Statistical Machine Learning

UoC Stats 37700, Winter quarter

Lecture 10: Learning theory IV: Randomized classifiers.

Going back to Occam's razor

- Remember Occam's razor principle: suppose $\widehat{f}_1, \widehat{f}_2, \ldots,$ is a countable family of functions and/or learning methods, such that we know a generalization error bound for each one taken individually, of the form: with probability $1-\delta$ over the draw of the sample S,
$$\mathcal{E}(\widehat{f}_k, \ell) \leq \mathcal{B}(\widehat{f}_k, S, \delta)\,.$$
- Then given a prior distribution π on $\{1, 2, \ldots\}$, it holds with probability $1-\delta$:
$$\forall i \geq 1\,,\; \mathcal{E}(\widehat{f}_k, \ell) \leq \mathcal{B}(\widehat{f}_k, \pi(k)\delta)\,.$$
- Despite its simplicity, this is a useful tool because it can apply "on top of" any other bound that we can have available for the single $\widehat{f}_i$'s: simple binomial tail bouns if the functions are fixed; VC or Rademacher bounds if the functions belong to a model of controlled complexity; etc.

- Remember also that Occam's razor readily implies the useful corollary: for any data-dependent choice $\widehat{k}(S)$ of a function among the family $\widehat{f}_1, \widehat{f}_2, \ldots$, with probability $1 - \delta$

$$\mathcal{E}(\widehat{f}_k, \ell) \leq \mathcal{B}(\widehat{f}_{k(S)}, \pi(\widehat{k}(S))\delta) .$$

- This formulation is actually equivalent to the previous formulation as a uniform bound.
- Occam's razor is a bound applying to any "rule" (or algorithm) for selecting an object from a countable class, when a probabilistic bound is known for each individual object.

How to generalize Occam's razor?

- It would be nice (!) to have a generalization of Occam's razor to continuous function classes.
- This is hopeless in general, unless there is some known structure over the function class (finite VC dimension, covering number entropy, control of Rademacher complexity etc.)
- However, even if there is no known structure, we can still obtain something interesting if we assume that the estimation process is randomized, i.e. that we choose the final $\widehat{f}$ from a fixed set $\mathcal{F}$ using some probability distribution Θ (that may depend on the observed data).

A dumb example

- Consider a “stupid” example where we suppose that we draw $\widehat{f}$ from $\mathcal{F}$ from a fixed distribution Θ, i.e. without looking at the data!
- Assume that for any fixed $f \in \mathcal{F}$ with probability $1 - \delta$ we have the known bound
$$\mathcal{E}(f, \ell) \leq \mathcal{B}(f, S, \delta)\,.$$
- Then, for any fixed ρ, with probability $1 - \delta$ over the draw of S and of $\widehat{f} \sim \Theta$ the same bound as above holds for $\widehat{f}$.

Setting

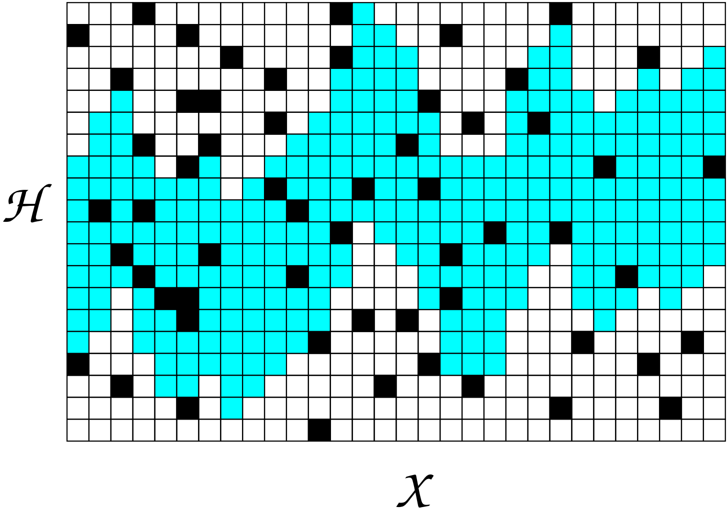
- Instead of considering a data-dependend choice $\widehat{f} \in \mathcal{F}$ of a function in $\mathcal{F}$, we consider the following ranfomized two-step procedure:
  - Choose a distribution $\Theta(S)$ on $\mathcal{F}$ from the data S (using some arbitrary "rule").
  - Pick at random an element $\widehat{f}$ from $\mathcal{F}$ by drawing according to $\Theta(S)$, and return $\widehat{f}$.
- Furthermore we will assume that $\Theta(S)$ admits a density $\theta(S, f)$ with respect to some fixed reference distribution μ on $\mathcal{F}$.

A slighlty less dumb example

- Assume that the learning procedure consists of a special case where a data-dependent subset $A(S) \subset \mathcal{F}$ is returned. The randomization step then picks a function at random from the distribution $\mu_{|A}$.
- Assume additionally that the refernce "volume" measure $\mu(A)$ is bounded from below by a constant $a > 0$.
- Then it holds with probability $1 - \delta$ over the draw of S and $\widehat{f} \sim \mu_{|A}$:

$$\mathcal{E}(\widehat{f}, \ell) \leq \mathcal{B}(\widehat{f}_k, S, a\delta)$$

A graphical representation of the set-output case



The general case

- Additional assumption: the single generalization bound $\mathcal{B}(f, S, \delta)$ is decreasing as a function of the level δ (this is a quite natural assumption).
- We consider two prior distributions: first, a prior Π on $\mathcal{F}$ with density π with respect to the reference μ.
- Secondly, let γ be a probability distribution function on $(0, +\infty)$. (a priori distribution on the inverse randomization density).
- Define $\beta(u) = \int_0^u x d\gamma(x)$.
- Occam's hammer bound: with probability $1 - \delta$ over the draw of S and $\widehat{f} \sim \Theta$

$$\mathcal{E}(\widehat{f}, \ell) \leq \mathcal{B}(\widehat{f}, S, \pi(\widehat{f})\beta(\theta(S, \widehat{f})^{-1})\delta)$$

The hammer and the razor

- A particular case: $\gamma = \delta_a \Rightarrow \beta(u) = a\mathbb{1}\{u \geq a\}$; we recover the result for the constant output subset case.
- A subcase of the above: $\mathcal{H}$ is discrete, μ the counting measure, $a = 1$, and the algorithm returns a single element $h_X \in \mathcal{H}$ $\longrightarrow$ we recover Occam's razor.

Application: randomized classifier choice

- As an example, consider some rule for picking randomly a classifier out of an arbitrary set $\mathcal{F}$. Take a "uniform" prior to simplify ($\pi \equiv 1$).
- For each single classifier, we can consider for example Hoeffing's bound.
- Consider the second prior $d\gamma = \alpha^{-1} x^{-\frac{1-\alpha}{\alpha}} dx$ on $[0,1]$ for some $\alpha > 0$; then $\beta(u) = (\alpha+1)^{-1} \min(x^{\frac{\alpha+1}{\alpha}}, 1)$ and we obtain that with probability at least $1 - \delta$:

$$\mathcal{E}(\widehat{f}) \leq \widehat{\mathcal{E}}(\widehat{f}, S) + \left(\frac{\log(\alpha+1)\delta^{-1} + (1+\alpha^{-1}) \log_+ \theta(\widehat{f}, S)}{2n} \right)^{\frac{1}{2}} .$$

Some conclusions

- Remember the latter inequality is valid for any choice of $\theta(\widehat{f}, S)$! We might want to choose θ to have the above bound as small as possible; a simple (approximate) solution is to choose uniformly from the set of classifiers having empirical error less that some (data-dependent) threshold $\widehat{t}$.
- One important point to note is that we can use an arbitrary randomization rule over an arbitrary space of classifiers, and that the role of "complexity" is then held by the log-randomization density (with respect to some reference measure).
- The tradeoff between empirical error and complexity is still present in this case since if we want to select with high probability classifiers with a lower empirical error, it entails choosing a high density for those classifiers, hence an increased complexity term.

Relation to False Discovery Rate in multiple testing

- Occam's razor, a.k.a. the union bound, is also used for multiple testing where it goes by the name of Bonferroni's correction.
- Assume $\mathcal{H}$ is a finite or countable set of null hypotheses about P.
- For any null hypothesis $h \in \mathcal{H}$ and level $\delta \in [0, 1]$, assume we know a test $T_h(\delta, X) \in \{0, 1\}$ with level (type I error) controlled by δ:

$$P \text{ satisfies null hypothesis } h \Rightarrow \mathbb{P}\left[T_h(\delta, X) = 1\right] \leq \delta.$$

- Let $\mathcal{H}_0 \subset \mathcal{H}$ the subset of null hypotheses actually satisfied by P, and $\mathcal{H}_1$ its complementary in $\mathcal{H}$.

Bonferroni's correction (with a prior)

- Let π be an "a priori" distribution on $\mathcal{H}$.
- Union bound with the a priori π: with probability at least $(1 - \pi(\mathcal{H}_0)\delta) \geq (1 - \delta)$, we have:

$$\forall h \in \mathcal{H}_0\,, \qquad T_h(\delta\pi(h), X) = 0\,.$$

- Thus, if we perform all tests T_h with a respective corrected level $\pi(h)\delta$, we control the probability of wrongly rejecting one or more hypotheses (Family-Wise Error , FWE).
- Referred to as *Bonferroni's correction* (generally with the uniform prior $\pi(h) = |\mathcal{H}|^{-1}$.)
- This is distribution-free bound – no assumption is made on the dependency structure of the family of tests.

The False Discovery Rate (FDR)

- Type I error control using FWE is too conservative (poor power).
- Benjamini and Hochberg (1995) propose a weaker form of type I error control, the False Discovery Rate:

  $$FDR = \mathbb{E}\left[\frac{V}{R}\mathbb{1}\{R > 0\}\right],$$

  where R = number of rejected hypotheses
  and V = number of wrongly rejected hypotheses.
- Define Θ_X the uniform distribution on the set of rejected hypotheses; then

  $$FDR = \mathbb{P}_{X\sim P;h\sim\Theta_X}\left[h \in \mathcal{H}_0\right].$$

Occam's hammer for FDR control

- $\mathcal{H}$ countable null hypotheses set; μ counting measure on $\mathcal{H}$. π and γ are arbitrary.
- Define the bad sets:

$$B(h,\delta) = \begin{cases} \{X : T_h(X,\delta) = 1\} & \text{si } h \in \mathcal{H}_0\,; \\ \emptyset & \text{otherwise.} \end{cases}$$

- Suppose the set of rejected null hypotheses A_X is such that

$$A_X \subset \{h \in \mathcal{H} : T_h(X, \delta\pi(h)\beta(|A_X|, X) = 1)\}\ ,$$

- Then

$$\mathbb{E}\left[\frac{|A_X \cap \mathcal{H}_0|}{|A_X|}\right] = \mathbb{P}_{\substack{X\sim P,\\ h\sim\mu_{|A_X}}}\left[X \in \mathcal{B}(h, \delta\pi(h)\beta(|A_X|))\right] \leq \pi(\mathcal{H}_0)\delta\,.$$

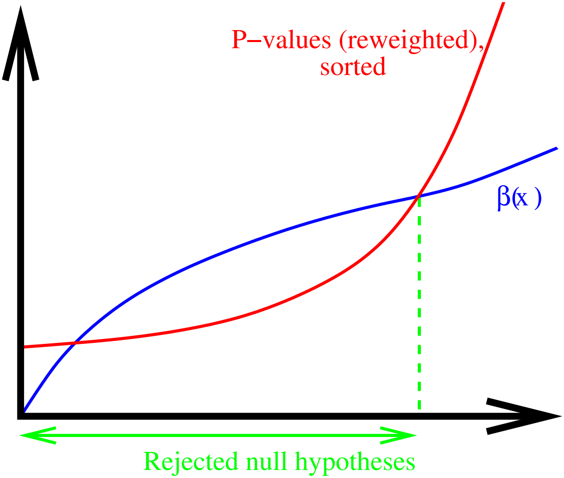
Step-up procedures for multiple testing

- We should preferrably choose the largest subset satisfying the previous condition:

$$A_X = \sup\{G \subset \mathcal{H} : \ \forall h \in G, \ T_h(X, \delta\pi(h)\beta(|G|)) = 1\} .$$

- $\Rightarrow$ "Step-up" procedure: denote $p_{(i)}$ the p-values reweighted by the prior π and sorted in increasing order; then we reject the $\widehat{k}$ hypotheses corresponding to the lowest eigenvalues, where

$$\widehat{k} = \sup\left\{k : p_{(k)} \leq \delta\beta(k)\right\} .$$



If $|\mathcal{H}| = M$, and one chooses $d\gamma = c\sum_{i=1}^{M} \frac{1}{i}\delta_{\frac{i}{M}}$, we get a linear "threshold" function $\beta(i) = ci$, with $c = \sum_{i=1}^{M} \frac{1}{i}$.

$\Rightarrow$ in this particular case we recover the distribution-free step-up procedure of Benjamini-Yekutieli (2001).

Some conclusions

- Occam's hammer makes sense for multiple testing for FDR control in a distribution-free point of view.
- Under an assumption of independence of the tests (or positive dependence), the original procedure of Benjamini-Hochberg (BH) is more powerful and uses other probabilistic tools.
- In the distribution-free point of view, Occam's hammer allows a more general approach and generalizes the Benjamini-Yekutieli procedure (BY) through the choice of γ and π. Also, theoretical possibility of considering continuous hypothesis space.