E0 270 Machine Learning **Lecture 24 (Apr 4, 2013)**

A Glimpse into Statistical Learning Theory: Statistical Consistency of Binary Classification Algorithms Based on Risk Minimization

Lecturer: Shivani Agarwal

Outline

- Statistical consistency
- Consistency of empirical risk minimization algorithms
- Consistency of empirical surrogate risk minimization algorithms
- Conclusion and pointers

1 Statistical Consistency

Let us recall the setting of binary classification from the first lecture. There is an instance space $\mathcal{X}$; the label and prediction spaces are $\mathcal{Y} = \widehat{\mathcal{Y}} = \{\pm 1\}$. We are given a training sample $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \{\pm 1\})^m$, and the goal is to learn from these examples a classifier $h_S : \mathcal{X} \to \{\pm 1\}$ which given a new instance $x \in \mathcal{X}$, predicts its label via $\widehat{y} = h_S(x)$.

How do we evaluate the quality of a classifier? As we discussed, it is common to assume that there is an underlying probability distribution D on $\mathcal{X} \times \{\pm 1\}$ from which the training examples are generated iid (i.e. $S \sim D^m$), and from which future test examples will also be generated; the quality of a classifier $h : \mathcal{X} \to \{\pm 1\}$ is then evaluated in terms of its expected 0-1 misclassification error according to D. Specifically, we defined the 0-1 loss function $\ell_{\text{0-1}} : \{\pm 1\} \times \{\pm 1\} \to \mathbb{R}_+$ as $\ell_{\text{0-1}}(y, \widehat{y}) = \mathbf{1}(\widehat{y} \neq y)$; the 0-1 error of h w.r.t. D is then

$$\mathrm{er}_D^{\text{0-1}}[h] = \mathbf{E}_{(x,y)\sim D}\big[\mathbf{1}(h(x) \neq y)\big] .$$

The Bayes error for D is the lowest possible error that can be achieved by any classifier:

$$\mathrm{er}_D^{\text{0-1},*} = \inf_{h:\mathcal{X}\to\{\pm 1\}} \mathrm{er}_D^{\text{0-1}}[h] .$$

How should we evaluate the quality of a binary classification *algorithm*, which given a training sample S, returns a classifier h_S? We would obviously like the algorithm to produce a classification model with low 0-1 error, preferable close to the Bayes error. In other words, we would like the *0-1 regret* of h_S, or *excess 0-1 error* or *excess 0-1 risk* of h_S,

$$\left(\mathrm{er}_D^{\text{0-1}}[h_S] - \mathrm{er}_D^{\text{0-1},*}\right) ,$$

to be small. We may even like it to be close to zero. But can we hope to achieve this for *all* training samples S? Perhaps, if the training sample is small, i.e. the algorithm has only seen a small number of examples, it would not be reasonable to expect it to learn an optimal model, but we would certainly want the model learned to approach the optimal performance as the number of examples m becomes large. Moreover, since the examples are drawn randomly, there is always a small possibility that even for large m, one gets an unrepresentative sample and is therefore unable to learn a good model. However, for large m, we would like the algorithm to learn a model with regret close to zero *with high probability.* The notion of statistical consistency captures this requirement.

Definition (Statistical consistency of a binary classification algorithm)**.** A learning algorithm $\mathcal{A}$ which, when given a training sample $S \in \cup_{m=1}^{\infty}(\mathcal{X} \times \{\pm 1\})^m$ returns a classifier $h_S : \mathcal{X} \to \{\pm 1\}$, is said to be *statistically consistent (or Bayes consistent) w.r.t. D* if, when examples are drawn from D, the 0-1 regret of the classifier learned by the algorithm converges in probability to zero, i.e. if for all $\epsilon > 0$,

$$\mathbf{P}_{S \sim D^m}\Big(\mathrm{er}_D^{0\text{-}1}[h_S] - \mathrm{er}_D^{0\text{-}1,*} \geq \epsilon\Big) \to 0 \quad \text{as } m \to \infty.$$

An algorithm is said to be *universally statistically consistent (or universally Bayes consistent)* if it is statistically consistent w.r.t. D for *all* distributions D on $\mathcal{X} \times \{\pm 1\}$.

Many of the algorithms we have seen learn a classifier from a fixed class of functions $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ (such as the class of linear classifiers when $\mathcal{X} \subseteq \mathbb{R}^d$). In this case, the best we can hope for is to achieve the lowest possible error within the class $\mathcal{H}$. It is then of interest to decompose the regret of the learned classifier into two terms:

$$\big(\mathrm{er}_D^{0\text{-}1}[h_S] - \mathrm{er}_D^{0\text{-}1,*}\big) \;=\; \underbrace{\big(\mathrm{er}_D^{0\text{-}1}[h_S] - \inf_{h \in \mathcal{H}} \mathrm{er}_D^{0\text{-}1}[h]\big)}_{\text{estimation error of } h_S \text{ in } \mathcal{H}} + \underbrace{\big(\inf_{h \in \mathcal{H}} \mathrm{er}_D^{0\text{-}1}[h] - \mathrm{er}_D^{0\text{-}1,*}\big)}_{\text{approximation error of } \mathcal{H}}.$$

The approximation error is an inherent property of the function class $\mathcal{H}$, and forms a lower bound on the regret of any function h_S learned from $\mathcal{H}$. If a function class $\mathcal{H}$ has zero approximation error for all distributions D, we say it is a *universal function class*. If the estimation error w.r.t. a distribution D of the classifier returned by an algorithm in $\mathcal{H}$ converges in probability to zero, we say the algorithm is *statistically consistent w.r.t. D within $\mathcal{H}$*; if this holds for all distributions D, we say it is*universally statistically consistent within $\mathcal{H}$*. Statistical consistency within specific function classes $\mathcal{H}$ has formed the basis for the notion of *probably approximately correct (PAC) learnability*, which has been studied extensively in the theoretical computer science community. Below we first look at consistency within a fixed class $\mathcal{H}$ for empirical risk minimization (ERM) algorithms that minimize the 0-1 loss on the training sample, and discuss how this can in principle be extended to yield Bayes consistent (but computationally infeasible) algorithms via structural risk minimization in a universal class. We then consider the question of Bayes consistency of various popular classification algorithms that minimize an empirical risk based on a surrogate loss.

2 Consistency of Empirical Risk Minimization Algorithms

Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$. Any algorithm $\mathcal{A}$ which, when given a training sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$, returns a classifier $h_S \in \mathcal{H}$ satisfying

$$h_S \in \mathrm{argmin}_{h \in \mathcal{H}} \underbrace{\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(h(x_i) \neq y_i)}_{\widehat{\mathrm{er}}_S^{0\text{-}1}[h]},$$

is said to be an *empirical (0-1) risk minimization* (ERM) algorithm in $\mathcal{H}$.

Theorem 1. Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ with $\mathrm{VCdim}(\mathcal{H}) = d < \infty$. Then any ERM algorithm in $\mathcal{H}$ is universally statistically consistent within $\mathcal{H}$.

Proof. Let $\mathcal{A}$ be an ERM algorithm in $\mathcal{H}$ which when given a training sample S returns $h_S \in \mathcal{H}$. Let D be any probability distribution on $\mathcal{X} \times \{\pm 1\}$. Then we have,

$$\begin{aligned}
\big(\mathrm{er}_D^{0\text{-}1}[h_S] - \inf_{h \in \mathcal{H}} \mathrm{er}_D^{0\text{-}1}[h]\big) &= \big(\mathrm{er}_D^{0\text{-}1}[h_S] - \widehat{\mathrm{er}}_S^{0\text{-}1}[h_S]\big) + \big(\widehat{\mathrm{er}}_S^{0\text{-}1}[h_S] - \inf_{h \in \mathcal{H}} \mathrm{er}_D^{0\text{-}1}[h]\big) \\
&\leq \big(\mathrm{er}_D^{0\text{-}1}[h_S] - \widehat{\mathrm{er}}_S^{0\text{-}1}[h_S]\big) + \sup_{h \in \mathcal{H}} \Big|\widehat{\mathrm{er}}_S^{0\text{-}1}[h] - \mathrm{er}_D^{0\text{-}1}[h]\Big| \\
&\leq 2 \sup_{h \in \mathcal{H}} \Big|\widehat{\mathrm{er}}_S^{0\text{-}1}[h] - \mathrm{er}_D^{0\text{-}1}[h]\Big|.
\end{aligned}$$

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

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

□

Note that the above result applies only to function classes $\mathcal{H}$ of finite VC-dimension. Since no such function class can be a universal function class, this does not give Bayes consistency for ERM. However it is possible to construct structural risk minimization algorithms that perform complexity-penalized ERM in a hierarchy of function classes and that can be Bayes consistent. Specifically, let $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \ldots$, where $\mathcal{H}_i \subseteq \{\pm 1\}^{\mathcal{X}}$. Given a training sample S, a *structural risk minimization* (SRM) algorithm in $(\mathcal{H}_i)_{i=1}^\infty$ returns a function h_S satisfying $h_S = h_S^{i_S}$ with

$$
i_S \;\in\; \arg\min_i \Big(\widehat{\mathrm{er}}_S^{0\text{-}1}[h_S^i] + \mathrm{penalty}(i,m)\Big)\,,
$$

where $h_S^i \in \mathcal{H}_i$ is the function returned by an ERM algorithm 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, we have the following result:

Theorem 2 (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 $d_i < d_{i+1}\ \forall i$. Then any SRM algorithm in $(\mathcal{H}_i)_{i=1}^\infty$ with penalties given by*

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

is universally statistically consistent within $\mathcal{H} = \cup_{i=1}^\infty \mathcal{H}_i$.

It is possible to construct sequences $(\mathcal{H}_i)_{i=1}^\infty$ satisfying the conditions of the above theorem 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. such that the approximation error of $\mathcal{H} = \cup_{i=1}^\infty \mathcal{H}_i$ is zero for all D); in this case, SRM in $(\mathcal{H}_i)_{i=1}^\infty$ as above is universally Bayes consistent. However, as we already know, ERM in most non-trivial function classes is not computationally feasible; this means that SRM in such classes is also not computationally feasible. This motivates us to consider whether any of the binary classification algorithms we have seen earlier, which minimize other 'surrogate' losses on the training sample but are computationally efficient, can be shown to be Bayes consistent.

3 Consistency of Empirical Surrogate Risk Minimization Algorithms

Recall that several of the binary classification algorithms we have seen before – including logistic regression, least squares regression, SVMs, and AdaBoost – can be viewed as learning a real-valued function $f_S:\mathcal{X}\to\mathbb{R}$ by minimizing a suitable loss $\ell:\{\pm1\}\times\mathbb{R}\to\mathbb{R}_+$ on the training sample S (possibly with some regularization) over some suitable class of functions $\mathcal{F}\subseteq\mathbb{R}^{\mathcal{X}}$, and then returning a classifier $h_S:\mathcal{X}\to\{\pm1\}$ given by $h_S(x) = \mathrm{sign}(f_S(x))$.[1] In particular, these algorithms minimize the following losses respectively:

$$
\begin{aligned}
\ell_{\log}(y,f) &= \log_2(1+e^{-yf}) \\
\ell_{\mathrm{sq}}(y,f) &= (y-f)^2 \;=\; (1-yf)^2 \\
\ell_{\mathrm{hinge}}(y,f) &= (1-yf)_+ \;=\; \max(1-yf,0) \\
\ell_{\exp}(y,f) &= e^{-yf}\,.
\end{aligned}
$$

[1] Least squares regression was introduced for regression problems where the input labels can be arbitrary real-valued numbers, but can also be applied to classification problems where the labels happen to be ±1.

Note that these losses can all be written as $\ell(y,f) = \phi(yf)$ for some function $\phi : \mathbb{R}\to\mathbb{R}_+$; such losses are often called *margin-based* losses. The 0-1 loss in this case, $\ell_{\text{0-1}} : \{\pm 1\} \times \mathbb{R}\to\mathbb{R}_+$, can also be written as a margin-based loss:

$$\ell_{\text{0-1}}(y,f) \;=\; \mathbf{1}\big(\text{sign}(f) \neq y\big) \;=\; \mathbf{1}\big(yf \leq 0\big)\,.$$

It is easy to see that all the four losses above form a convex upper bound on the 0-1 loss.

For any margin-based loss given by $\phi : \mathbb{R}\to\mathbb{R}_+$, define the ϕ-error of a function $f : \mathcal{X}\to\mathbb{R}$ w.r.t. a probability distribution D on $\mathcal{X} \times \{\pm 1\}$ and the Bayes ϕ-error for D as follows:

$$\begin{aligned}
\text{er}_D^{\phi}[f] \;&=\; \mathbf{E}_{(x,y)\sim D}\big[\phi(yf(x))\big] \\
\text{er}_D^{\phi,*} \;&=\; \inf_{f:\mathcal{X}\to\mathbb{R}} \text{er}_D^{\phi}[f]\,.
\end{aligned}$$

As before, we will say an algorithm which given a training sample S returns $f_S : \mathcal{X}\to\mathbb{R}$ is Bayes ϕ-consistent if the ϕ-regret of the learned function, $\big(\text{er}_D^{\phi}[f_S] - \text{er}_D^{\phi,*}\big)$, converges in probability to zero, and will say an algorithm which returns $f_S \in \mathcal{F}$ is ϕ-consistent within $\mathcal{F}$ if the ϕ-estimation error in $\mathcal{F}$, $\big(\text{er}_D^{\phi}[f_S] - \inf_{f\in\mathcal{F}} \text{er}_D^{\phi}[f]\big)$, converges in probability to zero.

For each of the four losses above, it is possible to show that minimizing the empirical ϕ-error in a function class $\mathcal{F}$ (of limited capacity, or with suitable regularization) is ϕ-consistent within $\mathcal{F}$ (similarly to the results for 0-1 consistency of 0-1 ERM), and moreover, that suitably regularized minimization of the empirical ϕ-error in a universal class $\mathcal{F}$ is Bayes ϕ-consistent. In particular, this means that for all the above losses ϕ, it is possible to construct efficient algorithms (which perform suitably regularized ERM using ϕ in a universal class $\mathcal{F}$) such that for all $\epsilon > 0$,

$$\mathbf{P}_{S\sim D^m}\Big(\text{er}_D^{\phi}[f_S] - \text{er}_D^{\phi,*} \geq \epsilon\Big) \to 0 \quad \text{as } m \to \infty.$$

The question of interest to us is: does minimizing such a surrogate loss also yield 0-1 Bayes consistency? In other words, does the above also imply

$$\mathbf{P}_{S\sim D^m}\Big(\text{er}_D^{\text{0-1}}[f_S] - \text{er}_D^{\text{0-1},*} \geq \epsilon\Big) \to 0 \quad \text{as } m \to \infty?$$

In order to answer this question, we will need some notation. For $\eta \in [0,1]$ and $\alpha \in \mathbb{R}$, let

$$\begin{aligned}
L_\phi(\eta,\alpha) \;&=\; \eta\phi(\alpha) + (1-\eta)\phi(-\alpha) \\
H_\phi(\eta) \;&=\; \inf_{\alpha\in\mathbb{R}} L_\phi(\eta,\alpha) \\
H_\phi^-(\eta) \;&=\; \inf_{\alpha\in\mathbb{R}:\alpha(\eta-\frac{1}{2})<0} L_\phi(\eta,\alpha)\,.
\end{aligned}$$

Definition (Classification-calibrated)**.** Say $\phi : \mathbb{R}\to\mathbb{R}_+$ is *classification-calibrated* if for all $\eta \neq \frac{1}{2}$,

$$H_\phi(\eta) < H_\phi^-(\eta)\,.$$

Theorem 3 (Bartlett et al., 2006)**.** Let $\phi : \mathbb{R}\to\mathbb{R}_+$ be classification-calibrated. Then there exists a strictly increasing, continuous function $g_\phi : \mathbb{R}_+\to[0,1]$ with $g_\phi(0) = 0$ such that for all $f : \mathcal{X}\to\mathbb{R}$ and all probability distributions D on $\mathcal{X} \times \{\pm 1\}$,

$$\big(\text{er}_D^{\text{0-1}}[f] - \text{er}_D^{\text{0-1},*}\big) \;\leq\; g_\phi\big(\text{er}_D^{\phi}[f] - \text{er}_D^{\phi,*}\big)\,.$$

Such a result is often called a *surrogate regret bound*, since it bounds the regret of a function w.r.t. the target loss of interest (in this case the 0-1 loss) in terms of the regret w.r.t. a surrogate loss that might be optimized by an algorithm (in this case the loss given by ϕ). In particular, this immediately yields that for classification-calibrated ϕ, any algorithm that is Bayes ϕ-consistent is also Bayes 0-1 consistent! For convex ϕ, one has the following useful characterization of classification-calibratedness:

Theorem 4 (Bartlett et al., 2006)**.** Let $\phi : \mathbb{R}\to\mathbb{R}_+$ be convex. Then ϕ is classification-calibrated if and only if ϕ is differentiable at 0 and $\phi'(0) < 0$.

Using this characterization, it is easy to verify that the four loss functions above are all classification-calibrated (show this!). In particular, this means each of the above four algorithms – when implemented with appropriate regularization in a universal function class – can be made universally Bayes (0-1) consistent!

4 Conclusion and Pointers

There has been much research in the last few years on statistical consistency and surrogate regret bounds, including derivation of regret bounds for binary classification in terms of broader families of surrogate losses than the margin-based losses considered above, as well as regret bounds and consistency results for a variety of other supervised learning problems with more complex label and prediction spaces and more complex loss structures. In general, statistical consistency of machine learning algorithms is currently an active research area and many questions remain to be understood. More on this topic will be covered in the course *E0 370: Statistical Learning Theory* to be offered during Aug–Dec 2013.