Exercise Sheet 8

Exercise 1

Identify and briefly explain the key steps and concepts that connect the goal of supervised learning to the VC Theorem. This includes providing definitions, mathematical formulations, and discussing the significance of each concept in the context of learning theory.

Exercise 2

Try to follow the proof of the Sauer-Shelah Lemma in Lecture vl06.

Exercise 3

In this exercise, we examine the VC generalization bound for the hypothesis class of linear classifiers. The bound is given by:

$$R(f) \leq R_n(f) + \sqrt{\frac{8}{n}\left(d \cdot \ln\left(\frac{2en}{d}\right) + \ln\left(\frac{4}{\delta}\right)\right)} = R_n(f) + \varepsilon.$$

Refer to page 19 of `vl06` for more details. The square root term on the right-hand side is referred to as the *error bound* ε.

(a) Construct a binary classification problem with a large ground truth dataset to determine the true risk of a linear classifier and to sample training sets of varying sizes n.

(b) Specify a sequence of training set sizes $n_1 < n_2 < \cdots < n_k$. For each training set size n:

- Calculate the error bound ε for $\delta = 0.05$.
- Conduct the following experiment T times for each training set size n:
  1. Draw a training set of size n.
  2. Fit a logistic regression model to the training data.
  3. Determine the empirical and true risk of the learned model.

(c) Plot the average true risk, empirical risk, and the VC generalization bound as a function of the training set size n.

(d) Let $\varepsilon = \delta = 0.05$. Compute the theoretical sample complexity to ensure $R(f) \leq R_n(f) + \varepsilon$ with confidence $1 - \delta$.

(e) Plot the estimated probability $\mathbb{P}(R(f) \leq R_n(f) + \varepsilon)$ as a function of n. Determine the empirical sample complexity to ensure $R(f) \leq R_n(f) + \varepsilon$ with confidence $1 - \delta$. Compare the theoretical and empirical sample complexity.