Statistical Machine Learning

UoC Stats 37700, Winter quarter

Lecture 2: Introduction to statistical learning theory.

Goals of statistical learning theory

- SLT aims at studying the performance of machine learning methods from a statistical perspective.
- One central goal is to control the generalization error $\mathcal{E}(\widehat{f})$ of a certain learning method.
- Several points of interest:
  - Can we have a practical estimation of $\mathcal{E}(\widehat{f})$ (confidence interval)?
  - Does $\mathcal{E}(\widehat{f})$ converge to the Bayes risk L^* as the sample size grows to infinity (consistency)?
  - How fast does the above convergence occur (convergence rates)?
  - Are there theoretical limits to how fast one can learn a certain type of function (lower bounds/minimax rates)?
  - Can we justify/design correct model or parameter selection procedures?

- If possible, we would like to consider a broader framework, not limited to classification; for example try to estimate a generalization loss

$$\mathcal{E}(\ell, f) := \mathbb{E}_{X,Y}\left[\ell(\widehat{f}, X, Y)\right] ,$$

  where $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a loss function.
- Example: classification with a misclassification cost function.

Estimating the generalization error

- Let us start with estimating in practice the generalization error of a given method.
- Standard procedure: use a fresh test sample $((X'_1, Y'_1), \ldots, (X'_m, Y'_m))$:

$$\begin{tikzcd}
S = ((X_i, Y_i)) \arrow[d] & & T = ((X'_i, Y'_i)) \arrow[d] & \\
\widehat{f} \arrow[rr] & & \widehat{\mathcal{E}}_{test}(T, \widehat{f}) & = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{\widehat{f}(X'_i) \neq Y'_i\} .
\end{tikzcd}$$

- If we consider the train sample S fixed or “frozen”, then $\widehat{f}$ is a fixed function, and $\mathbb{1}\{\widehat{f}(X) \neq Y\}$ is just a Bernoulli variable whose mean we want to estimate based on m observations;
- Assuming the test sample is i.d.d. $\widehat{\mathcal{E}}_{test}(T, \widehat{f})$ is just a (rescaled) Binomial variable.

Estimating the parameter of a binomial variable

- First, we want to build an upper confidence interval for the mean of the binomial: should be pretty standard ...
- The lower tail function of a binomial of parameters (n, p) is

$$F(p,n,k) = \mathbb{P}\left[B(n,p) < k\right] = \sum_{i \le k-1} \binom{n}{i} p^i (1-p)^{n-i} \,;$$

- Let us denote

$$\mathcal{B}(t,n,\alpha) = \sup\left\{p : F(p,n,nt) \ge \alpha\right\} \,.$$

- Then $\mathcal{B}(\widehat{\mathcal{E}}(T,\widehat{f}), m, \alpha)$ is an upper confidence bound for p based on the observed test error.

Bounding the binomial tail function

- For various reasons, one would like more explicit expressions than the “raw” binomial tail inversion. We are therefore looking for more explicit upper bounds.
- In principle, we know that the asymptotical behavior of the binomial distribution is given by the Central Limit Theorem. However, this is only asymptotical and the convergence is not uniform over possible values of the parameter p.
  (Can you prove this?)
- We will heavily use bounds based on the Cramér-Chernoff method, using the Laplace transform:

  $$F_X(\lambda) = \mathbb{E}\left[\exp(\lambda X)\right] ;$$

  then by Markov’s inequality for any $\lambda \geq 0$

  $$\mathbb{P}\left[X \geq t\right] = \mathbb{P}\left[\exp(\lambda X) \leq \exp(\lambda t)\right] \leq F_X(\lambda)\exp(-\lambda t)\,,$$

  and optimize the bound in λ.

Note on the Cramér-Chernoff method

If we denote

$$\psi_X(\lambda) = \log F_X(\lambda) = \log \mathbb{E}\left[\exp(\lambda X)\right] ,$$

then for any $t \geq \mathbb{E}\left[X\right]$:

$$\mathbb{P}\left[X \geq t\right] \leq \exp\left(\inf_{\lambda}\left(\psi_X(\lambda) - \lambda t\right)\right) = \exp(-\psi_X^*(t)) ,$$

where ψ_X^* is the Legendre transform of ψ_X (Cramér transform of X).
(note : why is is ok to include $\lambda < 0$ in the inf?)
Therefore

$$\mathbb{P}\left[X \geq (\psi_x^*)^{-1}(u)\right] \leq \exp(-u) .$$

Chernoff's bound

- Chernoff's bound is the best bound given by Chernoff's method for a binomial variable:

$$\mathbb{E}\left[\exp(-\lambda B_p)\right] = \exp(n \log((1-p) + pe^{-\lambda}))$$

Chernoff's method gives

$$\mathbb{P}\left[B_p \leq nt\right] \leq \exp\left(n\left(\log((1-p) + pe^{-\lambda}) + \lambda t\right)\right)$$

optimizing in λ gives $\lambda = -\log\left(\frac{t}{(1-t)}\frac{(1-p)}{p}\right)$ (nonnegative iff $t \leq p$) and the bound

$$\begin{aligned}\mathbb{P}\left[B_p \leq nt\right] &\leq \left(\frac{1-p}{1-t}\right)^{n(1-t)}\left(\frac{\mathrm{p}}{t}\right)^{nt} \\ &= \exp -\left(nD(t,p)\right) ,\end{aligned}$$

where $D(t,p) = t\log\frac{t}{p} + (1-t)\log\frac{1-t}{1-p}$.

- The inequality: for any $t \leq p$,

$$\mathbb{P}\left[B_p \leq nt\right] \leq \exp - \left(nD(t,p)\right) \ ,$$

  can be restated as: for any $\delta \in (0,1)$,

$$\mathbb{P}\left[D(\widehat{p},p)\mathbb{1}\{\widehat{p} \leq p\} \geq \frac{\log \delta^{-1}}{n}\right] \leq \delta \,,$$

  where $\widehat{p} = B_p/n$ is the oberved empirical mean.
- Note: the inequality can be made two-sided (i.e. without the $\mathbb{1}\{\widehat{p} \leq p\}$) if we replace δ by 2δ on the RHS.

Bounded loss

- Consider a more general situation where we want to estimate an average loss $\mathbb{E}\left[\ell(\widehat{f}, X, Y)\right]$ via its empirical counterpart on a test set.
- Assume ℓ is a function bounded by B; we still want to apply the Cramér-Chernoff method...
- Then we have Hoeffding's lemma:
  Let X be a random variable with $X \in [a, b]$ a.s.; then

$$\psi_X(\lambda) = \log \mathbb{E}\left[\exp \lambda(X - \mathbb{E}[X])\right] \leq \frac{(b-a)^2\lambda^2}{8}\,.$$

Hoeffding's inequality

- From the previous lemma, we deduce via the Cramér-Chernoff method: if (X_i) are independent random variables, $X_i \in [a_i, b_i]$ a.s., then

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[X_i\right] \geq t\right] \leq \exp\left(-\frac{2nt^2}{\frac{1}{n}\sum_i (b_i - a_i)^2}\right) ;$$

  in particular, for a bounded loss function $|\ell()| \leq B$, it holds with probability $1 - \delta$ that

$$\mathbb{E}\left[\ell(f, X, Y)\right] \leq \frac{1}{m}\sum_{i=1}^{m} \ell(f, X_i', Y_i') + B\sqrt{\frac{2\log\delta^{-1}}{m}} .$$

Bennett's inequality

- We might be interested in taking the variance into account.
- Assume (X_i) are square integrable independent random variables, with $X_i \leq b$ a.s. for some $b \geq 0$ and let $v = \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right]$.
- Let

$$S = \sum_{i=1}^{N} (X_i - \mathbb{E}[X_i]) \ ,$$

then

$$\psi_S(\lambda) \leq \frac{v}{b^2}\phi(b\lambda) \, ,$$

where $\phi(\lambda) = e^{\lambda} - \lambda - 1$ is the log-Laplace transform of a recentered Poisson variable. Hence for $t \geq 0$,

$$\mathbb{P}\left[\frac{1}{n}S \geq t\right] \leq \exp\left(-\psi_S^*(nt)\right) \leq \exp\left(-\frac{v}{b^2}h\left(\frac{bnt}{v}\right)\right) \ ,$$

where $h(u) = (1+u)\log(1+u) - u$.

Bernstein's inequality

- We can weaken Bennett's inequality while making it more explicit using the inequality

  $$h(u) = (1+u)\log(1+u) - u \geq 9\left(1 + \frac{u}{3} - \sqrt{1 + \frac{2u}{3}}\right) = h_2(u)\,,$$

  which has a nice inverse $h_2^{-1}(t) = \sqrt{2t} + t/3$, hence:
- Assume (X_i) are square integrable independent random variables, with $X_i \leq b$ a.s. for some $b \geq 0$ and let $\sigma = \frac{v}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right]$; then with probability at least $1 - e^{-t}$ the following holds:

  $$\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[X_i\right] \leq \sqrt{\frac{2\sigma t}{n}} + \frac{bt}{3n}\,.$$

  (Bernstein's inequality)

- We can apply Bernstein's inequality to a sum of i.i.d. Bernoulli(p) variables: with probability $1 - e^{-t}$ the following holds:

  $$p \leq \widehat{p} + \sqrt{2p(1-p)\frac{t}{n}} + \frac{t}{3n}\,;$$

  then we can invert this inequality p to get (after some upper bounding)

  $$p \leq \widehat{p} + \sqrt{\widehat{p}\frac{2t}{n}} + \frac{3t}{n}\,.$$

- Compare this to what we obtain with Chernoff's bound directly, using the inequality $D(p,q) \geq (p-q)^2/2q$ for $q \geq p$.

- Let us now consider a situation where we want to choose between a finite number k of different learning methods: for example because of some varying parameter.

$$
\begin{array}{ccccc}
S = ((X_i, Y_i)) & & T = ((X_i', Y_i')) & & \\
\downarrow & & \downarrow & & \\
\widehat{f}_1 & \rightarrow & \widehat{\mathcal{E}}_{test}(T, \widehat{f}_1) & = & \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{\widehat{f}_1(X_i') \neq Y_i'\} \\
\widehat{f}_2 & \rightarrow & \widehat{\mathcal{E}}_{test}(T, \widehat{f}_2) & = & \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{\widehat{f}_2(X_i') \neq Y_i'\} \\
\vdots & & \vdots & & \vdots \\
\widehat{f}_k & \rightarrow & \widehat{\mathcal{E}}_{test}(T, \widehat{f}_k) & = & \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{\widehat{f}_k(X_i') \neq Y_i'\}\,.
\end{array}
$$

- How reliable are now the test estimates of the errors? We need to correct for multiple estimation, the simplest is via the union bound:

$$
\begin{aligned}
&\mathbb{P}\,[\text{All bounds hold simultaneously}] \\
&\qquad = 1 - \mathbb{P}\,[\text{At least one bound is violated}] \\
&\qquad \geq 1 - \sum_{1 \leq i \leq k} \mathbb{P}\,[\text{Bound } i \text{ is violated}]\,.
\end{aligned}
$$

- This leads to a Bonferroni-type correction: in each bound replace δ (the initial upper bound on the probability of error) by δ/k .
- Let us apply this to our initial setting and Hoeffding's bound: with probability $1-\delta$, it holds for all $1 \leq j \leq k$ simultaneously:

$$\mathbb{E}\left[\ell(\widehat{f}_j, X, Y)\right] \leq \frac{1}{m}\sum_{i=1}^{m} \ell(\widehat{f}_j, X_i', Y_i') + B\sqrt{\frac{2(\log k + \log \delta^{-1})}{m}}\,.$$

- Important: the dependence is only logarithmic in k due to the exponential concentration property.
- In particular, if we decide (based on the data) to finally use $\widehat{f} = \widehat{f}_{\widehat{k}}$, then the above confidence bound holds for this choice.

Comparison in average performance

- From Hoeffding's lemma, we also can also deduce results in expectation over T:

$$\mathbb{E}_T \left[\sup_{1\leq i\leq k} \left| \widehat{\mathcal{E}}_{test}(T, \ell, \widehat{f}_i) - \mathcal{E}(\ell, \widehat{f}_i) \right| \right] \leq \sqrt{\frac{\log 2k}{2m}} .$$

- In particular, assume that we select the estimator index $\widehat{k}$ having the least test error, then it holds that

$$\mathbb{E}_T \left[\mathcal{E}(\ell, \widehat{f}_{\widehat{k}}) \right] \leq \inf_{1\leq i\leq k} \mathcal{E}(\ell, \widehat{f}_i) + \sqrt{\frac{2 \log 2k}{m}}$$

Hold-out and cross-validation

- If no dedicated test sample exist, we can arbitrarily divide our train sample S between a sub-train sample $\widetilde{S}$ of size m and a test sample of size $n-m$, (then generally called validation sample) and proceed with the previous methodology.
- This is called the Hold-out method.
- Practical disadvantages: throwing away part of the data for the “learning” stage; poor stability.
- V-fold Cross validation is often a better choice: divide S into V subsamples $(S^{(i)})$; denote $\widehat{f}_k^{(-i)}$ the estimator obtained by training on all datapoints except S_i; then select the index

$$\widehat{k} = \operatorname*{Arg\,Min}_j \sum_{i=1}^{V} \widehat{\mathcal{E}}(\widehat{f}_j^{(-i)}, S^{(i)}) \,.$$

…and re-use the whole sample to estimate $\widehat{f}_{\widehat{k}}$.

Union bound with a prior

- Let us now focus on the learning stage. Assume we want to choose from a finite set of fixed functions $\mathcal{F} = \{f_1, \dots, f_k\}$. Basically, we can apply the same principle as earlier
- This is unsatifying when k is very large, *a fortiori* infinite.
- Let us assume $\mathcal{C}$ is discrete, and choose a discrete "prior" probability π distribution on $\mathcal{C}$. Denote $\mathcal{B}(f, S, \delta)$ your favorite bound for estimating the error of a single f.
- Then the following holds: with probability at least $1 - \delta$, for all $f \in \mathcal{F}$ simultaneously,

$$\mathcal{E}(f) \leq \mathcal{B}(f, S, \pi(f)\delta)\,.$$

"Occam's razor"

- Let us apply again with Hoeffding's inequality: with confidence $1-\delta$, for all $f \in \mathcal{F}$:

$$\mathbb{E}\left[\mathcal{E}(\ell, f, X, Y)\right] \leq \frac{1}{n}\sum_{i=1}^{n} \ell(f, X_i, Y_i) + B\sqrt{\frac{2(\log \pi(f)^{-1} + \log \delta^{-1})}{n}} .$$

- From information theory, $-\log \pi(f)$ can be interpreted as a "coding length" for the function f.
- Example: the "bit bound": if f can be coded in a computer with $b(f)$ bits, then take $\pi(f) = 2^{-b(f)}$.
- The above can be interpreted as an approximation/complexity tradeoff.
- Gives a non-trivial way to pick $\widehat{f} \in \mathcal{F}$.
- Still depends on the choice of the prior π!

Returning to decision trees

- Assume the set of questions $\mathcal{Q}$ is finite;
- Define a prior on trees with:
  - For the tree structure, a binary branching process of parameter ρ ;
  - For the internal nodes' questions, a uniform prior on $\mathcal{Q}$;
  - For the external nodes' labels, a uniform prior on $\mathcal{Y}$;
- then it holds with probability $1 - \delta$ that for any tree T,

$$\mathcal{E}(T) \leq \widehat{\mathcal{E}}(T) + \sqrt{\frac{2(\lambda |T| + \log \delta^{-1})}{n}} \,,$$

with $\lambda = \log |\mathcal{Q}| + \log |\mathcal{Y}| + \log \rho(1 - \rho)$.

What is complexity?

- In the previous analysis “complexity” is represented by the $-$ log-prior or “description length.
- If we take a uniform prior π: complexity becomes the log-cardinality of the pool of functions we are considering for possible output: measure of size.
- A single function or classifier is not more or less “complex”. Only a set of such functions is.
- But the prior (or description length) is arbitrary... Is this notion of complexity arbitrary?
- A prior is a choice of relative priorities (or preferences).
- However, the choice of the “prior” (in a general sense) is still paramount for the final performance (in particular for data representation).