Exercise Sheet 5

Exercise 1

Suppose we have a lake with an infinite number of red and green fish. Red fish are disgusting (bad, bad) and green fish are tasty (good, good).

Run the following simulation a large number of time (say, $T = 10,000$ runs): There are $1,000$ fishermen each of whom has an unknown probability $\mu = 0.5$ of catching red fish. Each fisherman independently catches 10 fishes. We focus on the following fishing strategy:

1. f_1 is the first fisherman
2. f_r is a random fisherman
3. f_* is the fisher with the minimum frequency of red fish over all fishers.

(a) Record the T fractions μ_n^1, μ_n^r, and μ_n^* of red fish for the respective three fishermen. Plot the histograms of the distributions of μ_n^1, μ_n^r, and μ_n^*.

(b) Using (a), plot estimates for $\mathbb{P}(|\mu_n - \mu| \geq \varepsilon)$ as a function of $\varepsilon > 0$ together with the Hoeffding bound.

(c) Which fishermen obey the Hoeffding bound, and which ones do not? Explain why. Relate the results to learning.

Exercise 2

Consider the following setting: The input space is $\mathcal{X} = [0, 1]$ and the output space is $\mathcal{Y} = \{0, 1\}$. The input data $x \in \mathcal{X}$ is uniformly distributed. The Bayes classifier

$$f^*(x) = \begin{cases} 1 & : \quad x \in [0.25, 0.75] \\ 0 & : \quad \text{otherwise} \end{cases} .$$

has zero Bayes risk $R^* = R(f^*) = 0$. The hypothesis class

$$\mathcal{H} = \left\{ f_\theta \, : \, \theta \in [0, 1] \right\}$$

consists of threshold classifiers of the form

$$f_\theta(x) = \begin{cases} 0 & : \quad x < \theta \\ 1 & : \quad x \geq \theta \end{cases} .$$

(a) What is the true risk $R(f_{\mathcal{H}})$ of the best classifier from $\mathcal{H}$?

(b) For a given training set $\mathcal{S} = ((x_1, y_1), \ldots, (x_n, y_n))$, a learner A returns a model $f_n = A(\mathcal{S}) \in \mathcal{H}$. Implement two learners:

1. The first learner returns a hypothesis $f_n \in \mathcal{H}$ that minimizes the empirical risk on $\mathcal{S}$.
2. The second learner randomly picks an input example x_i from the training set and returns the hypothesis $f_n \in \mathcal{H}$ with threshold $\theta = x_i$.

(c) Consider the following experiment for a given learner A:

1. Sample a training set $\mathcal{S}$ of size n.
2. Fit a model f_n
3. Compute the empirical risk $R_n(f_n)$
4. Compute the true risk $R(f_n)$

Repeat the experiment T times for different sizes n and for both learners from part (b).

(d) Choose $\varepsilon = 0.1$ as error tolerance. Use the results from (c) to estimate the following probabilities for each size n and for both learners from part (b):

$$P(|R_n(f_n) - R(f_n)| \geq \varepsilon)$$
$$P(|R_n(f_n) - R(f_{\mathcal{H}})| \geq \varepsilon)$$
$$P(|R_n(f_n) - R(f^*)| \geq \varepsilon)$$

Plot the estimated probabilities together with Hoeffding's bound as a function of n. Discuss the results.

(e) Choose $\varepsilon = 0.1$ as error tolerance. Use the results from (c) to estimate the following probabilities for each size n and for both learners from part (b):

$$P(|R_n(f_n) - R(f_n)| \geq \varepsilon)$$
$$P(|R(f_n) - R(f_{\mathcal{H}})| \geq \varepsilon)$$
$$P(|R(f_n) - R(f^*)| \geq \varepsilon)$$

Plot the estimated probabilities as a function of n. Discuss the results.