Empirical Risk Minimization is an incomplete inductive principle

Thomas P. Minka

February 20, 2001

Abstract

Empirical Risk Minimization (ERM) only utilizes the loss function defined for the task and is completely agnostic about sampling distributions. Thus it only covers half of the story. Furthermore, ERM is equivalent to Bayesian decision theory with a particular choice of prior.

1 ERM is incomplete

Suppose we are given a data set $D = \{x_1, ..., x_N\}$ and we want to predict future data in a way that incurs minimum expected loss. That is, we want to pick a number θ to minimize

$$R = \int_x p(x|D) L(\theta, x) dx \tag{1}$$

where L is the loss in predicting θ when the real value is x and $p(x|D)$ represents what we know about future x after having observed D. This is the decision-theoretic method of estimation: (1) find the predictive density $p(x|D)$ and (2) choose the estimate which minimizes loss on that density. The first step employs a sampling model for the data and is irrespective of the loss. The second step employs a loss function for the task and is irrespective of the sampling model.

Empirical Risk Minimization instead chooses θ to minimize the average loss we would have incurred on the data set D:

$$ER = \frac{1}{N} \sum_i L(\theta, x_i) \tag{2}$$

which is seen as a Monte Carlo estimate of (1). The problem is that this method completely discards any information we may have about the sampling distribution of x. The following examples demonstrate.

1.1 Gaussian data with squared loss

Let the loss function be quadratic and let the data be independent Gaussian with unit variance and unknown mean:

$$
\begin{aligned}
L(\theta, x) &= (\theta - x)^2 && (3)\\
p(x|m) &\sim \mathcal{N}(m, 1) && (4)\\
&= \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x-m)^2) && (5)
\end{aligned}
$$

For the Bayesian approach, we also need a prior on m in order to get a predictive density $p(x|D)$. Let the prior be essentially uniform, e.g. a Gaussian with very large variance. Then the predictive density is

$$
\begin{aligned}
p(x|D) &= \int_m p(x|m)p(m|D)dm && (6)\\
&\approx \int_m p(x|m)\mathcal{N}(m; \bar{x}, \frac{1}{N})dx && (7)\\
&= \mathcal{N}(\bar{x}, 1 + \frac{1}{N}) && (8)\\
\bar{x} &= \frac{1}{N}\sum_i x_i && (9)
\end{aligned}
$$

Now, *after* we have computed the predictive density, we apply the loss function and minimize

$$R = \int_x p(x|D)(\theta - x)^2 dx \tag{10}$$

For any $p(x|D)$, the minimum loss is achieved by the estimate

$$\hat{\theta} = \int_x xp(x|D)dx = E[x|D] \tag{11}$$

which in this case is $\bar{x}$, the sample mean.

What about Empirical Risk Minimization? The empirical risk is

$$ER = \frac{1}{N}\sum_i (\theta - x_i)^2 \tag{12}$$

whose minimum is at $\hat{\theta} = \bar{x}$. So in this case the two methods coincide.

1.2 Gaussian data with absolute loss

Continue the previous setup but now let the loss function be

$$L(\theta, x) = |\theta - x| \tag{13}$$

What does decision theory tell us to do? The predictive density $p(x|D)$ is the same as before, since it doesn't depend on L. We just need to minimize

$$R = \int_x p(x|D)|\theta - x|dx \tag{14}$$

For any $p(x|D)$, the minimum loss is achieved at the *median* of the predictive distribution, i.e. the point where

$$p(x < \hat{\theta}|D) = p(x > \hat{\theta}|D) \tag{15}$$

Since the predictive density is Gaussian, the median is the same as the mean, and $\hat{\theta} = \bar{x}$ as before.

What about Empirical Risk Minimization? The empirical risk is

$$ER = \frac{1}{N} \sum_i |\theta - x_i| \tag{16}$$

whose minimum is at the sample *median* of the data. That is, $\hat{\theta}$ is chosen so that the number of samples less than $\hat{\theta}$ is equal to the number of samples greater than $\hat{\theta}$. For odd N, the median is unique, while for even N there is an interval of equally valid minima. The main point here is that the estimate *is not the same* as that prescribed by decision theory.

Which estimate is better? Decision theory says the sample mean must be optimal, and it is hard to doubt this considering Empirical Risk is only an approximation to the Bayes risk. Nevertheless we can confirm it with a simulation. We pick a value of m, draw N data points with the distribution $\mathcal{N}(m, 1)$, compute the two estimators and measure the error on new data. There is no need to actually sample new data since we know the expected loss will be

$$\begin{aligned} E[|x - \hat{\theta}|] &= 2\mathcal{N}(\hat{\theta}; m, 1) + 2(m - \hat{\theta})\phi(m - \hat{\theta}) - (m - \hat{\theta}) && (17) \\ \phi(y) &= \int_{-\infty}^{y} \mathcal{N}(x; 0, 1)dx && (18) \end{aligned}$$

We repeat the whole procedure many times (but holding m fixed), and plot a histogram of the losses. Figure 1 shows a clear win for the sample mean over the median. The average test-data loss over 5000 trials was 0.8176 for the sample mean and 0.8267 for the sample median. Changing the amount of training data changes the absolute loss values but doesn't change the overall shape of the curves.

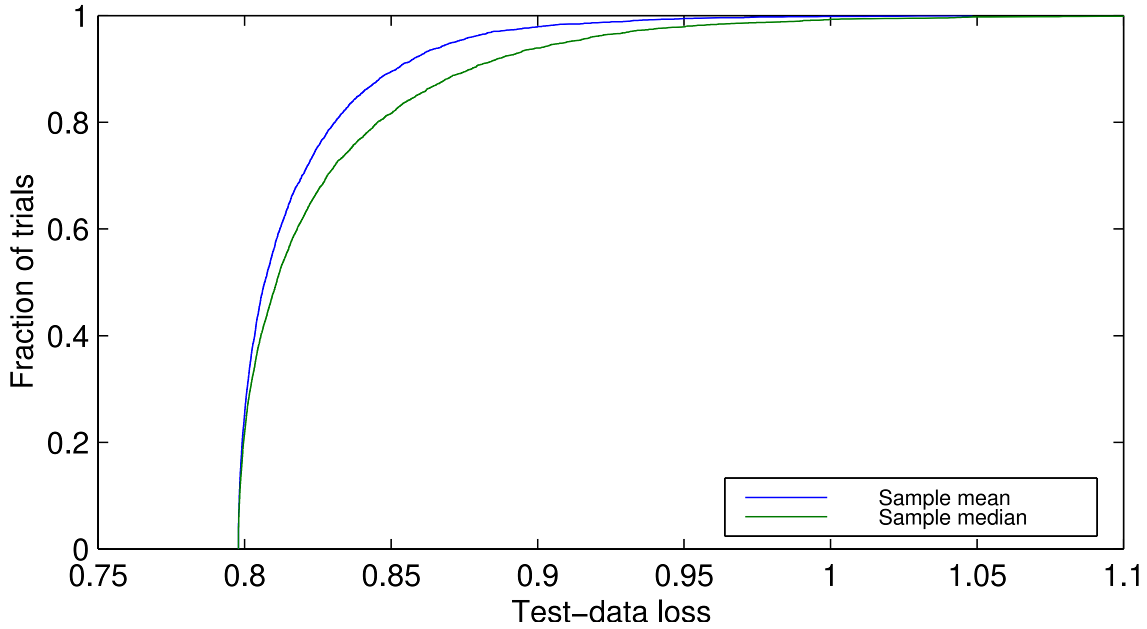

Figure 1: Cumulative histogram of the test-data loss for Bayesian decision theory (sample mean) and Empirical Risk Minimization (sample median) in Example 2. For each loss value, the curves show the percentage of trials (out of 5000) which had a test-data loss lower than that value. Higher curves are better. The true mean was $m = 5$ and there were $N = 20$ training points.

1.3 Laplacian data with squared loss

Instead of changing the loss function, now we change the sampling distribution. Let the loss function be quadratic and let the data be independent Laplacian with unit variance and unknown mean:

$$
\begin{aligned}
L(\theta, x) &= (\theta - x)^2 && (19)\\
p(x|m) &= \frac{1}{2}\exp(-|x-m|) && (20)
\end{aligned}
$$

Since the loss function is the same as Example 1, Empirical Risk Minimization will again pick the sample mean. What about decision theory? With a uniform prior on m, the predictive distribution is

$$
\begin{aligned}
p(x|D) &= \int_m p(x|m)p(m|D)dm && (21)\\
p(m|D) &= \frac{\exp(-\sum_{i=1}^N |x_i - m|)}{\int_{-\infty}^{\infty}\exp(-\sum_{i=1}^N |x_i - m|)dm} && (22)
\end{aligned}
$$

Since the loss function is quadratic, the minimum loss is achieved by the estimate $\hat{\theta} = E[x|D] = E[m|D]$ as before. This integral can be computed analytically as shown in Appendix A. With more than 10 data points, it is also reasonable to approximate $E[m|D]$ by the m which maximizes $p(m|D)$, i.e. the maximum-likelihood estimate. In this case, maximizing the likelihood is equivalent to minimizing $\sum_i |x_i - m|$, which we already saw is minimized by the *sample median.*

So now the tables are turned: Bayesian decision theory picks the sample median while ERM picks the sample mean. Which is better? Figure 2 shows the result of a simulation. The test-data loss is computed analytically as

$$E[(x-\hat{\theta})^2] = (m-\hat{\theta})^2 + 2 \qquad (23)$$

Even though the sample median is only an approximation to the Bayes optimum, it still does better than ERM. The average test-data loss over 5000 trials was 2.063 for exact Bayes, 2.066 for sample median, and 2.099 for sample mean.

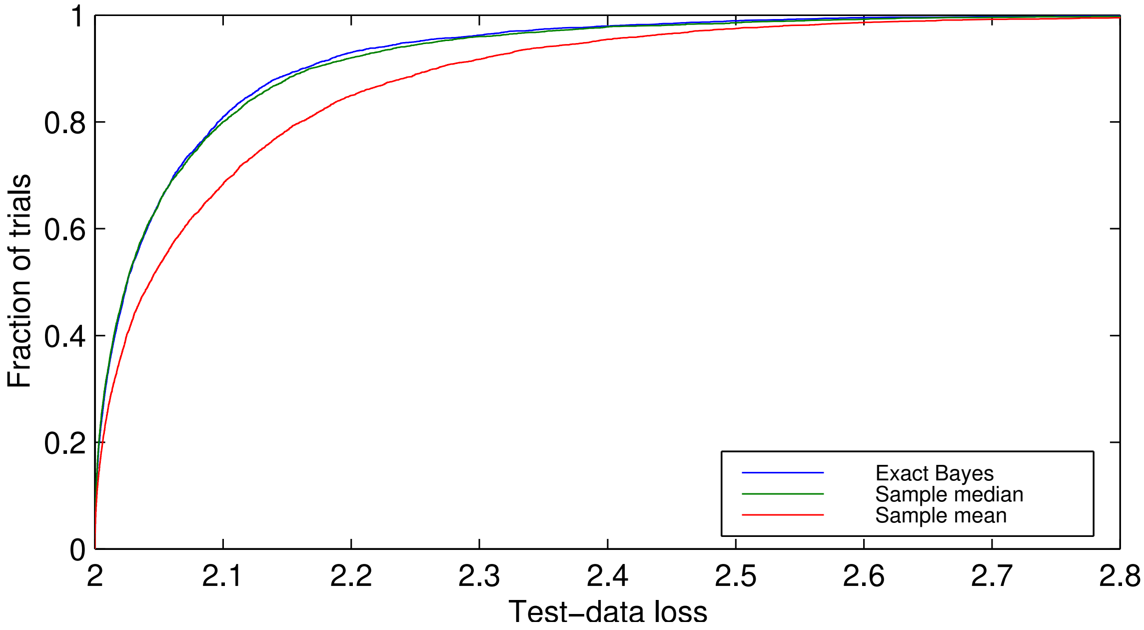

Figure 2: Cumulative histogram of the test-data loss for Bayesian decision theory (implemented exactly vs. approximately by the sample median) and Empirical Risk Minimization (sample mean) in Example 3. The true mean was $m = 5$ and there were $N = 20$ training points.

2 ERM is a special case of Bayesian decision theory

Is there a choice of sampling distribution which makes the Bayes risk (1) identical to the empirical risk (2)? Yes, all we need is for the predictive density $p(x|D)$ to be a sum of impulses at the training points:

$$p(x|D) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - x_i) \tag{24}$$

In other words, we have a maximally vague model, which assumes nothing beyond the training data. The only things we expect are things we have seen before. Such a model can be represented mathematically by a Dirichlet process (Ferguson, 1973) with parameter zero (which Ferguson calls the "noninformative Dirichlet prior").

3 Maximum-likelihood

This paper has focused on the deficiencies in ERM. Maximum-likelihood is also deficient, but in different ways. Maximum-likelihood does employ a sampling distribution, but it does not model uncertainty in the free parameters and it does not use a loss function. Maximum-likelihood does pretty well on the examples in this paper, matching the optimal result in the first two examples and approximately optimal in the third example. But because of these deficiencies one can construct counterexamples for maximum-likelihood as well.

References

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statisitcs*, *1*, 209–230.

A Posterior mean of the Laplacian likelihood

We want to compute

$$E[m|D] \quad = \quad \frac{\int_{-\infty}^{\infty} m \exp(-\sum_{i=1}^{N}|x_i - m|)dm}{\int_{-\infty}^{\infty} \exp(-\sum_{i=1}^{N}|x_i - m|)dm} \tag{25}$$

Renumber the data points so that they are sorted: $x_1 \leq x_2 \leq \cdots \leq x_N$. By splitting the range of integration into the separate intervals $(-\infty, x_1), (x_1, x_2), \cdots, (x_N, \infty)$, we get a series of simpler integrals. In general we get

$$\begin{aligned}
\int_{-\infty}^{\infty} f(m) \exp(-\sum_{i=1}^{N}|x_i - m|)dm \quad &= \quad \exp(-\sum_{i=1}^{N} x_i) \int_{-\infty}^{x_1} f(m) \exp(Nm)dm + &(26)\\
&\exp(x_1 - \sum_{i=2}^{N} x_i) \int_{x_1}^{x_2} f(m) \exp((N-2)m)dm + &(27)\\
&\exp(\sum_{i=1}^{2} x_i - \sum_{i=3}^{N} x_i) \int_{x_2}^{x_3} f(m) \exp((N-4)m)dm + &(28)\\
&\cdots + \exp(\sum_{i=1}^{N} x_i) \int_{x_N}^{\infty} f(m) \exp(-Nm)dm &(29)
\end{aligned}$$

For $f(m) = 1$ the inner integrals are

$$\int_{-\infty}^{x_1} \exp(Nm)dm \quad = \quad \exp(Nx_1)/N \tag{30}$$

$$\int_{x_i}^{x_{i+1}} \exp((N-2i)m)dm \quad = \quad \begin{cases} \frac{\exp((N-2i)x_{i+1}) - \exp((N-2i)x_i)}{N-2i} & \text{if } N - 2i \neq 0 \\ x_{i+1} - x_i & \text{if } N - 2i = 0 \end{cases} \tag{31}$$

$$\int_{x_N}^{\infty} \exp(-Nm)dm \quad = \quad \exp(-Nx_N)/N \tag{32}$$

For $f(m) = m$ the inner integrals are

$$\int_{-\infty}^{x_1} m \exp(Nm)dm \quad = \quad \exp(Nx_1)\frac{Nx_1 - 1}{N^2} \tag{33}$$

$$\int_{x_i}^{x_{i+1}} \exp((N-2i)m)dm \quad = \quad \begin{cases} \frac{\exp((N-2i)x_{i+1})((N-2i)x_{i+1}-1) - \exp((N-2i)x_i)((N-2i)x_i - 1)}{(N-2i)^2} & \text{if } N - 2i \neq 0 \\ \frac{x_{i+1}^2 - x_i^2}{2} & \text{if } N - 2i = 0 \end{cases} \tag{34}$$

$$\int_{x_N}^{\infty} \exp(-Nm)dm \quad = \quad \exp(-Nx_N)\frac{Nx_N - 1}{N^2} \tag{35}$$