Supervised Learning from the Bayesian Viewpoint

An informal overview

Claudio Mattiussi

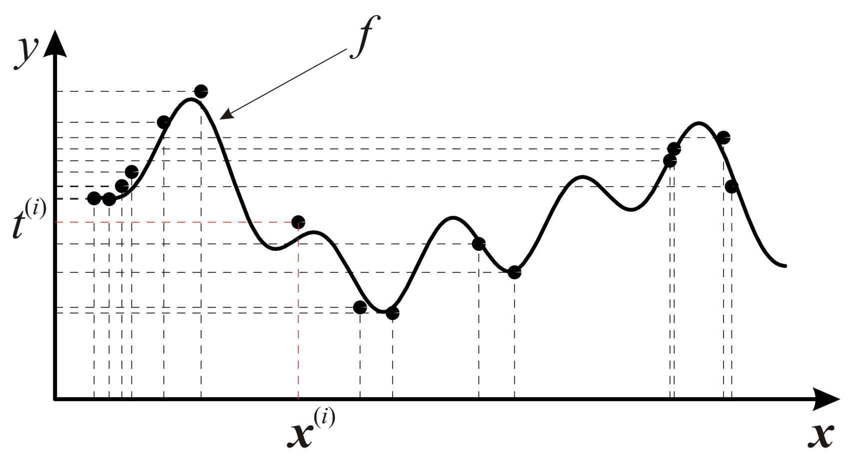
Laboratory of Intelligent Systems, EPFL

December 18, 2006

Supervised learning

A simple example (regression problem)

We would like to "learn" a function $f : \mathbf{x} \mapsto y$ given a *training set* $\mathcal{T} = \{(\mathbf{x}^{(i)}, t^{(i)})\}$, with $t^{(i)} = f(\mathbf{x}^{(i)}) + n^{(i)}$ (noisy samples of f)



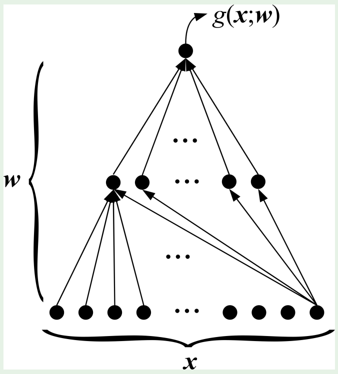
The conventional neural network (NN) viewpoint

The NN solution

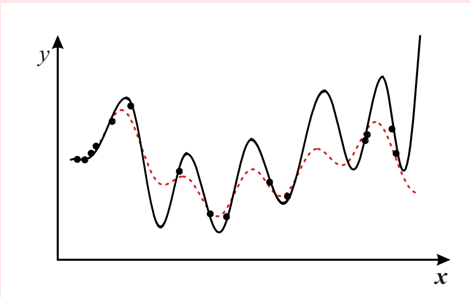
Define a NN with weights $\mathbf{w}$ which realizes the function $g(\mathbf{x};\mathbf{w})$, and define the *error function*

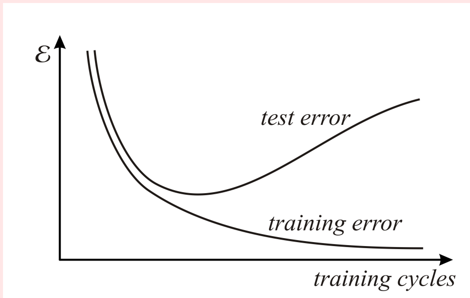
$$\mathcal{E}(\mathbf{w}) \propto \sum_i (t^{(i)} - g(\mathbf{x}^{(i)};\mathbf{w}))^2$$

then search for the vector of weights $\hat{\mathbf{w}}$ that minimizes the error and assume $g(\mathbf{x};\hat{\mathbf{w}})$ as the estimate of $f(\mathbf{x})$



but... beware of overfitting!





The conventional NN viewpoint (cont.)

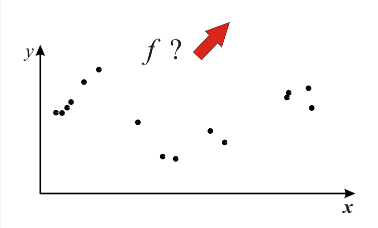
Some questions about the conventional approach:

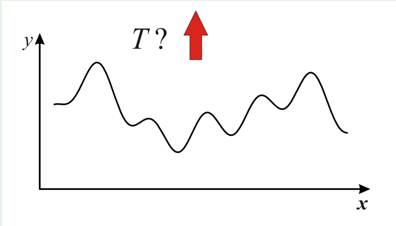
- Why this particular error function?
- Why the overfitting problem? What is "generalization"?
- How can you compare the performance of different network structures?

The Bayesian viewpoint

Use $\mathcal{T}$ to *update your degree of belief* about f

$$\underbrace{p(f\,|T,X,I)}_{\textit{posterior pdf}} \propto \underbrace{p(T\,|\,f,X,I)}_{\textit{likelihood}} \cdot \underbrace{p(f\,|I)}_{\textit{prior pdf}}$$





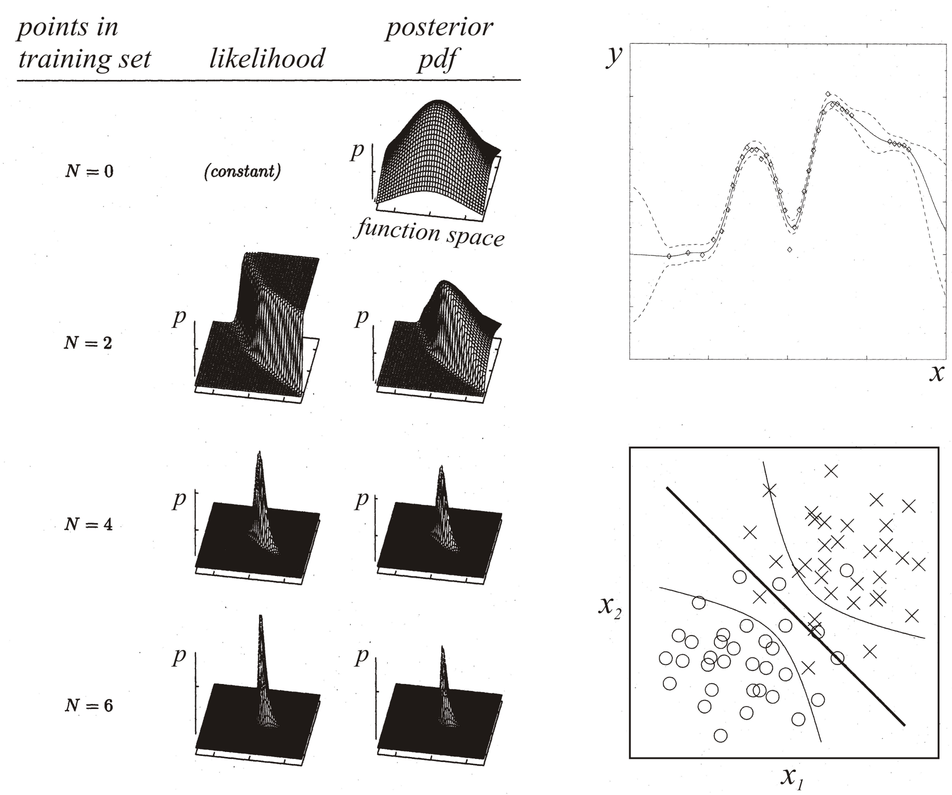
$$T = \{t^{(i)}\}$$
$$X = \{\mathbf{x}^{(i)}\}$$

Note that with this approach

- You need to make explicit your prior belief about f
- The result is a probability distribution over a space of functions, rather than a single function

The Bayesian viewpoint (cont.)

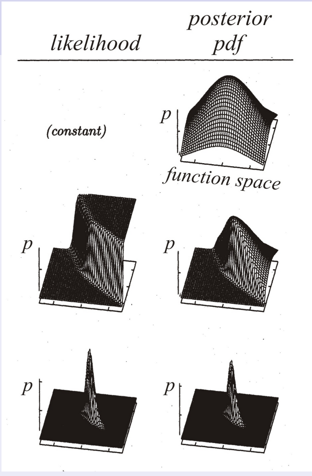
Examples (and discussion)



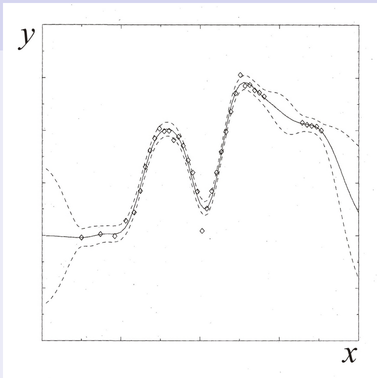
(adapted from [MacKay, 2003])

The Bayesian viewpoint (cont.)

Safeguard against "overfitting"



Better prediction model



(possibility of "active learning")

A digression on mathematical objects

There are typically two ways to define and consider a mathematical object:

- An intrinsic, coordinate/parameter-free way, that lets you *understand* what the object means and manipulate it *conceptually*
- A coordinate/parameter-based way that lets you represent and manipulate the object *practically* (but is seldom enlightening in itself)

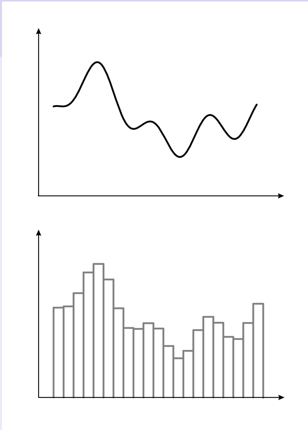
Examples

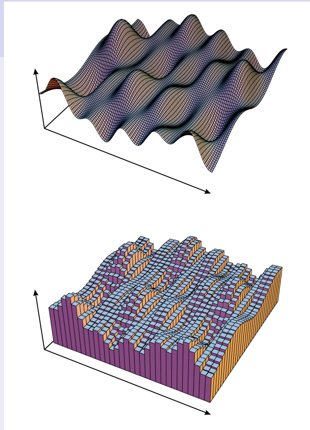
Vectors, matrices, determinants, the differential operators of mathematical physics (gradient, curl, divergence...), tensors, the objects of elementary geometry, the objects of the calculus of variations...

Representing functions

A naive approach

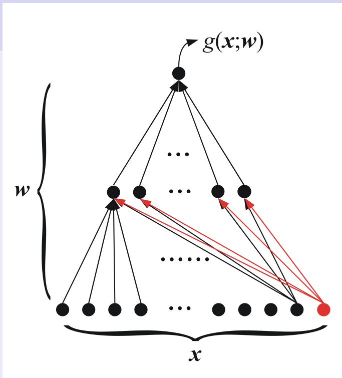
The "curse of dimensionality"





Neural networks

$$g(\mathbf{x};\mathbf{w}) = \varphi\Big(\sum_j w_{oj}\ \varphi(\sum_k w_{jk}\, x_k)\Big)$$



Supervised NN learning from the Bayesian viewpoint

Use $\mathcal{T}$ to update your degree of belief about $\mathbf{w}$

$$\underbrace{p(\mathbf{w}\,|T,X,I)}_{\textit{posterior pdf}} \;\propto\; \underbrace{p(T\,|\mathbf{w},X,I)}_{\textit{likelihood}} \;\cdot\; \underbrace{p(\mathbf{w}\,|I)}_{\textit{prior pdf}}$$

Likelihood for independent Gaussian noise

$$p(T\,|\mathbf{w},X,I) \;\propto\; \exp\left(-\sum_i \frac{(t^{(i)} - g(\mathbf{x}^{(i)};\mathbf{w}))^2}{2\sigma_i^2}\right)$$

Assigning the prior

$$p(\mathbf{w}\,|I) \;\propto\; \exp(-\beta \sum w^2)$$

(*weight decay* regularizer)

Other representations

- Use different sets or different combinations of "basis" functions
- Work directly in terms of probability distributions on the space of functions f

Gaussian Processes

It's one of the simplest types of probability distributions on spaces of functions (it generalizes the finite-dimensional Gaussian distribution)

The probability distribution $p(f(\mathbf{x})|..., I)$ is assigned by specifying the a *mean function* $\mu(\mathbf{x})$ and the *covariance function* $c(\mathbf{x}, \mathbf{x}')$

Many probability distributions on parametrized representations correspond to Gaussian processes

Model selection

Absolute plausibility

$$p(N_m \,|T, I) = \frac{p(T\,|N_m, I) \,\cdot\, p(N_m\,|I))}{P(T\,|I)}$$

$$p(N_m \,|T, I) = \frac{p(T\,|N_m, I) \,\cdot\, p(N_m\,|I))}{\sum_m P(T\,|N_m, I) \,\cdot\, p(N_m\,|I)}$$

Relative plausibility (model *comparison*)

$$\frac{p(N_{m1}\,|T, I)}{p(N_{m2}\,|T, I)} = \frac{p(T\,|N_{m1}, I)}{p(T\,|N_{m2}, I)} \cdot \frac{p(N_{m1}\,|I))}{p(N_{m2}\,|I))}$$

The evidence of the model

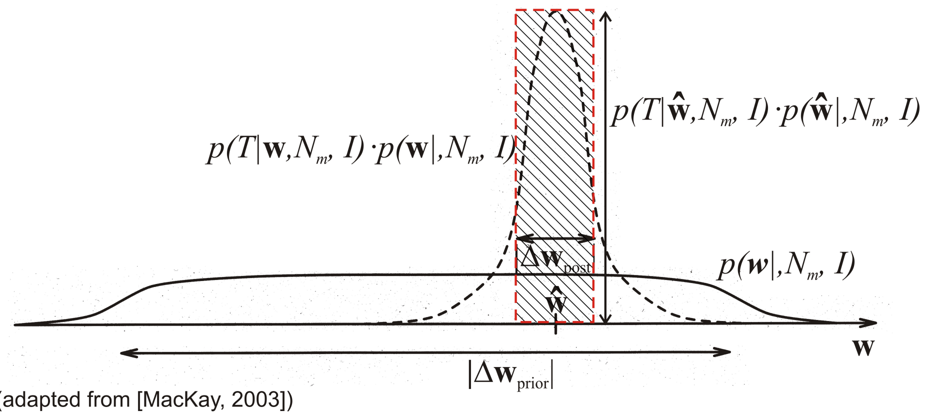
$$p(T\,|N_m, I) = \int_{\mathcal{W}} p(T\,|\mathbf{w}, N_m, I) \,\cdot\, p(\mathbf{w}\,|, N_m, I)\, \mathrm{d}\mathbf{w}$$

Approximating and understanding the evidence

$$p(T\,|N_m, I) = \int_{\mathcal{W}} p(T\,|\mathbf{w}, N_m, I)\,\cdot\, p(\mathbf{w}\,|, N_m, I)\, d\mathbf{w}$$

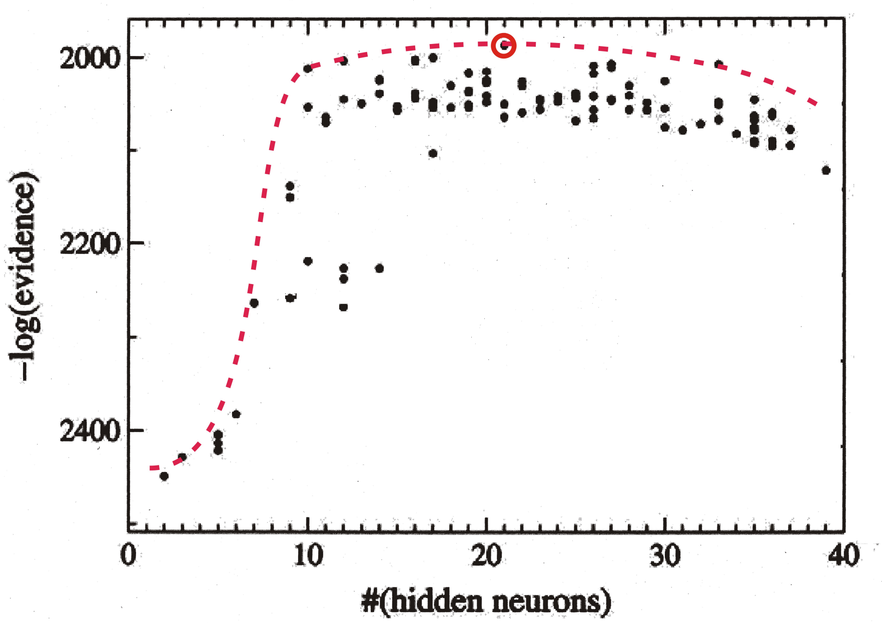
$$\underbrace{p(T\,|N_m, I)}_{\text{evidence}} \simeq \underbrace{p(T\,|\hat{\mathbf{w}}, N_m, I)}_{\text{best fit likelihood}} \cdot \underbrace{p(\hat{\mathbf{w}}\,|N_m, I) \cdot |\Delta\mathbf{w}_{post.}|}_{\text{Occam factor}}$$

$$\text{Occam factor} \simeq \frac{|\Delta\mathbf{w}_{post.}|}{|\Delta\mathbf{w}_{prior}|}$$



(adapted from [MacKay, 2003])

Example



(adapted from [Toussaint *et al*, 2006])

Closing comments...

The information paradox

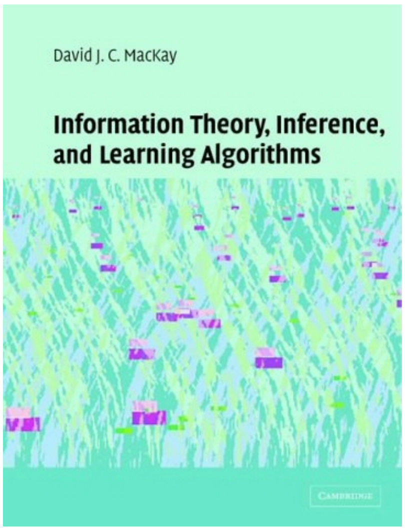
For a feed-forward NN with one hidden layer

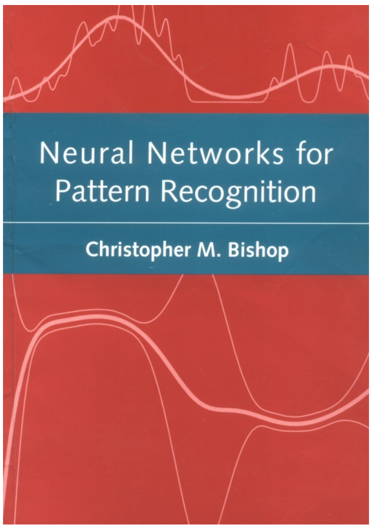
$$p(\mathbf{w}\,|T,X,\beta,I) \;\propto\; \exp\left(-\sum_i \frac{\left(t_i - \varphi\left(\sum_j w_{oj}\;\varphi\left(\sum_k w_{jk}\,x_k^{(i)}\right)\right)\right)^2}{2\sigma_i^2} - \beta\left(\sum_j w_{oj}^2 + \sum_{jk} w_{jk}^2\right)\right)$$

Further issues

- Handling complicated posterior pdfs (e.g., multiple peaks): numerical approximations, discarding negligible information...
- Probabilistic handling of hyperparameters
- Committees of networks
- Determination of input relevance
- ...

References





MacKay's book is available online at the author's website:
http://www.inference.phy.cam.ac.uk/itprnn/book.html