MATHEMATICS OF MACHINE LEARNING Part II
Example Sheet 1 (of 3) Lent 2025

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 it is specified that a regression setting is being considered, in which case the loss will typically be squared error. Assume that the computational complexity of inverting $M \in \mathbb{R}^{m\times m}$ is $O(m^3)$, and forming BC where $B \in \mathbb{R}^{a\times b}$ and $C \in \mathbb{R}^{b\times c}$ is $O(abc)$.

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

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. Let $\hat{h} = \hat{h}_D$ be a hypothesis trained on data $D = (X_i, Y_i)_{i=1}^n$ formed of iid copies of an independent random pair (X, Y). Define $\tilde{h}_{X_{1:n}}(x) := \mathbb{E}(\hat{h}_D(x) \mid X_{1:n})$.
   (a) Show that
$$\mathbb{E}[\{Y - \hat{h}_D(X)\}^2 \mid X = x] = \mathbb{E}\{\mathbb{E}(Y \mid X = x) - \tilde{h}_{X_{1:n}}(x)\}^2 + \mathbb{E}\{\hat{h}_D(x) - \tilde{h}_{X_{1:n}}(x)\}^2 + \mathrm{Var}(Y \mid X = x).$$
   (b) Show that considering squared error loss,
$$\mathbb{E}R(\hat{h}_D) - \mathbb{E}R(\tilde{h}_{X_{1:n}}) = \mathbb{E}\{\hat{h}_D(X) - \tilde{h}_{X_{1:n}}(X)\}^2.$$

4. Consider performing OLS regression using a set of d basis functions $(\varphi_1, \ldots, \varphi_d) := \varphi$ using data $(X_i, Y_i)_{i=1}^n$. Assume that the matrix $\Phi \in \mathbb{R}^{n\times d}$ with ith row $\varphi(X_i) \in \mathbb{R}^d$ has full column rank.
   (a) Show that the OLS coefficient vector $\hat{\beta} \in \mathbb{R}^d$ may be obtained in $O(nd^2)$ operations.
   (b) Show that the leave-one-out cross-validation score
$$\frac{1}{n}\sum_{i=1}^n \{Y_i - \varphi(X_i)^\top \hat{\beta}_{-i}\}^2$$
may be computed in $O(nd^2)$ operations. Here $\hat{\beta}_{-i} \in \mathbb{R}^d$ is the OLS coefficient vector when performing regression using a dataset with the ith point removed. [Use the matrix identity
$$(A - bb^\top)^{-1} = A^{-1} + \frac{A^{-1}bb^\top A^{-1}}{1 - b^\top A^{-1}b}$$

whenever $A \in \mathbb{R}^{p\times p}$ is invertible, $b \in \mathbb{R}^p$ and $b^\top A^{-1} b \neq 1$. Also assume $\varphi(X_i)^\top(\Phi^\top\Phi)^{-1}\varphi(X_i) < 1$, which holds provided each matrix formed of Φ with a row removed has full column rank.] [*Hint: Consider first computing* $(\Phi^\top\Phi)^{-1}\varphi(X_i) \in \mathbb{R}^d$ *for all* $i = 1, \ldots, n$.]

5. Consider a regression setting as in the previous question with $\Phi \in \mathbb{R}^{n\times d}$ and φ defined as above. For $\lambda \geq 0$, consider $\hat{h}_\lambda$ given by $\hat{h}_\lambda(x) = \varphi(x)^\top \hat{\beta}_\lambda$ with

$$\hat{\beta}_\lambda := \operatorname*{argmin}_{\beta\in\mathbb{R}^d} \{\|Y_{1:n} - \Phi\beta\|_2^2 + \lambda\|\beta\|_2^2\}.$$

   (a) Show that $\hat{\beta}_\lambda = (\Phi^\top\Phi + \lambda I)^{-1}\Phi^\top Y_{1:n}$.

   (b) Suppose $\mathrm{Var}(Y_1 \mid X_1 = x) > 0$ is constant in x and $\varphi(x)$ is not the zero vector. Show that for all x, $\lambda \mapsto \mathrm{Var}(\hat{h}_\lambda(x) \mid X_{1:n})$ is strictly decreasing.

6. In this question we investigate an alternative splitting criterion for a regression tree, based on maximising a likelihood assuming that the Y_i have a Poisson distribution conditional on X_i. Specifically, consider the first split and where $p = 1$ with $X_1 < \cdots < X_n$. Show that

$$\max_{\gamma_L,\gamma_R} \prod_{i\leq m}(\gamma_L^{Y_i} e^{-\gamma_L}) \times \prod_{i>m}(\gamma_R^{Y_i} e^{-\gamma_R})$$

   may be maximised over m with $O(n)$ computational cost.

7. The piecewise constant function produced by a regression tree may not always approximate the underlying true regression function well. Here we imagine we have an additional univariate predictor $T_1, \ldots, T_n \in \mathbb{R}$ which we permit to contribute to the fit in a linear fashion. Specifically, consider ERM with squared error loss over class

$$\mathcal{H} := \left\{ (t,x) \mapsto t\beta + \sum_{j=1}^J \gamma_j \mathbb{1}_{R_j}(x) \ : \beta \in \mathbb{R},\ \gamma \in \mathbb{R}^J \right\};$$

   here the R_j are fixed (for simplicity, unlike in the case of regression trees) and partition $\mathbb{R}^p$ and moreover all $I_j := \{i : X_i \in R_j\}$ are non-empty and have been pre-computed. Assume that $T_{1:n} \in \mathbb{R}^n$ is not in the span of $\{(\mathbb{1}_{R_j}(X_i))_{i=1}^n : j = 1, \ldots, J\}$. Show that the ERM may be computed in $O(n)$ time. [*Hint: Use the matrix identity that for* $M \in \mathbb{R}^{p\times p}$, $b \in \mathbb{R}^p$ *and* $a \in \mathbb{R}$,

$$\begin{pmatrix} a & b^\top \\ b & M \end{pmatrix}^{-1} = \begin{pmatrix} s^{-1} & -s^{-1}b^\top M^{-1} \\ -s^{-1}M^{-1}b & M^{-1} + s^{-1}M^{-1}bb^\top M^{-1} \end{pmatrix},$$

   where $s := a - b^\top M^{-1} b > 0$ *provided the matrix on the left is indeed invertible.*]

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

9. 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* $1 - \epsilon \leq e^{-\epsilon}$.]

10. Let random variable W be sub-Gaussian with parameter $\sigma > 0$.

   (a) Show that $\mathrm{Var}(W) \leq \sigma^2$. [*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 may prove this.*]

   (b) Suppose $\sigma_* = \inf\{\sigma > 0 : W \text{ is sub-Gaussian with parameter } \sigma\}$. Is it true that $\mathrm{Var}(W) = \sigma_*^2$?

11. This question applies concentration inequalities 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 $\hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^\top$.

   (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

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

   for $\alpha \in (-1/2, 1/2)$. [*Hint: Express* VW *as a difference of two independent scaled* χ_1^2 *random variables and use the fact that the mgf of a* χ_1^2 *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 $1 - 2/p$,

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