E0 370 Statistical Learning Theory **Lecture 10 (Sep 15, 2011)**

Excess Error, Approximation Error, and Estimation Error

Lecturer: Shivani Agarwal *Scribe: Shivani Agarwal*

1 Introduction

So far, we have considered the finite sample setting: given a finite sample $S \in (\mathcal{X} \times \mathcal{Y})^m$ drawn according to D^m, we have seen how to obtain (high confidence) bounds on the generalization error of a function learned from S, usually in terms of some empirical quantity that measures the performance of the function on S.

Another question of interest concerns the behaviour of a learning algorithm in the infinite sample limit: as it receives more and more data, does the algorithm converge to an optimal prediction rule, i.e. does the generalization error of the learned function approach the optimal error? Recall that for a distribution D on $\mathcal{X} \times \mathcal{Y}$ and a loss $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$, the optimal error w.r.t. D and ℓ is the lowest possible error achievable by any function $h : \mathcal{X} \to \mathcal{Y}$:

$$\mathrm{er}_D^{\ell,*} = \inf_{h:\mathcal{X}\to\mathcal{Y}} \mathrm{er}_D^{\ell}[h] \,. \tag{1}$$

For the 0-1 loss, the optimal error is known as the *Bayes error*.

To formalize the above, for any function $h : \mathcal{X} \to \mathcal{Y}$, define its *excess error* (w.r.t. D and ℓ) as

$$\left(\mathrm{er}_D[h] - \mathrm{er}_D^{\ell,*}\right). \tag{2}$$

We would like to study the behaviour of the excess error of the function learned by an algorithm from a training sample $S \sim D^m$ as $m \to \infty$.

As we have seen, since minimizing the error over all possible functions in $\mathcal{Y}^{\mathcal{X}}$ can be difficult, most learning algorithms select a function from some fixed function class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. In such cases, we can only hope to achieve generalization error close to the lowest possible within the class; we refer to this as the optimal error within $\mathcal{H}$ (w.r.t. D and ℓ):

$$\mathrm{er}_D^{\ell}[\mathcal{H}] = \inf_{h\in\mathcal{H}} \mathrm{er}_D^{\ell}[h] \,. \tag{3}$$

It is then useful to view the excess error of functions $h \in \mathcal{H}$ as a sum of the following two terms:

$$\left(\mathrm{er}_D[h] - \mathrm{er}_D^{\ell,*}\right) \;=\; \left(\mathrm{er}_D^{\ell}[h] - \mathrm{er}_D^{\ell}[\mathcal{H}]\right) + \left(\mathrm{er}_D^{\ell}[\mathcal{H}] - \mathrm{er}_D^{\ell,*}\right). \tag{4}$$

The first term is called the *estimation error*, and measures how far h is from the optimal within $\mathcal{H}$. The second term, called the *approximation error*, measures how close one can get to the optimal error using functions in $\mathcal{H}$; this is an inherent property of the function class, and forms a lower bound on the excess error of any function learned from $\mathcal{H}$.

In the following we will focus on the estimation error, which is what a learning algorithm learning from a function class $\mathcal{H}$ can hope to minimize. We first give a couple of definitions.

2 Statistical Consistency

Definition. Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. Let $\mathcal{A} : \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$ be a learning algorithm that given a training sample $S \in \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m$, returns a function $h_S \in \mathcal{H}$. Let D be a probability distribution on $\mathcal{X} \times \mathcal{Y}$ and

$\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. We say $\mathcal{A}$ is *(statistically) consistent in $\mathcal{H}$ w.r.t. D and ℓ* if the estimation error of the function learned by $\mathcal{A}$ from $S \sim D^m$ converges in probability to zero, i.e. if for all $\epsilon > 0$,

$$\mathbf{P}_{S \sim D^m} \left(\mathrm{er}_D^\ell[h_S] - \mathrm{er}_D^\ell[\mathcal{H}] \geq \epsilon \right) \longrightarrow 0 \quad \text{as } m \to \infty \,.$$

If $\mathcal{A}$ is consistent in $\mathcal{H}$ w.r.t. ℓ for *all* distributions D on $\mathcal{X} \times \mathcal{Y}$, we say $\mathcal{A}$ is *universally consistent in $\mathcal{H}$ w.r.t. ℓ*.[1]

Definition. Let $\mathcal{A} : \cup_{m=1}^\infty (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{Y}^\mathcal{X}$ be a learning algorithm that given a training sample $S \in \cup_{m=1}^\infty (\mathcal{X} \times \mathcal{Y})^m$, returns a function $h_S : \mathcal{X} \to \mathcal{Y}$. Let D be a probability distribution on $\mathcal{X} \times \mathcal{Y}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. We say $\mathcal{A}$ is *Bayes consistent w.r.t. D and ℓ* if the excess error of the function learned by $\mathcal{A}$ from $S \sim D^m$ converges in probability to zero, i.e. if for all $\epsilon > 0$,

$$\mathbf{P}_{S \sim D^m} \left(\mathrm{er}_D^\ell[h_S] - \mathrm{er}_D^{\ell,*} \geq \epsilon \right) \longrightarrow 0 \quad \text{as } m \to \infty \,.$$

If $\mathcal{A}$ is Bayes consistent w.r.t. ℓ for *all* distributions D on $\mathcal{X} \times \mathcal{Y}$, we say $\mathcal{A}$ is *universally Bayes consistent w.r.t. ℓ*.[2]

One can also define analogous notions of *strong* consistency, which require almost sure convergence instead of convergence in probability.

3 Consistency of Empirical Risk Minimization in $\mathcal{H}$

Let $\mathcal{H} \subseteq \mathcal{Y}^\mathcal{X}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. Consider the empirical risk minimization (ERM) algorithm in $\mathcal{H}$, which given a training sample $S \in \cup_{m=1}^\infty (\mathcal{X} \times \mathcal{Y})^m$ returns[3]

$$h_S \;\in\; \arg\min_{h \in \mathcal{H}} \mathrm{er}_S^\ell[h] \,. \tag{5}$$

Then for any distribution D on $\mathcal{X} \times \mathcal{Y}$, we can write the estimation error of h_S as

$$\begin{aligned}
\mathrm{er}_D^\ell[h_S] - \mathrm{er}_D^\ell[\mathcal{H}] &= \left(\mathrm{er}_D^\ell[h_S] - \mathrm{er}_S^\ell[h_S] \right) + \left(\mathrm{er}_S^\ell[h_S] - \mathrm{er}_D^\ell[\mathcal{H}] \right) && (6) \\
&\leq \left(\mathrm{er}_D^\ell[h_S] - \mathrm{er}_S^\ell[h_S] \right) + \sup_{h \in \mathcal{H}} \left| \mathrm{er}_S^\ell[h] - \mathrm{er}_D^\ell[h] \right| && (7) \\
&\leq 2 \sup_{h \in \mathcal{H}} \left| \mathrm{er}_S^\ell[h] - \mathrm{er}_D^\ell[h] \right| . && (8)
\end{aligned}$$

Therefore, uniform convergence of empirical errors in $\mathcal{H}$ implies consistency of ERM in $\mathcal{H}$! In particular, for binary classification, we immediately have the following:

Theorem 3.1. Let $\mathcal{H} \subseteq \{\pm 1\}^\mathcal{X}$ and $\ell = \ell_{0\text{-}1}$. If $\mathrm{VCdim}(\mathcal{H}) = d < \infty$, then ERM in $\mathcal{H}$ is universally consistent in $\mathcal{H}$ w.r.t. $\ell_{0\text{-}1}$.

Proof. Let D be any probability distribution on $\mathcal{X} \times \{\pm 1\}$. Let $\epsilon > 0$. Then

$$\begin{aligned}
\mathbf{P}_{S \sim D^m} \left(\mathrm{er}_D^{0\text{-}1}[h_S] - \mathrm{er}_D^{0\text{-}1}[\mathcal{H}] \geq \epsilon \right) &\leq \mathbf{P}_{S \sim D^m} \left(\sup_{h \in \mathcal{H}} \left| \mathrm{er}_D^{0\text{-}1}[h_S] - \mathrm{er}_D^{0\text{-}1}[\mathcal{H}] \right| \geq \frac{\epsilon}{2} \right) \quad \text{(by Eq. (8))} && (9) \\
&\leq 4 \left(\frac{2em}{d} \right)^d e^{-m\epsilon^2/32} \quad \text{(by previous results)} && (10) \\
&\longrightarrow 0 \text{ as } m \to \infty \,. && (11)
\end{aligned}$$

□

[1] Note that one could also define a notion of consistency in terms of convergence in *expectation*, which would require that $\mathbf{E}_{S \sim D^m} [\mathrm{er}_D^\ell[h_S] - \mathrm{er}_D^\ell[\mathcal{H}]] \longrightarrow 0$ as $m \to \infty$. It is easy to show that a sequence of bounded, non-negative random variables converges in probability if and only if it converges in expectation (show this!), and therefore when the loss function ℓ is bounded, consistency in terms of convergence in probability is equivalent to consistency in terms of convergence in expectation.

[2] Note that the term 'Bayes' consistency is usually used to refer to convergence to the optimal error for binary classification with the 0-1 loss; we will use the term for any learning problem/loss function to distinguish it from consistency within $\mathcal{H}$.

[3] We assume for simplicity that the minimum is achieved in $\mathcal{H}$; the results we discuss continue to hold if h_S is selected to be any function in $\mathcal{H}$ whose empirical error is within an appropriately small precision of $\inf_{h \in \mathcal{H}} \mathrm{er}_S^\ell[h]$.

Several remarks are in order:

1. As we have noted before, for binary classification, ERM is typically not computationally efficient, except for some simple classes $\mathcal{H}$. We will later discuss consistency of algorithms that minimize a convex upper bound on $\ell_{0\text{-}1}$.

2. Note that for any $0 < \delta \leq 1$, we have with probability at least $1 - \delta$ (over $S \sim D^m$),

$$\mathrm{er}_D^{0\text{-}1}[h_S] - \mathrm{er}_D^{0\text{-}1}[\mathcal{H}] \;\leq\; c\sqrt{\frac{d \ln m + \ln(\frac{1}{\delta})}{m}}\,.$$

   As a function of the sample size m, this gives a rate of convergence of $O\Big(\sqrt{\frac{\ln m}{m}}\Big)$ for the estimation error. For distributions D for which $\mathrm{er}_D[\mathcal{H}] = 0$ (so that there is a 'target function' $t \in \mathcal{H}$ such that with probability 1, the true label y of any instance x under D is given by $t(x)$, i.e. $\mathbf{P}_{(x,y)\sim D}\big(y = t(x)\big) = 1$), one can actually show a faster rate of convergence of $O\left(\frac{\ln m}{m}\right)$. This follows from a better uniform convergence bound for such distributions (with an $e^{-cm\epsilon}$ term in the bound rather than $e^{-cm\epsilon^2}$); we probably will not show this for the general case, but will show this for finite $\mathcal{H}$ in a later lecture. A derivation for the general case can be found for example in [1].

3. It is important to note that the above result applies only to classes of finite VC-dimension. Since no such class can have zero approximation error for all distributions D, ERM in such a class cannot achieve (universal) Bayes consistency.

4. For classes $\mathcal{H}$ of finite VC-dimension, the above result actually establishes that ERM in $\mathcal{H}$ is *strongly* universally consistent in $\mathcal{H}$, by virtue of the Borel-Cantelli lemma (see [1]).

4 Consistency of Structural Risk Minimization in $\mathcal{H} = \cup_i \mathcal{H}_i$

Let $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \ldots$, where $\mathcal{H}_i \subseteq \mathcal{Y}^{\mathcal{X}}$. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. Given a training sample $S \in \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m$, the structural risk minimization (SRM) algorithm in $(\mathcal{H}_i)_{i=1}^{\infty}$ returns

$$h_S \;\in\; \arg\min_i \Big(\mathrm{er}_S^{\ell}[h_S^i] + \mathrm{penalty}(i, m)\Big)\,, \tag{12}$$

where $h_S^i \in \mathcal{H}_i$ is the function returned by ERM in $\mathcal{H}_i$, and $\mathrm{penalty}(i, m)$ is a penalty term that increases with the complexity of $\mathcal{H}_i$. Under certain conditions, one can show that SRM in $(\mathcal{H}_i)_{i=1}^{\infty}$ is consistent in $\mathcal{H} = \cup_{i=1}^{\infty} \mathcal{H}_i$; if in addition the sequence $(\mathcal{H}_i)_{i=1}^{\infty}$ is such that $\mathcal{H} = \cup_{i=1}^{\infty} \mathcal{H}_i$ has zero approximation error, then SRM in $(\mathcal{H}_i)_{i=1}^{\infty}$ can also be Bayes consistent. For example, for binary classification, we have the following result:

Theorem 4.1 (Lugosi and Zeger, 1996)**.** Let $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \ldots$, where $\mathcal{H}_i \subseteq \{\pm 1\}^{\mathcal{X}}$, $\mathrm{VCdim}(\mathcal{H}_i) = d_i < \infty \; \forall i$, and $d_i < d_{i+1} \; \forall i$. Let $\ell = \ell_{0\text{-}1}$. Then SRM with penalties given by

$$\mathrm{penalty}(i, m) \;=\; \sqrt{\frac{8 d_i \ln(2em) + i}{m}}$$

is universally consistent in $\mathcal{H} = \cup_{i=1}^{\infty} \mathcal{H}_i$ w.r.t. $\ell_{0\text{-}1}$.

Proof. Let D be any probability distribution on $\mathcal{X} \times \{\pm 1\}$. Let $\epsilon > 0$. We can write the estimation error of h_S as

$$\begin{aligned} \mathrm{er}_D^{0\text{-}1}[h_S] - \mathrm{er}_D^{0\text{-}1}[\mathcal{H}] \;\;=\;\; & \Big(\mathrm{er}_D^{0\text{-}1}[h_S] - \inf_i \Big(\mathrm{er}_S^{0\text{-}1}[h_S^i] + \mathrm{penalty}(i, m)\Big)\Big) + \\ & \Big(\inf_i \Big(\mathrm{er}_S^{0\text{-}1}[h_S^i] + \mathrm{penalty}(i, m)\Big) - \mathrm{er}_D^{0\text{-}1}[\mathcal{H}]\Big)\,. \end{aligned} \tag{13}$$

Therefore we have

$$\mathbf{P}_{S\sim D^m}\left(\mathrm{er}_D^{0\text{-}1}[h_S] - \mathrm{er}_D^{0\text{-}1}[\mathcal{H}] \geq \epsilon\right) \;\leq\; \mathbf{P}_{S\sim D^m}\left(\mathrm{er}_D^{0\text{-}1}[h_S] - \inf_i \left(\mathrm{er}_S^{0\text{-}1}[h_S^i] + \mathrm{penalty}(i,m)\right) \geq \frac{\epsilon}{2}\right) + \mathbf{P}_{S\sim D^m}\left(\inf_i \left(\mathrm{er}_S^{0\text{-}1}[h_S^i] + \mathrm{penalty}(i,m)\right) - \mathrm{er}_D^{0\text{-}1}[\mathcal{H}] \geq \frac{\epsilon}{2}\right). \quad (14)$$

We will bound each probability in turn. For the first probability, we have

$$\begin{aligned}
&\mathbf{P}_{S\sim D^m}\left(\mathrm{er}_D^{0\text{-}1}[h_S] - \inf_i \left(\mathrm{er}_S^{0\text{-}1}[h_S^i] + \mathrm{penalty}(i,m)\right) \geq \frac{\epsilon}{2}\right) && (15)\\
&\leq\; \mathbf{P}_{S\sim D^m}\left(\sup_i \left(\mathrm{er}_D^{0\text{-}1}[h_S^i] - \left(\mathrm{er}_S^{0\text{-}1}[h_S^i] + \mathrm{penalty}(i,m)\right)\right) \geq \frac{\epsilon}{2}\right) && (16)\\
&\leq\; \sum_{i=1}^{\infty} \mathbf{P}_{S\sim D^m}\left(\mathrm{er}_D^{0\text{-}1}[h_S^i] - \mathrm{er}_S^{0\text{-}1}[h_S^i] \geq \frac{\epsilon}{2} + \mathrm{penalty}(i,m)\right) \quad \text{(by union bound)} && (17)\\
&\leq\; \sum_{i=1}^{\infty} 4\left(\frac{2em}{d_i}\right)^{d_i} e^{-m(\frac{\epsilon}{2}+\mathrm{penalty}(i,m))^2/8} && (18)\\
&\leq\; \sum_{i=1}^{\infty} 4(2em)^{d_i} e^{-m\epsilon^2/32} e^{-m(\mathrm{penalty}(i,m))^2/8} && (19)\\
&=\; 4e^{-m\epsilon^2/32} \sum_{i=1}^{\infty} (2em)^{d_i} e^{-(8d_i\ln(2em)+i)/8} && (20)\\
&=\; 4e^{-m\epsilon^2/32} \sum_{i=1}^{\infty} e^{-i/8} && (21)\\
&=\; \left(\frac{4}{1-e^{-1/8}}\right) e^{-m\epsilon^2/32}. && (22)
\end{aligned}$$

For the second probability, let i^* be such that

$$\mathrm{er}_D^{0\text{-}1}[\mathcal{H}_{i^*}] \;\leq\; \mathrm{er}_D^{0\text{-}1}[\mathcal{H}] + \frac{\epsilon}{4}, \quad (23)$$

and let m^* be such that for all $m \geq m^*$,

$$\mathrm{penalty}(i^*, m) \;\leq\; \frac{\epsilon}{8}. \quad (24)$$

Then we have

$$\begin{aligned}
&\mathbf{P}_{S\sim D^m}\left(\inf_i \left(\mathrm{er}_S^{0\text{-}1}[h_S^i] + \mathrm{penalty}(i,m)\right) - \mathrm{er}_D^{0\text{-}1}[\mathcal{H}] \geq \frac{\epsilon}{2}\right) && (25)\\
&\leq\; \mathbf{P}_{S\sim D^m}\left(\inf_i \left(\mathrm{er}_S^{0\text{-}1}[h_S^i] + \mathrm{penalty}(i,m)\right) - \mathrm{er}_D^{0\text{-}1}[\mathcal{H}_{i^*}] \geq \frac{\epsilon}{4}\right) && (26)\\
&\leq\; \mathbf{P}_{S\sim D^m}\left(\mathrm{er}_S^{0\text{-}1}[h_S^{i^*}] + \mathrm{penalty}(i^*,m) - \mathrm{er}_D^{0\text{-}1}[\mathcal{H}_{i^*}] \geq \frac{\epsilon}{4}\right) && (27)\\
&\leq\; \mathbf{P}_{S\sim D^m}\left(\mathrm{er}_S^{0\text{-}1}[h_S^{i^*}] - \mathrm{er}_D^{0\text{-}1}[\mathcal{H}_{i^*}] \geq \frac{\epsilon}{8}\right), \quad \text{for } m \geq m^* && (28)\\
&\leq\; \mathbf{P}_{S\sim D^m}\left(\sup_{h\in\mathcal{H}_{i^*}} \left|\mathrm{er}_S^{0\text{-}1}[h] - \mathrm{er}_D^{0\text{-}1}[h]\right| \geq \frac{\epsilon}{8}\right) && (29)\\
&\leq\; 4\left(\frac{2em}{d_{i^*}}\right)^{d_{i^*}} e^{-m\epsilon^2/512}. && (30)
\end{aligned}$$

Thus we have

$$\begin{aligned}
\mathbf{P}_{S\sim D^m}\left(\mathrm{er}_D^{0\text{-}1}[h_S] - \mathrm{er}_D^{0\text{-}1}[\mathcal{H}] \geq \epsilon\right) \;&\leq\; \left(\frac{4}{1-e^{-1/8}}\right) e^{-m\epsilon^2/32} + 4\left(\frac{2em}{d_{i^*}}\right)^{d_{i^*}} e^{-m\epsilon^2/512}, \quad \text{for } m \geq m^* && (31)\\
&\longrightarrow 0 \text{ as } m\to\infty. && (32)
\end{aligned}$$

□

A couple of remarks:

1. As noted above, if the sequence $(\mathcal{H}_i)_{i=1}^{\infty}$ is such that $\inf_i \inf_{h \in \mathcal{H}_i} \mathrm{er}_D^{0\text{-}1}[h] = \mathrm{er}_D^{0\text{-}1,*}$ for all distributions D on $\mathcal{X} \times \{\pm 1\}$ (i.e. if the approximation error of $\mathcal{H} = \cup_{i=1}^{\infty} \mathcal{H}_i$ is zero for all D), then SRM in $(\mathcal{H}_i)_{i=1}^{\infty}$ as above is universally Bayes consistent w.r.t. $\ell_{0\text{-}1}$.
2. Again, except for the simplest problems, SRM (particularly for binary classification) is often not computationally feasible; however it is useful as a theoretical tool for understanding model selection techniques and Bayes consistency, and can also serve as a guide for the development of approximate algorithms.

5 Consistency and Learnability: Two Sides of the Same Coin

In the next few lectures we will turn to learnability, and then return to a more detailed discussion of statistical consistency. As we will see, the two notions are closely related, although they arose in different communities and tend to emphasize somewhat different aspects:

Statistical Consistency	**Learnability**
• Origins in statistics	• Origins in theoretical computer science
• Starts with learning algorithm; asks if it is statistically consistent	• Starts with function class $\mathcal{H}$; asks if there is a learning algorithm that is statistically consistent in $\mathcal{H}$ (with an additional requirement we will see next time)
• Both consistency within $\mathcal{H}$ and Bayes consistency of interest	• By definition, interest is in consistency w.r.t. $\mathcal{H}$
• Mostly distribution-free; also interested in 'low-noise' settings	• Often assume $\mathrm{er}_D^{\ell}[\mathcal{H}] = 0$ ('target function' setting); mostly distribution-free otherwise, but sometimes interested in specific distributions (such as the uniform distribution over the Boolean cube $\mathcal{X} = \{0,1\}^n$)
• Focus on convergence rates $\epsilon(m, \delta)$	• Focus on sample complexity $m(\epsilon, \delta)$ and computational complexity

6 Next Lecture

In the next lecture we will introduce the notion of learnability, and will give a few basic results and examples to illustrate the concept. The next few lectures after that will discuss more results and examples related to learnability, before we return to talk more about statistical consistency.

References

[1] Luc Devroye, Laszlo Gyorfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, 1996.