2.3.1. Finite Hypothesis Classes

If $\mathcal{H}$ is a finite set, then $ERM_{\mathcal{H}}$ will not overfit provided a sufficiently large training sample.

Realizability Assumption There exists a hypothesis h^* such that the loss over the entire space is zero. This implies:

1. any random sample $S \sim D$ we have $L_S(h^*) = 0$.
2. for every ERM hypothesis $L_S(h^*) = 0$

i.i.d. assumption Examples in the training set S are independently and identically distributed from D. $S \sim D^m$ denotes probability over m-tuples induced by independently m times sampling from D.

It is always possible that S does not represent D "well", due to sampling being a random process. The probability of getting a nonrepresentative sample is denoted with δ. $(1 - \delta)$ is called the (prediction) confidence parameter.

Accuracy parameter, denoted by ϵ, reflects the quality of predictions. We compare the accuracy to the loss and distinguish between cases where the $L > \epsilon$ (failure of the learner) and $L \leq \epsilon$ (approximately correct). This accuracy is actually an error, not accuracy in a information retrieval context.

Define $\mathcal{H}_B$ as the set of bad hypotheses and M set of bad examples:

$$\mathcal{H}_B = \{h \in \mathcal{H}_B, L_S(h) = 0\}$$
$$M = \{S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0\}$$

We wish to use the realizability assumption to upper bound the probability of having a failure learner. Traning on the set of misleading examples yields zero loss for the sample with the resulting hypothesis h_S, which leads to failure on D: $L_{D,f}(h_S) > \epsilon$. So, the probabilty of a failure learner is less than the probability of drawing misleading examples:

$$D^m(\{S|_x : L_{(D,f)}(h_S) > \epsilon)\} \leq D^m(\cup_{h \in \mathcal{H}_B}\{S|_x : L_S(h) = 0\})$$

The right hand side can be replaced with a sum using the union bound rule and since i.i.d. sampling is done it can be broken down into indepedent probabilties

$$D^m(\{S|_x : L_{(D,f)}(h_S) > \epsilon)\} \leq \sum_{h \in \mathcal{H}_B} D^m(\{S|_x : L_S(h) = 0\})$$

$$D^m(\{S|_x : L_{(D,f)}(h_S) > \epsilon)\} \leq \sum_{h \in \mathcal{H}_B} \prod_{i=1}^{m} D(\{S|_x : h(x_i) = f(x_i)\})$$

Probability that a hypothesis on an instance is equal to the target label is:

$$D(\{x_i : h(x_i) = y_i\}) = 1 - L_{(D,f)}(h) \leq 1 - \epsilon$$

$$\prod_{i}^{m} D(\{x_i : h(x_i) = y_i\}) = (1-\epsilon)^m$$

$$\sum_{h \in \mathcal{H}_B} \prod_{i}^{m} D(\{x_i : h(x_i) = y_i\}) = |\mathcal{H}_B|(1-\epsilon)^m$$

Using this exponential inequality, we arrive at:

$$\sum_{h \in \mathcal{H}_B} D^m(\{S|_x : L_{(D,f)}(h) > \epsilon)\} \leq |\mathcal{H}_B| e^{-\epsilon m}$$

Said in plain words, for each individual bad hypothesis $h \in \mathcal{H}_B$ that probability of overfitting using ERM is at most $e^{-\epsilon m}$, which means the larger m, the smaller this probability. The larger ϵ (accuracy), the smaller this probability.

$$D^m(\{S|_x : L_{(D,f)}(h_S) > \epsilon)\} \leq \delta$$

Let $\mathcal{H}$ be a finite hypothesis size. Let $\delta \in (0,1), \epsilon > 0$ and m satisfy:

$$m \geq \frac{\log(|\mathcal{H}/\delta)}{\epsilon}$$

This means that for a sufficiently large sample size m, the $ERM_{\mathcal{H}}$ rule over a finite set of hypothesis will be *probably* (with confidence $1-\delta$) *approximately* (up to an error ϵ) correct.

2.4. Exercises

2.1. Overfitting of polynomial matching

Find a_i for which $a_0 + a_1 x_i + a_2 x_i^2 + ...a_n x_i^n \geq 0$ when $h(x) = y_i$. My solution here is something along $a_0 = -x_i$, $a_1 = 1$, which will when work $x \leq x_i$, which is good, but won't work when $x > x_i$.

Here, I will use the absolute value (not sure if it's allowed though to solve the problem):

$$h(x) = -1|x - x_i|$$

This formula means, if the datapoints are the same (only works for the train set), then they will output 0, which yields $h_s(x) = 1$, otherwise it will output a negative value.

2.2. Expectation of empirical risk equals true risk

$$\mathbb{E}_{S|x\sim D^m}[L_S(h)] = L_{(D,f)}(h)$$

Expectation of empirical risk is the true risk. Intuitively, if we sample an infinite number of times and calculate the empirical risk, we end up with the true risk, since the definition of expectation is a sum over all probabilities

This needs to be expressed mathematically.

2.3 Axis aligned rectangle

Part 1

Show that A is an ERM.

$$ERM = min_{h\in\mathcal{H}} L_D(h)$$

The loss with algorithm A will always be 0, therefore it's an ERM.

$$L_S(A) = 0$$

Part 2

RECAP QUESTIONS

1. We are trying to learn the discriminative function which maps data to labels. We assume there exists a perfect mapping from data to labels. What would a generative approach look like in this case? Is this entire book merely on discriminative approaches?
2. Which inductive bias do we have when relying on ERM algorithms? More inductive biases will be studied later (Occam's razor, independence in Naive Bayes)

Tom Mitchel's definition of inductive bias: an inductive bias of a learner is the set of additional assumptions sufficient to justify its inductive inferences as deductive inferences.

3. Look at the formula in Corollary 2.3. How come it does not depend at all on the domain space size?

Answer: it's all actually the delta parameter.