Pointwise Confidence scores with provable guarantees

Nivasini Ananthakrishnan,[1] **Shai Ben-David,** [1] **Tosca Lechner** [1]
[1] University of Waterloo
nanantha@uwaterloo.ca, bendavid.shai@gmail.com, tlechner@uwaterloo.ca

Abstract

Quantifying the probability of a machine learning prediction being correct on a given test point enables users to better decide how to use those predictions. Confidence scores are pointwise estimates of such probabilities. Ideally, these would be the label probabilities (a.k.a. the *labeling rule*) of the data generating distribution. However, in learning scenarios the learner does not have access to the labeling rule. The learner aims to figure out a best approximation of that rule based on training data and some prior knowledge about the task at hand, both for predicting a label and for possibly providing a confidence score. We formulate two goals for confidence scores, motivated by the use of these scores for safety-critical applications. First, they must not under-estimate the probability of error for any test point. Second, they must not be trivially low for most points. We consider a few common types of learner's prior knowledge and provide tools for obtaining pointwise confidence scores based on such prior knowledge and the relation between the test point and the training data. We prove that under the corresponding prior knowledge assumptions, our proposed tools meet these desired goals.

1 Introduction

The reliability of machine learnt programs is of course a major concern and has been the focus of much research. Theory offers quite a selection of tools for evaluating reliability, from generalization bounds to experimental result of test sets. However, most of those guarantees are statistical, in the sense that they only hold with high probability (over the generation of the training data and of the test points) and they provide no information about the correctness of prediction on any specific instance. In cases where an error on a specific instance may incur a very high cost, like in safety-critical applications, the common statistical guarantees do not suffice. We would also wish to be able to identify predictions with low confidence so that one could apply some safety procedures (such as a review by a human expert). Ideally, no low confidence prediction should go undetected, At the same time, since expert intervention could be expensive, one also wishes to minimize the occurrence of false positives in the predictions flagged as low confidence.

Can one do better than the overall statistical estimates when it comes to evaluating reliability on a given test case?

Arguably, the most common reason for an statistically reliable machine learning program to fail on a test point is that that point is an 'outlier', in the sense of not being well represented by the sample the program was trained on. This research aims to quantify this 'outlierness'. We propose theoretically founded confidence bounds that take into account the relation between the training sample and the specific test point in question (on top of the commonly used parameters of the learning algorithm, the size of training sample and assumptions about the processes generating both the training data and the test instance).

Clearly, the confidence of any prediction of an unknown label (or any piece of information) hinges upon some prior knowledge or assumptions. In this work we consider a few forms of prior knowledge that are commonly employed in machine learning theory, and develop and analyse confidence score for prediction of individual test points under such assumptions.

We consider the following types of prior knowledge:

Known hypothesis class with low approximation error: We discuss two cases - the realizable setting (i.e., when that approximation error is zero) and the agnostic setup (both in Section 2).

- In the realizable case, we show that there are indeed hypothesis classes for which it is possible to define a confidence score that does not overestimate confidences for any points, while providing high confidences to many points. However, there are also hypotheses classes, that do not allow non-trivial confidence scores fulfilling such a guarantee.
- For the agnostic setup, assuming the learner has knowledge of a hypothesis class with low (but not necessarily 0) approximation error, we show that in this case it is not possible to give any non-trivial con-

fidence score that does not overestimate confidence for some instances.

The data generating distribution is Lipschitz: We provide a an algorithm that calculates confidence scores under such an a Lipschitzness assumption. We show that with high probability over samples, the resulting confidence score of every point is an underestimate of its true confidence while the confidence score we obtain is non-trivial. We provide bounds on the probability (over points and samples) of assigning a low confidence score to a point with high true confidence that converge to zero as a function of the training sizes. For more details, see Section 5.

2 Related work

The closest previous work to ours is Jiang et al [5]. They consider a very similar problem to the one we address here - the problem of determining when can a classifier's prediction on a given point be trusted. For the sake of saving space, we refer the reader to that paper for a more extensive review of previous work on this topic. Their theoretical results differ from our work in two essential aspects. First, they consider only one setup - a data generating distribution that satisfies several technical assumptions. In particular they rely on the following strong condition: "for any point $x \in X$, if the ratio of the distance to one class's high-density-region to that of another is smaller by some margin γ, then it is more likely that x's label corresponds to the former class." We analyse our notion of confidence under several different incomparable assumptions, arguably, none of which is as strong as that. The second significant difference is that the main result on trust of labels there (theorem 4 of [5]) states that if a certain inequality holds then the predicted label agrees with that of the Bayes optimal predictor, and if another inequality holds, there is disagreement between them. However, those inequalities are not complementary. It may very well be that in many cases every domain point (or high probability of instances) fails both inequalities. For such points, that main result tells nothing at all. That paper does not offer any discussion of the conditions under which their main result is not vacuous in that respect. Additionally their result holds with high probability over the domain and is not a point-wise guarantee.

Selective Classification/ Classification with Abstension: One line of work that is related to our paper is learning with abstention. Similar to our setting, the classification problem does not only consist of the goal of classifying correctly, but to also allows the classifier to abstain from making a prediction, if the confidence of a prediction is too low. Many works in this line provide accuracy guarantees that hold with high probability over the domain ([1],[11], [3], [4], [6]). This is different from our goal of point-wise guarantees.

Point-wise guarantees are provided in [2] and [10]. El-Yaniv et al [2] gave a theoretical analysis of the selective classification setup in which a classification function and a selective function are learned simultaneously. The risk of a classification is then only accessed on the set of instances that was selected for classification. The selective function is evaluated by their coverage - how many instances in expectation are selected for classification. They analyse the trade-off between risk and coverage, and introduce the notion of "perfect classification" which requires risk 0 with certainty. This is similar to our requirements on a confidence score in the deterministic setting, where we require 0 risk with high probability. Their notion of coverage is similar to our notion of non-redundancy - in fact non-redundancy corresponds to worst-case coverage over a family of distributions. They provide an optimal perfect learning strategy in the realizable setting and show that there are hypothesis classes with a coverage that converges to 1 and hypothesis classes for which coverage is always 0 for some distributions. We use their results in our Section 4. In contrast to their paper our setup also considers probabilistic labeling functions and our analysis also considers other assumptions on the family of probability distributions, besides generalization guarantees for some fixed hypothesis space.

3 Problem definition

Let the domain of instances be X and the label set be $\{0, 1\}$. A learning task is determined by a probability distribution P over $X \times \{0, 1\}$. We denote the marginal distribution over the domain by P_X and the conditional labeling rule by ℓ_P (namely, for $x \in X$, $\ell_P(x) = \Pr_{(x',y')\sim P}[y' = 1|x' = x]$).

The Bayes classifier,

$$h^B_P(x) = 1 \text{ iff } \Pr_{(x',y')\sim P}[y' = 1|x' = x] \geq 0.5,$$

is the pointwise minimizer of the zero-one prediction loss. We sometime refer to its prediction $h^B_P(x)$ as the *majority label of a point* or the *Bayes label of a point.*

We are interested in point-wise confidence. For a point $x \in X$, the **confidence of a label** $y \in \{0, 1\}$ is

$$C_P(x, y) = \Pr_{(x',y')\sim P}[y' = y|x' = x].$$

Note that the label assigned by he Bayes predictor maximizes this confidence for every domain point x.

A **Confidence score** of the label confidence is an empirical estimate (based on some training sample S) of the true label confidence. Inevitably, the reliability of a confidence score is dependent on some assumptions on the data generating distribution (or, in other words, prior knowledge about the task at hand). Given a family of data generating distributions $\mathcal{P}$ (fulfilling some niceness assumption that reflect the learners prior knowledge or beliefs about the task), a training sample S, and a parameter δ, the *empirical confidence estimate for a point x and label y* is a function $C(x, y, S, \delta)$. We want the following property to hold: For every probability distribution $P \in \mathcal{P}$, with probability of at least

$1-\delta$ over an i.i.d. generation of S by P, we have

$$\Pr_{y' \sim \text{Bernoulli}(\ell_P(x))}[y' = y] \geq C(x, y, S, \delta).$$

That is, with high probability, we do not overestimate the probability of y being the correct label for x. Ideally, this should hold for every point x in the domain. Of course, there is a trivial solution for this - just let $C(\ ,\ ,\ ,\ )$ be the constant 0 function. The goal therefore is to get a confidence score that fulfils the condition above, while still being as high as possible for 'many' $x's$. That is, we aim for a confidence score, such that $\mathbb{E}_{x\sim P, S\sim P^m}[\max\{C(x,1,S,\delta), C(x,0,S,\delta)\}]$ is high. As mentioned above, given a data generating distribution P and a data representation available to the learner, the highest confidence on every instance $x \in X$ is achieved by the Bayes predictor $h^B_P(x)$ and it is easy to see that it is $\max\{\ell_P(x), (1-\ell_P(x))\}$.

In contrast with the common notion of a PAC style error bound is that confidence scores may vary over individual instances, capturing the heterogeneity of the domain and the specific training sample the label prediction relies on. To demonstrate this point, consider the following example:

Example 1. *Let X_1 be the 0.1 grid over $[0,1]^d$, let X_0 the 0.01 grid over $[0,0.1]^d$ and let our domain be $X = X_0 \cup X_1$ (for some large d). Consider the family $\mathcal{P}$ of all probability distributions over X that have a deterministic labeling rule satisfying the 10- Lipschitz condition (so points of distance 0.1 or more have no effect on each other). Assume further that all the distributions in $\mathcal{P}$ have half of their mass uniformly distributed over X_1 grid points and the other half of the mass uniformly distributed over X_0.*

Since outside the $[0,0.1]^d$ cube, every labeling is possible, for every learner there is a distribution $P \in \mathcal{P}$ w.r.t. which it errs on every domain point in $X_1 \setminus S_X$ (where $S_X = \{x : \exists y \in \{0,1\}$ s.t. $(x,y) \in S\}$). On the other hand, due to the Lipschitz condition, all the points in the $[0,0.1]^d$ grid, X_0, must get the same label. Therefore, given a sample S that includes a point in X_0, a learner that labels all the points in X_0 by the label of the sample points in it induces confidence 1 for all these points.

We conclude that, for sample sizes between 2 and, say, $10^d/2$, for most of the samples a learner can achieve confidence 1 for points in X_0 and no learner can achieve confidence above 0 for even a half of the domain points in X_1. Note also that the No Free Lunch theorem (as formulated in, e.g., [8]) implies that for sample sizes smaller than $10^d/2$, for every learner there exists some probability distribution $P \in \mathcal{P}$ for which its expected error is at least $1/8$ ($1/4$ over a subspace X_1 that has probability weight $1/2$).

4 Confidence scores for hypothesis classes

In the following we will analyze the point-wise confidence when all the prior knowledge available about the data generating distribution P is a bound on the approximation error of a class of predictors. We will distinguish two cases here,

1. The family $\mathcal{P}_{\mathcal{H},0}$ of distributions P which are realizable w.r.t. $\mathcal{H}$, i.e. $\inf_{h\in\mathcal{H}} L_P(h) = 0$ and
2. The family $\mathcal{P}_{\mathcal{H},\epsilon}$ of distributions P for which the approximation error of class $\mathcal{H}$ is low but not guaranteed to be zero, i.e. $\inf_{h\in\mathcal{H}} L_P(h) \leq \epsilon$, for some $\epsilon > 0$.

Note that, given a class of predictors, $\mathcal{H}$, the second family of possible data generating distributions is a superset of the first. Consequently, the pointwise error guarantees one can give in that non-necessarily-realizable case are weaker[1].

Definition 1 (Confidence Score, fulfilling the no-overestimation guarantee for all instances)**.** *We say a function C, that takes as input a sample S, a point x, a hypothesis h and a parameter δ and outputs a value in $[0,1]$. We say such a function C is a* confidence score fulfilling the no-overestimation guarantee for all instances *for a family of probability functions $\mathcal{P}$ if for every $P \in \mathcal{P}$ the probability over $S \sim P^m$ that there exists $x \in X$ with*

$$Pr_{y\sim Bernoulli(\ell_P(x))}[h(x) = y] < C(x, y, S, \delta)$$

is less than δ.

We say a function $C(x, y, S, \delta)$, is a *confidence score fulfilling the fulfilling the no-overestimation guarantee for positive mass instances* if the above guarantee holds not for all x, but for all x with $P(\{x\}) > 0$.

It is obvious, that this guarantee to achieve, if we give the confidence score 0 for all predictions. In order to measure the informativeness of a confidence score for a particular distribution we introduce the notion of non-triviality of a confidence score.

Definition 2 (Non-triviality)**.** *Given a confidence score C for some family of distributions $\mathcal{P}$, we define the non-triviality $nt_P(C, x, y)$ w.r.t. to a distribution $P \in P$ for a given sample size m and parameter δ, to be the extected difference between the estimated and the true confidence,i.e.:*

$$nt_P(C, m, \delta) = \mathbb{E}_{x\sim P, S\sim P^m}[1 - \min_{y\in\{0,1\}} |C_P(x, y) - C(x, y, S, \delta)|)]$$

Next we consider a specific confidence score that takes into account whether a hypothesis class is undecided on a point x given a sample S.

$$C_{\mathcal{H}}(x, y, S, \delta) = \begin{cases} 0 & \text{if there is } h \in \mathcal{H} \text{ s.t. } L_S(h) = 0 \text{ and } h(x) \neq y \\ 1 & \text{otherwise} \end{cases}$$

[1] To not have to deal with the ambiguity of the labeling function for points with mass 0, we will restrict this discussion to the family of distributions which have positive mass on all points. This implies that in the realizable setting all labeling function ℓ_P we consider are part of $\mathcal{H}$.

Proposition 1. *For the family of distributions P that are realizable with respect to $\mathcal{H}$, $C_{\mathcal{H}}$ is indeed a confidence score fulfilling the no-overestimation guarantee for all instances.*

This statement was made in a different setup by El-Yaniv et al [2]. We note that our confidence score $C_{\mathcal{H}}$ is equivalent to their notion of consistent selective strategy. Using our terminology, they show that if the realizability assumption holds, if an instance (x, y) is classified as 1 by $C_{\mathcal{H}}$ then x is guaranteed to have true label y (with probability 1). Furthermore, their Theorems 11 and Theorem 14 as well as their Corollory 28 give rise to the following observation about confidence scores.

Observation 1. *It turns out that $nt_P(C, m, \delta)$ for this confidence scoring rule $C_{\mathcal{H}}$ under the realizability assumption displays different behaviours for different classes (even when they have similar VC dimension):*

- *For some hypothesis classes, e.g. the class of thresholds on the real line $\mathcal{H}_{thres}$ or the class of linear separators in $\mathbb{R}^d$, $nt_P(C, m, \delta)$ converges to 1 for every $\delta > 0$ and every $P \in \mathcal{P}$ as the sample size go to infinity.*
- *On the other hand, for some hypothesis classes with finite VC-dimension for every $\epsilon > 0$ there exist $P \in \mathcal{P}_{\mathcal{H},0}$ with $nt_P(C, m, \delta) = 0$ for every sample size m and every $\delta < 1$. This phenomenon occurs for example for $\mathcal{H}$ being the class of singletons.*

For a more detailed analysis of which hypothesis classes have high non-triviality, we refer the reader to [2], noting that our notion of non-triviality corresponds to their notion of coverage.

We now look at the second case we wanted to address in this section: The family of probability distributions such that the approximation error of a class $\mathcal{H}$ is bounded by some ϵ. We fix a hypothesis class $\mathcal{H}$ and let $\mathcal{P}_{\mathcal{H},\epsilon}$ be the family of all probability distributions P such that $\mathcal{H}$ has approximation error at most ϵ w.r.t. P. We show that for for any (non-trivial) hypothesis class $\mathcal{H}$, it is not possible to find any satisfying confidence score for such a family.

Proposition 2. *Let $\mathcal{H}$ be any hypothesis over an infinite domain. Then there is no confidence score C such that the following two statements are true simultaneously:*

- *C fulfills the no-overestimation guarantee for all positive-mass instances w.r.t. $\mathcal{P}_{\mathcal{H},\epsilon}$*
- *there exists some $\eta > 0$ such that for every $P \in \mathcal{P}_{\mathcal{H},\epsilon}$ there are $\delta \in (0, 1)$, $m \in \mathbb{N}$ such that C has non-triviality $nt_P(C, m, \delta) > \eta$.*

This shows us that restricting ourselves to a family of probability distributions that allow for good generalization is not sufficient for allowing satisfying confidence scores. In the following section we will make stronger, more local assumptions and show that under these assumptions more satisfying confidence scores can be found.

5 Confidence scores under Lipschitz assumption

Lipschitz Assumption : We say that the probability distribution P over $X \times \{0, 1\}$ satisfies λ-Lipschitzness w.r.t. a metric $d(.,.)$ over X, if for every $x, x' \in X$, $|\ell_P(x) - \ell_P(x')| \leq \lambda d(x, x')$. When the domain is a subset of a Euclidean space, we will assume that d is the Euclidean distance unless we specify otherwise.

We provide an algorithm (1) to estimate the labelling probability of points using labelled samples. The algorithm partitions the space into cells. The algorithm outputs the same answer for points in the same cell. The input parameter r dictates the size of the cells. The algorithm estimates the average labelling probability for each cell. A confidence interval for this estimate is calculated based on the number of sample points in the cell. The interval is narrow when there are more sample points in the cell.

The following lemmas show how to estimate probability weights and average labelling probabilities of subsets of the domain:

Lemma 1. *Let P be a distribution over domain X. Let X' be a subset of X. Let S be an i.i.d. sample of size m drawn from the distribution P. Let $\hat{p}(X', S)$ be the fraction of the m samples that are in X'. For any $\delta > 0$, with probability $1 - \delta$ over the generation of the samples S,*

$$|P(X') - \hat{p}(X', S)| \leq w_p(m, \delta)$$

where

$$w_p(m, \delta) = \sqrt{\frac{1}{2m} \ln \frac{2}{\delta}}.$$

Lemma 2. *Let D be distribution over $X \times \{0, 1\}$. Let X' be a subset of X. Let S be an i.i.d. sample of size m drawn from D. Let $\hat{\ell}(X', S)$ be the fraction of the m labelled samples with label 1 in $S \cap X'$. For any $\delta > 0$, with probability $1-\delta$ over the generation of the samples S, if $\hat{p}(X', S) - w_p(m, \delta/2) > 0$, then*

$$|\bar{\ell}_P(X') - \hat{\ell}(X', S)| < w_\ell(m, \delta, \hat{p}(X', S))$$

$$w_\ell(m, \delta, \hat{p}(X', S)) = \frac{1}{\hat{p}(X', S) - w_p(m, \delta/2)} \cdot \left(w_p(m, \delta/2) + \sqrt{\frac{1}{2m} \ln \frac{4}{\delta}} \right),$$

where $\hat{p}(X', S)$ is the fraction of the samples from S in X' that have label 1, $w_p(m, \delta/2))$ is as defined in Lemma 1.

The following theorem shows that the true labelling probability of a point lies within the estimate interval provided by Algorithm 1, with high probability over the sample used to find the estimate.

Theorem 1. *Let the domain be $[0, 1]^d$. Suppose the data generating distribution P satisfies λ-Lipschitzness.*

Algorithm 1 Lipschitz labelling probability estimate

Input: Test point x, Labelled samples $S = (x_i, y_i)_{i=1}^m$,
Radius r, Estimation parameter δ,
Lipschitz constant λ
Output: Labelling probability estimate, confidence width of estimate
Split the domain $X = [0,1]^d$ into a grid of $(1/r)^d$ hypercube cells each of side length r.
Find the cell $\mathrm{t_x}$ containing the test point x.
$\hat{p}[\mathrm{t_x}] :=$ fraction of samples in $\mathrm{t_x}$.
$\hat{\ell}[\mathrm{t_x}] :=$ fraction of samples in the cell $\mathrm{t_x}$ with label 1.
$w[\mathrm{t_x}] := 1$
if $\hat{p}[\mathrm{t_x}] - w_p(m, \delta/2)$ **then**
  $w[\mathrm{t_x}] = w_\ell(m, \delta, \hat{p}[\mathrm{t_x}])$
end if
Return $\hat{\ell}[\mathrm{t_x}]$, $\min(1, w[\mathrm{t_x}] + r\lambda\sqrt{2})$

For any $r > 0, \delta > 0$, $m \in \mathbb{N}$, for any $x \in [0,1]^d$, define the confidence score based on Algorithm 1 as

$$\hat{C}^{r,\lambda}_{Lipschitz}(x, y; S, \delta) = \begin{cases} 1 - \hat{\ell}_S(x) - w_S(x) & \text{if } y = 1 \\ \hat{\ell}_S(x) - w_S(x) & \text{if } y = 0 \end{cases}$$

where $(\hat{\ell}_S(x), w_S(x))$ is the output of the Algorithm with input r, δ, λ. Then with probability $1 - \delta$ over samples S of size m,

$$\hat{C}^{r,\lambda}_{Lipschitz}(x, y; S, \delta) \leq C_P(x, y)$$

We now show that as sample size increases, for an appropriately chosen input parameter r, Algorithm 1 returns narrow estimate intervals for the labelling probabilities for most points. This implies that for most points, the confidence score is not much lower than the true confidence $(2|\ell_P(x) - \frac{1}{2}|)$

Theorem 2. *For every λ-Lipschitz distribution, for every $\epsilon_x, \epsilon_c, \delta > 0$, there is a sample size $m(\epsilon_x, \epsilon_c, \delta)$ such that with probability $1 - \delta$ over samples S of size $m(\epsilon_x, \epsilon_c, \delta)$,*

$$\Pr_{x \sim P}[w_S(x) > \epsilon_c] < \epsilon_x$$

where w_S is the width of labelling probability estimate obtained from Algorithm 1 with input parameter of grid size $r = 1/m^{\frac{1}{8d}}$

Note, that this theorem implies that the expected non-triviality is high.

6 Proofs

Useful lemmas

The following lemma appears as Theorem 2.2.6 of the book of [9], where the reader can find its proof.

Lemma 3 (Hoeffding's inequality for general bounded random variables)**.** *Let $X_1, \ldots, X_N$ be independent random variables. Assume that $X_i \in [m_i, M_i]$ for every i. Then, for any $t > 0$, we have*

$$\Pr\left[\sum_{i=1}^N (X_i - \mathbb{E}[X_i]) \geq t\right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^N (M_i - m_i)^2}\right).$$

Proof of Lemma 1. Let X_i be a random variable indicating if the i^{th} sample belongs to set X'. $X_i = 1$ if the i^{th} sample belongs to X' and zero otherwise. For each i, $\mathbb{E}[X_i] = P(X')$. $\hat{p}(X', S) = \frac{\sum_{i=1}^N X_i}{m}$. Applying Hoeffding's inequality, we get the inequality of the theorem. □

Proof of Lemma 2. Let X_i be a random variable such that

$$X_i = \begin{cases} 1 & \text{If } i^{\text{th}} \text{ sample is in } X' \text{ and has label one,} \\ 0 & \text{otherwise.} \end{cases}$$

$\mathbb{E}[X_i] = P(X')\bar{\ell}_P(X')$, for each i. $\sum_{i=1}^m X_i = m\hat{p}\hat{\ell}_P(X', S)$. Note that by triangle inequality,

$$\begin{aligned} &|P(X')\hat{\ell}_P(X', S) - P(X')\bar{\ell}_P(X')| \\ &\leq |\hat{p}\hat{\ell}_P(X', S) - P(X')\bar{\ell}_P(X')| + |\hat{p} - P(X')|\hat{\ell}_P(X', S) \\ &\leq |\hat{p}\hat{\ell}_P(X', S) - P(X')\bar{\ell}_P(X')| + w_p. \end{aligned}$$

For any $\epsilon > 0$,

$$\begin{aligned} &\Pr[|\bar{\ell}_P(X') - \hat{\ell}(X', S)| > \epsilon] \\ &= \Pr[P(X') \cdot |\bar{\ell}_P(X') - \hat{\ell}(X', S)| > P(X')\epsilon] \\ &\leq \Pr[|\hat{p}\hat{\ell}_P(X', S) - P(X')\bar{\ell}_P(X')| + w_p > (\hat{p} - w_p)\epsilon] \\ &= \Pr[|m\hat{p}\hat{\ell}(X', S) - mP(X')\bar{\ell}_P(X')| \\ &> m(\hat{p} - w_P)\epsilon - w_p] \\ &= \Pr[\sum_{i=1}^m |X_i - \mathbb{E}[X_i]| > m((\hat{p} - w_P)\epsilon - w_p)] \\ &\leq 2\exp\left(-2m((\hat{p} - w_P)\epsilon - w_p)^2\right) \end{aligned}$$

When $\hat{p} - w_p > 0$, choosing

$$w_l(m, \delta, \hat{p}) > \frac{w_p}{\hat{p} - w_p} + \frac{1}{\hat{p} - w_p}\sqrt{\frac{1}{2m}\ln\frac{4}{\delta}},$$

we get that with probability $1-\delta$, $|\bar{\ell}_P(X') - \hat{\ell}(X', S)| < w_\ell(m, \delta, \hat{p})$. □

Confidence scores for hypothesis classes

We start by recalling the definition of confidence scores fulfilling the no-overestimation guarantee for all instances:

Definition 1 (Confidence Score, fulfilling the no-overestimation guarantee for all instances)**.** *We say a function C, that takes as input a sample S, a point x, a hypothesis h and a parameter δ and outputs a value in $[0,1]$. We say such a function C is a* confidence score fulfilling the no-overestimation guarantee for all instances *for a family of probability functions $\mathcal{P}$ if for every $P \in \mathcal{P}$ the probability over $S \sim P^m$ that there exists $x \in X$*

$$Pr_{y \sim Bernoulli(\ell_P(x))}[h(x) = y] < C(x, y, S, \delta)$$

is less than δ.

6.1 Proof of Proposition 1

Proposition 1. *For the family of distributions P that are realizable with respect to $\mathcal{H}$, $C_{\mathcal{H}}$ is indeed a confidence score fulfilling the no-overestimation guarantee for all instances.*

We need to show that $C_{\mathcal{H}}$ fulfills Definition 1, that is, we need to show that for every hypothesis class $\mathcal{H}$ and every distribution P that fulfills the realizable condition w.r.t. $\mathcal{H}$, the probability over $S \sim P^m$ that there exists $x \in X$

$$Pr_{y \sim \text{Bernoulli}(\ell_P(x))}\left[h(x) = y\right] < C_{\mathcal{H}}(x, y, S, \delta)$$

is less than δ. Since $C_{\mathcal{H}}$ only assigns values 0 and 1 and the condition is trivially fulfilled for instances with confidence score 0, we will now only discuss the case where $C_{\mathcal{H}}$ assigns confidence score 1. Recall, that $C_{\mathcal{H}}$ only assigns confidence 1 if and only if there is no $h \in \mathcal{H}$ with $L_S(h) = 0$ and $h(x) \neq y$. Since we know, that realizability holds, we know that $\ell_P \in \mathcal{H}$ and since S is an i.i.d sample from P we know $L_S(\ell_P) = 0$. Now let $C_{\mathcal{H}}(x, y, S, \delta) = 1$, then by definition we know that $L_S(h) = 0$ implies $h(x) = y$. Thus $\ell_P(x) = y$. Thus, this confidence score does not overestimate the confidence of a point in any label.

6.2 Proof of Observation 1

We start by noting that our definition of the confidence score $C_{\mathcal{H}}$ is equivalent to the consistent selective strategy from [2]. In order to state their definition, we will first need to introduce some other concepts. First, we will state the definition of version space from [7].

Definition 3 (Version Space)**.** *Given a hypothesis class $\mathcal{H}$ and a labeled sample S, the* version space $\mathcal{H}_S$ *the set of all hypotheses in $\mathcal{H}$ that classify S correctly.*

Now, we can define the agreement set and maximal agreement set as in [2].

Definition 4 (agreement set)**.** *Let $\mathcal{G} \subset \mathcal{H}$. A subset $X \subset X$ is an* agreement set *with respect to $\mathcal{G}$ if all hypotheses in $\mathcal{G}$ agree on every instance in X', namely for all $g_1, g_2 \in \mathcal{G}$, $x \in X$*

$$g_1(x) = g_2(x).$$

Definition 5 (maximal agreement set)**.** *Let $\mathcal{G} \subset \mathcal{H}$. The* maximal agreement set *with respect to $\mathcal{G}$ is the union of all agreement sets with respect to $\mathcal{G}$.*

We can now state the definition of consistent selective strategy. Note, that a selective strategy is defined by a pair (h, g) of a classification function h and a selective function g. For our purposes, we will only need to look at the selective function g.

Definition 6 (consistent selective strategy (CSS))**.** *Given a labeled sample S, a* consistent selective strategy (CSS) *is a selective classification strategy that takes h to by any hypothesis in $\mathcal{H}_S$ (i.e., a consistent learner), and takes a (deterministic) selection function g that equals one for all points in the maximal agreement set with respect to $\mathcal{H}_S$ and zero otherwise.*

We now see that for any $\mathcal{H}$ and any labeled sample S the selected function g assigns one to x, if for every two $h_1, h_2 \in H$ with $L_S(h_1) = L_S(h_2) = 0$ implies $h_1(x) = h_2(x)$. Thus, $g(x) = C_{\mathcal{H}}(x, h(x), S, \delta)$ for any $h(x) \in \mathcal{H}_S$. In [2] the selection function is then analysed with respect to its *coverage*, which is defined by $\phi(h, g) = \mathcal{E}_{x \sim P}[g(x)]$ for a given distribution P. Note, that our notion of non-triviality corresponds to this notion of coverage for deterministic distributions and binary confidence scores. We can now use some of their results to show our observation.

Theorem 3 (non-achievable coverage, Theorem 14 from [2])**.** *Let m and $d > 2$ be given. There exist a distribution P, an infinite hypothesis class $\mathcal{H}$ with a finite VC-dimension d, and a target hypothesis in $\mathcal{H}$, such that $\phi(h, g) = 0$ for any selective classifier (h, g), chosen from $\mathcal{H}$ by CSS using a training sample S of size m drawn i.i.d. according to P.*

This directly implies the second part of our observation. For a more concrete example consider the class of singletons over the real line $\mathcal{H}_{sgl} = \{h_z : \mathbb{R} \to \{0, 1\} : h_z(x) = 1 \text{ if and only if } z = x\}$. We note that $C_{\mathcal{H}_{sgl}}$ only gives positive confidence scores to instances outside the sample if the sample contains a positively labeled instance. Let P be the uniform distribution over $[0, 1] \times \{0\}$. Obviously $P \in \mathcal{P}_{\mathcal{H},0}$. Furthermore any sample generated by P will not contain any positively labeled sample. Thus, $nt_P(C, m, \delta) = \mathbb{P}_{x \sim P_X, S \sim P^m}[x \in S_X] = 0$, where P_X and S_X denote the projections of P and S to the domain $X = \mathbb{R}$.

Let us consider the class of thresholds $\mathcal{H}_{thres} = \{h_a : \mathcal{R} \to \{0, 1\}, h_a(x) = 1 \text{ if and only if } x > a\}$. We can define the following two learning rules for thresholds:

$$A_1(S) = h_{a_1}, \text{ where } a_1 = \operatorname*{arg\,max}_{x_i \in \mathbb{R}:(x_i,0) \in S} x_i$$

$$A_2(S) = \bar{h}_{a_2}, \text{ where } \bar{h}_{a_2}(x) = 1 \text{ if and only if } x \geq a_2 := \operatorname*{arg\,min}_{x_i \in \mathbb{R}:(x_i,1) \in S} x_i.$$

Furthermore, The way $C_{\mathcal{H}_{thres}}$ is defined, for any S and $\delta \in (0, 1)$ we have $A_1(S)(x) = A_2(S)(x)$ if and only if there is $y \in \{0, 1\}$ with $C_{\mathcal{H}_{thres}}(x, y, S, \delta) = 1$. Furthermore, both of these learning rules are empirical risk minimizers for $\mathcal{H}_{thres}$ in the realizable case. Thus both of them are PAC-learners. Thus for any ϵ, δ, there is a $m(\epsilon, \delta)$, such that for any $m \geq m(, \delta)$ and any $P \in \mathcal{P}_{\mathcal{H}_{thres},0}$,

$$1 - \delta \leq Pr_{S \sim P^m, x \sim P}[A_1(S) = A_2(S)] =$$

$$Pr_{S \sim P^m, x \sim P}[\max_{y \in \{0,1\}} \{C_{\mathcal{H}_{thres}}(x, y, S, \delta)\}]$$

$$= nt_{\mathcal{P}_{\mathcal{H}_{thres},0}}(C, m, 0)$$

Thus $\lim_{m \to \infty} nt_{\mathcal{P}_{\mathcal{H}_{thres},0}}(C_{\mathcal{H}thres}, m, \delta) = 1$ for any $\delta \in (0, 1)$ and any $P \in \mathcal{P}_{\mathcal{H}_{thres},0}$. Furthermore note that we have uniform convergence.

6.3 Proof of Proposition 2

For every $\epsilon > 0$, every $x \in X'$ and every $n \in \mathbb{N}$ with $m > \frac{1}{\epsilon}$, we can construct a distribution $P_{x,n}$, such that $l_{P_{x,n}}(x') = h_1(x')$ for every $x' \in X \setminus \{x\}$ and $h_1(x') \neq l_{P_{x,n}}(x')$ and such that the marginal $P_{x,n,X}$ is uniform over some $X_n \subset X'$ with $|X_n| = n$. For a sample to distinguish between two such distributions $P_{x_1,n}$ and $P_{x_2,n}$ either the point x_1 or x_2 needs to be visited by the sample. Thus in order to give a point-wise guarantee for all instances with positive mass, only points in the sample can be assigned a positive confidence in this scenario. Thus any confidence score fulfilling this guarantee would have $nt_{P_{x,n}}(C, m, \delta) = \mathbb{P}_{x \sim P_X, S \sim P^m}[x \in S_X] \leq \frac{m}{n}$. For every $\eta > 0$ we can find n such that $\frac{m}{n} \leq \eta$, proving our claim.

Confidence scores using Lipschitz assumption

Proof of Theorem 1. The algorithm partitions the space into r^d cells. Let p_c be the probability weight of a cell c and let $\hat{p}_c$ be the estimate of p_c that is calculated based on a sample to be the fraction of sample points in the cell c. From Lemma 1 and a union bound, we know that with probability $1 - \frac{\delta}{2}$, for every cell c,

$$p_c \in [\hat{p}_c - w_p(c), \hat{p}_c + w_p(c)].$$

Here $w_p(c) = w_p(m, \delta/2r^d)$ (as defined in Lemma 1).

The algorithm also estimates the average label of a cell c - ℓ_c as $\hat{\ell}_c$. This is the fraction of the sample point in the cell that have the label one. This is the same as the labelling probability estimate defined in Lemma 2. When the true probability weights of cells lie within the calculated confidence interval, by Lemma 2, we know that with probability $1 - \frac{\delta}{2}$, for every cell c,

$$\hat{\ell}_c \in [\hat{\ell}_c - w_\ell(c), \hat{\ell}_c - w_\ell(c))].$$

Here $w_\ell(c) = w_\ell(m, \delta/2r^d, \hat{p}_c)$ (as defined in Lemma 2).

The maximum distance between any two points in any cell is $r\sqrt{2}$. By the λ-Lipshitz, any point in the cell has labelling probability within $\lambda r\sqrt{2}$ of the average labelling probability of the cell. Therefore, with probability $1 - \delta$, for each cell c, for every point x in the cell c, the labelling probability of x satisfies:

$$\ell_P(x) \in [\hat{\ell}_c - w_\ell(c) - \lambda r\sqrt{2}, \hat{\ell}_c + w_\ell(c) + \lambda r\sqrt{2}].$$

This is the interval returned by the algorithm. Now we lower bound true confidence based on the confidence interval of the labelling probability. For a point x, let $c(x)$ denote the cell containing the point.

$$\begin{aligned} C_P(x, 0) &= \ell_P(x) \\ &\geq \hat{\ell}_{c(x)} - w_\ell(c(x)) - \lambda r\sqrt{2} \\ C_P(x, 1) &= 1 - \ell_P(x) \\ &\geq 1 - \hat{\ell}_{c(x)} - w_\ell(c(x)) - \lambda r\sqrt{2}. \end{aligned}$$

□

Proof of Theorem 2. We choose the input to the algorithm to be $r = \frac{1}{m^{1/8d}}$. With probability $1 - \frac{\delta}{2}$, for all cells with probability weight greater than $\gamma = \frac{1}{m^{1/4}}$, the length of the confidence interval of the labelling probability is less than

$$\begin{aligned} &\frac{\frac{1}{m^{1/2}}}{\frac{1}{m^{1/4}} + \frac{1}{m^{1/2}}} - \frac{1}{\frac{1}{m^{1/4}} - \frac{1}{m^{1/2}}} \sqrt{\frac{1}{2m} \ln \frac{4m^{1/8}}{\delta}} + \frac{\lambda\sqrt{2}}{m^{1/8}} \\ &\leq \frac{1}{m^{1/4} - 1} + \frac{1}{m^{1/4} - 1} \sqrt{\frac{1}{16} \ln \frac{4m}{\delta}} + \frac{\lambda\sqrt{2}}{m^{1/8}}. \end{aligned}$$

This quantity decreases with increase in m and converges to zero. Therefore, for every $\epsilon_c > 0$, there is $M_1(\epsilon_c, \delta)$ such that this interval is less than ϵ_c. When sample size is larger than $M_1(\epsilon_c, \delta)$, with probability $1 - \frac{\delta}{2}$, the size of confidence intervals for labelling probabilities of cells with weights greater than $\gamma = \frac{1}{m^{1/4}}$, is smaller than ϵ_c.

The points for which we can't say anything about the interval lengths are points in cells with weight at most γ. The total weight of such points is at most $\gamma \frac{1}{r^d} = \frac{1}{m^{1/8}}$. For any $\epsilon_x > 0$, let $M_2(\epsilon_x)$ be such that $\frac{1}{M_2(\epsilon_x)^{1/8}} < \epsilon_x$.

Choosing a sample size M greater than $M_1(\epsilon_c, \delta)$ and $M_2(\epsilon_x)$, we get that

$$\Pr_{S \sim P^M}[w_\ell > \epsilon_c] < \epsilon_x.$$

□

References

[1] Bartlett, P. L.; and Wegkamp, M. H. 2008. Classification with a Reject Option using a Hinge Loss. *Journal of Machine Learning Research* 9(59): 1823–1840. URL http://jmlr.org/papers/v9/bartlett08a.html.

[2] El-Yaniv, R.; and Wiener, Y. 2010. On the Foundations of Noise-free Selective Classification. *J. Mach. Learn. Res.* 11: 1605–1641. URL http://portal.acm.org/citation.cfm?id=1859904.

[3] Freund, Y.; Mansour, Y.; Schapire, R. E.; et al. 2004. Generalization bounds for averaged classifiers. *The annals of statistics* 32(4): 1698–1722.

[4] Herbei, R.; and Wegkamp, M. H. 2006. Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* 709–721.

[5] Jiang, H.; Kim, B.; Guan, M.; and Gupta, M. 2018. To trust or not to trust a classifier. In *Advances in neural information processing systems*, 5541–5552.

[6] Kalai, A. T.; Kanade, V.; and Mansour, Y. 2012. Reliable agnostic learning. *Journal of Computer and System Sciences* 78(5): 1481–1495.

[7] Mitchell, T. M. 1977. Version Spaces: A Candidate Elimination Approach to Rule Learning. In Reddy, R., ed., *Proceedings of the 5th International Joint Conference on Artificial Intelligence. Cambridge, MA, USA, August 22-25, 1977*, 305–310. William Kaufmann. URL http://ijcai.org/Proceedings/77-1/Papers/048.pdf.

[8] Shalev-Shwartz, S.; and Ben-David, S. 2014. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press. ISBN 978-1-10-705713-5. URL http://www.cambridge.org/de/academic/subjects/computer-science/pattern-recognition-and-machine-learning/understanding-machine-learning-theory-algorithms.

[9] Vershynin, R. 2019. High-dimensional probability.

[10] Wiener, Y.; and El-Yaniv, R. 2015. Agnostic pointwise-competitive selective classification. *Journal of Artificial Intelligence Research* 52: 171–201.

[11] Yuan, M.; and Wegkamp, M. H. 2010. Classification Methods with Reject Option Based on Convex Risk Minimization. *J. Mach. Learn. Res.* 11: 111–130. URL https://dl.acm.org/citation.cfm?id=1756011.