Problem sheet 1: The theory of learning

Analisi dei Dati – CdS Matematica – 2024/25

Exercise 1.1. Assume $\mathcal{Y} = \{0, 1\}$ (binary classification). Given a training set $s = (x_i, y_i)_{i=1,2,\dots,n}$, consider the predictor

$$h_s^\circ(x) = \begin{cases} y_i & \text{if there is } i \text{ s.t. } x_i = x, \\ 1 & \text{otherwise.} \end{cases}$$

- Compute the empirical risk of h°.
- Show that given a training set s, there is a polynomial p_s such that $h_s^\circ(x) = 1$ if and only if $p_s(x) > 0$.

Exercise 1.2. Let $\mathcal{H}$ be the set of functions from $\mathcal{X}$ to $\{0, 1\}$ (binary classifiers). Show that $\mathbb{E}[\mathcal{R}_S(h)] = \mathcal{R}_{\mathscr{D}}(h)$.

Exercise 1.3. Let $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$ and

$$\mathcal{H} = \{h_r : h_r = \mathbb{1}_{\{|x| \leqslant r\}} : r > 0\}.$$

Under realizability prove that $\mathcal{H}$ can be approximately learned. Moreover,

$$m_{\mathcal{H}}(\epsilon, \delta) \leqslant \frac{1}{\epsilon} \log \frac{1}{\delta}.$$

Exercise 1.4. Show that the Bayesian predictor is optimal.

Exercise 1.5. Let X be a non-negative random variable. Let $a \geqslant e$ and $b > 0$ be two numbers such that for very $x \geqslant 0$,

$$\mathbb{P}[X \geqslant x] \leqslant a\, e^{-\frac{x^2}{b^2}}.$$

Then $\mathbb{E}[X] \leqslant b(1 + \sqrt{\log a})$.

Exercise 1.6. Prove that the VC-dimension is monotone by inclusion.

Exercise 1.7. Given a set $\mathcal{X}$ and an integer $k \leqslant \#\mathcal{X}$, find the VC-dmension of

$$\mathcal{H}_k = \big\{h : \mathcal{X} \to \{0, 1\} : \#\{x : h(x) = 1\} = k\big\}$$

Exercise 1.8. Given a set $\mathcal{X}$ and an integer $k \leqslant \#\mathcal{X}$, find the VC-dimension of

$$\mathcal{H}^k = \big\{h : \mathcal{X} \to \{0, 1\} : \#\{x : h(x) = 1\} \leqslant k \text{ or } \#\{x : h(x) = 0\} \leqslant k\big\}.$$

Exercise 1.9. Let $\mathcal{H} = \{\mathbb{1}_{B_r(v)} : v \in \mathbb{R}^d, r > 0\}$, and let $\phi : \mathbb{R}^d \to \mathbb{R}^{d+1}$ be the map defined by $\phi(x) = (x, \|x\|^2)$. Prove that, if $x_1, x_2, \dots, x_m$ are shattered by $\mathcal{H}$, then $\phi(x_1), \phi(x_2), \dots, \phi(x_m)$ are shattered by hyperplanes in $\mathbb{R}^{d+1}$ (take $\operatorname{sgn}(0) = 1$). Deduce a limitation for $\dim_{VC} \mathcal{H}$.

Exercise 1.10. Let $\mathcal{H}_1$, $\mathcal{H}_2$ be hypothesis classes in $\{0, 1\}^{\mathcal{X}}$. Show that there are $c_1, c_2 > 0$ such that

$$\dim_{VC}(\mathcal{H}_1 \cup \mathcal{H}_2) \leqslant c_1 M \log 2M + c_2,$$

where $M = \max(\dim_{VC} \mathcal{H}_1, \dim_{VC} \mathcal{H}_2)$.

Exercise 1.11. Prove the following elementary properties of Rademacher complexity,

- if $\mathcal{H}_1 \subset \mathcal{H}_2$, then $\mathscr{C}_n(\mathcal{H}_1) \leqslant \mathscr{C}_n(\mathcal{H}_2)$,
- $\mathscr{C}_n(\mathcal{H}_1 + \mathcal{H}_2) = \mathscr{C}_n(\mathcal{H}_1) + \mathscr{C}_n(\mathcal{H}_2)$, where $\mathcal{H}_1 + \mathcal{H}_2 = \{h_1 + h_2 : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$,
- $\mathscr{C}_n(c\mathcal{H}) = |c|\mathscr{C}_n(\mathcal{H})$, where $c\mathcal{H} = \{ch : h \in \mathcal{H}\}$,
- if $\#\mathcal{H} < \infty$, then $\mathscr{C}_n(\operatorname{co}\mathcal{H}) = \mathscr{C}_n(\mathcal{H})$, where $\operatorname{co}\mathcal{H}$ is the convex hull of $\mathcal{H}$.

Exercise 1.12. Generate data as $Y = f(X) + \epsilon$, where f is a linear function and ϵ is "noise".

- Fit prediction methods with increasing complexity (k-nearest-neighbour and polynomial regression).
- Measure the empirical mean[1] of the training error in dependence of the complexity parameters.
- Generate a test set with the same distribution, and compute the empirical mean of the test error in dependence of the complexity parameter.
- Repeat the previous task with f polynomial (with different degrees) and non-polynomial.

Represent also the results through one or more (summarising and explanatory) pictures.

Exercise 1.13. Consider independent random variables $X_1, X_2, \dots, X_k$, uniformly distributed on $[0, 1]^p$. If

$$D = \min_{i \neq j} |X_i - X_j|,$$

prove that

$$(1 - k v_p u^p)^{\frac{1}{2}(k-1)} \leqslant \leqslant \mathbb{P}[D \geqslant u] \leqslant \leqslant (1 - 2^{-2p} v_p u^p)_+^{\frac{1}{2}k(k-1)},$$

[1] The mean is computed over several samples with the same distribution.

where v_p is the volume of the p-dimensional unit ball. Deduce (or prove independently) that $D \to \infty$ in probability as $p \to \infty$.

Exercise 1.14. Generate a dataset with binary output from a suitable (and non-trivial) probability distribution. Determine empirically (and theoretically, if possible with the chosen distribution) the Bayes decision boundary. Compare the decision boundary obtained with those given by some of the known elementary methods (k-nearest-neighbours, naive Bayes, etc.) on a given dataset of the same kind (if suitable, try different values of the hyperparameters). Finally obtain a graphical representation on the plane of principal components (for instance as in *An introduction to statistical learning*, page 154).

Exercise 1.15. Implement the experiment at pages 161 (section 4.5.2) of the book *An introduction to statistical learning*.

Exercise 1.16. Analyse the dataset `MNIST`[2] through k-nearest neighbour, linear and quadratic discriminant analysis, naive Bayes classifier and logistic regression. If necessary due to computational constraints, implement a binary classification on a single digit (chosen a-priori).

Exercise 1.17. Implement tangent distance on the dataset `MNIST`. Compare the result obtained with a standard classification with a *dataset enlargement*, namely by adding (slightly) rotated images to the dataset.

[2]The dataset is available, for instance, through the package dslabs.