COMS 6998-4 Fall 2017 Presenter: Rong Zhou
November 8, 2017 Scribe: Wenxi Chen

Selective Prediction

1 Introduction

In our previous discussion on a variation on the Valiant Model [3], the described learner has the ability to output "I don't know" in addition to the regular binary outputs. This is the behavior of a selective classifier. In other words, a selective classifier is allowed to reject decision making without penalty. Such classifiers are useful in making medical predictions, because the cost of making a wrong prediction is usually much higher than refusing to make any decision in these situations.

1.1 Ideal Selective Classifier

What would an ideal selective classifier looks like? Intuitively, misclassification rate should not be the only measurement for selective classifiers. Imagine a selective classifier that refuses making any prediction, such classifier is useless, but it has a misclassification rate of 0. Therefore, it makes sense to evaluate a selective classifier not only by its misclassification rate, but also the probability it will refuse making prediction.

Let $\mathcal{C}$ be a selective classifier in binary classification setting, we define the following:

- Coverage ($\text{cover}(\mathcal{C})$): the probability that $\mathcal{C}$ predicts a label instead of refusing making prediction.
- Erorr ($\text{err}(\mathcal{C})$): the probability that the true label is different from $\mathcal{C}$'s prediction when $\mathcal{C}$ makes prediction.
- Risk: $\text{risk}(\mathcal{C}) = \frac{\text{err}(\mathcal{C})}{\text{cover}(\mathcal{C})}$

We seek to bound both error and coverage of an ideal classifier with high probability $(1-\delta)$ where $0 \leq \delta \leq 1$.

2 Realizable Setting

Let $\mathcal{X}$ denote the feature space, $\mathcal{D}$ over $\mathcal{X} \times \{-1, 1\}$ be the underlying unknown data distribution. Set $S = \{\{x_1, y_1\}, \{x_2, y_2\}, ..., \{x_n, y_n\}\}$ is a set of n labelled examples, and $U = \{x_{n+1}, x_{n+2}, ..., x_{n+m}\}$ is a set of m unlabelled examples, where $x_i \in \mathcal{X}$ and $y_j \in \{-1, 1\}$ for $1 \leq i \leq n+m, 1 \leq j \leq n$. Let $\mathcal{H}$ be a set of hypotheses, and $h^* \in \mathcal{H}$ be the target hypothesis such that the true label of x is the same as the prediction $h^*(x)$. In addition, we define version space V with respect to S to be the set of hypotheses that are consistent with the examples in S.

We introduce 3 selective classifiers: *Confidence-rated Predictor*, *CZ Selective Classifier* and *Selective Classifier Strategy*.

2.1 Confidence-rated Predictor

A confidence-rated predictor $\mathcal{C}$ maps from U to a set of m distributions over $\{-1, 0, 1\}$. If the j-th distribution is $[\beta_j, 1-\beta_j-\alpha_j, \alpha_j]$, then $\Pr\{\mathcal{C}(x_{n+j}) = -1\} = \beta_j$, $\Pr\{\mathcal{C}(x_{n+j}) = 1\} = \alpha_j$ and $\Pr\{\mathcal{C}(x_{n+j}) = 0\} = 1-\beta_j-\alpha_j$.

Algorithm 1 Confidence-rated Predictor

---

input Labelled data S, unlabelled data U, error bound ϵ.

1: Compute version space V with respect to S.

2: Solve the linear program:

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

subject to:

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

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

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

3: **return** the confidence-rated predictor $\{[\beta_i, 1-\beta_i-\alpha_i, \alpha_i], i = 1, 2, ..., m\}$.

Theorem 1. *A confidence-rated predictor produced by Algorithm 1 has an error guarantee ϵ with optimal coverage for the unlabelled examples in U.*

Proof. According the the constraints in the linear program, $0 \leq \alpha_i \leq 1, 0 \leq \beta_i \leq 1, 0 \leq 1 - \alpha_i - \beta_i \leq 1$ and the probability the predictor makes a wrong decision with respect to the uniform distribution over U is less than $\epsilon m/m = \epsilon$. Moreover, the linear program maximizes

$$\sum_{i=1}^{m} \alpha_i + \beta_i,$$

which is equivalent as minimizing

$$\frac{1}{m} \sum_{i=1}^{m} 1 - \alpha_i - \beta_i$$

that is the coverage of the predictor. □

Remark 1. *The feasible region of the linear program in Algorithm 1 is always non-empty.*

Proof. The candidate solution $\{[0, 1, 0], i = 1, 2, ..., 3\}$ will always be in the feasible solution set. □

2.2 CZ Selective Classifier

A CZ selective classifier $\mathcal{C}$ is defined by a tuple $(h, (\gamma_1, \gamma_2, ..., \gamma_m))$ where $h \in \mathcal{H}$ and $0 \leq \gamma_i \leq 1$ for all $i = 1, 2, ..., m$. Given unlabelled example $x_{n+i} \in U$, $\mathcal{C}$ predicts 0 with probability $1 - \gamma_i$ and predicts $\mathcal{C}(x_{n+i}) = h(x_{n+i})$ with probability γ_i.

Algorithm 2 CZ Selective Classifier

input Labelled data S, unlabelled data U, error bound ϵ.

1: Compute version space V with respect to S.
2: Randomly choose $h_0 \in V$
3: Solve the linear program:

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

subject to:

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

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

4: **return** the CZ selective classifier $(h_0, (\gamma_1, \gamma_2, ..., \gamma_m))$.

Theorem 2. *Let $\mathcal{P}$ be a confidence-rated predictor produced by Algorithm 1. A CZ selective classifier $\mathcal{C}$ produced by Algorithm 2 has an error guarantees ϵ with* $\text{cover}(\mathcal{C}) \geq \text{cover}(\mathcal{P}) - \epsilon$.

Intuitively, there exists $h_1 \in V$ that produces the most different predictions on U from h_0. If h_1 and h_0 lies in two opposite ends of the version space, $\text{cover}(\mathcal{C})$ would be worse than the average case.

2.3 Selective Classifier Strategy

There are a few drawbacks using confidence-rated predictors and CZ selective classifiers. First, the number of constraints can be infinite. Second, we need to input unlabelled dataset U and these two methods only produce prediction of examples in U. However, we want to generalize the problem and provide a solution under inductive setting. Thus, we are going to remove U from the input to the algorithm.

In order to use selective classifier strategy, we introduce a few additional definitions:

- Let d be the VC dimension of $\mathcal{H}$.
- True error rate of hypothesis h is:

$$\text{err}_P(h) = \Pr_{(X,Y)\sim\mathcal{D}}\{h(X) \neq Y\}.$$

- empirical error rate of hypothesis h is:

$$\mathrm{err}_S(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{h(x_i)\neq y_i\}}.$$

- For any hypothesis class $\mathcal{H}$, distribution $\mathcal{D}$ over $\mathcal{X}$, and real number $r > 0$,

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

$$\hat{\mathcal{V}}(h,r) = \{h' \in \mathcal{H} : \mathrm{err}_S(h') \leq \mathrm{err}_S(h) + r\}$$

  and

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

- Disagreement region of hypothesis class $\mathcal{H}$ is:

$$\mathrm{DIS}(H) = \{x \in \mathcal{X} : \exists h_1, h_2 \in \mathcal{H} \text{ s.t. } h_1(x) \neq h_2(x)\}$$

  for $G \subseteq \mathcal{H}$, let ΔG the volume of the disagreement region:

$$\Delta G = \Pr\{\mathrm{DIS}(G)\}.$$

- disagreement coefficient is:

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

- A new type of selective classifier $\mathcal{C}$ s.t.

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

  and

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

Algorithm 3 Selective Classifier Strategy

input n labelled data S, d, δ.

Output a selective classifier (h,g) s.t. $\mathrm{risk}(h,g) = \mathrm{risk}(h^*,g)$

1: Compute version space V with respect to S.
2: Randomly choose $h_0 \in V$.
3: Set $G = V$.
4: Construct g s.t. $g(x) = 1$ if and only if $x \in \{\mathcal{X} \setminus \mathrm{DIS}(G)\}$.
5: Set $h = h_0$.

Theorem 3. *(Consistent Hypothesis error rate bound in terms of VC dimension, theorem 2.15 from lecture notes). For any n and $\delta \in (0, 1)$, with probability at least $1 - \delta$, every hypothesis $h \in V$ has error rate*

$$\mathrm{err}_P(h) \leq \frac{4d \ln(2n+1) + 4 \ln \frac{4}{\delta}}{n}.$$

Theorem 4. *Let $\mathcal{C} = (h, g)$ be a selective classifier output by Algorithm 3, then h achieves 0 error rate when it makes prediction, and $\mathcal{C}$ achieves near optimal coverage with probability $1 - \delta$.*

Proof. Given $g(x) = 1$ if and only if x is not in the disagreement region of the version space V, when h makes a prediction, it will always agree with all hypotheses in V including h^*. Thus, h achieves 0 error rate, moreover $\mathrm{risk}(h, g) = \mathrm{risk}(h^*, g)$.

Now, we will show that $\mathcal{C}$ achieves near optimal coverage with probability $1 - \delta$. Let $r = \frac{4d \ln(2n+1) + 4 \ln \frac{4}{\delta}}{n}$, then for any $h \in V$, $h \in \mathcal{V}(h^*, r)$ and $V \subseteq \mathcal{V}(h^*, r)$. Furthermore, if $h \in \mathcal{V}(h^*, r)$, $h \in \mathcal{B}(h^*, r)$ in the realizable setting. Also, we know that $\forall r \in (0, 1), \Delta\mathcal{B}(h^*, r) \leq \theta r$. Therefore, with probability at least $1 - \delta$,

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

and

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

□

3 The Noisy Setting

In the noisy setting, the labels are corresponding to the prediction of target hypothesis h^* with noises. We will show that with high probability, the selective classifier produced by Algorithm 4 achieves same error rate as h^* when it makes prediction, and its coverage is near optimal.

Algorithm 4 Selective Classifier Strategy - Noisy

input n labelled data S, d, δ.

Output a selective classifier (h, g) s.t. $\mathrm{risk}(h, g) = \mathrm{risk}(h^*, g)$ with probability $1 - \delta$.

1: Set $\hat{h} = ERM(\mathcal{H}, S)$ so that $\hat{h}$ is any empirical risk minimizer from $\mathcal{H}$.
2: Set $G = \hat{\mathcal{V}}(\hat{h}, 4\sqrt{2\frac{d \ln(\frac{2ne}{d}) + \ln \frac{8}{\delta}}{n}})$.
3: Construct g s.t. $g(x) = 1$ if and only if $x \in \{\mathcal{X} \setminus \mathrm{DIS}(G)\}$.
4: Set $h = \hat{h}$.

Before we prove the error bound and coverage of Algorithm 4, let's introduce some new definitions:

- We would like to generalize the definition of risk using a loss fuction $\mathcal{L}(\mathcal{Y}, \mathcal{Y})$, then

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

- Excess loss class is defined as

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

- Given $0 \leq \beta \leq 1$ and $B \geq 1$, class $\mathcal{F}$ is said to be a (β, B)-Bernstein class with respect to $\mathcal{D}$, if every $f \in \mathcal{F}$ satisfies $\mathbb{E}[f^2] \leq B(\mathbb{E}[f])^\beta$.

Lemma 1. *If $\mathcal{F}$ is a (β, B)-Bernstein class with respect to $\mathcal{D}$, then for any $r > 0$,*

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

Proof. If $h \in \mathcal{V}(h^*, r)$, by definition of $\mathcal{V}$

$$\begin{aligned} \mathbb{E}[\mathbb{1}_{\{h(X)\neq Y\}}] &\leq \mathbb{E}[\mathbb{1}_{\{h^*(X)\neq Y\}}] + r \\ \mathbb{E}[\mathbb{1}_{\{h(X)\neq Y\}} - \mathbb{1}_{\{h^*(X)\neq Y\}}] &\leq r. \end{aligned}$$

Also

$$\begin{aligned} \mathbb{E}[\mathbb{1}_{\{h(X)\neq h^*(X)\}}] &= \mathbb{E}[|\mathbb{1}_{\{h(X)\neq Y\}} - \mathbb{1}_{\{h^*(X)\neq Y\}}|] \\ &= \mathbb{E}[(\mathbb{1}_{\{h(X)\neq Y\}} - \mathbb{1}_{\{h^*(X)\neq Y\}})^2] && (1) \\ &= \mathbb{E}[(\mathcal{L}(h(X), Y) - \mathcal{L}(h^*(X), Y))^2] && (2) \\ &= \mathbb{E}[f^2] \\ &\leq B(\mathbb{E}[f])^\beta \\ &= B(\mathbb{E}[\mathbb{1}_{\{h(X)\neq Y\}} - \mathbb{1}_{\{h^*(X)\neq Y\}}])^\beta \\ &\leq Br^\beta \end{aligned}$$

From (1) to (2) we assume using 0-1 loss function. □

Intuitively, we can imagine that $\mathcal{B}(h^*, Br^\beta)$ as having a bigger tolerance such that the ball includes $\mathcal{V}(h^*, r)$ as the 2D visualization Figure 1 shown below.

Theorem 5. *(Lemma 1 from [1]) For any $\delta > 0$, if $\mathcal{H}$ has VC dimension d, with probability at least $1 - \delta$*

$$\forall h \in \mathcal{H}, \text{err}_P(h) \leq \text{err}_S(h) + 2\sqrt{2\frac{d\ln(\frac{2ne}{d}) + \ln\frac{2}{\delta}}{n}}.$$

Similarly, under the same condition,

$$\forall h \in \mathcal{H}, \text{err}_S(h) \leq \text{err}_P(h) + 2\sqrt{2\frac{d\ln(\frac{2ne}{d}) + \ln\frac{2}{\delta}}{n}}.$$

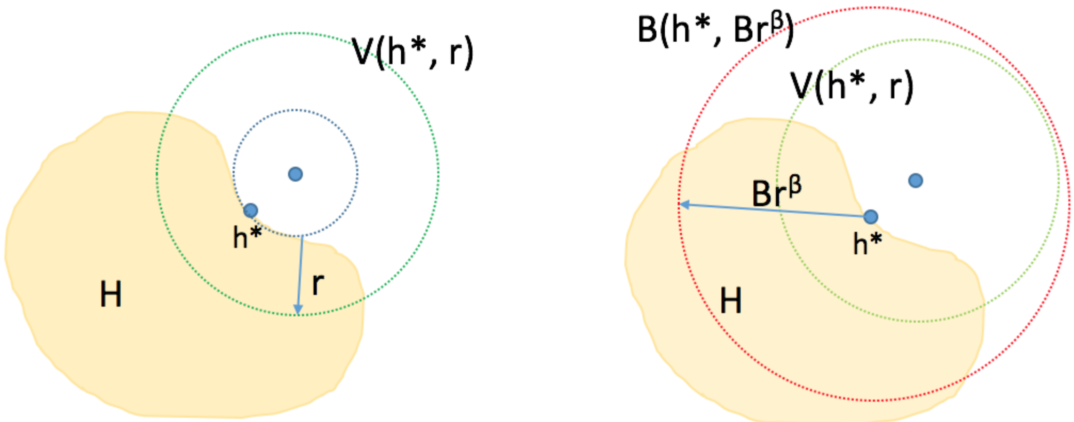

Figure 1: **2D visualization of intuition behind Lemma 1. Left:** The yellow region depicts the hypothesis class $\mathcal{H}$. The yellow region that lies inside the green dotted circle represents the hypotheses in $\mathcal{V}(h^*, r)$. **Right:** The yellow region inside the red dotted circle represents the hypotheses in $\mathcal{B}(h^*, Br^\beta)$. By Lemma 1 we show that the red dotted circle fully contains the green dotted circle. Also, because the difference in definition, the red dotted circle and green dotted circle have different centers.

Lemma 2. *For any $r > 0$, $0 < \delta < 1$, with probability at least $1 - \delta$,*

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

where

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

Proof. If $h \in \hat{\mathcal{V}}(\hat{h}, r)$, then

$$\mathrm{err}_S(h) \leq \mathrm{err}_S(\hat{h}) + r \tag{*}$$

$$\mathrm{err}_S(\hat{h}) \leq \mathrm{err}_S(h^*). \tag{**}$$

According to Theorem 5, with probability at least $1 - \delta$,

$$\mathrm{err}_P(h) \leq \mathrm{err}_S(h) + \sigma(n, \delta/2, d) \;\wedge\; \mathrm{err}_S(h^*) \leq \mathrm{err}_P(h^*) + \sigma(n, \delta/2, d).$$

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

$$\begin{aligned}
\mathrm{err}_P(h) + \mathrm{err}_S(h^*) &\leq \mathrm{err}_S(h) + \mathrm{err}_P(h^*) + 2\sigma(n, \delta/2, d) \\
\mathrm{err}_P(h) + \mathrm{err}_S(h^*) &\leq \mathrm{err}_S(\hat{h}) + r + \mathrm{err}_P(h^*) + 2\sigma(n, \delta/2, d) \quad \text{[applying (*)]} \\
\mathrm{err}_P(h) + \mathrm{err}_S(\hat{h}) &\leq \mathrm{err}_S(\hat{h}) + r + \mathrm{err}_P(h^*) + 2\sigma(n, \delta/2, d) \quad \text{[applying (**)]} \\
\mathrm{err}_P(h) &\leq \mathrm{err}_P(h^*) + 2\sigma(n, \delta/2, d) + r \\
\mathrm{err}_P(h) &\leq \mathcal{B}(h^*, 2\sigma(n, \delta/2, d) + r)
\end{aligned}$$

□

Similarly, we visualize the intuition below in Figure 2. Using Theorem 5, we construct a slightly bigger ball around h^* that includes the empirical error ball.

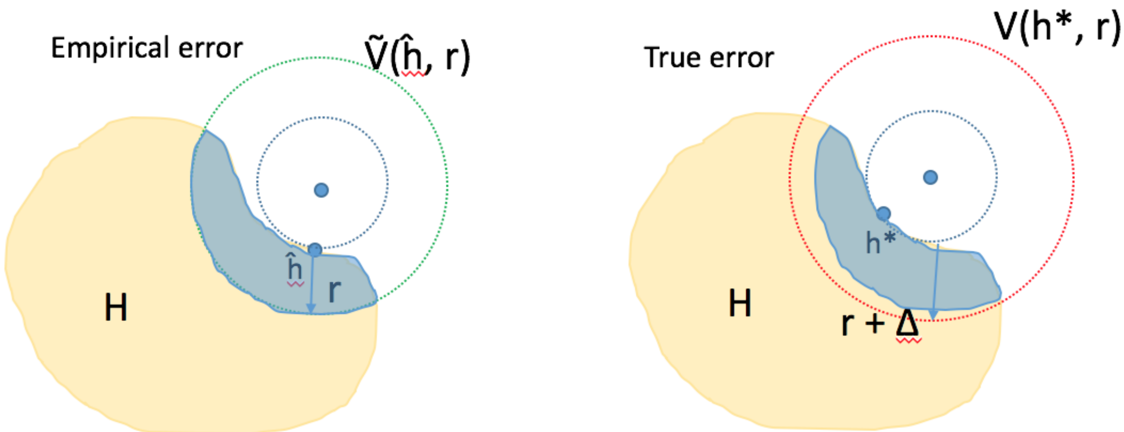

Figure 2: **2D visualization of intuition behind Lemma 2.** **Left:** The blue region, which is the intersection of the green dotted circle and $\mathcal{H}$, represents the hypotheses in $\hat{\mathcal{V}}(\hat{h}, r)$. **Right:** The intersection of $\mathcal{H}$ and the red dotted circle represents the hypotheses in $\mathcal{V}(h^*, r + \Delta)$. In Lemma 2, we show that when we set $\Delta = 2\sigma(n, \delta/2, d)$, with probability at least $1 - \delta$, hypotheses in $\hat{\mathcal{V}}(\hat{h}, r)$ are also in $\mathcal{V}(h^*, r + \Delta)$.

Lemma 3. *Assume that $\mathcal{H}$ has disagreement coefficient θ, and $\mathcal{F}$ is a (β, B)-Bernstein class with respect to $\mathcal{D}$, then for any $r > 0$ and $0 < \delta < 1$, with probability at least $1 - \delta$,*

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

Proof. Combine Lemma 1 and Lemma 2, we get with probability at least $1 - \delta$

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

Thus,

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

□

Theorem 6. *Assume that $\mathcal{H}$ has disagreement coefficient θ and that $\mathcal{F}$ is said to be a (β, B)-Bernstein class with respect to $\mathcal{D}$. Let (h, g) be selective classifier output by Algorithm 4 then for any $r > 0$ and $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\mathrm{cover}(h, g) \leq 1 - \theta B(4\sigma(n, \delta/4, d) + r)^\beta \;\wedge\; \mathrm{err}_P(h^*) = \mathrm{err}_P(h).$$

Proof. By Theorem 5, with probability at least $1-\delta/2$,

$$\mathrm{err}_S(h^*) \le \mathrm{err}_P(h^*) + \sigma(n, \delta/4, d) \ \wedge \ \mathrm{err}_P(\hat{h}) \le \mathrm{err}_S(\hat{h}) + \sigma(n, \delta/4, d).$$

Then, with probability at least $1-\delta/2$,

$$\mathrm{err}_S(h^*) \le \mathrm{err}_S(\hat{h}) + 2\sigma(n, \delta/4, d),$$

which implies $h^* \in \hat{\mathcal{V}}(\hat{h}, 2\sigma(n, \delta/4, d))$. So, for any $x \in \mathcal{X}$, if $g(x) = 1$, $\hat{h}(x) = h^*(x)$. Therefore, $\mathrm{err}_P(h^*) = \mathrm{err}_P(h)$.

Finally, using a union bound, we can applying Lemma 3, we get that with probability at least $1-\delta$,

$$\mathrm{cover}(\hat{h}, g) = 1 - \Delta G \ge 1 - \theta B(4\sigma(n, \delta/4, d))^\beta \ \wedge \ \mathrm{err}_P(h^*) = \mathrm{err}_P(h).$$

□

One of the concern with Algorithm 4 is that it is no so clear how to set β or B. We know that if the excess loss class is non-negative, then it is a $(1, B)$-Bernstein class for any B. However, it is unrealistic to assume non-negative excess loss class for arbitrary datasets.

Another area may be interesting for future research is to challenge the idea of using h^* for comparison in reliable learning under the noisy setting. Given that h^* still make mistakes, can we do better than h^* when we decide to make a decision given a reasonable coverage?

Bibliographic notes

Besides explicitly marked references, the confidence-rated predictor and CZ selective classifier are proposed by [2], and the selective classifier strategy in both realizable setting and noisy setting are proposed by [4].

References

[1] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

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

[3] R.L. Rivest and R. Sloan. A formal model of hierarchical concept-learning. *Inf. Comput.*, 114(1):88–114, October 1994.

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