Member-only story

Learning Theory: Empirical Risk Minimization

Marin Vlastelica · Follow
Published in Towards Data Science · 7 min read · Feb 27, 2019

Empirical Risk Minimization is a fundamental concept in machine learning, yet surprisingly many practitioners are not familiar with it. Understanding ERM is essential to understanding the limits of machine learning algorithms and to form a good basis for practical problem-solving skills. The theory behind ERM is the theory that explains the VC-dimension, Probably Approximately Correct (PAC) Learning and other fundamental concepts. In my opinion, anybody that is serious about machine learning should be comfortable with talking about ERM. I will try to explain the underlying concepts as simple, short and theoretically grounded as possible. This article is heavily based on the book Understanding Machine Learning by Schwartz and Ben-David, which I highly recommend for anyone interested in the fundamentals of learning theory.

Let's start with a simple supervised learning classification problem. Let us say that we want to classify spam emails, probably the most often used example in machine learning (note, this is not a post on naive Bayes). Each

email has a label 0 or 1, either spam or not spam. We denote the domain space with ***X*** and the label space with ***Y***, we also need a function for mapping the domain set space to the label set space, ***f: X -> Y***, this is just a formal definition of a learning task.

Now that we have our formal problem definition, we need a model that is going to make our predictions: spam or not spam. Coincidentally, the synonym for **model** is the hypothesis ***h***, which may be a bit confusing. The hypothesis, in this case, is nothing else than a function which takes input from our domain ***X*** and produces a label 0 or 1, i.e. a function ***h: X -> Y***.

In the end, we actually want to find the hypothesis that minimizes our error right? With this, we come to the term empirical risk minimization. The term empirical implies that we minimize our error based on a sample set ***S*** from the domain set **X**. Looking at it from a probabilistic perspective we say that we sample ***S*** from the domain set **X**, with ***D*** being the distribution over **X**. So when we sample from the domain, we express how likely a subset of the domain is sampled from the domain **X** by ***D*(S)**.

In the equation below, we can define the **true error**, which is based on the whole domain **X**:

$$L_{\mathcal{D},f}(h) \stackrel{\text{def}}{=} \mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq f(x)] \stackrel{\text{def}}{=} \mathcal{D}(\{x : h(x) \neq f(x)\}).$$

The error for hypothesis h. Starting from left to right, we calculate the error L based on a domain distribution D and a label mapping f. The error is equal to the probability of sampling x from d such that the label produced by the hypothesis is different from the actual label mapping.

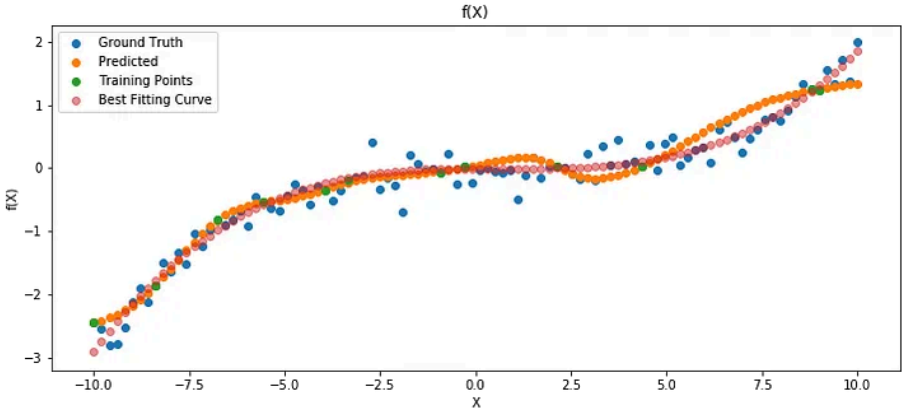
Since we only have access to ***S***, a subset of the input domain, we learn based on that sample of training examples. We don't have access to the **true error**, but to the **empirical error**:

$$L_S(h) \stackrel{\text{def}}{=} \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$$

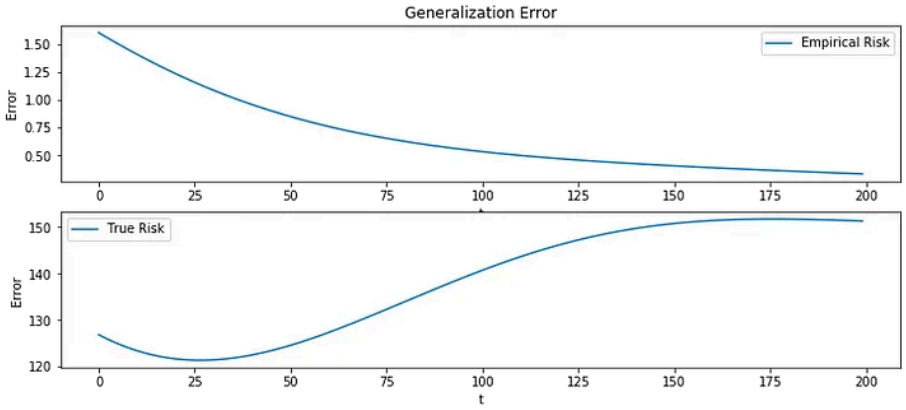
m denotes the number of training examples. You can see from the equation that we effectively define the empirical error as the fraction of misclassified examples in the set S.

The empirical error is also sometimes called the generalization error. The reason is that actually, in most problems, we don't have access to the whole domain ***X*** of inputs, but only our training subset ***S***. *W*e want to generalize based on ***S***, also called inductive learning. This error is also called the **risk**, hence the term risk in empirical risk minimization. If this reminded you of mini-batch gradient descent, you would be correct. This concept is basically ubiquitous in modern machine learning.

Now we can talk about the problem of **overfitting**. Namely, since we have only a subsample of the data it can happen that we minimize the **empirical error** but actually increase the **true error.** This result can be observed in a simple curve fitting problem. Let us imagine that we have some robot that we want to control, we want to map some sensor data **X** to torques. The sensor data has some kind of noise since sensors are never perfect, in this instance, we are going to use simple Gaussian noise to the sensor data. We fit a neural network to do this and we obtain the following result:



We can see this generalization error looking at another plot, notice how at some point the true error starts increasing while the empirical error decreases further. This is the consequence of the model overfitting on the training data.



Now that we defined our **empirical risk** and **actual risk**, the question arises if we can actually do anything useful with these? As it turns out, we can guarantee that ERM is going to work with a certain confidence. Phrased differently, we would like to produce an upper bound to the error of our model with a certain confidence. The meaning of upper bound is simply that

we can guarantee that the error is not going to get larger than this bound, hence the word bound.

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

This gives us the probability that a sampled set S (or training data in other words) is going to overfit the data, i.e. that the true error(L) is going to be larger than a certain number epsilon.

In the current case, we are going to operate under the **realizability assumption**. I am not going to write the formal definition, but in a nutshell, the assumption states that there exists a hypothesis ***h*** in the space of all possible hypotheses ***H*** that is optimal in the sense that it has a true risk of 0, this also implies that the found hypothesis on the subset ***S*** achieves a 0 empirical error. Of course, this mostly isn't true in real-world use-cases and there are learning paradigms that relax this assumption, but this I will leave perhaps for another post.

Let us define the set of the hypotheses for which the true error is higher than epsilon:

$$\mathcal{H}_B = \{h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \epsilon\}.$$

For this set of hypothesis, it is clear that either they received a non-representative set for learning, which has resulted in a low empirical error(risk) and high true error(risk) or they were not good enough to learn anything.

We want to separate the misleading training sets that lead to hypotheses that resulted in a low empirical error and high true error (the case of overfitting), which later on will be useful in deriving the upper bound:

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

The probability of sampling a specific subset ***S*** that is a non-representative sample is logically equal or lower than the probability of sampling ***M***, because ***S*** is a subset of ***M***. Thus we can write the following:

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

We apply the union bound lemma to the right-hand side of the equation, which states that a probability of sampling a union of two sets is lower than

the probability of sampling them individually. This is why we can write the sum on the right-hand side:

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

Also, we assume that the examples are independently and identically distributed (iid), therefore we can write the probability of the empirical error being zero as the product of probabilities of individual predictions being correct:

$$\begin{aligned}\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) &= \mathcal{D}^m(\{S|_x : \forall i, h(x_i) = f(x_i)\}) \\ &= \prod_{i=1}^{m} \mathcal{D}(\{x_i : h(x_i) = f(x_i)\}).\end{aligned}$$

The constant m denotes the number of training examples in the set S.

The probability of the hypothesis being correct at a certain datapoint can be written as 1 minus the **true risk**. This follows from the fact that we defined the risk as the fraction of misclassified examples. The inequality follows from the fact that we are assuming that the error is lower than or equal to the upper bound.

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

If we combine the previous two equations, we arrive at the following result:

$$\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) \leq (1 - \epsilon)^m \leq e^{-\epsilon m}$$

The exponential on the right comes from a simply provable inequality.

If we combine the upper equation with the previous one where we applied the union bound, we arrive at the following insightful result:

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

Here |H| is the cardinality of the hypothesis space, obviously, this formulation doesn't make sense in the case when we have an infinite number of hypotheses.

We can replace the left hand side with a certain constant 1-delta, where delta is how much confidence do we want that the error is not going to be higher than epsilon. We can simply rearrange the equation in order to state the following:

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

The final result tells us how many examples (m) do we need so that ERM does not result in an error higher than epsilon with a certain confidence delta, i.e. when we choose enough examples for ERM, it is probably not going to have an error higher than epsilon. I use the word probably here since it is subject to our confidence constant delta, which is between 0 and 1. It is an intuitive thing to say, but I think it is nice sometimes to look at the equations and see that it makes sense, mathematically.

In ERM, many assumptions are made. I mentioned the realizability assumption that states that there is an optimal hypothesis in our pool of hypotheses. Also the hypotheses space may not be finite as it is made to be here. In the future, I plan to go through paradigms that relax these assumptions. Nevertheless, ERM is a fundamental concept in learning theory and essential for any serious machine learning practitioner.

Machine Learning | Mathematics | Probability | Artificial Intelligence | Towards Data Science