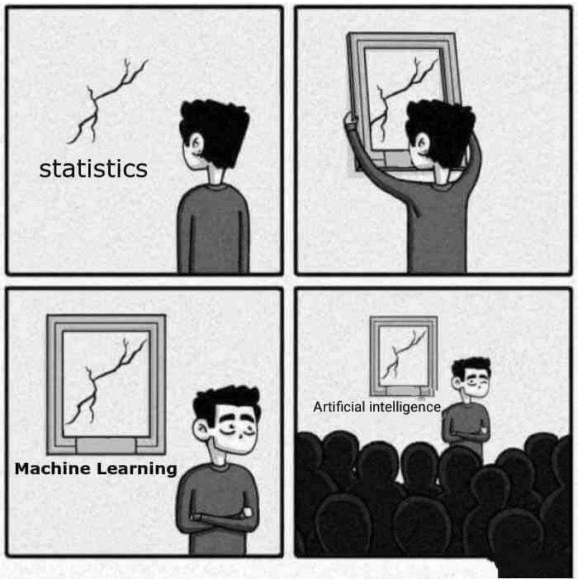
Chapter 1: The Learning Problem (II)

Let's see who you really are
machine learning



Section 1.3: Is Learning Feasible? and
Section 1.4: Error and Noise and
Section 2.2.3 The Test Set

The previous note packet focused on the *computational* component of learning. We used the Perceptron Learning Algorithm (PLA) as an example and derived its runtime.

This packet introduces the *statistical* component of learning. We will start by exploring the idea of training and testing sets, and see why they are used to evaluate model performance. Later notes will explore many other statistical components of learning.

It is very easy to get the computational and statistical aspects of learning confused. You should pay careful attention to the differences, which are highlighted in Figure 1.2, 1.9, and 1.11.

Problem 1. The PLA fails to converge (i.e. has an infinite runtime) if the input data is not linearly separable. Describe two scenarios that cause the PLA to fail to converge.

Problem 2. Define the in-sample error (page 21), out-of-sample error (page 21), true/bayes error, and generalization error (page 40).

Note 1. A fundamental goal of the machine learning discipline is to understand the out-of-sample error. One of the most powerful tools for doing this is Hoeffding's inequality. It is introduced informally in the textbook on page 19, Eq (1.4) in the context of a toy example involving marbles. The following theorem is a formally stated version of this inequality, and what you should rely on in this class.

Theorem 1 (Hoeffding Inequality)**.** Let $a_1, ..., a_N$ be N independent and identically distributed random variables satisfying $0 \leq a_i \leq 1$. Let $\nu = \frac{1}{n}\sum_{i=1}^{N} a_i$ be the empirical average and $\mu = \mathbb{E}\nu$ be the true mean of the underlying distribution. Then, for all $\epsilon > 0$,

$$\mathbb{P}\left(|\nu - \mu| \geq \epsilon\right) \leq 2\exp(-2\epsilon^2 N). \tag{1}$$

Problem 3. Describe the shape of the distribution defined by inequality 1 above.

Problem 4. It is common to divide our data points into a *training set* and a *test set.* Recall that we let N denote the size of the training set, $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)$ denote the training data points, and E_{in} the error on the training set. Similarly, we let N_{test} denote the size of the test set, $(\mathbf{x}_1^{\text{test}}, y_1^{\text{test}}), ..., (\mathbf{x}_{N_{\text{test}}}^{\text{test}}, y_{N_{\text{test}}}^{\text{test}})$ denote the test data points, and

$$E_{\text{test}}(h) = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} [\![h(\mathbf{x}_i) \neq y_i]\!]$$

denote the error on the test set.

Now let g be the result of running some learning algorithm on the training dataset. (The algorithm could be the PLA, or it could be some other algorithm that we haven't covered yet.) We would like to know $E_{\text{out}}(g)$, since this is the true performance of the hypothesis g on future data. We cannot calculate this quantity directly, but we can calculate $E_{\text{test}}(g)$ and use the Hoeffding inequality to bound $|E_{\text{out}}(g) - E_{\text{test}}(g)|$. We want this quantity to be small. The subproblems below explore how big the test set must be in order to ensure that E_{out} is close to E_{test}.

1. If you have a test set with 1000 samples, what bound does the Hoeffding inequality give on the probability that $|E_{\text{test}} - E_{\text{out}}| < 0.01$?

   HINT: Let $E_{\text{test}} = \nu$ and $E_{\text{out}} = \mu$. Then recall that $P(|E_{\text{test}} - E_{\text{out}}) \geq \epsilon) = 1 - P(|E_{\text{test}} - E_{\text{out}}| < \epsilon)$.

   ANSWER: -0.637

2. Why is this bound "trivial"?

   ANSWER: The bound above says that the probability is greater than a negative number. This is always true by definition because a probability is between 0 and 1. When a bound doesn't provide any new information, we call it *trivial.*

3. What if you change the accuracy to $|E_{\text{test}} - E_{\text{out}}| \leq 0.05$?
   ANSWER: 0.987

4. What if you use the original accuracy $|E_{\text{test}} - E_{\text{out}}| \leq 0.01$ but use 10000 samples in your test set?
   ANSWER: 0.729

Note 2. The calculations in the previous problem are all examples of how a statistician would apply Hoeffding's inequality. In particular, the properties of the learning problem (N, N_{test}) was fixed, and we computed confidence intervals based on the given data. These computations are closely related to p-values and statistical testing.

That is not how data miners operate. In the real world, problems are not statically given to us like they were in the problem above. Problems can be changed in order to make them easier to solve, and data miners use tools like Hoeffding's inequality to guide these changes. The following problem illustrates this concept. It is derived from a real capstone project.

Problem 5. You are a consultant for BigBank. BigBank has trained a model g_{house} for predicting whether to give a house loan to a client, and another model g_{car} for predicting whether to give a car loan to client. Experiments show that

$$\begin{aligned} E_{\text{test}}(g_{\text{house}}) &= 0.05 \\ E_{\text{test}}(g_{\text{car}}) &= 0.30 \end{aligned}$$

Unfortunately, the test sets for both problems are rather small, with only 100 data points each. You have been given an additional budget to collect data to expand one of these test sets to 10000 data points, but unfortunately budget constraints will only let you enlarge one test set and the other will remain small. Which test set should you enlarge and why?

Problem 6. Let g be the output of running the PLA on a dataset of size N. It is tempting to let $\nu = E_{\text{in}}(g)$ and $\mu = E_{\text{out}}(g)$, and then apply the Hoeffding inequality to state that the generalization error is bounded with high probability by

$$\mathbb{P}(|E_{\text{in}}(g) - E_{\text{out}}(g)| \geq \epsilon) \leq 2\exp(-2\epsilon^2 N).$$

Why is this not correct?

HINT: What assumption does Hoeffding require that we satisfied in the previous problem bounding $|E_{\text{test}} - E_{\text{out}}|$, but we do not satisfy in this problem bounding $|E_{\text{in}} - E_{\text{out}}|$?