State of the Art and Future Potential for Calibrated Fuzzy Classifiers

Luke S. Merrick[*]
Department of Systems and
Information Engineering
University of Virginia
Charlottesville, VA 22903
lsm3bn@virginia.edu

Mawulolo Ameko
Department of Systems and
Information Engineering
University of Virginia
Charlottesville, VA 22903
mka9db@virginia.edu

Nathan Ray
Department of Systems and
Information Engineering
University of Virginia
Charlottesville, VA 22903
nwr2gn@virginia.edu

Brett Libowitz
Department of Systems and
Information Engineering
University of Virginia
Charlottesville, VA 22903
bdl3sz@virginia.edu

Abstract

In this report we discuss the motivation, importance, and practical usefulness of calibrated fuzzy classifiers and the recent related advances in the theory of calibration. Further, we present a review of the current state of the art of both the abstract theory and practical algorithms that pertain to the calibration of fuzzy classifiers. As a case study, we focus on the SL-Isotron algorithm in the context of interpretable performance through simultaneous calibration and risk minimization, and we present detailed discussion on the limitations and potential extensions of the algorithm and the theory motivating it.

1 Introduction

1.1 Paper Outline

The structure of the paper is as follows: we begin with an introduction to the motivation for the development and use of calibrated fuzzy classification methods. Following this, we explore the theoretical measurements, tools, and bounds established to guide and analyze machine learning techniques that are developed in response to this motivation. In the latter portion of the paper, we explore and discuss the current state of the art of algorithms that have been developed to meet the needs motivating this field of theory, describe our attempts to create an extension to the state of the art, and finally we provide an outline of the future direction of this state of the art of both theory and learning algorithms.

1.2 Problem Setting

In many applications of machine learning classifiers, it is not possible to achieve perfect classification accuracy, and it becomes necessary to understand the limitations of the predictive power of the predictions output by the trained classifier. Since this paper is primarily concerned with the case of binary classification, we shall consider that the classifier in question is deciding between just two

[*] lukemerrick.com/about

possible labels. In these classifiers, the output may in the end be binary (from the set $\{\pm 1\}$), but it may not be possible to actually determine from the data passed into the classifier what the actual classification should be with high confidence. The limitations of the information provided to the classifier may result in situations for which the input is uninformative with respect to the predictand, and in these situations it is often crucial that the classifier signals this lack of confidence in its output.

Many binary classification algorithms, such as Naïve Bayes and Logistic Regression, produce confidence measures of probabilistic outcomes (which must be thresholded in order to achieve an actual binary output for which classification accuracy is calculated). While these confidence measures are typically interpreted as the conditional probability of the label $Y = 1$ given the feature vector x (a realization from a fixed distribution) that is input into the classifier, an important research topic is the quantification of the validity of such interpretation. When a classifier outputs a confidence measure, it is vital that the underlying theory provide justification the properties of the confidence measure and inform algorithmic improvements to the dependability of the results of such measures.

In a traditional statistical modeling approach, models are typically formulated with a realizable assumption that the data's underlying probability distribution has the same parametric form as the model. Statisticians normally create a parametric conditional distribution $P(Y|X, \theta)$ and assume that conditional distribution also coincides with this form. A problem with this assumption is that in modern data analysis often takes place in an agnostic setting, in which there is no knowledge of the underlying distribution. In an agnostic setting, conditional estimates can not necessarily serve as an analogue to the conditional probability, nor can they necessarily function as hypothesis testing methods which are designed to distinguish two parameter areas in the hypothesis space. Instead, the primary framework through which we understand the continuous-output binary classifiers in the field of machine learning is as "fuzzy classifiers" which constitute an active and important area of research.

1.3 Fuzzy Classifiers

Fuzzy classifiers, sometimes referred to as soft-labeling classifiers in the literature [12, 17], are a family of function approximators that assign labels to input data by using some probabilistic measure. Formally a fuzzy classifier K producing soft labels can be perceived as a function approximator $K : F \rightarrow [0, 1]^c$ where F is the feature space in which the object descriptions exist, and c is the number of classes. In this work the focus is on binary fuzzy classifiers, which represent a function approximation for conditional probability denoted $P(Y = 1|X)$, where Y is the label and X represent the input space.

1.4 Motivating Applications

There has been very early interest in fuzzy classifiers in the field of machine learning, and for a long time they have remained mainstream as pattern recognition methods [11, 14]. Fuzzy classifiers are desired for a number of reasons, namely *soft-labeling* (which provides a measure of confidence in the output of a classifier), *interpretablity* (because outputs, rather than giving an automatic class label, provide a better interpretation of labels, which is helpful in highly delicate setting like medical diagnosis and legal issues), and lastly *limited data and domain expertise* (in cases of rare events such as natural disasters, oil spills, etc. in which domain expertise needs to be combined with a classifier's output to achieve a reasonable decision).

A desire for interpretable fuzzy classifiers under an agnostic assumption is thus clearly not just an artifact of current trends in machine learning research, but rather a ubiquitous component of many popular applications of machine learning classification approaches. The creation of interpretable fuzzy classifiers is relevant in many agnostic settings including medical diagnosis and fraud detection. Often times in these applications the misclassification loss function is asymmetric, meaning that the risk associated with a false positive versus a false negative and how accurate the assumptions made can be critical. For example, a medical diagnostic system that produces a false negative could potentially be fatal for a patient, an outcome much more undesirable than the prescription of unnecessary testing or treatment under a false positive diagnosis.

To approach this problem, many researchers continue to explore the use and calibration of fuzzy classifiers. Essentially, the end goal is to make the predicted probability outputs of fuzzy classifiers agree with the relative frequency of correct predictions in an agnostic setting and, in agreeing, become interpretable and useful. Ideally, fuzzy calibration will be able to address the interpretability of conditional probability estimates and there various applications to help humans make informed and accurate decisions either by defining decision criteria based off of asymmetric loss formulations or through a more human process involving an interpretable probabilistic assessment of classification.

2 Theoretical Tools for Designing Solutions

In the pursuit of algorithms and methodologies for learning performant and useful fuzzy classifiers, several key theoretical frameworks have been developed in the field of supervised learning. In our research, we examined the cutting edge of the state of the art of this theory, assessing both its advantages and limitations in comparison to earlier methodologies.

2.1 Loss Calibration and Surrogate Loss Bounds

In terms of the "gold standard" theory for understanding and bounding the performance of a classifier, one needs to look no further than the work of Peter Bartlett in bringing the PAC learning excess risk bounds for empirical risk minimization into the realm of real classification algorithms. Given the computational intractability of directly optimizing the misclassification rate of a learner, a seminal work by Bartlett, Jordan, and McAuliffe in 2006 [1] established a robust framework for the calibration and risk bounding of so-called "surrogate" loss functions that can be efficiently minimized. Since most, if not all, of the learning algorithms popular today amount to some form of surrogate loss minimization, this theory creates a solid and widely applicable quantification of algorithmic performance. In their work on convexity-based surrogate risk bounds, Bartlett, Jordan, and McAuliffe showed that it is possible to bound the excess risk of a learner in terms of calibrated surrogate risk, and thus not only prove the convergence of a learning algorithm to the risk minimizer, but also to develop an upper bound on the rate of convergence.

The concept of a surrogate risk minimization is to train a classifier by minimizing, over a function class of hypotheses, some surrogate loss function $\phi(\alpha)$ which maps $\alpha = Yf(X)$, the product of label $Y \in \{\pm 1\}$ and classifier output, $f(X)$, to a real value. The hypothetical expression to be minimized, surrogate risk, is defined as $R_\phi(f) = \mathbb{E}[\phi(Yf(X))]$, while the actual expression to be minimized, empirical surrogate risk over a random sample, is defined as $\hat{R}_\phi(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i f(X_i))$.

When ϕ is chosen to be convex and the the function class $\mathcal{F}$ is chosen to consist of functions that are linear in a parameter vector θ, then efficient convex optimization methods can be used to find a θ that corresponds to the $f \in \mathcal{F}$ that minimizes empirical surrogate risk $\hat{R}_\phi(f)$. This is the basis of many popular learning algorithms such as SVM classification (in which surrogate loss is defined to be hinge loss $\phi(\alpha) = \max(0, 1 - \alpha)$. Accordingly, the theoretical guarantees of surrounding convergence to optimal performance in terms of 0-1 loss $R(f) = \mathbb{E}[\mathbb{1}_{Yf(X)\leq 0}]$ is very desirable, and the calibration of surrogate loss functions provides an important theoretical justification for many learning algorithms as well as inspiration for the creation of new algorithms.

Given the importance of surrogate loss calibration, it is important to understand the relation to fuzzy classifiers. The definition of surrogate loss calibration is given, in words, as the situation in which every surrogate loss minimizer agrees with the Bayes decision rule. In mathematical terms:

Definition 1 *Following the notation established above, we consider $\eta(x) = \mathbb{P}(Y = 1|X = x)$, surrogate loss $\phi(\alpha)$, and all surrogate loss minimizers $f^* = argmin_{f\in\mathcal{F}}\{R_\phi(f)\}$. ϕ is calibrated to the 0-1 loss if for all x such that $\eta(x) \neq 1/2$, the following condition holds:*

$$\mathbb{E}[\phi(Yf^*(X)|X] = \operatorname{sign}(2\eta(x) - 1)$$

(where we note that $\operatorname{sign}(2\eta(x) - 1)$ is the Bayes decision rule that minimizes R, so by extent all f^* also minimize misclassification risk)

We thus see that this type of calibration hearkens back to the posterior probability η insofar as it requires that the output of any surrogate-risk-minimization-based learning algorithm to match the Bayes decision rule for a specific decision threshold. Furthermore, while in the formal definition this threshold was defined to be at $\eta(x) = 1/2$, it has been shown that this agreement can easily be generalized to asymmetric loss [15], meaning that the decision threshold does not have to be maximum likelihood (as in the definition above) but can be a threshold corresponding to the Bayesian maximum expected utility for which more, less-costly mistakes will be made.

Using this theory, we can thus think of most traditional fuzzy classification algorithms as a search for the best hypothesis that agrees with the Bayes decision rule *for a specific asymmetric loss profile* (for which symmetric loss is simply a special case). For many uses of classifiers, this pointwise agreement is all that is necessary for optimal performance, and thus it represents a desirable classification goal.

While this backbone of theory is quite robust and provides a strong framework for analyzing, designing, and justifying learning classification algorithms, however, it is notable that loss calibration as a goal is expressed only in terms of the (potentially weighted) 0-1 loss function, without any notion of interpretability of the the classifier's fuzzy output or need for multiple loss interpretations of the same fuzzy output. In essence, the penalty for a false positive and false negative can be weighted asymmetrically in training and evaluation to allow for the minimization of a single, specific asymmetric 0-1 loss configuration, and the classifier output will be theoretically guaranteed to be good for this situation, but for values of $f(x)$ farther from the threshold, they may or may not necessarily match $\eta(x)$ well.

2.2 Prediction Calibration

To better address cases in which a single classifier will be used either for human interpretation or as the forecast informing multiple decisions with different asymmetric loss configurations, a recent paper by Gao, Parameswaran, and Peng [2] has proposed a novel notion of classifier calibration. This calibration is expressed in terms of intepretability of classifier output as a probabilistic measure of event outcome likelihood, and it allows for the analysis of the effect that asymmetic loss and intepretability have on the performance of a classifier. We state the definition here for the sake of convenience:

Definition 2 *Let $\mathcal{X}$ be the feature space, $\mathcal{Y} = \pm 1$ be the label space and $\mathcal{P}$ be the distribution over $\mathcal{X} \times \mathcal{Y}$. Let $f : \mathcal{X} \to [0,1]$ be a fuzzy classifier, then we say f is calibrated if for any $p_1 < p_2$, we have:*

$$\mathbb{E}_{\mathcal{X}\sim\mathcal{P}}[\mathbb{1}_{p_1<f(X)\leq p_2} f(X)] = \mathcal{P}(p_1 < f(X) \leq p_2, Y = 1)$$

In other words, a fuzzy classifier is calibrated if its output correctly reflects the relative frequency of labels among instances they believe to be similar. In this paper, an empirical variant of this definition has been defined for a finite data set of the closeness to calibration of the fuzzy classifier. This empirical calibration was also shown to converge to true calibraion. It is here stated:

Definition 3 *A fuzzy classifier f is ϵ-calibrated if*

$$c(f) = \sup_{p_1<p_2} |P(p_1 < f(X) \leq p_2, Y = 1) - \mathbb{E}_{X\sim P}[\mathbb{1}_{p_1<f(X)\leq p_2} f(X)]| \leq \epsilon$$

Definition 4 *A fuzzy classifier f is ϵ-empirically calibrated with respect to dataset D if*

$$c_{emp}(f, D) = \frac{1}{n} \sup_{p_1<p_2} |\sum_{i=1}^{n} \mathbb{1}_{p_1<f(X_i)\leq p_2, Y=1} - \sum_{i=1}^{n} \mathbb{1}_{p_1<f(X_i)\leq p_2} f(X_i)| \leq \epsilon$$

Where $D = \{(X_i, Y_i)\}_{i=1}^{n}$ is a dataset consisting of n i.i.d samples drawn from distribution P

With this new definition, any conditional probability estimation algorithm can be proven to be empirically calibrated or ϵ-calibrated, which in turn improves the interpretability of the output.

This measure of calibration does not directly indicate the performance at a specific task of classification, but rather to the meta-performance of how well the classifier outputs a reasonable degree of confidence in its output. The common baseline classification rule of assigning the relative frequency of label $Y = 1$ as the probability prediction for every input, regardless of the specifics of input vector x, is calibrated under this metric. While this regression to the mean provides the best guess if the input is completely uninformative, in situations in which the data set provides some partial predictive information regarding the correct classification, there is a need for both good misclassification accuracy *and* good agreement between the classifier's output and the relative frequency of labels in its output.

3 The State of the Art in Algorithms for Calibrated Fuzzy Classification

These important theoretical results have significant implications for the setting of general classifiers. Suppose we are given an uncalibrated fuzzy classifier denoted $f_o : \mathcal{X} \rightarrow [0, 1]$ and that we want to find a non-decreasing link or transfer function $g : [0, 1] \rightarrow [0, 1]$ such that the function composition $g \circ f_o$ represents a calibrated conditional probability estimate. This problem then amounts to solving the following optimization problem:

$$\min_{g \in \mathcal{G}} \max_{a,b} | \sum_{a<i\leq b} (\mathbb{1}_{y_i=1} - g(f_o(x_i)))| \tag{1}$$

for any data $D = \{(x_i, y_i)\}_{i=1}^{n}$ and $\mathcal{G}$ the function class of non-decreasing functions. It has also being shown that the optimal solution that minimizes the objective function defined by the squared loss denoted $\min\limits_{g\in\mathcal{G}} \sum\limits_{i=1}^{n} (\mathbb{1}_{y_i=1} - g(f_o(x_i)))^2$ also minimizes (1). The Pooled Adjacent Violator (PAV) algorithm for Isotonic regression which was first studied by Niculescu-Mizil et al [13] has been proven to be calibrated with respect to the above definition in [2]. Consequently, the SL-Isotron algorithm introduced by Sham Kakade et al [8], which uses a variant of the PAV algorithm to learn the link function for Single Index Models, offers an empirically calibrated probability estimate. In the section that follows, we study this algorithm in the context of calibrated conditional probability estimators.

3.1 The SL-Isotron Algorithm

Isotonic regression involves fitting an arbitrary one-dimensional non-decreasing function to a set of points. The perceptron algorithm on the other hand solves a binary classification problem where a margin is assumed between the positive and negative classes. The strengths of these have been combined into one elegant algorithm, called Isotron, for finding a nearly accurate estimation for the link function and the weight parameters in a single index model (SIM). By definition, for inputs $\{(x_i, y_i)\}_{i=1}^{n} \in \mathbb{R}^{\ltimes} \times \mathbb{R}$ drawn independently from a distribution D, the single index model represents the problem of finding the an accurate estimation for u and w such that $\mathbb{E}_{(x,y)\sim D}[y|x] = u(w.x)$ where $w \in \mathbb{R}^n$ and $u : \mathbb{R} \rightarrow \mathbb{R}$ a non-decreasing (Lipschitz continuous) function. This problem has been widely studied in the literature [4, 5, 6] using heuristics that are not guaranteed to converge to a global optimum. The first provably efficient method for learning SIM's has been proposed by Kalai and Satry [9] by the common assumption that u is monotonic Lipschitz and that the data is distributed according to a true u and w.

3.2 Extending SL-Isotron

In our research, we explored the possibilities of extending the SL-Isotron algorithm to minimize a loss function other than squared loss. To do this, we first re-interpreted the perceptron update step as a gradient descent update and achieved a surprising insight: unlike in a typical perceptron update step, the perceptron update step in SL-Isotron is slightly different than the gradient descent update for an unthresholded linear unit due to the chain rule application to the link function. This follows from the following derivation of the squared loss gradient:

$$\begin{aligned} \nabla \frac{1}{2}(y_i - u(\boldsymbol{w} \cdot \boldsymbol{x_i})^2 &= (y_i - u(\boldsymbol{w} \cdot \boldsymbol{x_i}) \cdot \nabla(y_i - u(\boldsymbol{w} \cdot \boldsymbol{x_i}) \\ &= (y_i - u(\boldsymbol{w} \cdot \boldsymbol{x_i}) \cdot (-u'(\boldsymbol{w} \cdot \boldsymbol{x_i}) \cdot \nabla \boldsymbol{w} \cdot \boldsymbol{x_i} \\ &= -(y_i - u(\boldsymbol{w} \cdot \boldsymbol{x_i}) \cdot u'(\boldsymbol{w} \cdot \boldsymbol{x_i}) \cdot \boldsymbol{x_i} \end{aligned} \tag{2}$$

Thus, to interpret the SL-Isotron's perceptron update step as a gradient descent update, it follows that the learning rate of the update step is not fixed at 1 as in the typical perceptron update, but rather is fixed at $\frac{1}{u'(\boldsymbol{w}\cdot\boldsymbol{x_i})}$. Further complicating the matter, $u'(\boldsymbol{w} \cdot \boldsymbol{x_i})$ could very well be equal to zero for some inputs, meaning that the perceptron update is making adjustments where the gradient is actually zeroed out due to the derivative of the link function.

Although we were unable to make a proven generalization of the SL-Isotron algorithm, it does seem that there is a possibility for future work in this area.

4 Discussion of Current Limitations and Opportunity for Future Progress

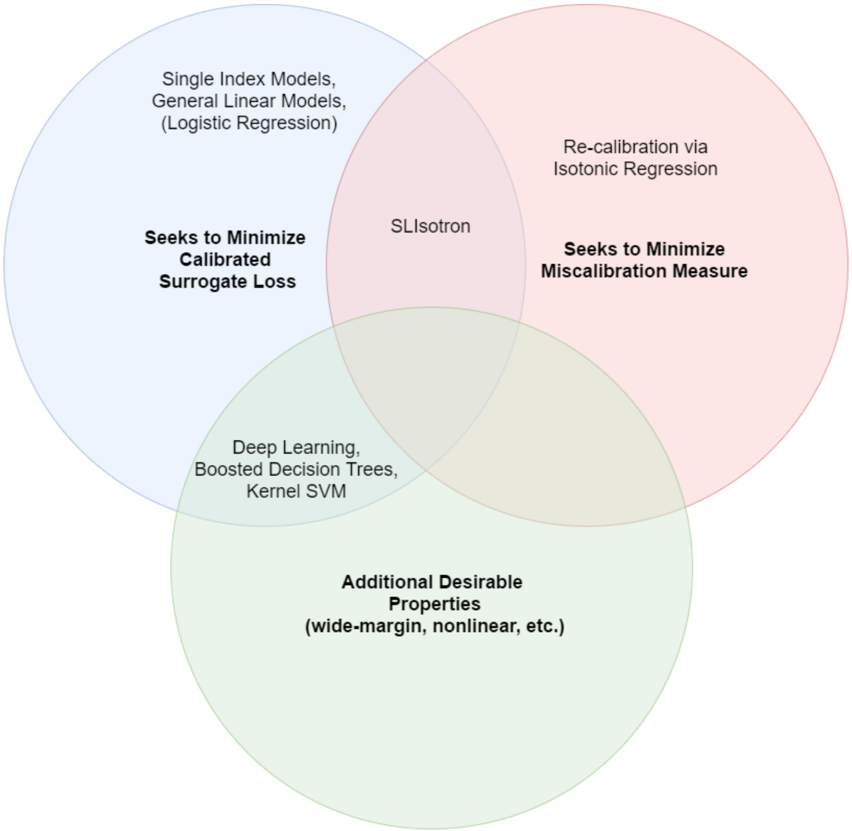

Figure 1: Current state of the art of algorithms related to calibrated fuzzy classification

4.1 Algorithmic Improvement

Even though the SL-Isotron algorithm is empirically calibrated as shown from [2], there remains the challenge of extending and proving this algorithm to be usable, intepretable, and PAC-learnable for a general class of loss functions. In addition, this algorithm has been proven only to a statistical rate slower than $\mathcal{O}\left(\sqrt{\frac{1}{n}}\right)$.

The first area of future direction is to investigate whether the square loss function is necessary to prove the ϵ-calibration of the SL-Isotron. If not, can a generalization results be obtained for a larger class of loss function, say exp-concave functions? Further, we believe that if this result is provable, then we can design an algorithm for a different exp-concave loss function which is not necessarily smooth even though most exp-concave functions have a smoothness property. Also, based on the recent results for using the average stability bound to achieve a faster statistical rate, we believe there is a possibility of achieving a similar rate for the SL-Isotron as the squared loss function is a particular case of exp-concave functions. In figure 1, this corresponds to moving SL-Isotron down into the section overlapped by all three areas, potentially in combination with some of the realted linear models like SVM's.

4.2 Extending Tighter Error Bounds to More Algorithms

Recent work in statistical learning theory have used arguments about the average algorithmic stability to achieve a faster estimation rate to the order of $\mathcal{O}\left(\frac{1}{n}\right)$ for a given function class. This theory has largely been based upon the results that algorithmic stability is a necessary and sufficient condition for learnability of the empirical risk minimizer ([10], [16]). This allows us to establish an equivalence in expectation for the average stability and generalization error based off of which we can provide a bound for excess risk for the empirical minimizer in a function class. A recent result by Gonen and Shalev-Shwartz [3], proves that the stability rate for a given algorithm is invariant to the coordinate system. In other words, a given data can be pre-conditioned by any positive definite matrix and the stabilty of the algorthm would not change. This result has led to an important implication for generalized linear regression problems where the empirical risk minimizer can achieve not only a faster estimation rate but also the bound term would be independent of the empirical condition number (proved by Kakade et al [7]). Instead, the empirical condition number which can go to infinity can now be replaced by d– the dimension of the data. For completeness, we state below this corollary:

Corollary 1 *Consider $\mathcal{X}$ be a any compact and convex subset of $\mathbb{R}^d$ and $\mathcal{Y}$ be an interval of the form $[-Y, Y]$. With domain $\mathcal{W}$ given by*

$\mathcal{W} = \{w \in \mathbb{R}^d : (\forall x \in \mathcal{X}) \quad |w^\top x| \leq Y\}$*, where for all $y \in \mathcal{Y}$, ϕ_y is ρ-Lipschitz and α-strongly convex. The expected excess risk of empirical risk minimization is bounded by*

$$\mathbb{E}[L(\hat{w}) - L(w^*)] \leq \mathbb{E}[\Delta(S)] \leq \frac{2\rho^2 d}{\alpha n} \tag{3}$$

Note: Δ denotes the average stability.

4.3 Future Opportunity for Calibration Theory

Beyond algorithmic advances and work to tighten statistical bound of learning algorithms, there is also ample opportunity for development in the calibration theories that provide insight and analysis to the algorithms discussed above. While surrogate loss calibration has been fleshed out to include generalization bounds and cover a robust array of learning algorithms, the theory of calibrated prediction output is lacking even a tie-in directly to the loss and risk of the learning methodologies for which it provides a hypothetical minimization goal.

Future theoretical advances in extending the goal of classification beyond 0-1 loss with a fixed asymmetry (via prediction calibration) will also allow for more robust linking of real-world motivation to the mathematical framework of learning algorithms. This, coupled with the equivalent of a generalization bound on the empirical calibration measure could allow for the bounding not only of traditional 0-1 risk, but also on the interpretability and performance in informing multiple deciders with different risk profiles.

These potentials for advances correspond to the Venn diagram in figure 1 as a movement from the fringes of the diagram to the center, in which complex and nonlinear classifiers can also output calibrated and interpretable results that perform well in both (thresholded) classification accuracy and in providing accurate and calibrated information to more complicated decision systems like human deciders or a set of deciders with varying risk profiles.

Acknowledgments

We would like to acknowledge the tireless contribution of Professor Quanquan Gu, who brought us from zero to sixty on statistical learning theory in just a single semester class.

References

[1] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. (Was Department of Statistics, U.C. Berkeley Technical Report number 638, 2003).

[2] Yihan Gao, Aditya Parameswaran, and Jian Peng. On the interpretability of conditional probability estimates in the agnostic setting. pages 1367–1374, 2017.

[3] Alon Gonen and Shai Shalev-Shwartz. Average stability is invariant to data preconditioning. implications to exp-concave empirical risk minimization. *arXiv preprint arXiv:1601.04011*, 2016.

[4] Wolfgang Hardle, Peter Hall, Hidehiko Ichimura, et al. Optimal smoothing in single-index models. *The annals of Statistics*, 21(1):157–178, 1993.

[5] Joel L Horowitz and Wolfgang Härdle. Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91(436):1632–1640, 1996.

[6] Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001.

[7] Daniel J Hsu, Sham M Kakade, John Langford, and Tong Zhang. Multi-label prediction via compressed sensing. In *Advances in neural information processing systems*, pages 772–780, 2009.

[8] Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pages 927–935, 2011.

[9] Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression.

[10] Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1):161–193, 2006.

[11] Detlef Nauck, Frank Klawonn, and Rudolf Kruse. *Foundations of neuro-fuzzy systems*. John Wiley & Sons, Inc., 1997.

[12] Quang Nguyen, Hamed Valizadegan, and Milos Hauskrecht. Learning classification with auxiliary probabilistic information. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 477–486. IEEE, 2011.

[13] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632. ACM, 2005.

[14] Johannes A Roubos, Magne Setnes, and Janos Abonyi. Learning fuzzy classification rules from labeled data. *Information Sciences*, 150(1):77–93, 2003.

[15] Clayton Scott. Calibrated asymmetric surrogate losses. *Electron. J. Statist.*, 6:958–992, 2012.

[16] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

[17] Yanbing Xue and Milos Hauskrecht. Learning of classification models from noisy soft-labels. 2016.