Problem Set 1 ⋄⋄ MLGGM Spring 2022

Deadline: April 18th 2022

Probability Basics

We recall Markov's inequality: For a non-negative random variable Y and $a \geq 0$, we have

$$\mathbb{P}(Y \geq a) \leq \frac{\mathbb{E}[Y]}{a}. \tag{1}$$

In the proof of the No-free-lunch theorem, we used the following extension of the Markov's Inequality, which we ask you to verify.

Exercise A1 Let Z be a random variable that takes values in $[0,1]$. Assume that $\mathbb{E}[Z] = \mu$. Show that for any $a \in (0,1)$,

$$\mathbb{P}[Z > 1-a] \geq \frac{\mu - (1-a)}{a}.$$

Hint: Use Markov inequality with suitable substitutions.

We can also derive the Chebyshev's inequality from the Markov's Inequality.

Exercise A2 (Chebyshev's inequality) Let X be a random variable with mean μ and variance σ^2. Show that for any $t > 0$, we have

$$\mathbb{P}\left(|X-\mu| \geq t\right) \leq \frac{\sigma^2}{t^2}.$$

Hint: As in the previous exercise, use Markov inequality with suitable substitutions.

Empirical risk minimization

Exercise B1 (Shalev-Shwartz & Ben-David, 2014, Exercise 2.2) Let $\mathcal{H}$ be a class of binary classifiers over a domain $\mathcal{X}$. Let $\mathcal{D}$ be an unknown distribution over $\mathcal{X}$, and let S be a set of training inputs sampled from $\mathcal{D}$. Let $f : \mathcal{X} \to \{0,1\}$ be the ground-truth labeling function in $\mathcal{H}$. Fix some $h \in \mathcal{H}$. The training error of h over S is

$$L_{(S,f)}(h) := \frac{1}{|S|} \sum_{x \in S} \mathbb{1}(h(x) \neq f(x)),$$

and the generalization error of h over $\mathcal{D}$ is

$$L_{(D,f)}(h) := \mathbb{E}_{x\sim\mathcal{D}}\Big[\mathbb{1}(h(x) \neq f(x))\Big].$$

Show that the expected value of $L_{(S,f)}(h)$ over the random choice of S equals $L_{(D,f)}(h)$, namely,

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_S(h)\right] = L_{(\mathcal{D},f)}(h).$$

Exercise B2 (Shalev-Shwartz & Ben-David, 2014, Exercise 3.7) Consider a regression problem, where the input domain $\mathcal{X}$ and the target domain $\mathcal{Y}$ are both unit intervals on a real line, i.e., $\mathcal{X} = \mathcal{Y} = [0, 1]$. Assume that the distribution $\mathcal{D}$ over $\mathcal{X}$ is a uniform distribution. Further, suppose the ground-truth function $f : \mathcal{X} \to \mathcal{Y}$ is an identity function, that is, $f(x) = x$ for all $x \in \mathcal{X}$. Suppose that we want to approximate f from a hypothesis class $\mathcal{H}$ of constant functions with respect to the squared loss $l(x, y) = (x-y)^2$. Compute the approximation error of the hypothesis class $\mathcal{H}$. That is, compute

$$\inf_{h \in \mathcal{H}} R(h),$$

where $R(h) = \mathbb{E}_{x \sim \mathcal{D}}\left[\left(h(x) - f(x)\right)^2\right]$.

Exercise B3 (Shalev-Shwartz & Ben-David, 2014, Exercise 3.7) Given any probability distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$, the predicting function

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } \mathbb{P}[Y = 1 \mid X = x] \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

is call the *Bayes predictor*. Show that this predictor is optimal, in the sense that for every classifier $g : \mathcal{X} \to \{0, 1\}$, we have

$$L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g),$$

where $L_{\mathcal{D}}$ is the generalization error: $L_{\mathcal{D}}(g) := \mathbb{E}_{(X,Y) \sim \mathcal{D}}[\mathbb{1}(g(X) \neq Y)]$.

Groups

We recall the definition of a group.

Definition 0.1. Let G be a non-empty set and $\star$ be an operation on G. We say that G with the operation $\star$ is a group if it satisfies the following conditions:

- **Associativity**: $a \star (b \star c) = (a \star b) \star c$, for all $a, b, c \in G$;
- **Existence of Identity**: There exists an element $e \in G$ called an identity such that $a \star e = e \star a = a$, for all $a \in G$;
- **Existence of Inverse**: For each $a \in G$, there exists an element $b \in G$ such that $a \star b = b \star a = e$.

Exercise C1 Consider the following sets with their associated operations. Indicate which are groups and which are not. For objects that are not groups, specify a condition (associativity, existence of identity, and existence of inverse) that it violates.

1. Real numberswith addition $x \star y = x + y$.
2. Real numbers with multiplication $x \star y = x \cdot y$.
3. Nonzero real numbers with multiplication $x \star y = x \cdot y$.

4. Integers with addition $x \star y = x + y$.

5. Integers with subtraction $x \star y = x - y$.

6. Nonnegative integers with addition $x \star y = x + y$.

7. Positive real numbers with the operation $x \star y = 2xy$.

8. Positive integers with addition $x \star y = x + y$.

9. All subsets of a set X with the operation $A \star B = A \cup B$.

Exercise C2 This is a variation on an example we worked out in class. Consider an integral operator $f : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ with a kernel Φ, defined as

$$f(\mathbf{x})(u) = \int_{\mathbb{R}^d} \Phi(u, v)\mathbf{x}(v)dv.$$

Assume that f is rotation equivariant: for any d-dimensional rotation $\rho \in \mathrm{SO}(d)$,

$$f(T(\rho)\mathbf{x}) = T(\rho)f(\mathbf{x}).$$

where T acts as

$$\left(T(\rho)\mathbf{x}\right)(u) \coloneqq \mathbf{x}\left(\rho^{-1}u\right). \tag{2}$$

Derive conditions on the kernel Φ that guarantee this equivariance (assuming all regularity you need).

Hints

- consider using polar coordinates in $\mathbb{R}^d$;
- consider the case $d = 2$ first if general d seems abstract.

Exercise C3 (A challenge problem!) Repeat Exercise C2 with the equivariance being the group of special Euclidean motions of the plane, SE(2), which consists of both translation in $\mathbb{R}^2$ and rotations in SO(2). That is, equivariance should hold for both rotations and translations.

Exercise C4 (*A coding counterpart of* ***Exercise C2***) Check out the Jupyter notebook (`https://sada.dmi.unibas.ch/download/35/notebook1.ipynb`) we uploaded: the second example shows benefits of enforcing symmetry in learning when we have prior knowledge that symmetry is present. The example focuses on translation equivariance.
In this exercise we ask you to code something similar, but with the input and output signals being images on a disc,

$$\begin{aligned}\Omega &= \left\{x \in \mathbb{R}^2 \ : \ \|x\| \leq 1\right\}\\ \mathcal{X}(\Omega) &= L^2(\Omega)\end{aligned}$$

The exercise is open-ended in the sense that you're supposed to dig a bit to make it happen. But here's a rough blueprint:

- Discretize the unit disc (make sure to choose an appropriate discretization), and store the coordinates of the grid points;
- Define some rotation-equivariant linear operator acting on this discretization;
- Generate a training set with, say, random functions; add noise to the "labels" (in this case also images on the disc);
- "Learn" a generic linear operator using least-squares;
- Do the same but in a class of operators that are constrained to be rotation-invariant

Hint: In polar coordinates rotation invariance becomes identical to periodic translation invariance from the notebook.

Exercise C5 Let G be a group and $g \in G$. Show that $gG = G$, where

$$gG = \{gh \ : \ h \in G\}.$$

(We used this fact to prove that $S_G f$ is G-invariant.)

Exercise C6 Recall from the lecture the definition of a G-smoothing operator for a finite group G,

$$S_G f = \frac{1}{|G|} \sum_{g \in G} f \circ g \qquad \text{and} \qquad S_G \mathcal{F} = \{S_G f \ : \ f \in \mathcal{F}\}.$$

Let $\Omega = \{1, \dots, d\}$, $\mathcal{X}(\Omega)$ the set of signals on Ω, G the group of cyclic shifts of Ω and

$$\mathcal{F} = \{\text{poly}_k(x_1, \dots, x_d)\}$$

the set of degree-k polynomials in d variables. Describe $S_G \mathcal{F}$. What if G is the symmetric group on Ω (the set of all permutations of d elemenets)?

Graphs

Recall that a graph $G = (V, E)$ is a tuple of a vertex set $V = \{v_1, \dots, v_n\}$ and an edge set E. We assume each edge between two vertices v_i and v_j carries a non-negative weight $w_{ij} \geq 0$; if $w_{ij} = 0$ this means that the vertices v_i and v_j are not connected by an edge. The *weight matrix* $\boldsymbol{W} \in \mathbb{R}^{n \times n}$ of the graph is defined to be a matrix whose (i, j)-th entry is w_{ij}. Additionally, we assume that G is undirected, meaning that $w_{ij} = w_{ji}$. The degree of a vertex $v_i \in V$ is defined as

$$d_i = \sum_{j=1}^{n} w_{ij}.$$

The *degree matrix* $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ is defined as the diagonal matrix with the degrees $d_1, \dots, d_n$ on the diagonal.

Exercise D1 Draw any graph with 5 vertices and at least 10 edges. Compute its weight matrix $\boldsymbol{W}$ and its degree matrix $\boldsymbol{D}$.

Exercise D2 Many properties of a graph can be characterized by its graph Laplacian matrix, which is defined as

$$\boldsymbol{L} \coloneqq \boldsymbol{D} - \boldsymbol{W}.$$

Let $\boldsymbol{L}$ be the Laplacian matrix of a graph that has n vertices. Show that $\boldsymbol{L}$ satisfies the following properties:

1. For every vector $\mathbf{z} = [z_1, \ldots, z_n]^\top \in \mathbb{R}^n$ we have

$$\mathbf{z}^\top \boldsymbol{L} \mathbf{z} = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (z_i - z_j)^2.$$

2. $\boldsymbol{L}$ is symmetric and positive semi-definite.

References

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014. (Cited on pages 1 and 2.)