Learning Curves of Classification Metrics based on Confusion Matrices

Yan Xue 202012407011@EMAIL.SXU.EDU.CN
School of Computer and Information Technology, Shanxi University, Taiyuan, China

Ruibo Wang WANGRUIBO@SXU.EDU.CN
School of Modern Educational Technology, Shanxi University, Taiyuan, China

Xuefei Cao CAOXUEFEI@SXU.EDU.CN
School of Automation and Software Engineering, Shanxi University, Taiyuan, China

Jing Yang 202223604008@EMAIL.SXU.EDU.CN
School of Automation and Software Engineering, Shanxi University, Taiyuan, China

Jihong Li LI_ML@SXU.EDU.CN
School of Modern Educational Technology, Shanxi University, Taiyuan, China

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Learning curves of classification metrics, including test error, precision (P), recall (R), F_1 score, with regard to training set sizes are a recent hot topic in the development of advanced methodologies for model selection and hyperparameter optimization. Existing studies concentrated on formulating the functional shapes of well-behaved learning curves of test error by using a normality assumption. However, the normality assumption is unreasonable for the learning curves of classification metrics because the distributions of most classification metrics, such as P, R, and F_1 score, are skewed, and interval estimations of these metrics based on the normality assumption may exceed [0,1]. In this study, considering that most classification metrics are obtained from confusion matrices, we develop a novel method to formulate the learning curves of classification metrics. Specifically, it is assumed that the four entries in a confusion matrix jointly follow a multi-nomial distribution rather than a normality distribution. Furthermore, the function of each entry in a confusion matrix with regard to training set sizes is formulated with an exponential form. Thus, the learning curves of a classification metric can be naturally obtained by transforming the functions of a confusion matrix in terms of the definition of the metric. Moreover, reasonable confidence bands of several popular metrics, including test error, P, R, and F_1 score, are derived in this study based on the assumption of the multi-nomial distribution of a confusion matrix. Extensive experiments are conducted on several synthetic and real-world datasets coupled with multiple typical non-neural and neural classification algorithms. Experimental results illustrate the improvements of the proposed learning curves of test error, P, R, and F_1 score and the superiority of the confidence bands.

Keywords: Learning Curve; Test Error; Confusion Matrix; Normality Assumption.

1. Introduction

Although the notion of learning curve was introduced nearly 30 years ago (Cortes et al., 1993), it has recently become an important tool to evaluate the performance of various classification algorithms and improve model selection of deep networks (Hoiem et al., 2021; Mohr and van Rijn, 2023; Dhamija et al., 2018; Hua et al., 2025; Wang et al., 2023, 2025). Learning curves formalize a classification metric as a function of training set size, thus overcoming the limitations of standard evaluation methods with fixed train-test splits and accelerating the development of an advanced methodology in model selection task. The model selection methodology depends not only on the forms of learning curves but also on the quality of the estimation of learning curves.

To date, many learning curves have been investigated both theoretically and empirically (Gu et al., 2001; Perlich et al., 2003; Kolachina et al., 2012; Li et al., 2023; Ruben and Pehlevan, 2023; Adriaensen et al., 2023) and surveyed in several comprehensive studies (Mohr and van Rijn, 2022; Viering and Loog, 2023). Intuitively, the performance of a classification algorithm is frequently improved with more training data, and the corresponding learning curves are considered well-behaved.

For well-behaved learning curves, a dozen parametric forms have been proposed (Frey and Fisher, 1999; Boonyanunta and Zeephongsekul, 2004; Singh, 2005; Brumen et al., 2014; Viering and Loog, 2023). The parametric forms can be categorized into two clusters of power-law and exponential shapes (Viering and Loog, 2023). These learning curves frequently concentrated on the metric of test error rather than other classification metrics, including precision, recall, and F_1 score. Moreover, the parameters in these forms are typically estimated with a least square method and a Levenberg-Marquardt method (Moré, 1977), and a normal distribution is frequently used as an assumption in the estimation method (Hoiem et al., 2021).

However, the assumption of a normal distribution is not suitable for the commonly-used classification metrics. In practice, the popular classification metrics, such as test error, precision, recall, F_1 score, and so on, take values within the range of $[0, 1]$, but many learning curves of these metrics based on the normal distribution may exceed $[0, 1]$. For example, considering a widely recommended inverse power law learning curve of test error, i.e., $e_{ij} = \alpha + \eta n_i^{\gamma} + \epsilon_{ij}$ with $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ (Hoiem et al., 2021), if the parameters α and η are estimated improperly, the estimation of test error has a large chance to be out of $[0, 1]$.

Another important reason to illustrate the irrationality of the normality assumption in learning curves is that many classification metrics are distributed in a skewed manner rather than a normal manner Qi et al. (2021). For example, Goutte and Gaussier (2005) showed that precision, as well as recall, follows a Beta distribution. Wang et al. (2014) proved that the distribution of F_1 score is a reciprocal form with regard to a Beta-prime distribution. Thus, novel parametric forms of learning curves and the corresponding estimation methods should be developed based on a reasonable assumption rather than a normality assumption.

In this study, for a binary classification task, considering that popular classification metrics are obtained on a confusion matrix (Table 1), we introduce an assumption about the distribution of a confusion matrix. Specifically, the four counts in a confusion matrix, namely, true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) are assumed to jointly follow a multi-nomial distribution (Goutte and Gaussier, 2005).

Furthermore, the logits of the parameters of the multi-nomial distribution are regarded to be linear with respect to the power law of training set size. Thus, closed-form expressions for the learning curves of the parameters are obtained and similar to those of ordinary logistic regression forms. Then, a novel method of maximizing regularized likelihood objective is proposed for estimating the learning curves. Correspondingly, learning curves of classification metrics can be obtained by functionally transforming the learning curves of the parameters of the multi-nomial distribution in terms of the definition of the metrics over a confusion matrix.

We also develop reasonable confidence bands of the learning curves of test error, precision, recall, and F_1 score. On the basis of the multi-nomiality assumption, from a Bayesian perspective, the posterior distributions of test error, precision, recall, and F_1 score can be naturally obtained (Goutte and Gaussier, 2005; Wang et al., 2014). Then, the corresponding confidence bands can be derived from the posterior distributions. The confidence bands are consistently located within the range of $[0, 1]$ and relatively conservative with a sufficient confidence degree. In contrast, the confidence bands induced from a normality assumption (Hoiem et al., 2021) may exceed $[0, 1]$ and possess lower confidence degrees that would tend to produce false positive decisions in a model selection task.

We compare our proposed learning curves with the majority of existing parametric learning curves on sufficient numerical experiments. The experiments involve multiple synthetic and real-world datasets as well as multiple non-neural and neural classification algorithms. The root mean squared error (RMSE) is used as a measure to evaluate the learning curves, and the measures of degree of confidence (DoC) and interval length (IL) are employed to assess the quality of the proposed confidence bands. Experimental results show that our learning curves achieve an obvious improvement in fitting the learning behavior of a classification algorithm with regard to training set size, and the proposed confidence bands are superior to the confidence bands obtained from a normality assumption.

In summary, our contributions in this study are listed as follows.

1. We obtain novel parametric forms of learning curves of test error, precision, recall, and F_1 score based on a multi-nomiality assumption over a confusion matrix.
2. We formulate a proper method for estimating the parameters of learning curves based on a regularized optimization of maximizing likelihood objective.
3. We develop reasonable confidence bands with regard to the learning curves of test error, precision, recall, and F_1 score.

2. Preliminaries and Our Method

In this section, we mainly consider the binary classification setting and four popular classification metrics, namely, test error, precision, recall, and F_1 score. Hence, we concentrate on the question of **how to approximate the learning curves of these four metrics in a supervised binary classification task**? For multi-class classification task and other classification metrics, an extension of the proposed method will be investigated in future work. Considering that the four metrics are computed on a confusion matrix, we first introduce the definition and probabilistic interpretation of a confusion matrix and then

		Gold labels	
		+	-
Predictions	+	True Positives	False Positives
	-	False Negatives	True Negatives

Table 1: Illustration of a typical confusion matrix.

derive the theoretical forms and estimations of the learning curves of the metrics through investigating the functional relationships of a confusion matrix with regard to training set sizes.

2.1. Confusion Matrix

In a binary classification task, we assume that a dataset D_n consists of n samples, namely, $D_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where the samples are independently sampled from unknown distribution $\mathcal{P}$, $\mathbf{x}_i$ is the i-th numerical object, and $y_i \in \mathscr{Y}$ is the corresponding class label of $\mathbf{x}_i$ such that $\mathscr{Y} = \{+, -\}$. Let $\mathbb{A}$ be a binary classification algorithm, and an implementation of algorithm $\mathbb{A}$ trained on dataset D_n is denoted as a model $\mathbb{A}(D_n)$. Furthermore, a classification metric is introduced to assess the true performance of model $\mathbb{A}(D_n)$. Several commonly-used metrics, including test error, precision, recall, and F$_1$ score, are considered in this section.

On the basis of a test set $T_{n'} \sim \mathcal{P}^{n'}$, a typical classification metric of model $\mathbb{A}(D_n)$ can be estimated from a confusion matrix. As illustrated in Table 1, a conventional confusion matrix $\boldsymbol{C}$ consists of four counts, namely, TP, FP, FN, and TN, such that $\text{TP} + \text{FP} + \text{FN} + \text{TN} = n'$. Define $n'_+ = \text{TP} + \text{FN}$ is the number of positive samples in $T_{n'}$ and $n'_- = \text{TN} + \text{FP}$ is the number of negative samples in $T_{n'}$. Then, the estimators of test error (e_n), precision (p_n), recall (r_n), and F$_1$ score ($f_{1,n}$) are computed through Eq. (1) where the subscript n indicates the size of D_n used in a training process of algorithm $\mathbb{A}$.

$$\hat{e}_n = \frac{\text{TP} + \text{TN}}{n'}, \quad \hat{p}_n = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \hat{r}_n = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{n'_+}, \quad \hat{f}_{1,n} = \frac{2\hat{p}\hat{r}}{\hat{p} + \hat{r}}. \tag{1}$$

It is noted that the accuracy of model $\mathbb{A}(D_n)$ is computed as $a_n = 1 - e_n$, and thus our proposed method can be naturally generalized to the metric of accuracy. Formally, our goal in this study is to obtain closed-form expressions of the learning curves of e_n, p_n, r_n, and $f_{1,n}$ with regard to n.

2.2. Probabilistic Interpretation of a Confusion Matrix

A conventional confusion matrix $\boldsymbol{C} = (\text{TP}, \text{FP}, \text{FN}, \text{TN})$ can be regarded as a sample of counts drawn from a multi-nomial distribution (Goutte and Gaussier, 2005) as follows.

$$\boldsymbol{C} \sim \mathcal{M}(n'; \boldsymbol{\pi}_n), \tag{2}$$

where $\mathcal{M}$ stands for a multi-nomial distribution parameterized by n' and a probabilistic vector $\boldsymbol{\pi}_n = (\pi_{\text{TP},n}, \pi_{\text{FP},n}, \pi_{\text{FN},n}, \pi_{\text{TN},n})$ such that $\pi_{\text{TP},n} + \pi_{\text{FP},n} + \pi_{\text{FN},n} + \pi_{\text{TN},n} = 1$. Let

$\pi_+ = \pi_{\mathrm{TP},n} + \pi_{\mathrm{FN},n}$ be the prior probability of positive class and $\pi_- = \pi_{\mathrm{FP},n} + \pi_{\mathrm{TN},n} = 1 - \pi_+$ be the prior probability of negative class. Therefore, we obtain that $n'_+ = n'\pi_+$ and $n'_- = n'\pi_-$.

In this study, we assume that π_+ and π_- are two constants conditioned on $\mathcal{P}$ and unrelated to n. Moreover, the subscription n used in $\boldsymbol{\pi}_n$ is used to indicate that different training set sizes of D_n could induce different values of $\boldsymbol{\pi}_n$.

Conditioned on priors of π_+ and π_-, four normalized probabilities $\mathbf{\Pi}_n = (\Pi_{\mathrm{TP},n}, \Pi_{\mathrm{FP},n}, \Pi_{\mathrm{FN},n}, \Pi_{\mathrm{TN},n})$ can be defined in Eq. (3).

$$\Pi_{\mathrm{TP},n} = \frac{\pi_{\mathrm{TP},n}}{\pi_+}, \quad \Pi_{\mathrm{FP},n} = \frac{\pi_{\mathrm{FP},n}}{\pi_-}, \quad \Pi_{\mathrm{FN},n} = \frac{\pi_{\mathrm{FN},n}}{\pi_+}, \quad \Pi_{\mathrm{TN},n} = \frac{\pi_{\mathrm{TN},n}}{\pi_-}. \tag{3}$$

Obviously, the normalized probabilities satisfy the constraints in Eq. (4).

$$\Pi_{\mathrm{TP},n} + \Pi_{\mathrm{FN},n} = 1, \qquad \Pi_{\mathrm{TN},n} + \Pi_{\mathrm{FP},n} = 1. \tag{4}$$

From the perspective of the multi-nomial distribution in Eq. (2), several conditional distribution with regard to confusion matrix $\boldsymbol{C}$ can be naturally obtained as showed in Eqs. (5) and (6).

$$\mathrm{TP}|n'_+ \sim \mathcal{B}(n'_+; \Pi_{\mathrm{TP},n}), \qquad \mathrm{FP}|n'_- \sim \mathcal{B}(n'_-; \Pi_{\mathrm{FP},n}), \tag{5}$$

$$\mathrm{FN}|n'_+ \sim \mathcal{B}(n'_+; \Pi_{\mathrm{FN},n}), \qquad \mathrm{TN}|n'_- \sim \mathcal{B}(n'_-; \Pi_{\mathrm{TN},n}), \tag{6}$$

where $\mathcal{B}$ stands for a binomial distribution. In particular, the estimator of test error, namely, $\hat{e}_n$ in Eq. (1), follows the binomial distribution in Eq. (7).

$$\hat{e}_n \sim \mathcal{B}(n'; \pi_{\mathrm{FP},n} + \pi_{\mathrm{FN},n}). \tag{7}$$

On the basis of $\boldsymbol{\pi}_n$, the metrics of test error, precision, recall, and F_1 score are expressed in Eq. (8).

$$e_n = \pi_{\mathrm{FP},n} + \pi_{\mathrm{FN},n}, \quad p_n = \frac{\pi_{\mathrm{TP},n}}{\pi_{\mathrm{TP},n} + \pi_{\mathrm{FP},n}}, \quad r_n = \Pi_{\mathrm{TP},n}, \quad f_{1,n} = \frac{2p_n r_n}{p_n + r_n}. \tag{8}$$

From a Bayesian perspective, assuming the priors of precision and recall are a Beta distribution, namely, $p_n \sim \mathcal{B}e(\lambda, \lambda)$ and $r_n \sim \mathcal{B}e(\lambda, \lambda)$, then it has been proved that the posterior distributions of precision and recall are the Beta distributions in Eq. (9) (Goutte and Gaussier, 2005).

$$p_n \sim \mathcal{B}e(\mathrm{TP}_n + \lambda, \mathrm{FP}_n + \lambda), \qquad r_n \sim \mathcal{B}e(\mathrm{TP}_n + \lambda, \mathrm{FN}_n + \lambda). \tag{9}$$

Analogously, for the metric of test error, assuming that $e_n \sim \mathcal{B}e(\lambda, \lambda)$ and considering the likelihood of test error in Eq. (7), we obtain the posterior distribution of test error in Eq. (10).

$$e_n \sim \mathcal{B}e(\mathrm{FP}_n + \mathrm{FN}_n + \lambda, \mathrm{TP}_n + \mathrm{TN}_n + \lambda). \tag{10}$$

For the metric of F_1 score, it has been proved that the form of its posterior distribution is a closed-form expression in Eq. (11) (Wang et al., 2014).

$$P(t) = \frac{2^a (1-t)^{a-1} (2-t)^{-a-b} t^{b-1}}{\mathcal{B}e(a, b)}, \tag{11}$$

where $0 \leq t \leq 1$ and $a = \mathrm{FP}_n + \mathrm{FN}_n + 2\lambda$ and $b = \mathrm{TP}_n + \lambda$.

Two popular Beta priors can be considered: Jeffrey's non-informative prior with $\lambda = 1/2$, and the uniform prior with $\lambda = 1$.

2.3. Closed Forms of the Proposed Learning Curves

Previous work frequently use a normality distribution as an assumption to build the closed forms of learning curves of test error (Hoiem et al., 2021). However, Eq. (7) shows that the test error is not normally distributed. This observation motivates us to develop novel learning curves for the metrics of test error, precision, recall, and F_1 score under a more reasonable assumption.

Eq. (8) illustrates that the considered classification metrics can be analytically expressed with $\boldsymbol{\pi}_n$. Therefore, an intuition for obtaining closed-form expressions of the learning curves of the metrics is to first develop the functions of $\boldsymbol{\pi}_n$ with regard to n. These functions are also named as learning curves of $\boldsymbol{\pi}_n$ in this study.

Considering that the normalized probabilities of $\boldsymbol{\pi}_n$, i.e., $\boldsymbol{\Pi}_n$, are the parameters of the binomial distributions in Eqs. (5) and (6), a natural idea to formalize the relationship of $\boldsymbol{\Pi}_n$ with n is to use logit functions. Specifically, we express the logit function of $\Pi_{\mathrm{TP},n}$ in Eq. (12).

$$logit(\Pi_{\mathrm{TP},n}) = \ln \frac{\Pi_{\mathrm{TP},n}}{1 - \Pi_{\mathrm{TP},n}} = \alpha_{\mathrm{TP}} + \eta_{\mathrm{TP}} n^{\gamma}. \tag{12}$$

Then, a learning curve of $\Pi_{\mathrm{TP},n}$ is expressed as a logistic regression form in Eq. (13).

$$\Pi_{\mathrm{TP},n} = \frac{e^{\alpha_{\mathrm{TP}} + \eta_{\mathrm{TP}} n^{\gamma}}}{1 + e^{\alpha_{\mathrm{TP}} + \eta_{\mathrm{TP}} n^{\gamma}}}, \tag{13}$$

where α_{TP}, η_{TP}, and γ are three parameters to be estimated on test dataset $T_{n'}$. Furthermore, in terms of the constraint in Eq. (4), the learning curve of $\Pi_{\mathrm{FN},n}$ is expressed in Eq. (14).

$$\Pi_{\mathrm{FN},n} = \frac{1}{1 + e^{\alpha_{\mathrm{TP}} + \eta_{\mathrm{TP}} n^{\gamma}}}. \tag{14}$$

Analogously, the learning curves of $\Pi_{\mathrm{TN},n}$ and $\Pi_{\mathrm{FP},n}$ can be expressed in Eq. (15).

$$\Pi_{\mathrm{TN},n} = \frac{e^{\alpha_{\mathrm{TN}} + \eta_{\mathrm{TN}} n^{\gamma}}}{1 + e^{\alpha_{\mathrm{TN}} + \eta_{\mathrm{TN}} n^{\gamma}}}, \quad \Pi_{\mathrm{FP},n} = \frac{1}{1 + e^{\alpha_{\mathrm{TN}} + \eta_{\mathrm{TN}} n^{\gamma}}}. \tag{15}$$

According to the definitions of $\boldsymbol{\Pi}_n$ in Eq. (3), the learning curves of $\boldsymbol{\pi}_n$ are provided in Eqs. (16) and (17).

$$\pi_{\mathrm{TP},n} = \frac{\pi_{+} e^{\alpha_{\mathrm{TP}} + \eta_{\mathrm{TP}} n^{\gamma}}}{1 + e^{\alpha_{\mathrm{TP}} + \eta_{\mathrm{TP}} n^{\gamma}}}, \quad \pi_{\mathrm{FN},n} = \frac{\pi_{+}}{1 + e^{\alpha_{\mathrm{TP}} + \eta_{\mathrm{TP}} n^{\gamma}}}, \tag{16}$$

$$\pi_{\mathrm{TN},n} = \frac{\pi_{-} e^{\alpha_{\mathrm{TN}} + \eta_{\mathrm{TN}} n^{\gamma}}}{1 + e^{\alpha_{\mathrm{TN}} + \eta_{\mathrm{TN}} n^{\gamma}}}, \quad \pi_{\mathrm{FP},n} = \frac{\pi_{-}}{1 + e^{\alpha_{\mathrm{TN}} + \eta_{\mathrm{TN}} n^{\gamma}}}. \tag{17}$$

On the basis of the definitions in Eq. (8), our proposed learning curves of the classification metrics can be obtained and showed in Eqs. (18), (19), and (20).

$$\begin{aligned} e_n &= \frac{\pi_+}{1+e^{\alpha_{\mathrm{TP}}+\eta_{\mathrm{TP}}n^\gamma}} + \frac{\pi_-}{1+e^{\alpha_{\mathrm{TN}}+\eta_{\mathrm{TN}}n^\gamma}}, && (18)\\ p_n &= \frac{1}{1+\frac{\pi_-(1+e^{\alpha_{\mathrm{TP}}+\eta_{\mathrm{TP}}n^\gamma})}{\pi_+ e^{\alpha_{\mathrm{TP}}+\eta_{\mathrm{TP}}n^\gamma}(1+e^{\alpha_{\mathrm{TN}}+\eta_{\mathrm{TN}}n^\gamma})}}, && (19)\\ r_n &= \frac{e^{\alpha_{\mathrm{TP}}+\eta_{\mathrm{TP}}n^\gamma}}{1+e^{\alpha_{\mathrm{TP}}+\eta_{\mathrm{TP}}n^\gamma}}. && (20) \end{aligned}$$

In addition, the learning curve of F_1 score can be obtained by substituting Eqs. (19) and (20) into $f_n = 2p_n r_n/(p_n + r_n)$.

It is noted that the parameters of the proposed learning curves in Eqs. (18)-(20) are α_{TP}, α_{TN}, η_{TP}, η_{TN}, and γ. The estimation methods of these parameters are discussion in the next subsection.

2.4. Estimating the Proposed Learning Curves

We can consider the forms of $\Pi_{\mathrm{TP},n}$ and $\Pi_{\mathrm{TN},n}$ in Eqs. (13) and Eqs. (15) as two logistic regression models. Hence, the parameters of α_{TP}, α_{TN}, η_{TP}, and η_{TN} are regarded as the weights of the models. Furthermore, given a value of γ, the maximum likelihood estimations (MLEs) of the four parameters can be obtained.

Formally, in order to train algorithm $\mathbb{A}$ with different sizes of training sets, we generate K subsets of training sets from dataset D_n and denote the subsets as D_{N_i} where $i = 1, \ldots, K$ and N_i is the size of D_{N_i}. The K subsets satisfy $D_{\mathrm{N}_1} \subset D_{\mathrm{N}_2} \subset \cdots \subset D_{\mathrm{N}_K}$. Without loss of generality, we define $N_K = n$. Furthermore, we use dataset D_{N_i} to train algorithm $\mathbb{A}$ and induce the model $\mathbb{A}(D_{\mathrm{N}_i})$ and then use a hold-out validation set $V_{n^{(v)}}$ to evaluate model $\mathbb{A}(D_{\mathrm{N}_i})$ to obtain confusion matrices $\boldsymbol{C}_i^{(v)} = (\mathrm{TP}_i^{(v)}, \mathrm{FP}_i^{(v)}, \mathrm{FN}_i^{(v)}, \mathrm{TN}_i^{(v)})$. The subscription i of $\boldsymbol{C}_i^{(v)}$ indicates that the training set size of $\boldsymbol{C}_i^{(v)}$ is N_i, and the superscript of $\boldsymbol{C}_i^{(v)}$ indicates that the confusion matrix is obtained on a validation set rather than a test set. The numbers of positive and negative samples in $V_{n^{(v)}}$ are denoted as $n_+^{(v)} = \mathrm{TP}_i + \mathrm{FN}_i$ and $n_-^{(v)} = \mathrm{TN}_i + \mathrm{FP}_i$ such that $n_+^{(v)} + n_-^{(v)} = n^{(v)}$.

Based on the confusion matrices of $\boldsymbol{C}_i^{(v)}$ with $i = 1, \ldots, K$, two log likelihood function are defined in Eqs. (21) and (22).

$$\begin{aligned} L_{\mathrm{TP}}^{(i)} &= \mathrm{TP}_i(\alpha_{\mathrm{TP}} + \eta_{\mathrm{TP}}\mathrm{N}_i^\gamma) - n_+^{(v)} \ln(1 + e^{\alpha_{\mathrm{TP}}+\eta_{\mathrm{TP}}\mathrm{N}_i^\gamma}), && (21)\\ L_{\mathrm{TN}}^{(i)} &= \mathrm{TN}_i(\alpha_{\mathrm{TN}} + \eta_{\mathrm{TN}}\mathrm{N}_i^\gamma) - n_-^{(v)} \ln(1 + e^{\alpha_{\mathrm{TN}}+\eta_{\mathrm{TN}}\mathrm{N}_i^\gamma}). && (22) \end{aligned}$$

Furthermore, the parameters in the proposed learning curves, namely, $\{\alpha_{\mathrm{TP}}, \alpha_{\mathrm{TN}}, \eta_{\mathrm{TP}}, \eta_{\mathrm{TN}}\}$ is estimated by optimizing a log likelihood objective in Eq. (23).

$$L(\gamma) = \max_{\alpha_{\mathrm{TP}},\alpha_{\mathrm{TN}},\eta_{\mathrm{TP}},\eta_{\mathrm{TN}}} \sum_{i=1}^{K} (L_{\mathrm{TP}}^{(i)} + L_{\mathrm{TN}}^{(i)}). \qquad (23)$$

In order to search an optimal value of γ, we maximize a log likelihood objective with an L_1 prior that slightly regularizes values to close to -0.5. The L_1 prior is heuristic and borrowed from Hoiem et al. (2021). The regularized likelihood objective is showed in Eq. (24).

$$\max_{\gamma\in[-1,1]} L(\gamma) - \tau|\gamma + 0.5|. \tag{24}$$

The estimators of $\{\alpha_{\mathrm{TP}}, \alpha_{\mathrm{TN}}, \eta_{\mathrm{TP}}, \eta_{\mathrm{TN}}, \gamma\}$ based on Eq. (24) are denoted as $\{\hat{\alpha}_{\mathrm{TP}}, \hat{\alpha}_{\mathrm{TN}}, \hat{\eta}_{\mathrm{TP}}, \hat{\eta}_{\mathrm{TN}}, \hat{\gamma}\}$. Substituting the parameters in Eqs. (18)-(20) with the corresponding estimators, we can obtain the estimation versions of the learning curves of test error, precision, recall, and F_1 score.

2.5. Confidence bands of the Classification Metrics

According to the posterior distributions of test error, precision, recall, and F_1 score in Eqs. (10)-(11), we can naturally obtain the confidence bands of these four metrics.

Specifically, we substitute $\{\hat{\alpha}_{\mathrm{TP}}, \hat{\alpha}_{\mathrm{TN}}, \hat{\eta}_{\mathrm{TP}}, \hat{\eta}_{\mathrm{TN}}, \hat{\gamma}\}$ into the learning curves of $\boldsymbol{\pi}_n$ in Eqs. (16) and (17) and derive the estimated learning curves denoted as $\hat{\boldsymbol{\pi}}_n = (\hat{\pi}_{\mathrm{TP,n}}, \hat{\pi}_{\mathrm{FP,n}}, \hat{\pi}_{\mathrm{FN,n}}, \hat{\pi}_{\mathrm{TN,n}})$. We further define a virtual confusion matrix $\hat{C}_n = n^{(v)}\hat{\boldsymbol{\pi}}_n = (\widehat{\mathrm{TP}}_n, \widehat{\mathrm{FP}}_n, \widehat{\mathrm{FN}}_n, \widehat{\mathrm{TN}}_n)$ where $n^{(v)}$ is the size of a validation set.

Then, the confidence bands of test error, precision, recall, and F_1 score with a probability $1-\alpha$ are expressed in Eqs. (25)-(28).

$$\begin{aligned}\mathrm{CI}_e = [&\mathcal{B}e_{\frac{\alpha}{2}}(\widehat{\mathrm{FP}}_n + \widehat{\mathrm{FN}}_n + \lambda, \widehat{\mathrm{TP}}_n + \widehat{\mathrm{TN}}_n + \lambda),\\ &\mathcal{B}e_{\frac{2-\alpha}{2}}(\widehat{\mathrm{FP}}_n + \widehat{\mathrm{FN}}_n + \lambda, \widehat{\mathrm{TP}}_n + \widehat{\mathrm{TN}}_n + \lambda)],\end{aligned} \tag{25}$$

$$\mathrm{CI}_p = [\mathcal{B}e_{\frac{\alpha}{2}}(\widehat{\mathrm{TP}}_n + \lambda, \widehat{\mathrm{FP}}_n + \lambda), \quad \mathcal{B}e_{1-\frac{\alpha}{2}}(\widehat{\mathrm{TP}}_n + \lambda, \widehat{\mathrm{FP}}_n + \lambda)], \tag{26}$$

$$\mathrm{CI}_r = [\mathcal{B}e_{\frac{\alpha}{2}}(\widehat{\mathrm{TP}}_n + \lambda, \widehat{\mathrm{FN}}_n + \lambda), \quad \mathcal{B}e_{1-\frac{\alpha}{2}}(\widehat{\mathrm{TP}}_n + \lambda, \widehat{\mathrm{FN}}_n + \lambda)], \tag{27}$$

$$\mathrm{CI}_{f_1} = [\frac{1}{1+\mathcal{B}e'_{1-\frac{\alpha}{2}}(\bar{a},\bar{b})}, \quad \frac{1}{1+\mathcal{B}e'_{\frac{\alpha}{2}}(\bar{a},\bar{b})}], \tag{28}$$

where $\mathcal{B}e'_{\alpha}$ is the α quantile of a beta-prime distribution with parameters of $\bar{a} = \widehat{\mathrm{FP}}_n + \widehat{\mathrm{FN}}_n + 2\lambda$ and $\bar{b} = \widehat{\mathrm{TP}}_n + \lambda$. In this study, $\lambda = 1$ of an uniform prior is used.

2.6. Baseline

To date, multiple parametric learning curves have been developed and well investigated in recent survey papers (Mohr and van Rijn, 2022; Viering and Loog, 2023). Most of the parametric forms of learning curves considered the metric of test error. To our best knowledge, there is no relevant research available about the learning curves of precision, recall, and F_1 score.

There are 14 types of the existing parametric learning curves that are frequently clustered into two categories: power-law shape and exponential shape. The existing learning curves have been summarized in Table 1 of (Viering and Loog, 2023) and are not repeated here due to limited space.

Most of the existing learning curves were compared by Gu et al. (2001) and Kolachina et al. (2012) with sufficient empirical experiments over multiple kinds of datasets and algorithms. These studies consistently recommended the inverse power law form of $an^{-\gamma} + c$

as the best learning curves. The form is named as POW3 in (Viering and Loog, 2023). Moreover, in recent years, Hoiem et al. (2021) further used POW3 to analyze the learning behavior of deep neural networks and Mohr and van Rijn (2023) studied the fitting problems in the curves of POW3. Therefore, we use POW3 as a strong baseline in this paper to illustrate the improvements in our proposed method. Moreover, most of the 14 types of the existing learning curves, as well as our proposed curves, are also numerically compared with our experimental settings. Due to the limited space, the corresponding results are given in the supplemental materials (SM).

3. Experiments

3.1. Datasets and Algorithms

Two synthetic, four real-world datasets, six non-neural classifiers, and two neural classifiers are considered.

Synthetic datasets include a **simple** dataset that has a binary class label and ten predictors jointly drawn from two conditional Gaussian distributions, and the **Exp2** dataset that is a binary-class variant of the EXP6 dataset frequently used in the algorithm comparison task (Dietterich, 1998; Wang and Li, 2023). The sizes of the training set D_n and test set $T_{n'}$ are set to 1,280 and 320, respectively. Multiple popular non-neural classification algorithms are used on the two synthetic datasets, including logistic regression (LR), linear discriminant analysis (LDA), naive bayes (NB), decision tree (TREE), random forest (RF), and multi-layer perceptron (MLP).

Four real-world datasets are Cat VS Dog dataset, Cifar10 dataset, COVID-19 Chest X-Ray dataset, binary gender classification dataset. Cat VS Dog dataset has 12,500 images per class, where 10,000 images per class are IID drawn without replacement as a training set D_n and the remaining 2500 images per class are used as a test set $T_{n'}$. For Cifar 10 dataset, the first five classes are combined as positive class, and the other five classes are combined as negative class. From the 50,000 images in the original training set of Cifar10 dataset, 40,000 images are randomly drawn in a stratified manner from a training set D_n and the remaining 10,000 images are used as a test set $T_{n'}$. For COVID-19 Chest X-Ray dataset, the COVID class and NORMAL class are regarded as the positive and negative classes, respectively. For binary gender classification dataset, the female class and male class are regarded as the positive and negative classes, respectively. In COVID-19 Chest X-Ray dataset and binary gender classification dataset, 1,000 images per class are IID drawn without replacement as a training set D_n and the remaining 200 images per class are used as a test set $T_{n'}$. On the real-world datasets, two deep neural networks, i.e., ResNet18 and DenseNet, are used.

For the synthetic datasets, six types of sizes of the training subsets D_{N_i} are used, namely, $N_i = 40 \times 2^i$ with $i = 0, 1, \ldots, 5$. For Cat VS Dog dataset, $N_i = 625 \times 2^i$ with $i = 0, 1, \ldots, 4$ are used. For Cifar10 dataset, $N_i = 1250 \times 2^i$ with $i = 0, 1, \ldots, 5$. For COVID-19 Chest X-Ray dataset, $N_i = 125 \times 2^i$ with $i = 0, 1, \ldots, 4$ are used. For binary gender classification dataset, $N_i = 125 \times 2^i$ with $i = 0, 1, \ldots, 5$. A validation set $V_{n^{(v)}}$ with size $n^{(v)} = N_i/3$ is randomly sampled from D_{N_i} for estimating the proposed learning curves.

The technical details of the above datasets and the settings of the non-neural and neural classification algorithms are elaborated in SM.

3.2. Evaluation Protocols

The estimation methods of the baseline method POW3 and the corresponding confidence bands are similar with Hoiem et al. (2021). Although Hoiem et al. (2021) merely considered the metric of test error, we also use their methods coupled with the a normality assumption to produce the baseline results with regard to precision, recall, and F_1 score for a thorough comparison.

To validate the quality of the estimated learning curves, a measure of RMSE and leave-one-size-out evaluation strategy are used. When fitting a learning curve, for each training set size of N_i, different fractions of dataset D_n are used to train multiple classifiers when $N_i < n$. When training a classifier with D_{N_i}, 2/3 of D_{N_i} is used in a training process, and the remaining 1/3 of D_{N_i} is used as a validation set $V_{n^{(v)}}$ for inducing a confusion matrix $C_i^{(v)}$. Moreover, considering training set sizes of $\{N_1, N_2, \ldots, N_K\}$, a leave-one-size-out evaluation uses the training sets of sizes in $\{N_1, \ldots, N_{i-1}, N_{i+1}, \ldots, N_K\}$ to fit a learning curve and compute the fitting value of the learning curve with regard to the training set size of N_i. Moreover, parameter γ is discretely searched over $[-1, 1]$ with a step of 0.01, and $\tau = 5$ is empirically set.

In order to compute an RMSE value, taking test error as an example, denote the fitting value of a learning curve with regard to a training set size N_i as $\hat{e}_{N_i}$. Moreover, we divide D_n into multiple disjoint training sets of size N_i and use each training set to training a classifier and evaluate the test error of the classifier on test set $T_{n'}$. Then, multiple numerical values of test error on $T_{n'}$ are obtained, and these values are averaged to approximate the true value of test error that is denoted as $\bar{e}_{N_i}$. Furthermore, for each experimental dataset, we use different random seeds to sample J times of different D_n and $T_{n'}$ and correspondingly perform J times of the above protocols. Then, we obtain J pairs of fitting values and true values, denoted as $\{(\hat{e}_{N_i,j}, \bar{e}_{N_i,j})\}_{j=1}^{J}$. Then, a numerical RMSE is computed in Eq. (29).

$$\text{RMSE}_e(N_j) = \sqrt{\sum_{j=1}^{J} (\hat{e}_{N_i,j} - \bar{e}_{N_i,j})^2}. \tag{29}$$

The RMSE values of precision, recall, and F_1 score, namely, RMSE_p, RMSE_r, RMSE_{f_1}, can be calculated in a similar manner. We set $J = 1,000$ for synthetic datasets and $J = 10$ for real-world datasets.

To assess the quality of the proposed confidence bands, The typical measures of DoC and IL are employed. DoC is the probability of the inclusion of the true value of a metric in the confidence band, and IL is the expected width of a confidence band at a training set size of N_i. A higher value of DoC with a shorter IL indicates a better confidence band. In our experiments, we use the empirical proportion over the inclusions of the J times of true metric values to approximate the true value of DoC and the averaged length of confidence bands of J simulations of a dataset as the true value of IL. Moreover, $\alpha = 0.05$ is used to indicate the level of confidence.

3.3. Results and Analysis

Due to the limited space, all experimental results are included in SM. In the paper, we list a fraction of typical results of the comparisons between POW3 baseline and the proposed

Table 2: Comparison of RMSE values of learning curves on the simple dataset.

Classifier	POW3						Our proposed method					
	40	80	160	320	640	1280	40	80	160	320	640	1280
							Test error					
LDA	3.14	2.46	2.57	2.70	2.96	3.81	**2.80**	**2.29**	**2.47**	**2.65**	**2.88**	**3.19**
NB	3.00	2.26	2.26	2.29	2.56	3.19	**2.73**	**2.12**	**2.16**	**2.26**	**2.49**	**2.69**
TREE	3.06	2.27	**2.26**	2.47	3.01	5.05	**2.96**	**2.23**	2.28	**2.44**	**2.95**	**3.57**
MLP	3.23	2.34	2.06	2.23	2.75	4.00	**2.68**	**1.91**	**1.94**	**2.19**	**2.56**	**3.55**
							Precision					
LDA	4.27	3.06	2.98	3.19	3.57	4.43	**3.39**	**2.77**	**2.91**	**3.17**	**3.45**	**3.75**
NB	3.96	2.73	2.73	2.82	3.26	4.08	**3.34**	**2.57**	**2.68**	**2.82**	**3.15**	**3.37**
TREE	3.84	3.21	2.56	**2.84**	**3.67**	5.89	**3.49**	**2.98**	**2.52**	2.87	3.78	**4.19**
MLP	4.76	2.92	2.44	2.59	3.13	4.32	**3.42**	**2.37**	**2.39**	**2.57**	**3.04**	**3.81**
							Recall					
LDA	5.10	3.88	4.00	4.18	4.67	5.93	**4.56**	**3.53**	**3.86**	**4.12**	**4.45**	**4.82**
NB	4.41	3.53	3.31	**3.47**	3.89	4.76	**3.90**	**3.11**	**3.24**	**3.47**	**3.74**	**4.06**
TREE	5.83	5.53	4.35	**4.48**	5.60	7.95	**5.46**	**4.64**	**3.86**	4.67	**5.08**	**5.99**
MLP	5.49	**3.87**	**3.45**	3.46	**4.47**	**6.47**	**5.07**	4.05	3.60	**3.45**	4.49	6.67
							F_1 score					
LDA	3.67	2.87	2.96	3.05	3.31	4.26	**3.34**	**2.62**	**2.81**	**2.98**	**3.21**	**3.54**
NB	3.09	2.42	2.40	2.45	2.72	3.40	**2.75**	**2.22**	**2.28**	**2.41**	**2.65**	**2.88**
TREE	3.70	3.05	2.66	2.81	3.25	4.95	**3.38**	**2.28**	**2.52**	**2.78**	**3.00**	**3.59**
MLP	4.00	2.76	2.44	2.50	3.04	4.52	**3.63**	**2.66**	**2.41**	**2.45**	**2.91**	**4.19**

curve lines on the simple dataset, Cifar10 dataset, and COVID19 dataset in terms of the measures of RMSE, DoC and IL in Tables 2, 3, 4 and 5. In the tables, lower values of RMSE are labeled in bold font to indicate better performance. Moreover, the values of DoC higher than $1 - \alpha = 0.95$ are listed in bold font to indicate the corresponding confidence bands possess sufficient degree. Comprehensively analyzing the results in Tables 2, 3, 4, 5 and SM, we can obtain several observations in the following sections.

3.3.1. Analysis about RMSE of learning curves.

In Tables 2, 3, 4 and 5, for the metric of test error, most values of RMSE of our proposed learning curves are lower than those in POW3 baselines, which indicates that our proposed learning curves achieve an obvious improvement in the fitting of learning curves. Similar conclusions can be obtained from the remaining results in SM.

Moreover, for both POW3 baseline and the proposed method, as a training set takes a small size and a large size, the corresponding RMSE values are slightly higher than those with moderate training set sizes. It illustrates that a challenge of fitting a learning curve occurs when a training set size is small or large. Similar phenomenon is observed on Cifar10

Table 3: Comparison of confidence bands of test error on the simple dataset.

Measure	40	80	160	320	640	1280	40	80	160	320	640	1280
	LDA											
DoC	0.18	0.15	0.11	0.15	0.18	0.17	**1.00**	**1.00**	**1.00**	**1.00**	**0.95**	0.81
IL($\times 10^{-1}$)	0.15	0.10	0.08	0.10	0.13	0.15	4.53	3.25	2.32	1.58	1.12	0.81
	NB											
DoC	0.19	0.13	0.13	0.14	0.20	0.16	**1.00**	**1.00**	**1.00**	**1.00**	**0.95**	0.79
IL($\times 10^{-1}$)	0.13	0.08	0.07	0.09	0.12	0.13	3.82	2.95	2.08	1.39	0.98	0.70
	TREE											
DoC	0.19	0.20	0.14	0.16	0.16	0.12	**1.00**	**1.00**	**1.00**	**1.00**	**0.96**	0.74
IL($\times 10^{-1}$)	0.17	0.11	0.08	0.09	0.13	0.17	4.72	3.37	2.41	1.68	1.18	0.87
	MLP											
DoC	0.24	0.20	0.18	0.17	0.20	0.17	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**	0.78
IL($\times 10^{-1}$)	0.16	0.11	0.08	0.10	0.13	0.17	4.53	3.37	2.37	1.68	1.19	0.83

Table 4: Comparison of learning curves and confidence bands on Cifar10 dataset and ResNet Classifier.

Metric	POW3						Our proposed method					
$N_i(\times 10)$	125	250	500	1000	2000	4000	125	250	500	1000	2000	4000
Test error	3.16	0.77	2.49	**0.41**	1.73	**2.00**	**2.37**	**0.34**	**1.50**	1.08	**1.62**	5.69
Precision	3.66	0.95	2.28	**0.53**	1.87	**1.88**	**2.74**	**0.38**	**1.15**	1.13	**1.86**	5.32
Recall	1.48	1.40	3.33	**0.91**	1.42	**3.19**	**0.60**	**1.01**	**2.43**	1.41	**1.30**	6.14
F_1 score	2.36	0.98	2.69	**0.48**	1.60	**2.17**	**1.57**	**0.53**	**1.75**	1.10	**1.51**	5.74
Measure	Evaluation of the confidence bands of test error.											
DoC	0.0	0.0	0.0	0.1	0.0	0.1	**1.0**	**1.0**	**1.0**	**1.0**	0.9	0.0
IL($\times 10^{-1}$)	0.03	0.02	0.02	0.03	0.04	0.05	2.87	1.98	1.33	0.86	0.53	0.21

dataset in Table 4. In particular, on Cifar10 dataset, when the training set size is 40,000, our proposed method is slightly worse than POW3 method that is as similar as the results on the Cat VS Dog dataset when a training set size achieves its maximum. In contrast, when training set size is small, our method frequently outperforms POW3 baseline. These observations demonstrate that our proposed learning curves have a large strengthens in fitting the curves on a training set with a relatively small size and moderate size. We will polish the proposed method on large training sets in future work.

For the metrics of precision, recall, and F_1 score, Tables 2, 3, 4 and 5 also illustrate that our proposed learning curves that are transformed from the learning curves of a confusion matrix are slightly better than the POW3 baselines. It indiates that multi-nomial distribu-

Table 5: Comparison of learning curves and confidence bands on COVID19 dataset and ResNet Classifier.

Metric	POW3					Our proposed method				
N_i	125	250	500	1000	2000	125	250	500	1000	2000
Test error	2.74	2.06	1.07	0.78	1.01	**2.66**	2.10	1.09	**0.77**	**0.88**
Precision	4.29	3.46	2.41	1.72	2.51	**4.17**	3.49	2.43	1.73	**2.39**
Recall	1.31	0.54	0.46	0.54	0.67	**1.23**	0.60	**0.42**	**0.53**	0.84
F_1 score	2.69	2.00	1.04	0.76	1.00	**2.61**	2.05	1.07	0.77	**0.87**
Measure	Evaluation of the confidence bands of test error.									
DoC	0.0	0.0	0.0	0.0	0.0	**1.0**	**1.0**	**1.0**	**1.0**	0.5
IL($\times 10^{-1}$)	0.34	0.21	0.30	0.37	0.35	9.11	5.49	3.17	1.80	0.88

tions and beta distributions are regarded as more reasonable assumptions than a normality distribution.

Besides the POW3 baseline, we also compare our method with other nine types of parametric learning curves. The numerical results in SM also show that the proposed learning curves are frequently better than the other existing learning curves in most experimental settings regardless of the considered classification metrics.

3.3.2. Analysis about confidence bands.

The comparison results of confidence bands of POW3 baseline and our method are also listed in Tables 2, 3, 4 and 5. A reasonable confidence band should possess a high DoC value first and then a small IL value. It is noted that the DoC values in Table 4 and 5 merely keep one decimal place because the simulation count J of a real-world dataset is $J = 10$.

All experimental results demonstrate that the numerical DoC values of POW3 baseline are constantly far lower than the nomial degree $1 - \alpha = 0.95$ although POW3 possesses a substantially smaller IL than our method. From the unpromising DoC values in POW3 baseline, we can infer that a model selection method based on the confidence band in POW3 easily produces liberal false positive decisions. Nevertheless, it is difficult to obtain an universal unbiased variance estimator in a confidence band because multiple estimators of model performance obtained on a dataset are frequently correlated and the corresponding correlation coefficients can not be well estimated (Bengio and Grandvalet, 2004). Thus, a relative conservative confidence band is preferred in model selection task to reduce false positive decisions.

In contrast to POW3 baseline, our proposed confidence bands frequently possess DoC values higher than the desired degree $1 - \alpha = 0.95$. It illustrate that the proposed confidence bands is more suitable to a model selection task than the confidence bands of POW3 baseline. The results of confidence bands in SM also show the confidence bands of the other existing curves are liberal too, and the proposed confidence bands are more reasonable than other methods. It is noted that majority of the DoC values are very close to one, which indicates the proposed confidence bands are over-conservative. Correspondingly, the IL in the proposed method becomes substantially larger than POW3. The possible reason of the over-conservative property is that the confidence bands uses a smaller validate set size $n^{(v)}$. However, a larger value of $n^{(v)}$ leads to a smaller training set and would produce underfitting models that harms the estimation of learning curves. Moreover, different from the normal distribution, the Beta distributions in Eqs. (25)-(28) make the proposed confidence bands not exceeding the range of [0,1], which is another strength of our method.

3.3.3. Summary.

Compared our proposed learning curves with POW3 baseline and the other existing learning curves, the following two conclusions are obtained.

- The proposed learning curves can achieve smaller RMSE values in the majority of experimental settings.
- The proposed confidence bands which are relative conservative frequently own slightly high confidence degrees and wide interval ranges.

4. Conclusion

In this study, we obtained the closed-form expressions of the learning curves of test error, precision, recall, and F_1 score over confusion matrices based on an assumption of multinomial distribution instead of the widely-used normality assumption. Furthermore, from a Bayesian perspective, reasonable confidence bands are derived from the posterior distributions of test error, precision, recall, and F_1 score. Experimental results illustrate the improvements in the proposed method and the superiority of the confidence bands among the existing learning curves.

In future, we will refine the optimization method to improve the proposed learning curves over large training sets and investigate the effect of the partitioning of training and validation sets on the confidence bands for developing better confidence bands. Moreover, we will generalize our method to the scenario of multi-class classification task and other classification metrics.

Acknowledgments

This work was supported by the Fundamental Research Program of Shanxi Province (202403 021221028) and the National Natural Science Foundation of China (61806115). The experiments are supported by High Performance Computing System of Shanxi University. The corresponding author is Jihong Li (email: li_ml@sxu.edu.cn).

References

Steven Adriaensen, Herilalaina Rakotoarison, Samuel Müller, and Frank Hutter. Efficient bayesian learning curve extrapolation using prior-data fitted networks. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, 2023.

Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2004.

Natthaphan Boonyanunta and Panlop Zeephongsekul. Predicting the relationship between the size of training sample and the predictive power of classifiers. In *Knowledge-Based Intelligent Information and Engineering Systems: 8th International Conference, KES 2004*, pages 529–535, 2004.

Bostjan Brumen, Ivan Rozman, Marjan Hericko, Ales Cernezel, and Marko Hölbl. Best-fit learning curve model for the C4.5 algorithm. *Informatica*, 25(3):385–399, 2014.

Corinna Cortes, Lawrence D. Jackel, Sara A. Solla, Vladimir Vapnik, and John S. Denker. Learning curves: Asymptotic values and rate of convergence. In *Advances in Neural Information Processing Systems 6, [7th NIPS Conference]*, pages 327–334. Morgan Kaufmann, 1993.

Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, 10(7):1895–1923, 1998.

Lewis J. Frey and Douglas H. Fisher. Modeling decision tree performance with the power law. In *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics, AISTATS 1999*. Society for Artificial Intelligence and Statistics, 1999.

Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359, 2005.

Baohua Gu, Feifang Hu, and Huan Liu. Modelling classification performance for large data sets. In *Advances in Web-Age Information Management, Second International Conference, WAIM 2001*, volume 2118 of *Lecture Notes in Computer Science*, pages 317–328, 2001.

Derek Hoiem, Tanmay Gupta, Zhizhong Li, and Michal Shlapentokh-Rothman. Learning curves for analysis of deep networks. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 4287–4296, 2021.

Cong Hua, Qianqian Xu, Zhiyong Yang, Zitai Wang, Shilong Bao, and Qingming Huang. Openworldauc: Towards unified evaluation and optimization for open-world prompt tuning. In *The Forty-second International Conference on Machine Learning*, 2025.

Prasanth Kolachina, Nicola Cancedda, Marc Dymetman, and Sriram Venkatapathy. Prediction of learning curves in machine translation. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 22–30. The Association for Computer Linguistics, 2012.

Yicheng Li, Haobo Zhang, and Qian Lin. On the asymptotic learning curves of kernel ridge regression under power-law decay. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, 2023.

Felix Mohr and Jan N. van Rijn. Learning curves for decision making in supervised machine learning - A survey. *CoRR*, abs/2201.12150, 2022.

Felix Mohr and Jan N. van Rijn. Fast and informative model selection using learning curve cross-validation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(8):9669–9680, 2023.

Jorge J Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis: proceedings of the biennial Conference held at Dundee*, pages 105–116, 1977.

Claudia Perlich, Foster J. Provost, and Jeffrey S. Simonoff. Tree induction vs. logistic regression: A learning-curve analysis. *J. Mach. Learn. Res.*, 4:211–255, 2003.

Qi Qi, Youzhi Luo, Zhao Xu, Shuiwang Ji, and Tianbao Yang. Stochastic optimization of areas under precision-recall curves with provable convergence. *Advances in Neural Information Processing Systems*, 34:1752–1765, 2021.

Benjamin S. Ruben and Cengiz Pehlevan. Learning curves for noisy heterogeneous feature-subsampled ridge ensembles. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, 2023.

Sameer Singh. Modeling performance of different classification methods: deviation from the power law. *Project Report, Department of Computer Science, Vanderbilt University, USA*, 7, 2005.

Tom J. Viering and Marco Loog. The shape of learning curves: A review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(6):7799–7819, 2023.

Ruibo Wang and Jihong Li. Block-regularized 5×2 cross-validated mcnemar's test for comparing two classification algorithms. *CoRR*, abs/2304.03990, 2023.

Yu Wang, Jihong Li, Yanfang Li, Ruibo Wang, and Xingli Yang. Confidence interval for f_1 measure of algorithm performance based on blocked 3×2 cross-validation. *IEEE Transactions on Knowledge and Data Engineering*, 27(3):651–659, 2014.

Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. Optimizing partial area under the top-k curve: Theory and practice. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4):5053–5069, 2023.

Zitai Wang, Qianqian Xu, Zhiyong Yang, Peisong Wen, Yuan He, Xiaochun Cao, and Qingming Huang. Top-k pairwise ranking: Bridging the gap among ranking-based measures for multi-label classification. *Int. J. Comput. Vis.*, 133(1):211–253, 2025.