MATHEMATICS OF MACHINE LEARNING Part II
Example Sheet 2 (of 3) RDS/Lent 2024

In the following questions, where appropriate, suppose $(X_1, Y_1), \dots, (X_n, Y_n)$ are i.i.d. and consider loss ℓ to be misclassification loss, unless specified otherwise.

1. Show that if $|\mathcal{H}|$ is finite, then
$$\mathbb{E}R(\hat{h}) - R(h^*) \le \sqrt{\frac{\log|\mathcal{H}|}{2n}}.$$

2. N participants of a machine learning competition are given training data with which to develop classifiers. To decide the winner, the classifiers are applied to n new i.i.d. datapoints (the so-called test data). Give a value of n such that we can be sure with probability at least $1-\delta$ that the risk of the winning classifier is within ϵ of the minimum risk across the submitted classifiers.

3. Let $\mathcal{F}$ be the set of all polynomials of degree at most 2 on $\mathcal{X} = \mathbb{R}^p$. Show that $\mathrm{VC}(\mathcal{H}) \le \binom{p+2}{2}$, where $\mathcal{H} = \{\mathrm{sgn} \circ f : f \in \mathcal{F}\}$.

4. Given a collection of sets $\mathcal{A}$, let $\mathcal{H} = \{\mathbb{1}_A : A \in \mathcal{A}\}$.
   (a) Show that $\mathrm{VC}(\mathcal{H}) \le 6$ when $\mathcal{A}$ is the set of (filled) ellipses in $\mathbb{R}^2$.
   (b) Show that $\mathrm{VC}(\mathcal{H}) = 2p$ when $\mathcal{A} = \left\{ \prod_{j=1}^p [a_j, b_j] : a_1, b_1, \dots, a_p, b_p \in \mathbb{R} \right\}$.
   (c) Show that $\mathrm{VC}(\mathcal{H}) = \infty$ when $\mathcal{A}$ is the set of (filled) convex polygons in $\mathbb{R}^2$ and $\mathcal{H} = \{\mathbb{1}_A : A \in \mathcal{A}\}$.

5. Let $\mathcal{H} = \{x \mapsto \mathrm{sgn}(\beta^\top x) : \beta \in \mathbb{R}^p\}$. Show that $\mathrm{VC}(\mathcal{H}) = p$.

6. Let $\mathcal{H}_1, \mathcal{H}_2$ be classes of functions $f : \mathcal{X} \to \{a, b\}$ where $a \ne b$. Show that $s(\mathcal{H}_1 \cup \mathcal{H}_2, n) \le s(\mathcal{H}_1, n) + s(\mathcal{H}_2, n)$.

7. Let $W_1, \dots, W_n$ be i.i.d. $\mathbb{R}^p$-valued random vectors and let $F : \mathbb{R}^p \to [0,1]$ be given by
$$F(t_1, \dots, t_p) = \mathbb{P}(W_{11} \le t_1, \dots, W_{1p} \le t_p).$$
Define function $\hat{F} : \mathbb{R}^p \to [0,1]$ by
$$\hat{F}(t_1, \dots, t_p) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_A(W_i)$$
where $A = \prod_{j=1}^p (-\infty, t_j]$. Show that
$$\mathbb{E} \sup_{t \in \mathbb{R}^p} |F(t) - \hat{F}(t)| \le 2\sqrt{\frac{2\{p \log(n+1) + \log 2\}}{n}}.$$
[*Hint: Consider* $\mathcal{H} := \{\mathbb{1}_A : A = \prod_{j=1}^p (-\infty, t_j],\ t_j \in \mathbb{R}\}$, *and* $\mathcal{H}_- := \{-h : h \in \mathcal{H}\}$, *and use question 6.*]

8. Let $\varphi_1, \ldots, \varphi_d : \mathcal{X} \to \mathbb{R}$ be functions and let $\mathcal{H}$ be the class of all hypotheses $h : \mathcal{X} \to \{-1, 1\}$ of the form
$$h(x) = \operatorname{sgn}\left(\sum_{j \in A} \beta_j \varphi_j(x)\right)$$
where $A \subseteq \{1, \ldots, d\}$ with $|A| = s$ and $\beta_j \in \mathbb{R}$ for all j. Show that
$$\mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{2s}{n}} \sqrt{\log(n+1) + \log d}$$
where $\mathcal{F} = \{(x, y) \mapsto \ell(h(x), y) : h \in \mathcal{H}\}$ and ℓ is misclassification loss.

9. (a) Let $f, g : C \to \mathbb{R}$ be convex functions. Then if $a, b \geq 0$, $af + bg$ is a convex function.
   (b) Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function and fix $A \in \mathbb{R}^{d \times m}$ and $b \in \mathbb{R}^d$. Then $g : \mathbb{R}^m \to \mathbb{R}$ given by $g(x) = f(Ax - b)$ is a convex function.
   (c) Let $C_\alpha \subseteq \mathbb{R}^d$ be convex for all $\alpha \in I$ where I is some index set. Then $\cap_{\alpha \in I} C_\alpha$ is convex.
   (d) If $f : C \to \mathbb{R}$ is convex, then for each $M \in \mathbb{R}$, $D := \{x \in C : f(x) \leq M\}$ is convex.
   (e) Suppose $f_\alpha : C \to \mathbb{R}$ is convex for all $\alpha \in I$ where I is some index set, and define $g(x) := \sup_{\alpha \in I} f_\alpha(x)$. Then
      i. $D := \{x \in C : g(x) < \infty\}$ is convex and
      ii. function g restricted to D is convex.

10. Let $S \subseteq \mathbb{R}^d$ be a set of points.
   (a) Show that if D is the set of convex combinations of sets of points in S, then $D \supseteq \operatorname{conv} S$.
   (b) Let $S \subseteq C \subseteq \mathbb{R}^d$ be a convex set and let $f : C \to \mathbb{R}$ be convex. Show that $\sup_{x \in \operatorname{conv} S} f(x) = \sup_{x \in S} f(x)$. [*Hint: Use Qu. 9 (d).*]

11. Use 10 (b) to prove that for any $A \subseteq \mathbb{R}^n$, $\hat{\mathcal{R}}(A) = \hat{\mathcal{R}}(\operatorname{conv} A)$.

12. (Harder) Suppose function $\phi : \mathbb{R} \to [0, \infty)$ is convex and differentiable at 0 with $\phi'(0) < 0$. This question quantifies the fact that if for $f : \mathcal{X} \to \mathbb{R}$ the ϕ-risk is small, then the misclassification risk of $h := \operatorname{sgn} \circ f$ will be small.
   (a) Let $C_\eta(\alpha) := \eta\phi(\alpha) + (1 - \eta)\phi(-\alpha)$ and define
   $$H(\eta) := \inf_{\alpha \in \mathbb{R}} C_\eta(\alpha) \quad \text{for } \eta \in [0, 1].$$
   Show that
   $$\mathbb{E}H(\eta(X)) \geq \inf_{g \in \mathcal{G}} R_\phi(g)$$
   where $\eta(x) := \mathbb{P}(Y = 1 \,|\, X = x)$ is the regression function and $\mathcal{G}$ is the set of all functions $g : \mathcal{X} \to \mathbb{R}$ (you may ignore any measurability issues). (In fact equality holds in the display above.) [*Hint: Show that given $\epsilon > 0$ there exists $g \in \mathcal{G}$ such that $\mathbb{E}H(\eta(X)) + \epsilon \geq R_\phi(g)$.*]
   (b) Show that $\phi(0) = \inf_{\alpha : \alpha(2\eta - 1) \leq 0} C_\eta(\alpha)$.

(c) Define

$$\psi(\theta) := \phi(0) - H((1+\theta)/2) \quad \text{for } \theta \in [0,1].$$

Show that $\psi(0) = 0$ and ψ is convex. [*Hint: For the last part use Qu. 9 (e).*]

(d) Show that

$$\psi(|2\eta - 1|) = \phi(0) - H(\eta).$$

(e) Let h_0 be a Bayes classifier. Show that

$$\psi(R(h) - R(h_0)) \leq \mathbb{E}\{\mathbb{1}_{\{h(X) \neq h_0(X)\}} \psi(|2\eta(X) - 1|)\}.$$

[*Hint: Use Qu. 1 of Ex. Sheet 1.*]

(f) Show finally that

$$\psi(R(h) - R(h_0)) \leq R_\phi(f) - \inf_f R_\phi(f).$$

[*Hint: Argue that* $\mathbb{1}_{\{h(X) \neq h_0(X)\}} \inf_{\alpha:\alpha(2\eta(X)-1)\leq 0} C_{\eta(X)}(\alpha) \leq C_{\eta(X)}(f(X))$.]