Aalto University

Probabilistic machine learning | Intro (D)

Introduction to machine learning

Francesco Corona

Chemical and Metallurgical Engineering
School of Chemical Engineering

Curve fitting

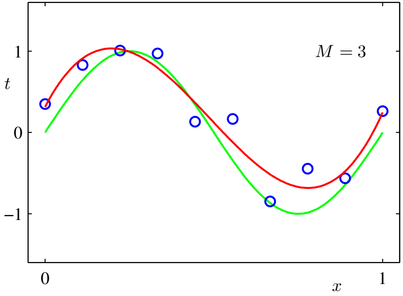
Intro (D)

Curve fitting, reloaded

The objective of regression is to estimate the value of one or more **continuous target variables** t, given the value of some D-dimensional vector x of **continuous input variables**

We assumed that the training data consists of $N = 10$ pairs (x_n, t_n) generated by a deterministic function ($\sin(2\pi x)$, green curve), plus a small amount of Gaussian noise

$$t = \sin(2\pi x) + \text{noise}$$



We fit the data $\{x_n, t_n\}$ using a polynomial

$$y(x|w) = \sum_{n_m=0}^{N_m} w_{n_m} x^{n_m}$$

The parameters $\{w^\star_{n_m}\}_{n_m=1}^{N_M}$ were obtained by minimising the usual sum of the squares

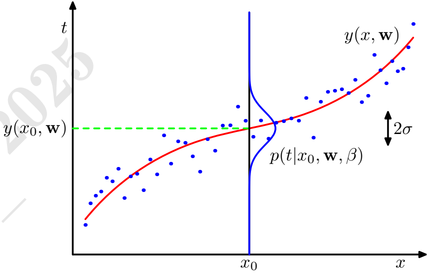
$$E(w) = (1/2) \sum_{n=1}^{N} \Big(y(x_n|w) - t_n \Big)^2$$

Implicitly, our regression model accepted the existence of an **uncertainty over the targets**

Curve fitting, reloaded (cont.)

Given x, t is assumed to be a Gaussian with mean $y(x|w)$ and some precision $\beta = 1/\sigma^2$

$$p(t|x, w, \beta) = \frac{1}{(2\pi\beta^{-1})^{1/2}} \exp\left(-\frac{\beta}{2}(t - y(x|w))^2\right)$$



Assuming our data $\{t_n\}$ are independent draws from $p(t|x, w, \beta)$, the **likelihood function**

$$\begin{aligned} p(\{t_n\}|\{x_n\}, w, \beta) &= \prod_{n=1}^{N} p(t_n|y(x_n|w), \beta^{-1}) \\ &= \prod_{n=1}^{N} \frac{1}{(2\pi\beta^{-1})^{1/2}} \exp\left(-\frac{\beta}{2}(t_n - y(x_n|w))^2\right) \end{aligned}$$

Notice that this is also the (conditional) distribution of the data, given the parameters

- This is a product of (independent) Gaussians and thus it is also a Gaussian

Curve fitting, reloaded (cont.)

To estimate the coefficients $\{w_{n_m}\}$ of the polynomial, we can minimise the negative log-likelihood with respect to w, under the assumption that the precision β is known

$$-\ln\left(p(\{t_n\}|\{x_n\}, w, \beta)\right) = \underbrace{\frac{\beta}{2}\sum_{n=1}^{N}\left(y(x_n|w) - t_n\right)^2}_{E(w)} + \underbrace{\frac{N}{2}\left(\ln(2\pi) - \ln(\beta) +\right)}_{\text{constant}}$$

We obtain a maximum likelihood estimate w_{ML}, regardless of the assumed precision β

We can also use minimise the negative log-likelihood to estimate the data precision β

$$\frac{1}{\beta^{ML}} = \frac{1}{N}\sum_{n=1}^{N}\left(t_n - y(x_n|w_{ML})\right)^2$$

Curve fitting, reloaded (cont.)

Having determined estimates of w and β, we can make predictions t for new values x

$$y(x|w^{ML}) = \sum_{n_m=0}^{N_m} w_{n_m}^{ML} x^{n_m}$$

However, we have a first probabilistic model, predictions can be more comprehensive

Rather than being content with point-estimates of t only, the **predictive distribution**

$$p(t|x, w, \beta) = \frac{1}{(2\pi\beta_{ML}^{-1})^{1/2}} \exp\left(-\frac{\beta_{ML}}{2}(t - y(x|w_{ML}))^2\right)$$

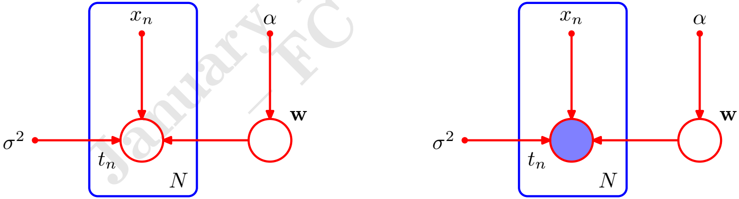
Curve fitting | Bayesian treatment

We can make a step towards a Bayesian treatment of curve fitting, by introducing a **prior distribution** over the parameters w of the polynomial model chosen for regression

Before seeing the data $\{x_n, t_n\}$, let us assume w are from a round zero-mean Gaussian

$$p(w|0, \alpha^{-1}I) = \left(\frac{\alpha}{2\pi}\right)^{(N_w+1)/2} \exp\left(-\frac{\alpha}{2}\, w^T w\right)$$

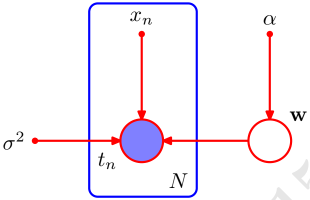
The probabilistic model for curve fitting contains N Gaussian random variables $\{t_n\}$ which are independent and N_M random variables $\{w_{n_m}\}$ which are jointly Gaussian



The conditional probability of $\{t_n\}$ given the hidden parameters $\{w\}$, or the likelihood

$$p(\{t_n\}|\{x_n\}, \beta, w) = \prod_{n=1}^{N} \frac{1}{(2\pi\beta^{-1})^{1/2}} \exp\left(-\frac{\beta}{2}(t_n - y(x_n|w))^2\right)$$

Curve fitting | Bayesian treatment (cont.)



Random variables

- $\{t_n\}$, i.i.d. from a Gaussian
- $\{w_{n_m}\}$, jointly Gaussian

Hyper-parameters

- $\beta = 1/\sigma^2$
- α

Thence, the probabilistic model for curve regression is the joint distribution $p(\{t_n\}, w)$

$$p(\{t_n\}, w) = \underbrace{\prod_{n=1}^{N} \frac{1}{(2\pi\beta^{-1})^{1/2}} \exp\left(-\frac{\beta}{2}(t_n - y(x_n|w))^2\right)}_{p(\{t_n\}|\{x_n\},\beta,w)} \underbrace{\left(\frac{\alpha}{2\pi}\right)^{(N_w+1)/2} \exp\left(-\frac{1}{2}\, w^T \alpha w\right)}_{p(w|\alpha)}$$

Using the Bayes rule, we can establish the **posterior probability** of w, given data $\{t_n\}$

$$p(w|\{t_n\}, \{x_n\}, \beta, \alpha) = \frac{p(\{t_n\}|\{x_n\}, \beta, w)p(w|\alpha)}{p(\{t_n\}|\{x_n\}, \beta, \alpha)}$$

Curve fitting | Bayesian treatment | Maximum a posteriori

$$p(w|\{t_n\},\{x_n\},\beta,\alpha) = \frac{p(\{t_n\}|\{x_n\},\beta,w)p(w|\alpha)}{p(\{t_n\}|\{x_n\},\beta,\alpha)}$$

$$= \frac{\underbrace{p\{t_n\}|\{x_n\},\beta,w)}_{\text{likelihood}}\underbrace{p(w|\alpha)}_{\text{prior}}}{\underbrace{\int p\{t_n\}|\{x_n\},\beta,w)p(w|\alpha)dw}_{\text{marginal likelihood}}}$$

For arbitrary distributions, computing **marginal likelihoods** $p(\{t_n\})$ not straightforward

- This, in general, prevents us from determining the posterior distribution of w
- (Not true with our Gaussians, though we save this for the next slides)

For the moment, we can be content with an (point-) estimate $w^\star$ which is the most probable value of w given the data $\{t_n\}$, the maximiser of the un-normalised posterior

Or, equivalently, the minimiser of the negative of the log of the likelihood-prior product

$$\underbrace{\frac{\beta}{2}\sum_{n=1}^{N}\Big(y(x_n|w)-t_n\Big)^2+\frac{\alpha}{2}w^Tw}_{\widetilde{E}(w|\lambda)\text{ with }\lambda=\alpha/\beta}+\text{constant}(\beta)$$

The technique is known as **maximum a posteriori** (**MAP**) and we already encountered it

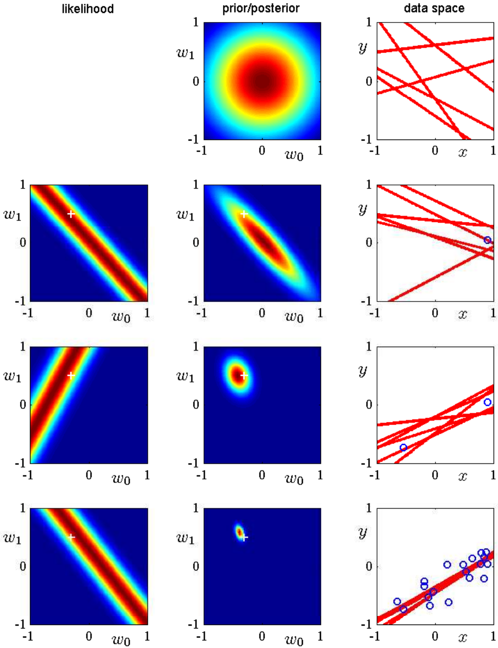
Curve fitting | Bayesian | Posterior

Due to the choice of prior distribution $p(w|0,\alpha^{-1}I)$, a Gaussian, and the assumed likelihood function $p(\{t_n\}|\{x_n\},\beta,w)$, also the **posterior** $p(w|\{t_n\},\{x_n\},\beta,\alpha)$ is a Gaussian

$$p(w|\{t_n\},\{x_n\},\beta,\alpha) = \frac{1}{(2\pi)^{N_w/2}}\frac{1}{|\Sigma_N|^{1/2}}\exp\left(-\frac{1}{2}(w-\mu_N)^T\frac{1}{\Sigma_N}(w-\mu_N)\right)$$

The mean vector $\mu_N = \beta\Sigma_N\underline{\mathrm{x}}^T\underline{\mathrm{t}}$ and covariance matrix $\Sigma_N^{-1} = \alpha I + \beta\underline{\mathrm{x}}^T\underline{\mathrm{x}}$ of the posterior are given as function of the observed data $\underline{\mathrm{x}} = (x_1,\ldots,x_N)$ and $\underline{\mathrm{t}} = (t_1,\ldots,t_N)$

Had we selected an arbitrary Gaussian $p(w|\mu_0,\Sigma_0)$ as parameter prior, the posterior Gaussian would have had mean $\mu_N = \Sigma_N\left(\Sigma_0^{-1}\mu_0 + \beta\underline{\mathrm{x}}^T\underline{\mathrm{t}}\right)$ and $\Sigma_N^{-1} = \Sigma_0^{-1} + \beta\underline{\mathrm{x}}^T\underline{\mathrm{x}}$



Example

$$y(x|w) = w_0 + w_1 x$$

Fixed variance of the noise

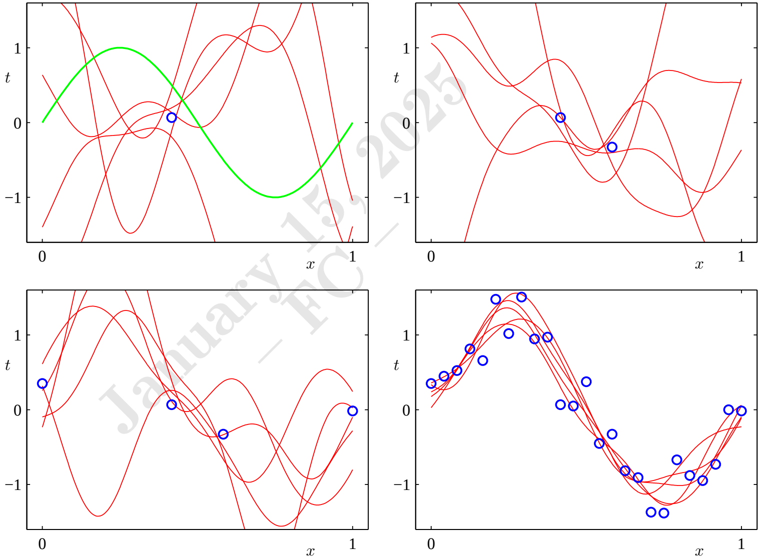
$$\beta = (1/0.2)^2$$

Gaussian prior over w

$$p(w|0, \alpha^{-1} I)$$

with fixed precision $\alpha = 2$

Curve fitting | Bayesian treatment | Posteriors (cont.)



Draws from the posterior distribution of w, for data sets $\{(x_n, t_n)\}_{n=1}^{N}$ of varying size

Curve fitting | Bayesian treatment | Posterior (cont.)

In a fully Bayesian treatment of curve fitting, we could introduce priors also over hyper-parameters α and β and make predictions by marginalising with respect to (w, α, β)

$$p(w, \beta, \alpha | \{t_n\}, \{x_n\}, \theta) = \frac{p(\{t_n\} | \{x_n\}, \beta, w, \alpha, \theta) p(w, \alpha, \beta | \theta)}{p(\{t_n\} | \{x_n\}, \theta)}$$

This can be achieved by introducing two **hyper-priors** $p(\alpha | \theta_\alpha)$ and $p(\beta | \theta_\beta)$ for α and β

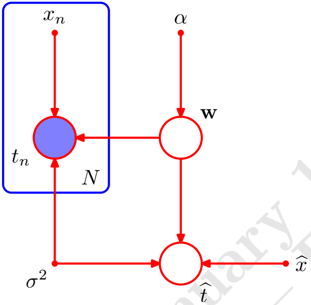
- Marginalising parameters w and hyper-parameters α and β is un-tractable

We must to resort to approximate inference using **sampling** or **empirical Bayes** methods

- Empirical Bayes is also known as type-2, or generalised, maximum likelihood
- The same technique is also often denoted to as evidence approximation

Curve fitting | Bayesian treatment | Predictive distribution

While having at our disposal a posterior for w is fundamental and allows us to extract point-estimates $w^\star$, the ultimate goal is to make predictions $\widehat{t}$ for yet unseen values $\widehat{x}$



Random variables

- $\{t_n\}$ and $\widehat{t}$, i.i.d. from a Gaussian
- $\{w_{n_m}\}$ are jointly Gaussian

Hyper-parameters

- $\beta = 1/\sigma^2$
- α

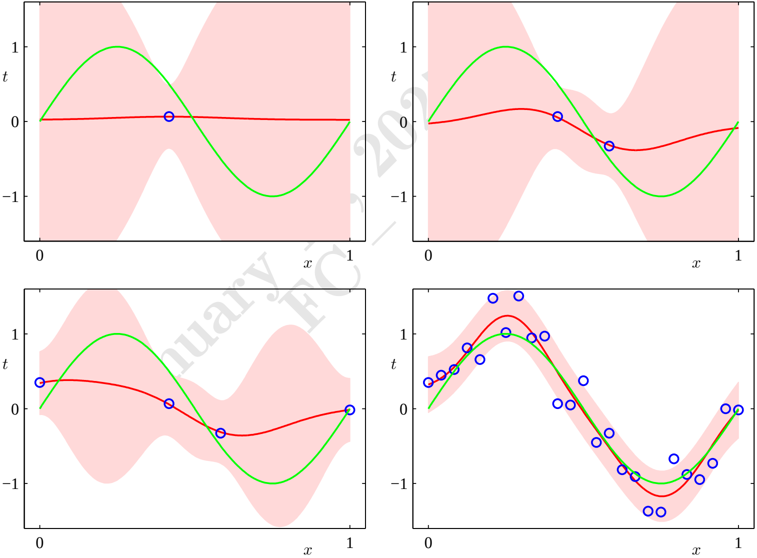
$$p(\widehat{t}, w) = p(\widehat{t}|\widehat{x}, \beta, w)p(w|\{t_n\}, \{x_n\}, \alpha, \beta)$$

This requires determining the **predictive distribution** $p(\widehat{t}|\widehat{x}, \alpha, \beta)$ at the new unseen $\widehat{x}$

$$p(\widehat{t}|\{t_n\}, \{x_n\}, \widehat{x}, \alpha, \beta) = \int p(\widehat{t}|\widehat{x}, w, \beta)p(w|\{t_n\}, \{x_n\}, \alpha, \beta)dw$$
$$= \frac{1}{(2\pi\beta_N^{-1}(\widehat{x}))^{1/2}}\exp\left(-\frac{\beta_N(\widehat{x})}{2}(\widehat{t} - \mu_N^T\widehat{x})^2\right)$$

This is yet another Gaussian, now with mean $\mu_N^T\widehat{x}$ and variance $\sigma_N^2(\widehat{x}) = 1/\beta + \widehat{x}^T\Sigma_N\widehat{x}$

Curve fitting | Bayesian treatment | Predictive distribution (cont.)



The predictive distribution (shaded area, one standard deviation) and its mean function

Linear models for regression

Intro (D)

Linear models for regression

Simple linear models for regression involve a linear combination of inputs $x \in \mathcal{X} \subseteq \mathcal{R}^{N_x}$

Linear models for regression

$$
\begin{aligned}
y(x|w) &= w_0(1) + w_1 x_1 + \cdots + w_{n_x} x_{n_x} + \cdots + w_{N_x} x_{N_x} \\
&= w_0(1) + \sum_{n_x=1}^{N_x} w_{n_x} x_{n_x}
\end{aligned}
$$

The central property of this class of model is its linearity in the parameters $\{w_{n_x}\}_{n_x=0}^{N_x}$

- However, it is also linear in the inputs $\{x_{n_x}\}_{n_x=1}^{N_x}$ and thus potentially limited

Linear basis function models

We can extend this simple class of models while, importantly, retaining its main properties, by introducing linear combinations of N_b nonlinear functions $\varphi : \mathcal{X} \rightarrow \mathcal{T}$ ($\mathcal{T} \subseteq \mathcal{R}$)

- They render the model $y(x|w)$ a nonlinear function of the inputs
- These functions $\{\varphi_{n_b}\}_{n_b=1}^{N_b}$ are often denoted as **basis functions**

$$
\begin{aligned}
y(x|w) &= w_0 \underbrace{(1)}_{\varphi_0} + \sum_{n_b=1}^{N_b} w_{n_b} \varphi_{n_b}(x_1, x_2, \ldots, x_{N_x}) \\
&= w^T \varphi(x)
\end{aligned}
$$

With $\varphi = (\varphi_0, \ldots, \varphi_{N_b})$, it is linear function in the parameters $w = (w_0, w_1, \ldots, w_{N_b})$

Linear models for regression (cont.)

The **polynomials** chosen for modelling the curve fitting task are an example of univariate basis functions in which the basis functions are given by the powers of x, $\varphi_{n_b}(x) = x^{n_b}$

$$y(x|w) = w_0 \underbrace{x^0}_{\varphi_0} + w_1 \underbrace{x^1}_{\varphi_1} + \cdots + w_{n_b} \underbrace{x^{n_b}}_{\varphi_{n_b}} + \cdots + w_{N_b} \underbrace{x^{N_b}}_{\varphi_{N_b}}$$

The model extends to inputs x of any dimensionality N_x, polynomials of any order N_m

$$\begin{aligned} y(x|w) &= w_{00} x_1^0 x_2^0 + w_{10} x_1^1 x_2^0 + w_{01} x_1^0 x_2^1 + w_{21} x_1^2 x_2^1 + w_{12} x_1^1 x_2^2 + w_{22} x_1^2 x_2^2 \\ &= w_{00} \underbrace{(1)}_{\varphi_0} + w_{10} \underbrace{x_1}_{\varphi_1} + w_{01} \underbrace{x_2}_{\varphi_2} + w_{21} \underbrace{x_1^2 x_2}_{\varphi_3} + w_{12} \underbrace{x_1 x_2^2}_{\varphi_4} + w_{22} \underbrace{x_1^2 x_2^2}_{\varphi_5} \end{aligned}$$

One limitation of polynomials is that they are global functions of x, they lack locality

- Locality can be embedded by using local polynomials like **spline functions**

There exist many possible alternatives for the basis functions, some are commonly used

Gaussian bumps

$$\varphi_{n_b}(x|\mu_{n_b}, \Sigma_{n_b}) = \exp\Big(-(x - \mu_{n_b})^T \Sigma_{n_b}^{-1} (x - \mu_{n_b})\Big)$$

Sigmoids

$$\varphi_{n_b}(x_{n_x}|\mu_{n_b}, \sigma_{n_b}) = \frac{1}{1 + \exp\big(-\dfrac{x_{n_x} - \mu_{n_b}}{\sigma_{n_b}}\big)}$$

Linear models for regression (cont.)

Independent of the choice of basis functions, our discussion on curve fitting allows us to seamless export that theory to the treatment of arbitrary linear models for regression

$$t = \underbrace{w^T\varphi(x)}_{y(x|w)} + \varepsilon$$

Since we assumed that the noise $\varepsilon \sim p(\varepsilon|0, \beta^{-1})$, we have $p(t|x, w, \beta) = p(t|y(x|w), \beta^{-1})$

Given sample $\{(x_n, t_n)\}$ of examples assumed to be independent draws from $p(t|x, w, \beta)$, we determine the conditional probability of the data $\{t_n\}$ given model parameters w, β

$$p(\{t_n\}|\{x_n\}, w, \beta) = \prod_{n=1}^{N} p(t_n|w^T\varphi(x_n), \beta^{-1})$$

Viewing this conditional as likelihood function yields us point-estimates w_{ML} and β_{ML}

$$w_{ML} = \left(\Phi^T\Phi\right)^{-1}\Phi^T\underline{t}$$

$$\frac{1}{\beta_{ML}} = \frac{1}{N(-1)}\sum_{n=1}^{N}\left(t_n - w_{ML}^T\varphi(x_n)\right)^2$$

in which we introduced Φ to be the usual $(N \times N_b)$ design matrix $\Phi = (\varphi_{n_b}(x_n))_{\substack{n=1,\dots,N \\ n_b=1,\dots,N_b}}$

Linear models for regression (cont.)

By introducing of a Gaussian prior distribution $p(w|\mu_0, \Sigma_0)$ over parameters w yield us both its maximum a posteriori point-estimate w_{MAP} and its posterior distribution:

$$w_{MAP} = \left(\Phi^T \Phi + \underbrace{\lambda}_{(\alpha/\beta}\ I\right)^{-1} \Phi^T \underline{t}, \qquad \text{(only if } \mu_0 = 0 \text{ and } \Sigma_0 = \alpha^{-1} I)$$

$$p(w|\{t_n\}, \{x_n\}\beta, \alpha) = \frac{1}{(2\pi)^{N_b/2}} \frac{1}{|\Sigma_N|^{1/2}} \exp\Big(- \frac{1}{2}(w - \mu_N)^T \Sigma_N^{-1} (w - \mu_N)\Big)$$

where mean $\mu_N = \Sigma_N \left(\Sigma_0^{-1}\mu_0 + \beta \Phi^T \underline{t}\right)$ and covariance matrix $\Sigma_N^{-1} = \Sigma_0^{-1} + \beta\Phi^T\Phi$

What remains to be defined is the predictive distribution over t for unseen values of x

$$p(\widehat{t}|\{t_n\}, \{x_n\}, \widehat{x}, \alpha, \beta) = \frac{1}{(2\pi\beta_N^{-1}(\widehat{x}))^{1/2}} \exp\left(-\frac{\beta_N(\widehat{x})}{2}(\widehat{t} - \mu_N^T\phi(\widehat{x}))^2\right)$$

in which mean equals to $\mu_N^T \phi(\widehat{x})$ and the variance equals to $\sigma_N^2(\widehat{x}) = 1/\beta + \phi(\widehat{x})^T \Sigma_N \phi(\widehat{x})$