MATHEMATICS OF MACHINE LEARNING Part II
Example Sheet 1 (of 3) RDS/Lent 2022

In the following questions, where appropriate, suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. and take values in $\mathcal{X} \times \mathcal{Y}$. We will take $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \{-1, 1\}$ and the loss ℓ will be misclassification loss, unless stated otherwise. For a hypothesis class $\mathcal{H}$, we set $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} R(h)$ (we will assume this minimiser exists).

1. Show that
$$R(h) - R(h_0) = \mathbb{E}\{\mathbb{1}_{\{h(X) \neq h_0(X)\}} |2\eta(X) - 1|\}$$
where
$$h_0(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ -1 & \text{otherwise.} \end{cases}$$

2. In each of the settings below, find a classifier that minimises the risk corresponding to the loss functions given.
   (a) Consider the weighted misclassification loss $\ell : \{-1, 1\}^2 \to \mathbb{R}$ given by $\ell(-1, -1) = \ell(1, 1) = 0$ and $\ell(-1, 1) = \alpha$, $\ell(1, -1) = \beta$ where $\alpha, \beta > 0$.
   (b) Suppose $\mathcal{Y} = \{1, \ldots, K\}$ and loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfies
$$\ell(y', y) = \begin{cases} 0 & \text{if } y = y' \\ 1 & \text{otherwise.} \end{cases}$$

3. Consider the regression setting with squared error loss and let $\mathcal{H} = \{x \mapsto \beta^T x : \beta \in \mathbb{R}^p\}$. Let $\Sigma_{XX} := \operatorname{Var}(X) \in \mathbb{R}^{p \times p}$ and $\Sigma_{XY} = \operatorname{Cov}(X, Y) \in \mathbb{R}^p$. Suppose Σ_{XX} is positive definite, $\mathbb{E}X = 0$ and $\mathbb{E}Y^2 < \infty$. Show that $h^* := \operatorname{argmin}_{h \in \mathcal{H}} R(h)$ is given by $h^*(x) = x^T \beta^*$ where $\beta^* = \Sigma_{XX}^{-1} \Sigma_{XY}$.

4. Suppose $|\mathcal{H}|$ is finite and there exists $h^* \in \mathcal{H}$ with $R(h^*) = 0$. Show that with probability at least $1 - \delta$, every empirical risk minimiser $\hat{h}$ satisfies
$$R(\hat{h}) \leq \frac{\log |\mathcal{H}| + \log(1/\delta)}{n}.$$
   [*Hint: Argue that* $\hat{R}(\hat{h}) = 0$ *and use that* $1 - \epsilon \leq e^{-\epsilon}$.]

5. Let random variable W be sub-Gaussian with parameter $\sigma > 0$.
   (a) Show that $\operatorname{Var}(W) \leq \sigma^2$. [*You may use the fact that* $\mathbb{E}(\sum_{r=3}^{\infty} \alpha^{r-2} W^r / r!) \to 0$ *as* $\alpha \to 0$. *If you took Probability & Measure, you can prove this.*]
   (b) Suppose $\sigma_* = \inf\{\sigma > 0 : W \text{ is sub-Gaussian with parameter } \sigma\}$. Is it true that $\operatorname{Var}(W) = \sigma_*^2$?

6. This question applies concentration ineqalities to study the problem of (potentially high-dimensional) covariance matrix estimation. Suppose $Z_i \overset{\text{i.i.d.}}{\sim} N_p(0, \Sigma)$ for $i = 1, \ldots, n$ where $\Sigma \in \mathbb{R}^{p \times p}$ is a covariance matrix with $\Sigma_{jj} = 1$ for $j = 1, \ldots, p$. The maximum likelihood estimate of Σ is $\hat{\Sigma} := \frac{1}{n} \sum_{i=1}^n Z_i Z_i^T$.

(a) Suppose V and W are mean-zero and jointly Gaussian with $\mathrm{Var}(V) = \mathrm{Var}(W) = 1$ and $\mathrm{Cov}(V, W) = \rho$. Show that

$$\mathbb{E}e^{\alpha VW} = [\{1 - \alpha(1+\rho)\}\{1 + \alpha(1-\rho)\}]^{-1/2}$$

for $\alpha \in (-1/2, 1/2)$. [*Hint: Express VW as a difference of two independent scaled χ^2_1 random variables and use the fact that the mgf of a χ^2_1 random variable is $1/\sqrt{1-2\alpha}$ for $\alpha < 1/2$.*]

(b) Using the fact that

$$e^{-\alpha\rho}[\{1 - \alpha(1+\rho)\}\{1 + \alpha(1-\rho)\}]^{-1/2} \leq e^{2\alpha^2}$$

whenever $|\alpha| < 1/4$ and $\rho \in [-1, 1]$, show that for fixed $j, k \in \{1, \ldots, p\}$ and $t \in (0, 1)$,

$$\mathbb{P}(|\hat{\Sigma}_{jk} - \Sigma_{jk}| \geq t) \leq 2e^{-nt^2/8}.$$

Conclude that with probability at least $1 - 2/p$,

$$\max_{j,k} |\hat{\Sigma}_{jk} - \Sigma_{jk}| \leq 5\sqrt{\frac{\log(p)}{n}}.$$

7. (Hoeffding's lemma.) Suppose random variable W takes values in $[a, b]$.

(a) Show that $\mathrm{Var}(W) \leq (b-a)^2/4$.

(b) Now assume $\mathbb{E}W = 0$ and define $\psi(\alpha) = \log \mathbb{E}e^{\alpha W}$. Show that $\psi'(0) = 0$. Show further that $\psi''(\alpha) = \mathrm{Var}_\alpha(W)$ where $\mathrm{Var}_\alpha(W) = \mathbb{E}_\alpha(W^2) - (\mathbb{E}_\alpha W)^2 = \mathbb{E}_\alpha\{(W - \mathbb{E}_\alpha W)^2\}$ and $\mathbb{E}_\alpha(g(W)) := \mathbb{E}(g(W)e^{\alpha W})/\mathbb{E}(e^{\alpha W})$ for any (measurable) $g : [a, b] \to \mathbb{R}$. Conclude that $\psi''(\alpha) \leq (b-a)^2/4$.

(c) Show that W is sub-Gaussian with parameter $(b-a)/2$. [*Hint: Consider a Taylor expansion of $\psi(\alpha)$ about 0.*]

8. Show that if $|\mathcal{H}|$ is finite, then

$$\mathbb{E}R(\hat{h}) - R(h^*) \leq \sqrt{\frac{\log|\mathcal{H}|}{2n}}.$$

9. 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.

10. (a) Suppose $\mathcal{H} := \{h_1, h_2, \ldots\}$ is countable. Let $\delta_1, \delta_2, \ldots$ be such that $\sum_{j=1}^\infty \delta_j =: \delta < 1$ and each $\delta_j > 0$. Show that with probability at least $1 - \delta$, the following holds for all $j = 1, 2, \ldots$:

$$R(h_j) \leq \hat{R}(h_j) + \sqrt{\frac{\log(1/\delta_j)}{2n}}.$$

(b) Suppose the input space $\mathcal{X} = [0, 1)^2$. Recall the histogram classifier from lectures: fix $m \in \mathbb{N}$ and partition $[0, 1)^2$ into m^2 disjoint squares $R_1^{(m)}, \ldots, R_{m^2}^{(m)} \subset [0, 1)^2$ of the form $[r/m, (r+1)/m) \times [s/m, (s+1)/m)$. Then for $r, s = 0, \ldots, m-1$, let

$$\bar{Y}_j^{(m)} := \mathrm{sgn}\Big(\sum_{i: X_i \in R_j^{(m)}} Y_i\Big)$$

and define

$$\hat{h}_m(x) := \sum_{j=1}^{m^2} \bar{Y}_j^{(m)} \mathbb{1}_{R_j^{(m)}}(x).$$

Show first that given $0 < \delta < 1$, for each fixed $m \in \mathbb{N}$, with probability at least $1 - \delta$, we have that

$$R(\hat{h}_m) \leq \hat{R}(\hat{h}_m) + \sqrt{\frac{m^2 \log 2 + \log(1/\delta)}{2n}}.$$

Next show that with probability at least $1 - \delta$, we have that for all $m \in \mathbb{N}$,

$$R(\hat{h}_m) \leq \hat{R}(\hat{h}_m) + \sqrt{\frac{m^2 \log 2 + \log\{m(m+1)\} + \log(1/\delta)}{2n}}.$$

Hint: Consider $\sum_{m=1}^{\infty} \frac{1}{m(m+1)}$.

11. Let $Z_1, \ldots, Z_n$ be an i.i.d. sequence of real-valued random variables with density f. A standard estimate of f is a so-called *kernel density estimate* given by

$$\hat{f}(z) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{z - Z_i}{h}\right),$$

where $K : \mathbb{R} \to [0, \infty)$ is a kernel function satisfying $\int_{-\infty}^{\infty} K(u)du = 1$, and $h > 0$ is known as a bandwidth parameter. A common approach to assessing the quality of $\hat{f}$ is using the L_1-norm $\|\hat{f} - f\|_1 := \int_{-\infty}^{\infty} |\hat{f}(u) - f(u)|du$. Prove that

$$\mathbb{P}\left(\|\hat{f} - f\|_1 \geq \mathbb{E}\|\hat{f} - f\|_1 + t\right) \leq e^{-nt^2/2}.$$

12. Let $\mathcal{F}_1, \ldots, \mathcal{F}_m$ be classes of functions $f : \mathcal{Z} \to \mathcal{D} \subseteq \mathbb{R}$.

    (a) Let $\mathcal{G} = \{f_1 + f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$. Show that $\mathcal{R}_n(\mathcal{G}) = \mathcal{R}_n(\mathcal{F}_1) + \mathcal{R}_n(\mathcal{F}_2)$.

    (b) Suppose $\mathcal{D} = [0, M]$. Show that $\mathcal{R}_n(\cup_{j=1}^m \mathcal{F}_j) \leq \max_{j=1,\ldots,m} \mathcal{R}_n(\mathcal{F}_j) + M\sqrt{2\log(m)/n}$. [*Hint: Argue that* $G(V_1, \ldots, V_n) := \sup_{f \in \mathcal{F}_1} \sum_{i=1}^n \varepsilon_i f(Z_i)/n$, *where* $V_i = (\varepsilon_i, Z_i) \in \{-1, 1\} \times \mathcal{Z}$, *satisfies a bounded differences property.*]