Problem 0. *Error bound for binary classification*

Part I: Hoeffding inequalities.

Let Y be a zero mean random variable, such that $Y \in [a,\, b]$ almost surely.

(i) Using convex properties of the exponential on the interval $[a,\, b]$, show that for $s > 0$,

$$\mathbf{E}\, e^{sY} \leq \left\{1 - \mu + \mu\, e^{s(b-a)}\right\} e^{-s\mu(b-a)} = e^{\varphi(u)}\,,$$

where $\mu = -a/(b-a)$, and $\varphi(u) = -\mu u + \log(1 - \mu + \mu e^u)$.

(ii) Using a second order Taylor expansion of φ around 0, show that

$$\varphi(u) \leq \frac{s^2(b-a)^2}{8}\,,$$

and conclude that $\mathbf{E}\, e^{sY} \leq \exp\left(\frac{s^2(b-a)^2}{8}\right)$.

(iii) Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed bounded random variables, such that $X_i \in [a_i,\, b_i]$ with probability 1, and put $S_n = \sum_{i=1}^n X_i$. Making use of Markov inequality, show that for any $s, \epsilon > 0$,

$$\mathbf{P}\left(\sum_{i=1}^n (X_i - \mathbf{E}\, X_i) \geq \epsilon\right) \leq e^{-s\epsilon} \prod_{i=1}^n \mathbf{E}\, e^{s(X_i - \mathbf{E}\, X_i)}\,.$$

(iv) Making use of the bounds derived in *(ii)* and *(iii)*, conclude that

$$\mathbf{P}(S_n - \mathbf{E}\, S_n \geq \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{\sum_i (b_i - a_i)^2}\right)\,.$$

(v) Derive similar bounds for $\mathbf{P}(S_n - \mathbf{E}\, S_n \leq -\epsilon)$ and $\mathbf{P}(|S_n - \mathbf{E}\, S_n| \geq \epsilon)$.

Part II: Error bound

Consider the problem of binary classification, with bounded loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0,\, 1]$. Consider a finite collection of models $\mathcal{F}$, with cardinal $|\mathcal{F}|$. As usual, the risk of a fixed classifier $f \in \mathcal{F}$ is denoted $\mathcal{R}(f) = \mathbf{E}\, \ell(Y, f(X))$, and the empirical risk is denoted $\hat{\mathcal{R}}_n(f) = n^{-1} \sum_{i=1}^n \ell(Y_i, f(X_i))$, based on a learning sample $\mathcal{L}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.

(vi) Making use of the bound derived in *(iv)*, show that

$$\mathbf{P}(\mathcal{R}(f) - \hat{\mathcal{R}}_n(f) \geq \epsilon) \leq e^{-2n\epsilon^2}\,. \tag{1}$$

(vii) Conclude that for all $f \in \mathcal{F}$ and $0 < \delta < 1$, with probability at least $1 - \delta$, holds

$$\mathcal{R}(f) \leq \hat{\mathcal{R}}_n(f) + \sqrt{\frac{\log|\mathcal{F}| - \log\delta}{2n}}\,. \tag{2}$$

Problem 1. *Bound for countably infinite $\mathcal{F}$*
We derive a bound similar to (2) in Problem1 in the case of a countable class of models $\mathcal{F}$. A possible approach is to assign a positive number $c(f)$ for each $f \in \mathcal{F}$, such that

$$\sum_{f \in \mathcal{F}} e^{-c(f)} = 1 \,.$$

The number $c(f)$ can be interpreted as a measure of complexity of f, or as the logarithm of a prior probability attached to f.

(i) Make sure you understand the interpretations given for $c(f)$.

(ii) Show that the bound (1) in Problem 1 can be rewritten

$$\mathbf{P}\left(\mathcal{R}(f) - \hat{\mathcal{R}}_n(f) \geq \sqrt{\frac{-\log \eta}{2n}}\right) \leq \eta \,, \tag{3}$$

for any $\eta > 0$.

(iii) We let η introduced in (3) depend on the model y. Specifically, replace η with $\delta(f) = \delta e^{-c(f)}$, for some $\delta > 0$. Conclude that with probability at least $1 - \delta$, for all $f \in \mathcal{F}$,

$$\mathcal{R}(f) \leq \hat{\mathcal{R}}_n(f) + \sqrt{\frac{c(f) - \log \delta}{2n}} \,. \tag{4}$$

(iv) By considering a uniform prior in the case of a finite $\mathcal{F}$, show that the bound (4) coincides with (2).

Problem 2. *Sauer's Lemma*
Sauer's Lemma in its original form states that for a model class $\mathcal{F}$ with finite VC dimension d,

$$\mathcal{S}(\mathcal{F}, n) \leq \sum_{i=0}^{d} \binom{n}{i} \,.$$

Show that this implies $\mathcal{S}(\mathcal{F}, n) \leq (n+1)^d$ for all n

Problem 3. *VC dimension*
Derive the VC dimension of the following classes of functions. For part *(i)*, you will in addition derive the shattering number $\mathcal{S}(\mathcal{F}, n)$, for any $n \geq 1$.

(i) $\mathcal{F} = \{f : \mathbb{R} \to \{0,1\} \,|\, f(x) = \mathbf{1}(x \leq a),\, a \in \mathbb{R}\}$

(ii) $\mathcal{F} = \{f : \mathbb{R}^2 \to \{0,1\} \,|\, f(x) = \mathbf{1}(x_i \leq a) \text{ or } \mathbf{1}(x_i \geq a)\,, i = 1 \text{ or } 2\,, \text{ where } x = (x_1, x_2)\}$

(iii) $\mathcal{F} = \{f : \mathbb{R}^2 \to \{0,1\} \,|\, f(x) = \mathbf{1}(x \in C)\,,\ C \subset \mathbb{R}^2 \text{ convex}\}$.

Problem 4.
We denote by $\mathcal{X}$ the input space, and let $A_1, \ldots, A_m$ be non empty subsets of $\mathcal{X}$, such that $A_1, \ldots, A_m$ form a partition of $\mathcal{X}$. Let $\mathcal{F}$ be the set of functions $\mathcal{X} \to \mathbb{R}$, constant on each set A_j, so that a function $f \in \mathcal{F}$ can be written

$$f(x) = \begin{cases} y_1 & \text{if } x \in A_1 \\ \vdots & \vdots \quad \vdots \\ y_m & \text{if } x \in A_m \,, \end{cases}$$

where $y_1, \ldots, y_m$ are real numbers. Put

$$\mathcal{G} := \{g : \mathcal{X} \to \{-1,\, 1\} \,|\, g(x) = \operatorname{sign}[f(x)] \,,\ \text{for } f \in \mathcal{F}\} \,,$$

where $\operatorname{sign}(u) := \mathbf{1}(u \geq 0) - \mathbf{1}(u < 0)$.

(i) How many elements are in $\mathcal{G}$?

(ii) What is the Vapnik Chervonenkis dimension of $\mathcal{G}$?

Consider a binary classification problem, under a 0/1 loss $\ell(y, g(x)) = \mathbf{1}(y \neq g(x))$, where $y \in \{-1, 1\}$. The risk of $g \in \mathcal{G}$ is denoted $\mathcal{R}(g) := \mathbf{E}\{\mathbf{1}(Y \neq g(X))\}$. Let $\bar{g}$ be the best classifier in $\mathcal{G}$, and $\hat{g}_n$ the empirical risk minimizer, based on a training sample of size n.

(iii) Using an argument established during the lecture, show that with probability larger than $1 - \delta$,

$$\mathcal{R}(\hat{g}_n) \leq \mathcal{R}(\bar{g}) + \sqrt{\frac{2[(m+1)\ln(2) - \ln(\delta)]}{n}} \,.$$

(iv) Deduce from *(iii)* a bound for

$$\mathbf{E}\left\{\mathcal{R}(\hat{g}_n) - \mathcal{R}(\bar{g})\right\}^2 - \frac{2(m+1)\ln 2}{n} \,.$$

(v) Deduce from *(iv)* a bound for $\mathbf{E}\{\mathcal{R}(\hat{g}_n)\} - \mathcal{R}(\bar{g})$.