Selective Prediction

Binary classifications

Rong Zhou

November 8, 2017

Table of contents

What are selective classifiers?

Introduction

Selective classifiers are:

- allowed to reject making predictions without penalty.
- compelling with applications where wrong classifications are not welcomed and partial domain for predictions is allowed.

Introduction

From Hierarchical Concept Learning:
A variation on the Valiant Model [2]:

. . . the learner is (instead) supposed to give a program taking instances as input, and having three possible outputs: 1,0, and "I don't know".

. . . Informally we call a learning algorithm **useful** *if the program outputs "I don't know" on at most a fraction ϵ of all instances . . .*

What is an ideal selective classifier?

Suppose we are given training examples labelled -1 or 1, and the goal is to design an algorithm to find a good selective classifier.

- The *misclassification rate* should not be the only measurement for selective classifiers.
- A selective classifier with zero *misclassification rate* can be a very "bad" classifier. Examples?

Notations and Definitions

For a selective classifier/predictor $\mathcal{C}$ in a binary classification problem where $x_i \in \mathcal{X}$ and $y_i \in \{-1, 1\}$.

- Coverage ($cover(\mathcal{C})$) : the probability that $\mathcal{C}$ predicts a label instead of 0.
- Error ($err(\mathcal{C})$): the probability that the true label is the opposite of what $\mathcal{C}$ predicts [Note: 0 is not counted as errors].
- Risk ($risk(\mathcal{C})$):

$$risk(\mathcal{C}) = \frac{err(\mathcal{C})}{cover(\mathcal{C})}$$

An ideal classifier/predictor should have both error and coverage guarantees with high probability $(1 - \delta)$.

Forms of selective predictors/classifiers

For a specific sample x:

- Confidence-rated Predictor

$$[p_{-1}, p_0, p_1]$$

- Selective Classifier
    - $$(h, \gamma_x), \text{ where } 0 \leq \gamma_x \leq 1, h \in H$$
    - $$(h, g(x)) \text{ where } g(x) = 0 \text{ or } 1 \text{ and } h \in H$$

.

The Realizable Setting

The Realizable Setting

In the realizable setting, our target hypothesis h^* is in our hypothesis class H and the labels are corresponding to what h^* predicts.

An Optimization Problem

We are given:

- a set of n labelled examples $S = \{\{x_1, y_1\}, \{x_2, y_2\}, \dots, \{x_n, y_n\}\}$
- a set of m unlabelled examples $U = \{x_{n+1}, x_{n+2}, \dots, x_{n+m}\}$
- a set of hypotheses H

Goal: learn a selective classifier/predictor with an error guarantee ϵ, and the best possible coverage **for the unlabelled examples in U**.

An Optimization Problem

Confidence-rated predictor: A confidence-rated predictor ($\mathcal{C}$) is a mapping from U to a set of m distributions over {-1,0,1}. For example, if the i-th distribution is $[\beta_i, 1 - \beta_i - \alpha_i, \alpha_i]$, then

$$Pr(\mathcal{C}(x_i) = -1) = \beta_i$$

$$Pr(\mathcal{C}(x_i) = 1) = \alpha_i$$

$$Pr(\mathcal{C}(x_i) = 0) = 1 - \beta_i - \alpha_i$$

Recall that the version space V is a candidate set of hypotheses in the hypothesis class H.

An Optimization Problem

Algorithm 1: Confidence-rated Predictor [1]

Inputs: Labelled data S, unlabelled data U, error bound ϵ.

Compute version space V with respect to S.

Solve the linear program:

$$\max \sum_{i=1}^{m} (\alpha_i + \beta_i)$$

subject to:

$$\forall i, \alpha_i + \beta_i \leq 1$$

$$\forall i, \alpha_i, \beta_i \geq 0$$

$$\forall h \in V, \sum_{i:h(x_{n+i})=1} \beta_i + \sum_{i:h(x_{n+i})=-1} \alpha_i \leq \epsilon m$$

Output the confidence-rated predictor:

$$\{[\beta_i, 1 - \beta_i - \alpha_i, \alpha_i], i = 1, 2, \ldots, m\}$$

An Optimization Problem

Let a selective classifier ($\mathcal{C}$) defined by a tuple $(h, (\gamma_1, \gamma_2, \ldots, \gamma_m))$ where $h \in H, 0 \leq \gamma_i \leq 1$ for all $i = 1, 2, \ldots m$.

For any x_i, $\mathcal{C}(x_i) = h(x_i)$ with probability γ_i, and 0 with probability $1 - \gamma_i$.

An Optimization Problem

Algorithm 2: Selective Classifier [1]

Inputs: Labelled data S, unlablelled data U, error bound ϵ.

Compute version space V with respect to S. Pick an arbitrary $h_0 \in V$

Solve the linear program:

$$\max \sum_{i=1}^{m} \gamma_i$$

subject to:

$$\forall i, 0 \leq \gamma_i \leq 1$$

$$\forall h \in V, \sum_{i : h(x_{n+i}) \neq h_0(x_{n+i})} \gamma_i \leq \epsilon m$$

Output the selective classifier:

$$(h_0, (\gamma_1, \gamma_2, \ldots, \gamma_m))$$

.

Optimization Problems

Both algorithms can guarantee the ϵ error with optimal/ "almost optimal" coverage.

Some drawbacks using the optimization algorithms:

- Only work for those m unlabelled samples.
- Number of constraints can be infinite.

A More General Problem

Now let's generalize the problem:
We are given:

- a set of n labelled examples $S = \{\{x_1, y_1\}, \{x_2, y_2\}, \dots, \{x_n, y_n\}\}$
- a set of hypotheses H with VC dimension d

Goal: learn a selective classifier/predictor with **zero error** over the distribution $\mathcal{X}$ and the largest possible coverage with high probability $1 - \delta$.

Notations and Definitions

Let the selective classifier be:

$$\mathcal{C}(x) = (h, g)(x) = \begin{cases} h(x) & \text{if } g(x) = 1 \\ 0 & \text{if } g(x) = 0 \end{cases}$$

$$cover(h, g) = \mathbb{E}[g(X)]$$

Let $\hat{h}$ be the empirical error minimizer. Define the true error:

$$err_P(h) = Pr_{(X,Y)\sim P}(h(X) \neq Y)$$

Notations and Definitions

With respect to the hypothesis class H, distribution P over $\mathcal{X}$, and real number $r > 0$, define a true error ball:

$$\mathcal{V}(h, r) = \{h' \in H : err_P(h') \leq err_P(h) + r\}$$

and

$$\mathcal{B}(h, r) = \{h' \in H : Pr_{X \sim P}\{h'(X) \neq h(X)\} \leq r\}$$

Notations and Definitions

Define the disagreement region of a hypotheses set H:

$$\mathbb{DIS}(H) = \{x \in \mathcal{X} : \exists h_1, h_2 \in H \text{ such that } h_1(x) \neq h_2(x)\}$$

For $G \subseteq H$, let ΔG denotes the volume of the disagreement region. Specifically,

$$\Delta G = Pr\{\mathbb{DIS}(G)\}$$

Learning a Selective Classifier

Algorithm 3: Selective Classifier Strategy

Inputs: n labelled data S, d, δ.
Output: a selective classifier (h,g) such that $risk(h, g) = risk(h^*, g)$
Compute version space V with respect to S. Pick an arbitrary $h_0 \in V$
Set $G = V$
Construct g such that $g(x) = 1$ if and only if $x \in \{\mathcal{X} \setminus \mathbb{DIS}(G)\}$
$h = h_0$

Learning a Selective Classifier

Analysis of the Strategy

$\forall x \in \mathcal{X}$, when $g(x) = 1$, the target hypothesis h^* agrees with h.

$$\Rightarrow risk(h, g) = risk(h^*, g)$$

Learning a Selective Classifier

(thm 2.15: Consistent Hypothesis error rate bound in terms of VC dimension) For any n and $\delta \in (0,1)$, with probability at least $1-\delta$, every hypothesis $h \in V$ has error rate

$$err_P(h) \leq \frac{4d\ln(2n+1) + 4\ln\frac{4}{\delta}}{n}$$

Let $r = \frac{4d\ln(2n+1)+4\ln\frac{4}{\delta}}{n}$, we know that if $h \in V$, $h \in \mathcal{V}(h^*, r)$

$$\Rightarrow V \subseteq \mathcal{V}(h^*, r)$$

Learning a Selective Classifier

Now, if $h \in \mathcal{V}(h^*, r)$

$$\mathbb{E}[1_{h(X) \neq h^*(X)}] = \mathbb{E}[1_{h(X) \neq Y}] \leq r$$

By definition, $h \in \mathcal{B}(h^*, r)$.
Thus, with probability $1 - \delta$

$$V \subseteq \mathcal{V}(h^*, r) \subseteq \mathcal{B}(h^*, r)$$

$$\Delta V \leq \Delta \mathcal{B}(h^*, r)$$

Learning a Selective Classifier

Recall the definition of *disagreement coefficient*:

$$\theta = sup_{r>0} \frac{\Delta\mathcal{B}(h^*, r)}{r}$$

we have:

$$\forall r \in (0, 1), \Delta\mathcal{B}(h^*, r) \leq \theta \cdot r$$

Therefore, with probability at least $1 - \delta$,

$$\Delta V \leq \Delta\mathcal{B}(h^*, r) \leq \theta \cdot r$$

$$cover(h, g) = 1 - \Delta V \geq 1 - \theta \cdot r = 1 - \theta \frac{4d \ln(2n + 1) + 4 \ln \frac{4}{\delta}}{n}$$

The Noisy Setting

The Noisy Setting

In the noisy setting, our target hypothesis h^* is in our hypothesis class H but the labels are corresponding to the prediction of h^* with noises.

Learning a Selective Classifier - the Noisy Setting

Algorithm 4: Selective Classifier Strategy - Noisy [3]

Inputs: n labelled data S, d, δ.

Output: a selective classifier (h,g) such that $risk(h, g) = risk(h^*, g)$ with probability $1 - \delta$

Set $\hat{h} = ERM(H, S)$ so that $\hat{h}$ is any empirical risk minimizer from H.

Set $G = \hat{\mathcal{V}}(\hat{h}, 4\sqrt{2\frac{d \ln(\frac{2ne}{d}) + \ln \frac{8}{\delta}}{n}})$

Construct g such that $g(x) = 1$ if and only if $x \in \{\mathcal{X} \setminus \mathbb{DIS}(G)\}$

$h = \hat{h}$

Learning a Selective Classifier - the Noisy Setting

Consider a loss function $\mathcal{L}(\mathcal{Y}, \mathcal{Y})$.

$$risk(h, g) = \frac{\mathbb{E}[\mathcal{L}(h(X), Y)) \cdot g(X)]}{cover(h, g)}$$

Let h^* be the true risk minimizer, we define the *excess loss class* as:

$$\mathcal{F} = \{\mathcal{L}(h(x), y) - \mathcal{L}(h^*(x), y) : h \in H\}$$

Learning a Selective Classifier - the Noisy Setting

Class $\mathcal{F}$ is said to be a (β, B)-*Bernstein* class with respect to P (where $0 \leq \beta \leq 1$ and $B \geq 1$), if every $f \in \mathcal{F}$ satisfies

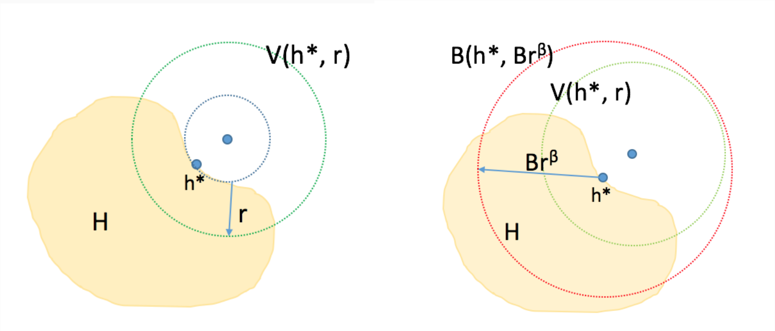
$$\mathbb{E}f^2 \leq B(\mathbb{E}f)^\beta$$

Learning a Selective Classifier - the Noisy Setting

We will proof the following lemmas to show the error guarantee and the coverage guarantee. [Note: The following proofs define the loss function to be 0/1 loss].

- If $\mathcal{F}$ is said to be a (β, B)-*Bernstein* class with respect to P, then for any $r > 0$:

$$\mathcal{V}(h^*, r) \subseteq \mathcal{B}(h^*, Br^\beta)$$



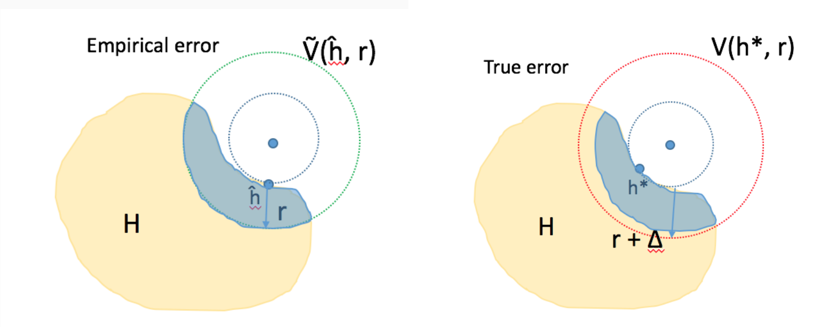
Learning a Selective Classifier - the Noisy Setting

Let

$$\sigma(n,\delta,d) = 2\sqrt{2\frac{d\ln(\frac{2ne}{d}) + \ln\frac{2}{\delta}}{n}}$$

- For any $0 < \delta < 1$, and $r > 0$, with probability of at least $1-\delta$,

$$\hat{\mathcal{V}}(\hat{h}, r) \subseteq \mathcal{V}(h^*, 2\sigma(n, \delta/2, d) + r)$$



Learning a Selective Classifier - the Noisy Setting

- Assume that H has disagreement coefficient θ and that $\mathcal{F}$ is said to be a (β, B)-*Bernstein* class with respect to P, then for any $r > 0$ and $0 < \delta < 1$, with probability of at least $1 - \delta$:

$$\Delta\hat{\mathcal{V}}(\hat{h}, r) \leq B\theta(2\sigma(n, \delta/2, d) + r)^{\beta}$$

Learning a Selective Classifier - the Noisy Setting

- Assume that H has disagreement coefficient θ and that $\mathcal{F}$ is said to be a (β, B)-*Bernstein* class with respect to P, then for any $r > 0$ and $0 < \delta < 1$, with probability of at least $1 - \delta$:

  $$\mathit{cover}(h, g) \geq 1 - B\theta(2\sigma(n, \delta/2, d) + r)^{\beta} \quad \wedge \quad \mathit{risk}(h, g) = \mathit{risk}(h^*, g)$$

References

Kamalika Chaudhuri and Chicheng Zhang.
Improved algorithms for confidence-rated prediction with error guarantees.
2013.

Ronald L Rivest and Robert Sloan.
A formal model of hierarchical concept-learning.
Information and Computation, 114(1):88–114, 1994.

Yair Wiener and Ran El-Yaniv.
Agnostic selective classification.
In *Advances in neural information processing systems*, pages 1665–1673, 2011.