Study on overtraining

Luca
<piero.mana@ntnu.no>

Draft of 10 August 2018 (first drafted 6 August 2018)

1 Synopsis

The literature usually associates overtraining with overfitting and too large number of parameters.

The relation between these aspects is more subtle than that. In fact, overtraining can be a consequence of *underfitting*. We have a clearer view approaching this question from the point of view of probability theory.

Consider a classification problem: given input $x \in \{1, \dots, M\}$, we want to predict output $y \in \{1, \dots, N\}$. Let's consider both discrete.

Two cases must be distinguished: whether every input has several possible outputs, or just one. The second case is a special instance of the first, and could be simply treated as the prediction of a 'function' $x \mapsto y$. But both cases are instances of inference with a *partially exchangeable* model.

Our assumptions are summarized in the proposition I. We assume that the probability of y, for each kind of input x, is exchangeable. Thus the probability for several pairs

$$\mathrm{p}(y^{(T)}, \dots, y^{(1)} | \, x^{(T)}, \dots, x^{(1)}, I) \tag{1}$$

must have a partially exchangeable form [refs]: for every kind of input x we consider the N relative frequencies $\boldsymbol{f}_x \coloneqq (f_{1|x}, \dots, f_{N|x})$ of the N possible outputs. The probability above is given by this integral:

$$\mathrm{p}(y^{(T)}, \dots, y^{(1)} | \, x^{(T)}, \dots, x^{(1)}, I) = \int \cdots \int \Bigl[\prod_{x=1}^{M} \prod_{t}^{x^{(t)}=x} f_{y^{(t)}|x} \Bigr] \, \mathrm{p}(\boldsymbol{f}_1, \dots, \boldsymbol{f}_M | \, I) \, \mathrm{d}\boldsymbol{f}_1 \cdots \mathrm{d}\boldsymbol{f}_M . \tag{2}$$

The complicated look of this formula doesn't do justice to its simple and intuitive interpretation:

Choose a kind of input x. We're judging that the probability for a very large sequence of outputs generated with this input doesn't depend on their order. Suppose we knew the relative frequencies of each kind of output in this sequence, $(f_{1|x}, \ldots, f_{N|x}) =: \boldsymbol{f}_x$, and knew nothing else. Then out of symmetry reasons we should give a probability $f_{y|x}$ of observing outcome y at the next observation. This is just a 'drawing without replacement' problem. At the subsequent observations with the same input we give again the same probabilities to each y, because the sequence is so large that we approximate 'drawing without replacement' with 'drawing with replacement'. The probability for a sequence of outputs $y^{(1)}, \ldots, y^{(T)}$ from the same input x is then

$$\mathrm{p}(y^{(T)}, \ldots, y^{(1)} | \text{ same input } x, \boldsymbol{f}_x, I) = f_{y^{(T)}|x} \times \cdots \times f_{y^{(1)}|x}. \tag{3}$$

Within a larger sequence of outputs from all possible inputs, the outputs $y^{(()}t)$ coming from input x are those for which $x^{(()}t) = x$. Thus we can write the formula above more generally as

$$\prod_t^{x^{(t)}=x} f_{y^{(t)}|x} \tag{4}$$

The above reasoning applies for each kind of input x. Thus the probability for a sequence of outputs coming from different inputs is the product of the probabilities above for all different x:

$$\mathrm{p}(\text{all outputs} | \text{ inputs}, \boldsymbol{f}_1, \ldots, \boldsymbol{f}_M, I) = \prod_{x=1}^{M} \prod_t^{x^{(t)}=x} f_{y^{(t)}|x}. \tag{5}$$

Now what if we don't know the relative frequencies $\boldsymbol{f}_x$, for any input? Then we assign a probability distribution over their possible valuesand use the law of total probability:

$$\mathrm{p}(\text{all outputs} | \text{ inputs}, I) =$$

$$\int \cdots \int \mathrm{p}(\text{all outputs} | \text{ inputs}, \boldsymbol{f}_1, \ldots, \boldsymbol{f}_M, I)\, \mathrm{p}(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M | I)\, \mathrm{d}\boldsymbol{f}_1, \ldots, \mathrm{d}\boldsymbol{f}_M. \tag{6}$$

Substituting the explicit expression (5) in this formula we obtain formula (6). In summary,

$$\int\cdots\int \Bigg[\prod_{x=1}^{M} \prod_{t}^{x^{(t)}=x} f_{y^{(t)}|x} \Bigg] \, \mathrm{p}(f_1,\dots,f_M|\, I)\,\mathrm{d}f_1\cdots\mathrm{d}f_M \quad . \tag{7}$$

product over all inputs

probability for outputs from same input

uncertainty over all frequencies

The possible frequencies give one input, $f_x \equiv (f_{1|x},\dots,f_{N|x})$, belong to the $(N-1)$-dimensional simplex

$$\Delta := \{(f_1,\dots,f_N) \mid f_i \geqslant 0, \textstyle\sum_i f_i = 1\}, \tag{8}$$

and the collection of possible frequencies $(f_1,\dots,f_M)$ belongs to the M-fold Cartesian product Δ^M. From now on we denote $f := (f_1,\dots,f_M)$.

The probability for a new sequence of T' outputs given their inputs and given that we've learned a previous sequence of T input-output pairs is determined by Bayes's theorem:

$$\mathrm{p}(y^{(T')},\dots,y^{(T+1)}|\, x^{(T')},\dots,x^{(T+1)},\; y^{(T)},x^{(T)},\dots,y^{(1)},x^{(1)},\; I) = \frac{\displaystyle\int \Bigg[\prod_{x=1}^{M} \prod_{t=T+1,\dots,T'}^{x^{(t)}=x} f_{y^{(t)}|x}\Bigg]\, \mathrm{p}(f|\, I)\,\mathrm{d}f}{\displaystyle\int \Bigg[\prod_{x=1}^{M} \prod_{t=1,\dots,T}^{x^{(t)}=x} f_{y^{(t)}|x}\Bigg]\, \mathrm{p}(f|\, I)\,\mathrm{d}f}. \tag{9}$$

This formula is equivalent to (2) with and updated distribution for the frequencies:

$$\mathrm{p}(f|\, y^{(T)},x^{(T)},\dots,y^{(1)},x^{(1)},\; I)\,\mathrm{d}f = \frac{\big[\prod_{x=1}^{M} \prod_{t=1,\dots,T}^{x^{(t)}=x} f_{y^{(t)}|x}\big]\, \mathrm{p}(f|\, I)}{\int \big[\prod_{x=1}^{M} \prod_{t=1,\dots,T}^{x^{(t)}=x} f_{y^{(t)}|x}\big]\, \mathrm{p}(f|\, I)\,\mathrm{d}f}\,\mathrm{d}f. \tag{10}$$

The form of this updated distribution has important consequences for our learning process.

If the number of learned data is enough large compared with the numbers M, N of possible inputs and outputs and with the magnitude

of the initial distribution for the frequencies, and if the latter is strictly positive, then the updated distribution becomes very peaked on the collection of relative frequencies $(\boldsymbol{q}_1, \dots, \boldsymbol{q}_M)$ of the learned outputs for all input values. This can be seen from the asymptotic expression in terms of the relative entropy (Kullback-Leibler divergence) D,

$$\mathrm{p}(\boldsymbol{f}|\, y^{(T)}, x^{(T)}, \dots, y^{(1)}, x^{(1)},\ I) \propto \exp\bigl[-\textstyle\sum_x T_x D(\boldsymbol{q}_x|\, \boldsymbol{f}_x)\bigr]\, \mathrm{p}(\boldsymbol{f}|\, I), \quad (11)$$

where T_x is the number of observations with input x, with $\sum_x T_x = T$.

If the number of learned data is small compared with the dimensions of the input and output spaces, then the initial distribution for the frequencies $\mathrm{p}(\boldsymbol{f}|\, I)\,\mathrm{d}\boldsymbol{f}$ greatly influence our inference (9). This distribution determines two important ***

2 Utility functions and probabilities

Utility of behaving as if proposition $B \in X$ is true given that proposition $A \in X$ is true: $c(B|\, A)$. Probability for A given D, I: $\mathrm{P}(A|\, D, I)$. Optimal decision is B that maximizes

$$\sum_A c(B|\, A)\, \mathrm{P}(A|\, D, I). \quad (12)$$

Now consider different probabilities for all A given the same data D: $\mathrm{P}(A|\, D, I')$. The decision B will still be the same if we use a new utility

$$c'(B|\, A) \coloneqq c(B|\, A)\, \frac{\mathrm{P}(A|\, D, I)}{\mathrm{P}(A|\, D, I')}. \quad (13)$$

So the same choice can be made with a different probability, if the utility is appropriately changed, provided $p(A|\, D, I') > 0$ for all A.

This leads to a slightly more general view than Tishby et al.'s (**tishbyetal1989**; **levinetal1990**) and Mackay's (**mackay1992**; **mackay1992b**)