Linear regression

Course of Machine Learning
Master Degree in Computer Science

University of Rome “Tor Vergata”

a.a. 2020-2021

Giorgio Gambosi

Linear models

- Linear combination of input features

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_D x_D$$

  with $\mathbf{x} = (x_1, \ldots, x_D)$
- Linear function of parameters $\mathbf{w}$
- Linear function of features $\mathbf{x}$

More compactly,

$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \overline{\mathbf{x}}$$

where $\overline{\mathbf{x}} = (1, x_1, \ldots, x_D)$

Base functions

- Extension to linear combination of base functions $\phi_1, \ldots, \phi_M$ defined on $\mathbb{R}^D$

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{M} w_j \phi_j(\mathbf{x})$$

- Each vector $\mathbf{x}$ in $\mathbb{R}^D$ is mapped to a new vector in $\mathbb{R}^M$, $\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_M(\mathbf{x}))$
- the problem is mapped from a D-dimensional to a M-dimensional space (usually with $M > D$)

Base functions

- Many types:
  - Polynomial (global functions)

$$\phi_j(x) = x^j$$

  - Gaussian (local)

$$\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{2s^2}\right)$$

  - Sigmoid (local)

$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right) = \frac{1}{1+e^{-\frac{x-\mu_j}{s}}}$$

  - Hyperbolic tangent (local)

$$\phi_j(x) = \tanh(x) = 2\sigma(x) - 1 = \frac{1-e^{-\frac{x-\mu_j}{s}}}{1+e^{-\frac{x-\mu_j}{s}}}$$

Base functions

Observe that a set of items (extended by 1 values)

$$\overline{\mathbf{X}} = \begin{pmatrix} - & \overline{\mathbf{x}}_1 & - \\ & \vdots & \\ - & \overline{\mathbf{x}}_2 & - \end{pmatrix} \overline{\mathbf{x}}_N = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1D} \\ 1 & x_{21} & \cdots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{ND} \end{pmatrix}$$

is transformed into

$$\mathbf{\Phi} = \begin{pmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \cdots & \phi_M(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \cdots & \phi_M(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_N) & \phi_2(\mathbf{x}_N) & \cdots & \phi_M(\mathbf{x}_N) \end{pmatrix}$$

Maximum likelihood and least squares

- Assume an additional gaussian noise

$$t = y(\mathbf{x}, \mathbf{w}) + \varepsilon$$

  with

$$p(\varepsilon) = \mathcal{N}(\varepsilon | 0, \sigma^2)$$

- Then,

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \sigma^2)$$

  and the expectation of the conditional distribution is

$$E[t|\mathbf{x}] = \int tp(t|\mathbf{x})dt = y(\mathbf{x}, \mathbf{w})$$

Maximum likelihood and least squares

- The likelihood of a given training set $\mathbf{X}, \mathbf{t}$ is

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(t_i|\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i), \sigma^2)$$

- The corresponding log-likelihood is then

$$\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \sum_{i=1}^{N} \ln \mathcal{N}(t_i|\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i), \sigma^2) = N \ln \sigma - \frac{N}{2} \ln(2\pi) - \frac{1}{\sigma^2} E_D(\mathbf{w})$$

  where

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} \Big(t_i - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i)\Big)^2 = \frac{1}{2} (\mathbf{\Phi w} - \mathbf{y})^T (\mathbf{\Phi w} - \mathbf{y})$$

Maximum likelihood and least squares

- Maximizing the log-likelihood w.r.t. $\mathbf{w}$ is equivalent to minimizing the error function $E_D(\mathbf{w})$
- Maximization performed by setting the gradient to 0

$$\mathbf{0} = \frac{\partial}{\partial \mathbf{w}} \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \sum_{i=1}^{N} \left(t_i - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i)\right) \boldsymbol{\phi}(\mathbf{x}_i)^T$$

$$= \sum_{i=1}^{N} t_i \boldsymbol{\phi}(\mathbf{x}_i)^T - \mathbf{w}^T \left(\sum_{i=1}^{N} \boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_i)^T\right)$$

- Which results into the normal equations for least squares

$$\mathbf{w}_{ML} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{t}$$

Gradient descent

- The minimum of $E_D(\mathbf{w})$ can be computed numerically, by means of gradient descent methods
- Initial assignment $\mathbf{w}^{(0)} = (w_0^{(0)}, w_1^{(0)}, \ldots, w_D^{(0)})$, with a corresponding error value

$$E_D(\mathbf{w}^{(0)}) = \frac{1}{2}\sum_{i=1}^{N}\left(t_i - (\mathbf{w}^{(0)})^T\boldsymbol{\phi}(\mathbf{x}_i)\right)^2$$

- Iteratively, the current value $\mathbf{w}^{(i-1)}$ is modified in the direction of steepest descent of $E_D(\mathbf{w})$, that is the one corresponding to the negative of the gradient evaluated at $\mathbf{w}^{(i-1)}$
- At step i, $w_j^{(i-1)}$ is updated as follows:

$$w_j^{(i)} := w_j^{(i-1)} - \eta \left.\frac{\partial E_D(\mathbf{w})}{\partial w_j}\right|_{\mathbf{w}^{(i-1)}}$$

Gradient descent

- In matrix notation:

$$\mathbf{w}^{(i)} := \mathbf{w}^{(i-1)} - \eta \frac{\partial E_D(\mathbf{w})}{\partial \mathbf{w}}\Big|_{\mathbf{w}^{(i-1)}}$$

- By definition of $E_D(\mathbf{w})$:

$$\mathbf{w}^{(i)} := \mathbf{w}^{(i-1)} - \eta(t_i - \mathbf{w}^{(i-1)}\boldsymbol{\phi}(\mathbf{x}_i))\boldsymbol{\phi}(\mathbf{x}_i)$$

Regularized least squares

- Regularization term in the cost function

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

  $E_D(\mathbf{w})$ dependent from the dataset (and the parameters), $E_W(\mathbf{w})$ dependent from the parameters alone.
- The regularization coefficient controls the relative importance of the two terms.

Regularized least squares

- Simple form

$$E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} = \frac{1}{2}\sum_{i=0}^{M-1} w_i^2$$

- Sum-of squares cost function: weight decay

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\{t_i - \mathbf{w}^T\phi(\mathbf{x}_i)\}^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} = \frac{1}{2}(\mathbf{\Phi w} - \mathbf{y})^T(\mathbf{\Phi w} - \mathbf{y}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

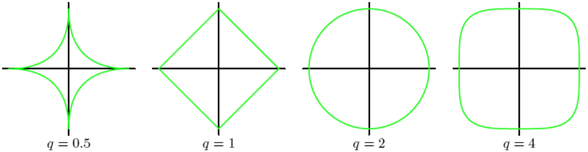
with solution

$$\mathbf{w} = (\lambda\mathbf{I} + \mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{t}$$

Regularization

- A more general form

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\{t_i - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)\}^2 + \frac{\lambda}{2}\sum_{j=0}^{M-1}|w_j|^q$$



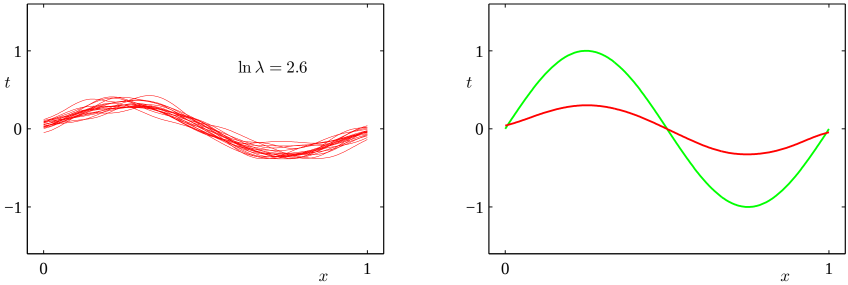
- The case $q = 1$ is denoted as lasso: sparse models are favored

Bias vs variance: an example

- Consider the case of function $y = \sin 2\pi x$ and assume $L = 100$ training sets $\mathcal{T}_1, \ldots, \mathcal{T}_L$ are available, each of size $n = 25$.
- Given $M = 24$ gaussian basis functions $\phi_1(x), \ldots, \phi_M(x)$, from each training set $\mathcal{T}_i$ a prediction function $y_i(x)$ is derived by minimizing the regularized cost function

$$E_D(\mathbf{w}) = \frac{1}{2}(\mathbf{\Phi w} - \mathbf{t})^T(\mathbf{\Phi w} - \mathbf{t}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

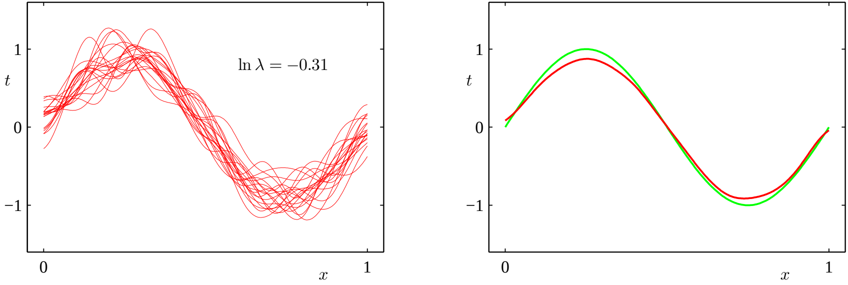
An example



Left, a possible plot of prediction functions $y_i(\mathbf{x})$ $(i = 1, \ldots, 100)$, as derived, respectively, by training sets $\mathcal{T}_i, i = 1, \ldots, 100$ setting $\ln\lambda = 2.6$. Right, their expectation, with the unknown function $y = \sin 2\pi x$.

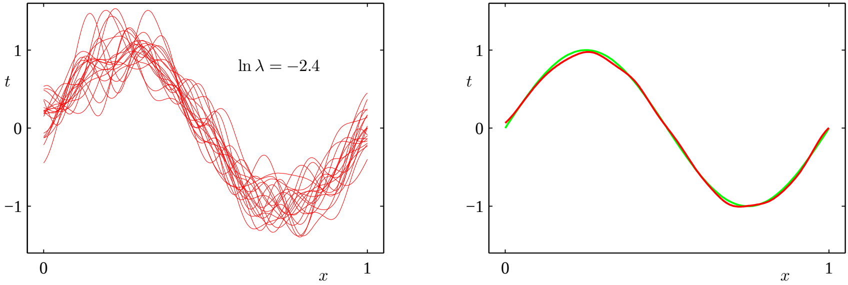
The prediction functions $y_i(\mathbf{x})$ do not differ much between them (small variance), but their expecation is a bad approximation of the unknown function (large bias).

An example



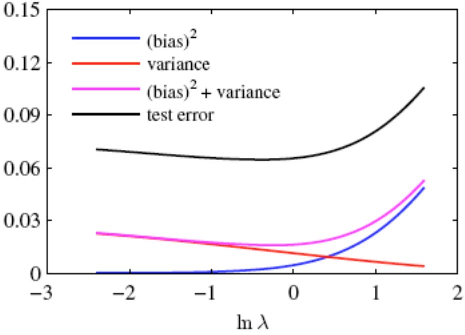
Plot of the prediction functions obtained with $\ln \lambda = -0.31$.

An example



Plot of the prediction functions obtained with $\ln\lambda = -2.4$. As λ decreases, the variance increases (prediction functions $y_i(\mathbf{x})$ are more different each other), while bias decreases (their expectation is a better approximation of $y = \sin 2\pi x$).

An example



- Plot of (bias)^2, variance and their sum as unctions of λ: las λ increases, bias increases and varinace decreases. Their sum has a minimum in correspondance to the optimal value of λ.
- The term $E_{\mathbf{x}}[\sigma^2_{y|\mathbf{x}}]$ shows an inherent limit to the approximability of $y = \sin 2\pi x$.

Bayesian approach to regression

- Applying maximum likelihood to determine the values of model parameters is prone to overfitting: need of a regularization term $\mathcal{E}(\mathbf{w})$.
- In order control model complexity, a bayesian approach assumes a prior distribution of parameter values.

Prior distribution

Posterior proportional to prior times likelihood: likelihood is gaussian (gaussian noise).

$$p(\mathbf{t}|\mathbf{\Phi}, \mathbf{w}, \beta) = \prod_{i=1}^{n} \mathcal{N}(t_i|\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i), \beta^{-1})$$

Conjugate of gaussian is gaussian: choosing a gaussian prior distribution of $\mathbf{w}$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$$

results into a gaussian posterior distribution

$$p(\mathbf{w}|\mathbf{t}, \mathbf{\Phi}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \propto p(\mathbf{t}, \mathbf{\Phi}|\mathbf{w})p(\mathbf{w})$$

where

$$\begin{aligned}
\mathbf{m}_N &= \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^T\mathbf{t}) \\
\mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^T\mathbf{\Phi}
\end{aligned}$$

Prior distribution

A common approach: zero-mean isotropic gaussian prior distribution of $\mathbf{w}$

$$p(\mathbf{w}|\alpha) = \prod_{i=0}^{M-1} \left(\frac{\alpha}{2\pi}\right)^{1/2} e^{-\frac{\alpha}{2} w_i^2}$$

- Parameters in $\mathbf{w}$ are assumed independent and identically distributed, according to a gaussian with mean $\mathbf{0}$, uniform variance $\sigma^2 = \alpha^{-1}$ and null covariance.
- Prior distribution defined with a hyper-parameter α, inversely proportional to the variance.

Posterior distribution

Given the likelihood

$$p(\mathbf{t}|\boldsymbol{\Phi}, \mathbf{w}, \beta) = \prod_{i=1}^{n} e^{-\frac{\beta}{2}(t_i - \mathbf{w}^T \boldsymbol{\phi}(x_i))^2}$$

the posterior distribution for $\mathbf{w}$ derives from Bayes' rule

$$p(\mathbf{w}|\mathbf{t}, \boldsymbol{\Phi}, \alpha, \sigma) = \frac{p(\mathbf{t}|\boldsymbol{\Phi}, \mathbf{w}, \sigma)p(\mathbf{w}|\alpha)}{p(\mathbf{t}|\boldsymbol{\Phi}, \alpha, \sigma)} \propto p(\mathbf{t}|\boldsymbol{\Phi}, \mathbf{w}, \sigma)p(\mathbf{w}|\alpha)$$

In this case

It is possible to show that, assuming

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \qquad p(\mathbf{t}|\mathbf{w}, \boldsymbol{\Phi}) = \mathcal{N}(\mathbf{t}|\mathbf{w}^T\boldsymbol{\Phi}, \beta^{-1}\mathbf{I})$$

the posterior distribution is itself a gaussian

$$p(\mathbf{w}|\mathbf{t}, \boldsymbol{\Phi}, \alpha, \sigma) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

with

$$\mathbf{S}_N = (\alpha\mathbf{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1} \qquad \mathbf{m}_N = \beta\mathbf{S}_N\boldsymbol{\Phi}^T\mathbf{t}$$

In this case

Note that as $\alpha \to 0$ the prior tends to have infinite variance, and we have minimum information on $\mathbf{w}$ before the training set is considered. In this case,

$$\mathbf{m}_N \to (\mathbf{\Phi}^T \beta \mathbf{I} \mathbf{\Phi})^{-1} (\mathbf{\Phi}^T \beta \mathbf{I} \mathbf{t}) = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} (\mathbf{\Phi}^T \mathbf{t})$$

that is $\mathbf{w}_{ML}$, the ML estimation of $\mathbf{w}$.

Maximum a Posteriori

- Given the posterior distribution $p(\mathbf{w}|\Phi, \mathbf{t}, \alpha, \beta)$, we may derive the value of $\mathbf{w}_{MAP}$ which makes it maximum (the mode of the distribution)
- This is equivalent to maximizing its logarithm

$$\log p(\mathbf{w}|\Phi, \mathbf{t}, \alpha, \beta) = \log p(\mathbf{t}|\mathbf{w}, \Phi, \beta) + \log p(\mathbf{w}|\alpha) - \log p(\mathbf{t}|\Phi, \beta)$$

and, since $p(\mathbf{t}|\Phi, \beta)$ is a constant wrt $\mathbf{w}$

$$\begin{aligned} \mathbf{w}_{MAP} &= \underset{\mathbf{w}}{\operatorname{argmax}} \log p(\mathbf{w}|\Phi, \mathbf{t}, \alpha, \beta) \\ &= \underset{\mathbf{w}}{\operatorname{argmax}} \left(\log p(\mathbf{t}|\mathbf{w}, \Phi, \beta) + \log p(\mathbf{w}|\alpha)\right) \end{aligned}$$

that is,

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\operatorname{argmin}} \left(-\log p(\mathbf{t}|\boldsymbol{\Phi}, \mathbf{w}, \beta) - \log p(\mathbf{w}|\alpha)\right)$$

Derivation of MAP

By considering the assumptions on prior and likelihood,

$$\begin{aligned}
w_{MAP} &= \operatorname*{argmin}_{\mathbf{w}} \left(\frac{\beta}{2} \sum_{i=1}^{n} (t_i - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i))^2 + \frac{\alpha}{2} \sum_{i=0}^{M-1} w_i^2 + \text{constants} \right) \\
&= \operatorname*{argmin}_{\mathbf{w}} \left(\sum_{i=1}^{n} (t_i - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i))^2 + \frac{\alpha}{\beta} \sum_{i=0}^{M-1} w_i^2 \right)
\end{aligned}$$

this is equivalent to considering a cost function

$$E_{MAP}(\mathbf{w}) = \sum_{i=1}^{n} (y_i - \mathbf{w}^T \boldsymbol{\phi}(x_i)) + \frac{\alpha}{\beta} \mathbf{w}^T \mathbf{w}$$

that is to a regularized min square function with $\lambda = \dfrac{\alpha}{\beta}$

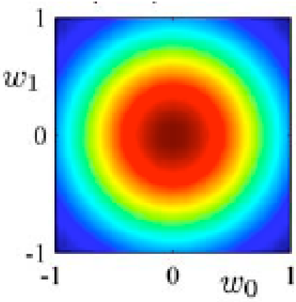
Sequential learning

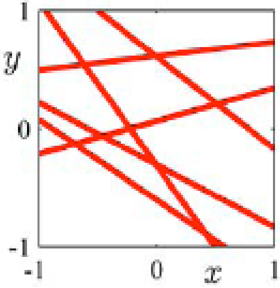
- The posterior after observing T_1 can be used as a prior for the next training set acquired.
- In general, for a sequence $T_1, \ldots, T_n$ of training sets,

$$
\begin{aligned}
&p(\mathbf{w}|T_1, \ldots T_n) \propto p(T_n|\mathbf{w})p(\mathbf{w}|T_1, \ldots T_{n-1})\\
&p(\mathbf{w}|T_1, \ldots T_{n-1}) \propto p(T_{n-1}|\mathbf{w})p(\mathbf{w}|T_1, \ldots T_{n-2})\\
&\ldots\\
&p(\mathbf{w}|T_1) \propto p(T_1|\mathbf{w})p(\mathbf{w})
\end{aligned}
$$

Example

- Input variable x, target variable t, linear regression $y(x, w_0, w_1) = w_0 + w_1 x$.
- Dataset generated by applying function $y = a_0 + a_1 x$ (with $a_0 = -0.3$, $a_1 = 0.5$) to values uniformly sampled in $[-1, 1]$, with added gaussian noise ($\mu = 0$, $\sigma = 0.2$).
- Assume the prior distribution $p(w_0, w_1)$ is a bivariate gaussian with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I} = 0.04\mathbf{I}$

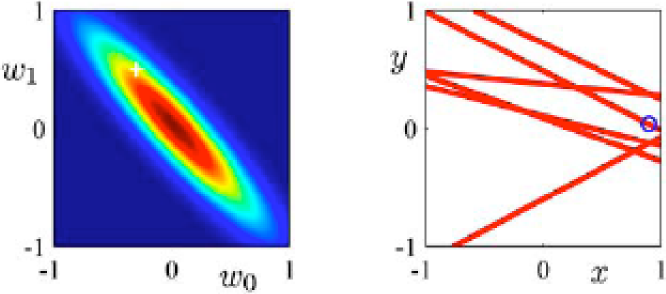




Left, prior distribution of w_0, w_1; right, 6 lines sampled from the distribution.

Example

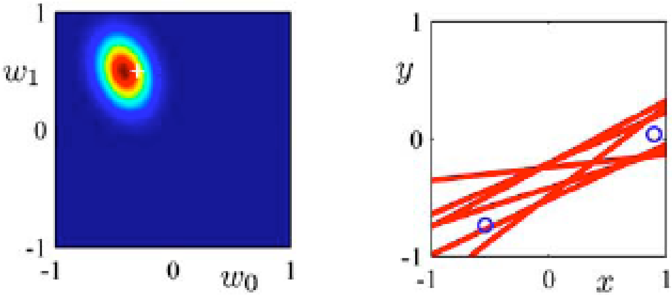
After observing item (x_1, y_1) (circle in right figure).



Left, posterior distribution $p(w_0, w_1 | x_1, y_1)$; right, 6 lines sampled from the distribution.

Esempio

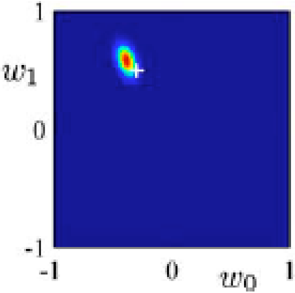
After observing items $(x_1, y_1), (x_2, y_2)$ (circles in right figure).

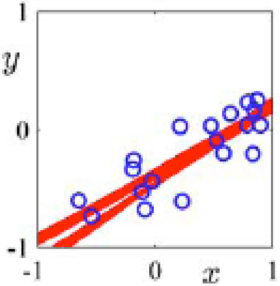


Left, posterior distribution $p(w_0, w_1 | x_1, y_1, x_2, y_2)$; right, 6 lines sampled from the distribution.

Example

After observing a set of n items $(x_1, y_1), \ldots, (x_n, y_n)$ (circles in right figure).





Left, posterior distribution $p(w_0, w_1 | x_i, y_i, i = 1, \ldots, n)$; right, 6 lines sampled from the distribution.

Example

- As the number of observed items increases, the distribution of parameters w_0, w_1 tends to concentrate (variance decreases to 0) around a mean point a_0, a_1.
- As a consequence, sampled lines are concentrated around $y = a_0 + a_1 x$.

Approaches to prediction in linear regression

Classical

- A value $\mathbf{w}_{LS}$ for $\mathbf{w}$ is learned through a point estimate, performed by minimizing a quadratic cost function, or equivalently by maximizing likelihood (ML) under the hypothesis of gaussian noise; regularization can be applied to modify the cost function to limit overfitting
- Given any $\mathbf{x}$, the obtained value $\mathbf{w}_{LS}$ is used to predict the corresponding t as $y = \overline{\mathbf{x}}^T\mathbf{w}_{LS}$, where $\overline{\mathbf{x}}^T = (1, \mathbf{x})^T$, or, in general, as $y = \phi(\mathbf{x})^T\mathbf{w}_{LS}$

Approaches to prediction in linear regression

Bayesian point estimation

- The posterior distribution $p(\mathbf{w}|\mathbf{t}, \mathbf{\Phi}, \alpha, \beta)$ is derived and a point estimate is performed from it, computing the mode $\mathbf{w}_{MAP}$ of the distribution (MAP)
- Equivalent to the classical approach, as $\mathbf{w}_{MAP}$ corresponds to $\mathbf{w}_{LS}$ if $\lambda = \dfrac{\alpha}{\beta}$
- The prediction, for a value $\mathbf{x}$, is a gaussian distribution $p(y|\boldsymbol{\phi}(\mathbf{x})^T\mathbf{w}_{MAP}, \beta)$ for y, with mean $\boldsymbol{\phi}(\mathbf{x})^T\mathbf{w}_{MAP}$ and variance β^{-1}
- The distribution is not derived directly from the posterior $p(\mathbf{w}|\mathbf{t}, \mathbf{\Phi}, \alpha, \beta)$: it is built, instead, as a gaussian with mean depending from the expectation of the posterior, and variance given by the assumed noise.

Approaches to prediction in linear regression

Fully bayesian

- The real interest is not in estimating $\mathbf{w}$ or its distribution $p(\mathbf{w}|\mathbf{t}, \boldsymbol{\Phi}, \alpha, \beta)$, but in deriving the predictive distribution $p(y|\mathbf{x})$. This can be done through expectation of the probability $p(y|\mathbf{x}, \mathbf{w}, \beta)$ predicted by a model instance wrt model instance distribution $p(\mathbf{w}|\mathbf{t}, \boldsymbol{\Phi}, \alpha, \beta)$, that is

$$p(y|\mathbf{x}, \mathbf{t}, \boldsymbol{\Phi}, \alpha, \beta) = \int p(y|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \boldsymbol{\Phi}, \alpha, \beta) d\mathbf{w}$$

- $p(y|\mathbf{x}, \mathbf{w}, \beta)$ is assumed gaussian, and $p(\mathbf{w}|\mathbf{t}, \boldsymbol{\Phi}, \alpha, \beta)$ is gaussian by the assumption that the likelihood $p(\mathbf{t}|\mathbf{w}, \boldsymbol{\Phi}, \beta)$ and the prior $p(\mathbf{w}|\alpha)$ are gaussian themselves and by their being conjugate

$$p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}), \beta)$$
$$p(\mathbf{w}|\mathbf{t}, \boldsymbol{\Phi}, \alpha, \beta) = \mathcal{N}(\mathbf{w}|\beta \mathbf{S}_N \boldsymbol{\Phi}^T \mathbf{t}, \mathbf{S}_N)$$

where $\mathbf{S}_N = (\alpha \mathbf{I} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1}$

Approaches to prediction in linear regression

Fully bayesian

Under such hypothesis, $p(y|\mathbf{x})$ is gaussian

$$p(y|\mathbf{x}, \mathbf{y}, \mathbf{\Phi}, \alpha, \beta) = \mathcal{N}(y|m(\mathbf{x}), \sigma^2(\mathbf{x}))$$

with mean

$$m(\mathbf{x}) = \beta \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \mathbf{\Phi}^T \mathbf{t}$$

and variance

$$\sigma^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})$$

Approaches to prediction in linear regression

Fully bayesian

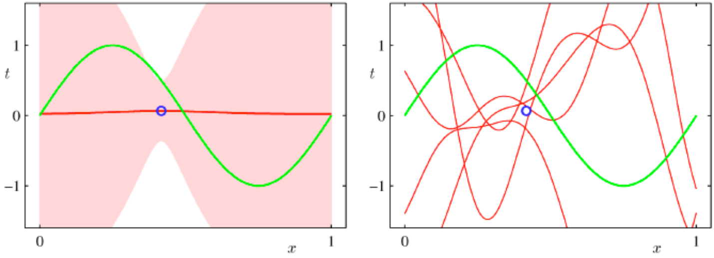
- $\frac{1}{\beta}$ is a measure of the uncertainty intrinsic to observed data (noise)
- $\boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})$ is the uncertainty wrt the values derived for the parameters $\mathbf{w}$
- as the noise distribution and the distribution of $\mathbf{w}$ are independent gaussians, their variances add
- $\boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}) \rightarrow 0$ as $n \rightarrow \infty$, and the only uncertainty remaining is the one intrinsic into data observation

Example

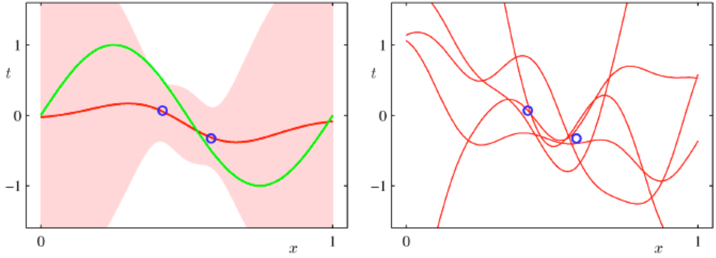
- predictive distribution for $y = \sin 2\pi x$, applying a model with 9 gaussian base functions and training sets of 1, 2, 4, 25 items, respectively
- left: items in training sets (sampled uniformly, with added gaussian noise); expectation of the predictive distribution (red), as function of x; variance of such distribution (pink shade within 1 standard deviation from mean), as a function of x
- right: items in training sets, 5 possible curves approximating $y = \sin 2\pi x$, derived through sampling from the posterior distribution $p(\mathbf{w}|\mathbf{t}, \mathbf{\Phi}, \alpha, \beta)$

Example

$n = 1$

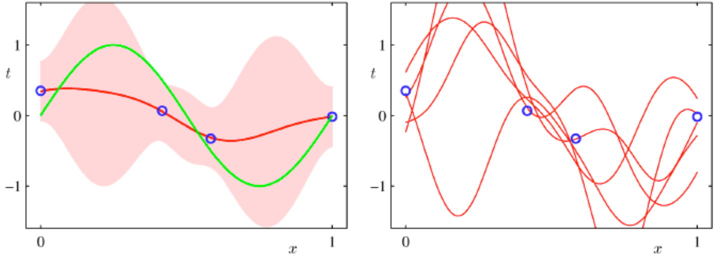


$n = 2$

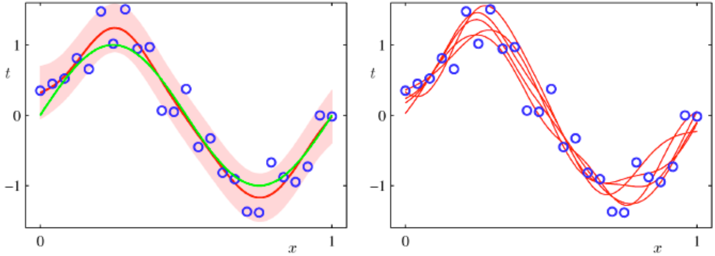


Example

$n = 4$



$n = 25$



Fully bayesian regression and hyperparameter marginalization

- In a fully bayesian approach, also the hyper-parameters α, β are marginalized

$$p(t|\mathbf{x}, \mathbf{t}, \mathbf{\Phi}) = \int\int\int p(t|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\mathbf{t}, \mathbf{\Phi}, \alpha, \beta)p(\alpha, \beta|\mathbf{t}, \mathbf{\Phi})d\mathbf{w}d\alpha d\beta$$

  this marginalization wrt $\mathbf{w}, \alpha, \beta$ is analytically intractable
- we may consider an approximation where hyperparameter values are derived by maximizing $p(\alpha, \beta|\mathbf{t}, \mathbf{\Phi})$
- since $p(\alpha, \beta|\mathbf{t}, \mathbf{\Phi}) \propto p(\mathbf{t}|\mathbf{\Phi}, \alpha, \beta)p(\alpha, \beta)$, if we assume that $p(\alpha, \beta)$ is relatively flat, then

$$\underset{\alpha,\beta}{\operatorname{argmax}}\, p(\alpha, \beta|\mathbf{t}, \mathbf{\Phi}) \simeq \underset{\alpha,\beta}{\operatorname{argmax}}\, p(\mathbf{t}|\mathbf{\Phi}, \alpha, \beta)$$

  and we may consider the maximization of the marginal likelihood (marginal wrt to coefficients $\mathbf{w}$)

$$p(\mathbf{t}|\mathbf{\Phi}, \alpha, \beta) = \int p(\mathbf{t}|\mathbf{w}, \mathbf{\Phi}, \beta)p(\mathbf{w}|\alpha)d\mathbf{w}$$

Marginal likelihood maximization

The marginal log-likelihood can be proved to be

$$\log p(\mathbf{t}|\mathbf{\Phi}, \alpha, \beta) = \frac{M}{2}\log\alpha - \frac{N}{2}\log\beta - E(\mathbf{m}_N) - \frac{1}{2}\log|\mathbf{S}_N^{-1}| - \frac{N}{2}\log(2\pi)$$

where M is the dimensionality, N the dimension of the training set, and

$$E(\mathbf{m}_N) = \frac{\beta}{2}||\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N||^2 + \frac{\alpha}{2}\mathbf{m}_N^T\mathbf{m}_N$$

$\mathbf{S}_N = (\alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi})^{-1}$ and $\mathbf{m}_N = \beta\mathbf{S}_N\mathbf{\Phi}\mathbf{t}$ are, respectively, the covariance matrix and the expectation vector of the posterior distribution $p(\mathbf{w}|\mathbf{t}, \mathbf{\Phi}, \alpha, \beta)$ of parameters.

Maximization of marginal likelihood wrt α

It can be shown that the value $\hat{\alpha}$ which maximizes the marginal likelihood verifies the equality

$$\frac{M}{2\hat{\alpha}} - \frac{1}{2}\mathbf{m}_N^T\mathbf{m}_N - \frac{1}{2}\sum_{i=1}^{M}\frac{1}{\lambda_i + \hat{\alpha}} = 0$$

where $\lambda_1, \ldots, \lambda_M$ are the eigenvalues of $\beta\mathbf{\Phi}^T\mathbf{\Phi}$.
That is,

$$\hat{\alpha}\mathbf{m}_N^T\mathbf{m}_N = M - \hat{\alpha}\sum_{i=1}^{M}\frac{1}{\lambda_i + \hat{\alpha}} = \sum_{i=1}^{M}\left(1 - \frac{\hat{\alpha}}{\lambda_i + \hat{\alpha}}\right) = \sum_{i=1}^{M}\frac{\lambda_i}{\lambda_i + \hat{\alpha}} = \gamma$$

and

$$\hat{\alpha} = \frac{\gamma}{\mathbf{m}_N^T\mathbf{m}_N}$$

This is an implicit solution for $\hat{\alpha}$, since both γ and $\mathbf{m}_N$ depend on α, and some iterative procedure should be applied.

Maximization of marginal likelihood wrt β

Here, it can be proved that the value $\hat{\beta}$ which maximizes the marginal likelihood verifies the equality

$$\frac{N}{2\beta} - \frac{1}{2}\sum_{i=1}^{N}\left(t_i - \mathbf{m}_N^T \boldsymbol{\phi}(\mathbf{x}_i)\right)^2 - \frac{\gamma}{2\beta} = 0$$

that is,

$$\frac{1}{\hat{\beta}} = \frac{1}{N-\gamma}\sum_{i=1}^{N}\left(t_i - \mathbf{m}_N^T \boldsymbol{\phi}(\mathbf{x}_i)\right)^2$$

Again, this is an implicit solution since both $\mathbf{m}_N$ and γ depend on β and an iterative method should be applied also in this case.

Equivalent kernel

- The expectation of the predictive distribution can be written also as

$$y(\mathbf{x}) = \beta\boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}_N\boldsymbol{\Phi}^T\mathbf{t} = \sum_{i=1}^{n}\beta\boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_i)t_i$$

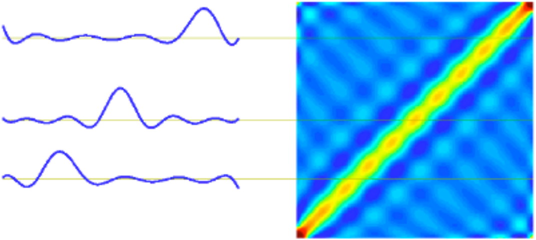
- The prediction can then be seen as a linear combination of the target values t_i of items in the training set, with weights dependent from the item values $\mathbf{x}_i$ (and from $\mathbf{x}$)

$$y(\mathbf{x}) = \sum_{i=1}^{n}\kappa(\mathbf{x}, \mathbf{x}_i)t_i$$

The weight function $\kappa(\mathbf{x}, \mathbf{x}') = \beta\boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}')$ is said *equivalent kernel* or linear smoother

Equivalent kernel

Right: plot on the plane (x, x_i) of a sample equivalent kernel, in the case of gaussian basis functions.
Left: plot as a function of x_i for three different values of x



In deriving y, the equivalent kernel tends to assign greater relevance to the target values t_i corresponding to items x_i near to x.