ECE 901
Lecture 6: Introduction to PAC Learning

R. Nowak

5/17/2009

1 Key Aspect of Learning Problems

The fundamental problem in learning from data is proper model selection. As we have seen in the previous lectures, a model that is too complex could overfit the training data (causing an estimation error) and a model that is too simple could be a bad approximation of the function that we are trying to estimate (causing an approximation error). The estimation error arises because of the fact that we do not know the true joint distribution of data in the input and output space. When using the empirical risk as a surrogate of the true risk we incur in some error, that will critically after the performance of our learning procedure. The approximation error measures how well the functions in the chosen model space can approximate the underlying relationship between the output space on the input space, and in general improves as the "size" of our model space increases.

In the preceding lectures, we looked at some solutions to deal with the overfitting problem. The basic approach followed was the *Method of Sieves*, in which the complexity of the model space was chosen as a function of the number of training samples. In particular, both the denoising and classification problems we looked at consider estimators based on partitions of the feature space. The size of the partition was an increasing function of the number of training samples. In this following lectures, we refine our learning methods to introduce model selection procedures that automatically adapt to the distribution of the training data, rather than basing the model class solely on the number of samples. This sort of adaptivity will play a major role in the design of more effective classifiers and denoising methods. The key to designing data-adaptive model selection procedures is obtaining useful upper bounds on the estimation error. To this end, we will introduce the idea of "Probably Approximately Correct" learning methods.

2 Recap: Method of Sieves

The method of Sieves underpinned our approaches in the denoising problem and in the histogram classification problem. Recall that the basic idea is to define a sequence of model spaces $\mathcal{F}_1, \mathcal{F}_2, \ldots$ of increasing complexity, and then given the training data $\{X_i, Y_i\}_{i=1}^n$ select a model according to

$$\hat{f}_n = \arg \min_{f \in \mathcal{F}_{m_n}} \hat{R}_n(f) \ ,$$

where m_n is an increasing sequence. The choice of the model space $\mathcal{F}_{m_n}$ (and hence the model complexity and structure) is determined completely by the sample size n, and does not depend on the (empirical) distribution of training data. This is a major limitation of the sieve method. In a nutshell, the method of sieves tells us to average the data in a certain way, e.g. over a partition of $\mathcal{X}$) based on the sample size, independent on the sample values themselves.

In general, learning basically comprises of two things:

1. Averaging data to reduce variability

2. Deciding *where (or how)* to average

Sieves basically force us to deal with (2) *a priori* (before we analyze the training data). This will lead to suboptimal classifiers and estimators, in general. Indeed deciding where/how to average is the really interesting and fundamental aspect of learning; once this is decided we have effectively solved the learing problem. There are at least two possibilities for breaking the rigidity of the method of sieves, as we shall see in the following section.

3 Data Adaptive Model Spaces

3.1 Structural Risk Minimization (SRM)

This uses a similar setting to the one used in the method of sieves, but we choose the model class according to the distribution of the data. The basic idea is to select the model space $\mathcal{F}_k$ from a large collection, based on the training data themselves. Again, let $\mathcal{F}_1$, $\mathcal{F}_2$, ... be a sequence of model spaces of increasing sizes/complexities.

Let

$$\hat{f}_{n,k} = \arg\min_{f \in \mathcal{F}_k} \hat{R}_n(f)$$

be the element of $\mathcal{F}_k$ that minimizes the empirical risk. This gives us a sequence of selected models $\hat{f}_{n,1}, \hat{f}_{n,2}, \ldots$ Also associate with each set $\mathcal{F}_k$ a value $C_{n,k} > 0$ that measures the complexity or "size" of the set $\mathcal{F}_k$. Typically, $C_{n,k}$ is monotonically increasing with k (since the sets are of increasing complexity) and decreasing with n (since we become more confident with more training data). More precisely, suppose that the $C_{n,k}$ chosen so that

$$P\left(\sup_{f \in \mathcal{F}_k} |\hat{R}_n(f) - R(f)| > C_{n,k}\right) < \delta \tag{1}$$

for some small $\delta > 0$. In words, for **any** model $f \in \mathcal{F}_k$ we have $R(f) \leq \hat{R}_n(f) + C_{n,k}$ with probability at least $1 - \delta$. This type of bound suffices to bound the estimation error (variance) of the model selection process of the form $R(f) \leq \hat{R}_n(f) + C_{n,k}$, and SRM selects the final model by minimizing this bound over all functions in $\bigcup_{k \geq 1} \mathcal{F}_k$. The selected model is given by $\hat{f}_{n,\hat{k}}$, where

$$\hat{k} = \arg\min_{k \geq 1} \left\{ \hat{R}_n(\hat{f}_{n,k}) + C_{n,k} \right\} .$$

A typical example could be the use of VC dimension to characterize the complexity of the collection of model spaces *i.e.,* $C_{n,k}$ is derived from a bound on the estimation error.

3.2 Complexity Regularization

Consider a very large class of candidate models $\mathcal{F}$. To each $f \in \mathcal{F}$ assign a complexity value $C_n(f)$. Assume that the complexity values chosen so that

$$P\left(\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| > C_n(f)\right) < \delta \tag{2}$$

This probability bound also implies an upper bound on the estimation error and complexity regularization is based on the criterion

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \left\{ \hat{R}_n(f) + C_n(f) \right\} \tag{3}$$

Complexity Regularization and SRM are very similar in spirit, and equivalent in certain instances. In both procedures the complexity and structure of the model is not fixed prior to examining the data; the data aid in the selection of the best complexity. In fact, the key difference compared to the method of sieves is that these techniques can allow the data to play an integral role in deciding where and how to average the data.

4 Probably Approximately Correct (PAC) learning

Probability bounds of the forms in (1) and (2) are the foundation for SRM and complexity regularization techniques. The simplest of these bounds are know as PAC bounds in the machine learning community.

4.1 Approximation and Estimation Errors

Recall the approximation/estimation error decomposition: Let $\mathcal{F}$ be a class of candidate models and $\hat{f}_n \in \mathcal{F}$.

$$R(\hat{f}_n) - R^* \;=\; \underbrace{R(\hat{f}_n) - \inf_{f\in\mathcal{F}} R(f)}_{\text{estimation Error}} \;+\; \underbrace{\inf_{f\in\mathcal{F}} R(f) - R^*}_{\text{approximation error}} \quad .$$

The approximation error depends both on the characteristics of f^* and our choice of model class $\mathcal{F}$. If no assumptions are made about f^* we have no control over this term. On the other hand the estimation error is generally quantifiable without making extra assumptions on the distribution P_{XY}, and depends on the complexity or size of $\mathcal{F}$.

Probability bounds of the forms in (1) and (2) guarantee that the empirical risk is uniformly close to the true risk, and using (1) and (2) it is possible to show that with high probability the selected model $\hat{f}_n$ satisfies

$$R(\hat{f}_n) - \inf_{f\in\mathcal{F}_k} R(f) \le C(n,k)$$

or

$$R(\hat{f}_n) - \inf_{f\in\mathcal{F}_k} R(f) \le C_n(f) \; .$$

4.2 The PAC Learning Model (Valiant '84)

The estimation error will be small if $R(\hat{f}_n)$ is close to $\inf_{f\in\mathcal{F}} R(f)$. PAC learning expresses this as follows. We want $\hat{f}_n$ to be a "probably approximately correct" (PAC) model from $\mathcal{F}$. Formally, we say that $\hat{f}_n$ is ε accurate with confidence $1-\delta$, or $(\varepsilon,\delta)-$PAC for short, if

$$P\left(R(\hat{f}_n) - \inf_{f\in\mathcal{F}} R(f) > \varepsilon\right) < \delta \; .$$

In other words $R(\hat{f}_n) - \inf_{f\in\mathcal{F}} R(f) \le \varepsilon$ with probability at least $1-\delta$. We will frequently abbreviate "with probability at least $1-\delta$" by "w.p.$\ge 1-\delta$".

4.3 A Simple PAC Bound

We will now construct a PAC bound for a very particular setting. The setting is quite restrictive, but nonetheless gives us a good introduction to the type of results we are seeking. Consider the classification setting, with $\mathcal{Y} = \{0,1\}$ and the 0/1-loss function. Let $\mathcal{F}$ consist of a finite number of models, and let $|\mathcal{F}|$ denote that number. Furthermore, assume that $\min_{f\in\mathcal{F}} R(f) = 0$. It is important to notice that this is a very strong assumption, implying the Bayes classifier makes no errors and also that there is a perfect classifier in our model class under consideration.

Example 1 *$\mathcal{F}$ = set of all histogram classifiers with M bins* $\implies |\mathcal{F}| = 2^M$

$$\min_{f\in\mathcal{F}} R(f) = 0 \implies \exists \text{ a classifier in } \mathcal{F} \text{ that has a zero probability of error}$$

Theorem 1 *In the setting above let $\hat{f}_n = \arg\min_{f\in\mathcal{F}} \widehat{R}_n(f)$, where $\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{f(X_i) \ne Y_i\}$. Then for every n and $\varepsilon > 0$,*

$$P\left(R(\hat{f}_n) > \varepsilon\right) \;\le\; |\mathcal{F}| e^{-n\varepsilon} \;\equiv\; \delta \; .$$

Proof: Recall that $R(f) = P(f(X) \neq Y)$. Note that since $\min_{f\in\mathcal{F}} R(f) = 0$, it follows that $\hat{R}_n(\hat{f}_n) = 0$. In fact, there may be several $f \in \mathcal{F}$ such that $\hat{R}_n(f) = 0$. Let $\mathcal{G} = \{f : \hat{R}_n(f) = 0\}$, denote the set of possibilities (clearly $\hat{f}_n \in \mathcal{G}$).

$$\begin{aligned}
P(R(\hat{f}_n) > \varepsilon) &\leq P\left(\bigcup_{f\in\mathcal{G}} \{R(f) > \varepsilon\}\right) \\
&= P\left(\bigcup_{f\in\mathcal{F}} \{R(f) > \varepsilon,\ \hat{R}_n(f) = 0\}\right) \\
&= P\left(\bigcup_{f\in\mathcal{F}:R(f)>\varepsilon} \{\hat{R}_n(f) = 0\}\right) \\
&\leq \sum_{f\in\mathcal{F}:R(f)>\varepsilon} P(\hat{R}_n(f) = 0) \\
&\leq \sum_{f\in\mathcal{F}:R(f)>\varepsilon} (1-\varepsilon)^n \\
&\leq |\mathcal{F}|.(1-\varepsilon)^n \\
&\leq |\mathcal{F}|e^{-n\varepsilon} \ .
\end{aligned}$$

The third inequality follows since $\hat{R}_n(f) = 0$ implies that $f(X_i) = Y_i$ for all $i \in 1, \ldots, n$, and so

$$\begin{aligned}
P(\hat{R}(f) = 0) &= P(\cap_{i=1}^n \{f(X_i) = Y_i\}) \\
&= \prod_{i=1}^n P(f(X_i) = Y_i) \\
&= (1 - R(f))^n \leq (1-\varepsilon)^n \ ,
\end{aligned}$$

where the second equality follows from the fact that the data are *independent* samples from P_{XY}.

For the last step of the proof we just need to recall that $1 - x \leq e^{-x}$. ■

Note that for n sufficiently large, $\delta = |\mathcal{F}|e^{-n\varepsilon}$ is arbitrarily small. To achieve a (ε, δ)-PAC bound for a desired $\varepsilon > 0$ and $\delta > 0$ we require at least $n = \frac{\log|\mathcal{F}| - \log\delta}{\varepsilon}$ training examples.

Corollary 1 *Assume that* $|\mathcal{F}| < \infty$ *and* $\min_{f\in\mathcal{F}} R(f) = 0$. *Then for every* n

$$E[R(\hat{f}_n)] \leq \frac{1 + \log|\mathcal{F}|}{n}$$

Proof: Recall that for any non-negative random variable Z with finite mean, $E[Z] = \int_0^\infty P(Z > t)dt$. This follows from an application of integration by parts. Now let $u > 0$ be an arbitrary number. Then

$$\begin{aligned}
E[R(\hat{f}_n)] &= \int_0^\infty P(R(\hat{f}_n) > t)dt \\
&= \int_0^u \underbrace{P(R(\hat{f}_n) > t)}_{\leq 1} dt + \int_u^\infty P(R(\hat{f}_n) > t)dt, \quad \text{for any } u > 0 \\
&\leq u + |\mathcal{F}| \int_u^\infty e^{-nt} dt \\
&= u + \frac{|\mathcal{F}|}{n} e^{-nu}
\end{aligned}$$

Minimizing with respect to u produces the smallest upper bound with $u = \frac{\log |\mathcal{F}|}{n}$ ■