresponding LDL decision function approaching the optimal classifier. Furthermore, six commonly used LDL measures are also shown to be sufficient for classification, for the reason that all six measures bound absolute loss somehow. Besides, data-dependent error probability bounds for LDL are given to demonstrate feasibility of LDL to perform classification.

Appendices

Details of Proof of Theorem 11

For discrete probability distribution $\mathbf{p}, \mathbf{q} \in \mathbb{R}^m$, the missing proof for relation between measures in Table 1 and absolute loss is as following.

Intersection Similarity. For intersection similarity and absolute distance, it satisfies $1 - \text{Sim}_2 = \text{Dis}/2$.

Proof. Firstly for $a, b \in \mathbb{R}$, we have $\min(a, b) = (a+b)/2 - |a - b|/2$, and it follows that $\text{Sim}_2(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{m}[(p_i + q_i)/2 - |p_i - q_i|/2] = 1 - \text{Dis}/2$ □

Cosine Similarity. For cosine similarity and absolute distance, it satisfies $1 - m\text{Sim}_1 \geq \text{Dis}^2/2$.

Proof. According to cosine law, we have

$$|||\mathbf{p}||_2^2 + ||\mathbf{q}||_2^2 - 2||\mathbf{p}||_2||\mathbf{q}||_2\text{Cosine}(\mathbf{p}, \mathbf{q}) = ||\mathbf{p} - \mathbf{q}||_2^2.$$

Notice that $||\mathbf{p}||_1 = 1$ and $||\mathbf{q}||_1 = 1$. Apply Cauchy-Schwarz's inequality to both sides of above equation, and it follows that $m/2 - 2\text{Cosine}(\mathbf{p}, \mathbf{q}) \geq \text{Dis}^2/m$. □

Kullback-Leibler Distance. For Kullback-Leibler distance and absolute distance, it satisfies $\sqrt{\text{Dis}_4} \geq \text{Dis}/2$.

Proof. This proof is according to (Cover and Thomas 2012), where the probability distribution is continuous, and here KL divergence is applied to discrete probability distribution. Observing $-\text{Dis}_4(\mathbf{p}, \mathbf{q}) = \sum_i p_i \ln \frac{q_i}{p_i}$, which equals

$$\begin{aligned}
&\sum_i p_i \left\{\ln \min(q_i/p_i, 1) + \ln \max(q_i/p_i, 1)\right\} \\
&\leq \ln \sum_i p_i \min(\frac{q_i}{p_i}, 1) + \ln \sum_i p_i \max(\frac{q_i}{p_i}, 1) \\
&\leq \ln \sum_i \min(p_i, q_i) + \ln \sum_i \max(p_i, q_i) \\
&\leq \ln \sum_i \frac{p_i + q_i - |p_i - q_i|}{2} + \ln \sum_i \frac{p_i + q_i + |p_i - q_i|}{2} \\
&\leq \ln \left\{1 - \text{Dis}(p, q)/2\right\} + \ln \left\{1 + \text{Dis}(p, q)/2\right\} \\
&\leq \ln \left\{1 - \text{Dis}^2/4\right\},
\end{aligned}$$

where the third inequality is according to Jensen's inequality. Arrange items, and it follows that $(\frac{1}{2}\text{Dis})^2 \leq 1 - \exp(-\text{Dis}_4) \leq \text{Dis}_4$, where the second inequality is according to the trivial inequality $1 - \mathrm{e}^{-x} \leq x$ for any $x \geq 0$, which concludes the proof. □

Proof of Equation (3)

The proof is according to the work of (Shalev-Shwartz and Ben-David 2014), which bounds expected distance between a random $\mathbf{x}$ and its closest neighbor in the training set.

Proof. For $k = 1$, according to (Shalev-Shwartz and Ben-David 2014),

$$\mathbb{E}_{\mathbf{x},S}\left[||\mathbf{x}_{N_1(\mathbf{x})} - \mathbf{x}||_2\right] \leq 4\sqrt{d}n^{-\frac{1}{d+1}},$$

which satisfies Eq. (3). One can easily check that when $k = 1$, the right-hand side of Eq. (3) is $2^{1/(d+1)}4\sqrt{d}n^{-1/(d+1)} > 4\sqrt{d}n^{-1/(d+1)}$, thus Eq. (3) holds. Furthermore for $k \geq 2$,

Lemma 15. (Shalev-Shwartz and Ben-David 2014) *Let $C_1, C_2, \ldots, C_r$ be a collection of subsets of $\mathcal{X}$, then for any $k \geq 2$,*

$$\mathbb{E}_S\left[\sum_{i:|C_i \cap S| < k} \mathbb{P}[C_i]\right] \leq \frac{2rk}{n}.$$

Firstly, let $C_1, C_2, \ldots, C_r$ be the cover of $\mathcal{X} = [0, 1]^d$ using boxes of length λ. Then for $\mathbf{x}$ and its k nearest neighbors in the same box, we have $\sum_{j \in N_k(\mathbf{x})} ||\mathbf{x} - \mathbf{x}_j||_2 \leq k\sqrt{d}\lambda$, otherwise $\sum_{j \in N_k(\mathbf{x})} ||\mathbf{x} - \mathbf{x}_j||_2 \leq k\sqrt{d}$. Therefore,

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},S}\left[\sum_{j \in N_k(\mathbf{x})} ||\mathbf{x}_j - \mathbf{x}||_2\right] &\leq k\sqrt{d}\mathbb{E}_S\left[\sum_{i:|C_i \cap S| < k} \mathbb{P}[C_i]\right] \\
&+ k\sqrt{d}\lambda\mathbb{E}_S\left[\sum_{i:|C_i \cap S| \geq k} \mathbb{P}[C_i]\right].
\end{aligned}$$

By Lemma 15 and trivial bound $\sum_{i:|C_i \cap S| \geq k} \mathbb{P}(C_i) \leq 1$, it follows that

$$\mathbb{E}_{\mathbf{x},S}\left[\sum_{j \in N_k(\mathbf{x})} ||\mathbf{x}_j - \mathbf{x}||_2\right] \leq k\sqrt{d}(\frac{2rk}{n} + \lambda).$$

Since number of boxes $r = (1/\lambda)^d$, set $\lambda = (2kd/n)^{(1/d+1)}$ to maximize the right-hand side of the preceding equation,

$$\mathbb{E}_{\mathbf{x},S} \sum_{j \in N_k(\mathbf{x})} ||\mathbf{x}_j - \mathbf{x}||_2 \leq \sqrt{d}k\left(\frac{2k}{n}\right)^{\frac{1}{d+1}}(d^{\frac{-d}{d+1}} + d^{\frac{1}{d+1}}).$$

Notice that function $d^{-d/d+1} + d^{1/d+1}$ is monotone decreasing with maximum 2 when $d = 1$. Therefore,

$$\mathbb{E}_{\mathbf{x},S}\left[\sum_{j \in N_k(\mathbf{x})} ||\mathbf{x}_j - \mathbf{x}||_2\right] \leq 2\sqrt{d}k\left(\frac{2k}{n}\right)^{\frac{1}{d+1}},$$

and Eq. (3) holds as well, which concludes the proof. □

Acknowledgments

This research was supported by the National Key Research & Development Plan of China (No. 2017YFB1002801), the National Science Foundation of China (61622203), the Jiangsu Natural Science Funds for Distinguished Young Scholar (BK20140022), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the Collaborative Innovation Center of Wireless Communications Technology.

References

Bartlett, P. L., and Mendelson, S. 2003. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* 3:463–482.

Berger, A. L.; Pietra, V. J. D.; and Pietra, S. A. D. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist.* 22(1):39–71.

Biau, G., and Devroye, L. 2015. *Lectures on the nearest neighbor method*. Springer.

Borchani, H.; Varando, G.; Bielza, C.; and Larrañaga, P. 2015. A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5(5):216–233.

Cha, S.-H. 2007. Comprehensive survey on distance/similarity measures between probability density functions.

Cover, T. M., and Thomas, J. A. 2012. *Elements of information theory*. John Wiley & Sons.

Devroye, L., and Krzyźak, A. 1989. An equivalence theorem for l1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference* 23(1):71 – 82.

Devroye, L.; Györfi, L.; and Lugosi, G. 1996. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Gao, W., and Zhou, Z.-H. 2016. Dropout rademacher complexity of deep neural networks. *Science China Information Sciences* 59(7):072104.

Geng, X., and Hou, P. 2015. Pre-release prediction of crowd opinion on movies by label distribution learning. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, 3511–3517.

Geng, X., and Ling, M. 2017. Soft video parsing by label distribution learning. In *AAAI Conference on 31th AAAI Conference on Artificial Intelligence, AAAI'17*, 1331–1337.

Geng, X., and Xia, Y. 2014. Head pose estimation based on multivariate label distribution. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR'14*, 1837–1842.

Geng, X.; Yin, C.; and Zhou, Z. 2013. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(10):2401–2412.

Geng, X. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering* 28(7):1734–1748.

Hou, P.; Geng, X.; Huo, Z.-W.; and Lv, J.-Q. 2017. Semi-supervised adaptive label distribution learning for facial age estimation. In *AAAI Conference on 31th AAAI Conference on Artificial Intelligence, AAAI'17.*

Jia, X.; Li, W.; Liu, J.; and Zhang, Y. 2018. Label distribution learning by exploiting label correlations. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI'18.*

Kakade, S. M.; Sridharan, K.; and Tewari, A. 2009. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems, NIPS'09*, 793–800.

Li, Y.; Zhang, M.; and Geng, X. 2015. Leveraging implicit relative labeling-importance information for effective multi-label learning. In *2015 IEEE International Conference on Data Mining*, 251–260.

Maurer, A. 2016. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory*, 3–17.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2012. *Foundations of machine learning*. MIT press.

Ren, Y., and Geng, X. 2017. Sense beauty by label distribution learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI'17*, 2648–2654.

Shalev-Shwartz, S., and Ben-David, S. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Stone, C. J. 1977. Consistent nonparametric regression. *The Annals of Statistics* 5(4):595–620.

Villani, C. 2008. *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Xu, M., and Zhou, Z.-H. 2017. Incomplete label distribution learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI'17*, 3175–3181.

Xu, N.; Tao, A.; and Geng, X. 2018. Label enhancement for label distribution learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI'18*, 2926–2932.

Zhang, M., and Zhou, Z. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26(8):1819–1837.

Zhao, P., and Zhou, Z.-H. 2018. Label distribution learning by optimal transport. In *AAAI Conference on 32th AAAI Conference on Artificial Intelligence, AAAI'18.*

Zheng, X.; Jia, X.; and Li, W. 2018. Label distribution learning by exploiting sample correlations locally. In *AAAI Conference on 32th AAAI Conference on Artificial Intelligence, AAAI'18.*

Zhou, Y.; Xue, H.; and Geng, X. 2015. Emotion distribution recognition from facial expressions. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, 1247–1250. New York, NY, USA: ACM.