Bayesian Inference and Minimun Description Length

Michele Caselle – University of Torino and INFN
caselle@to.infn.it

Bayesian Inference

Bayesain inference is based on the Bayes Theorem

Posterior probability — Likelihood — Prior probability

$$p(A|B) = \frac{p(B|A)\,p(A)}{p(B)}$$

Which states that our belief (the prior probability) is updated to the posterior probability after we observe the data (likelihood)

Bayesian Inference

In machine learning the same theorem can be translated as follows

$$P(h|D) = \frac{P(D|h).P(h)}{P(D)}$$

Which states that we try to find which hypothesis h describes the data D, given the data.
In general the space of possible hypothesis is infinite and we want to maximize the probability that one particular h is most likely to originate the given data

Bayesian Inference

$$P(h|D) = \frac{P(D|h).P(h)}{P(D)}$$

This amount to find the "Maximum a Posteriori" value of h

$$h \;/\; \text{argmax}\,\{P(h|D)\}$$

Since P(D) does not depend on h we have

$$h_{MAP} = \text{argmax}\{P(D|h)P(h)\}$$

Minimum Description Length

From which we have:

$$
\begin{aligned}
h_{MAP} &= \arg\max P(D|h).P(h) \\
&= \arg\max \log_2(P(D|h).P(h)) \\
&= \arg\max [\log_2 P(D|h) + \log_2 P(h)] \\
&= \arg\min [-\log_2 P(D|h) - \log_2 P(h)]
\end{aligned}
$$

Where $\log_2 P(h)$ is the length of a code which describes the hypothesis (i.e fixes all the free parameters which define our model)

Minimum Description Length

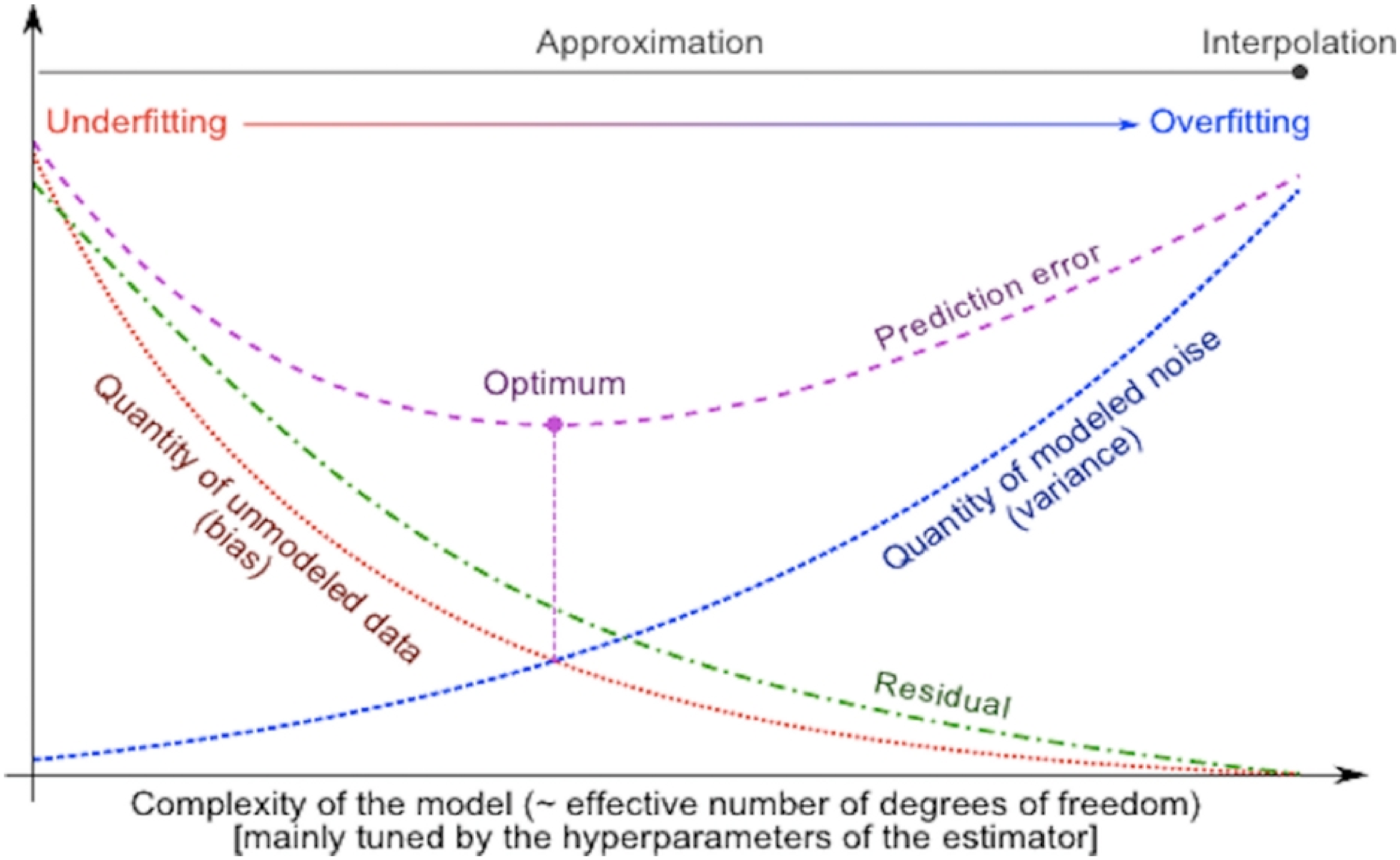
$$h_{MAP} = \arg\max P(D|h).P(h)$$
$$= \arg\max \log_2(P(D|h).P(h))$$
$$= \arg\max [\log_2 P(D|h) + \log_2 P(h)]$$
$$= \arg\min [- \log_2 P(D|h) - \log_2 P(h)]$$

While $\log_2 P(D|h)$ is the length of a code which describes the ability of the hypothesis to describe the data (i.e. the number of errors that we make when we use h to describe D).

There is a trade off between log2 P(h) and log2 P(D|h)

Minimum Description Length

Trade off between log2 P(h) and log2 P(D|h)



Minimum Description Length

The goal of typical inference tools is to minimize the description length

- when P(D|h) can be evaluated exactly using a variational method
- otherwise by using a Markov Chain Montecarlo Process

It can be used to perform community detection (INFOMAP) or topic modelling (LDA)