Machine learning

Linear regression

Corso di Laurea Magistrale in Informatica

Università di Roma Tor Vergata

Giorgio Gambosi

a.a. 2024-2025

Linear models

Linear models are based on a linear combination of input features

$$h(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

More compactly,

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

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

- Linear function of parameters $\mathbf{w}$
- Linear function of features $\mathbf{x}$

Base functions

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

$$h(\boldsymbol{\phi}(\mathbf{x}), \mathbf{w}) = \sum_{j=1}^{m} w_j \phi_j(\mathbf{x})$$

- 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 an m-dimensional space (usually with $m > d$)

Base functions

- Many types:
  - Polynomial (global functions)

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

  - Gaussian (local)

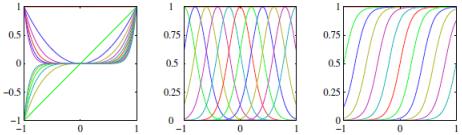
$$\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{2s^2}\right)$$

  - 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

$$\mathbf{X} = \begin{pmatrix} - & \mathbf{x}_1 & - \\ & \vdots & \\ - & \mathbf{x}_n & - \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ x_{21} & \cdots & x_{2d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{pmatrix}$$

is transformed into

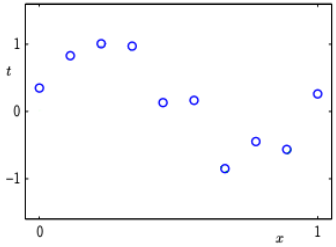
$$\mathbf{\Phi} = \begin{pmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \cdots & \phi_m(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \cdots & \phi_m(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_n) & \phi_2(\mathbf{x}_n) & \cdots & \phi_m(\mathbf{x}_n)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

Example

Problem

- A set of n observations of two variables $x, t \in \mathbb{R}$: $(x_1, t_1), \ldots, (x_n, t_n))$ is available. We wish to exploit these observations to predict, for any value $\tilde{x}$ of x, the corresponding unknown value of the target variable t
- The training set is a pair of vectors $\mathbf{x} = (x_1, \ldots, x_n)^T$ and $\mathbf{t} = (t_1, \ldots, t_n)^T$, related through an unknown rule (function)

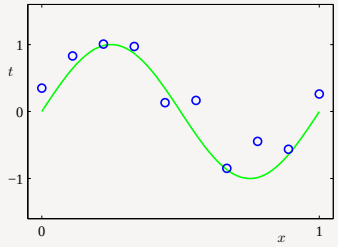
Example of a training set.



Example

Training set

In this case, we assume that the (unknown) relation between x and t in the training set is provided by the function $t = \sin(2\pi x)$, with an additional gaussian noise with mean 0 and given variance σ^2. Hence, $t_i = \sin(2\pi x_i) + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.



Purpose

Guessing, or approximating as well as possible, the deterministic relation $t = \sin(2\pi x)$, on the basis of the analysis of data in the training set.

Example: polynomial regression

Approach

Let us approximate the unknown function through a suitable polynomial of given degree $m > 0$

$$h(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_m x_m = \sum_{j=0}^{m} w_j x^j$$

whose coefficients $\mathbf{w} = (w_0, w_1, \ldots, w_m)^T$ are to be computed.

Base functions

This corresponds to applying a set of $m + 1$ base functions $\phi_j(x) = x^j, j = 0, \ldots, m$ to the unique feature x

$$h(x, \mathbf{w}) = \sum_{j=0}^{m} w_j \phi_j(x)$$

Regression loss

Base functions and linear models

When base functions are applied, $h(x, \mathbf{w})$ is a nonlinear function of x, but it is still a linear function (model) of $\mathbf{w}$.

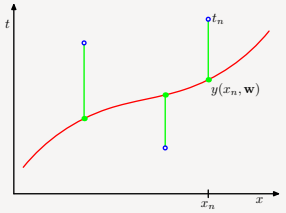
Parameter estimation

The values assigned to coefficients should minimize the empirical risk computed wrt some error function (a.k.a. cost function)

Regression loss

Least squares

A most widely adopted error function is the quadratic loss $(h(\phi(\mathbf{x}_i)) - t_i)^2$, which results into the least quares approach



$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{n} r_i(\mathbf{w})^2$$

where

$$r_i(\mathbf{w}) = h(\phi(\mathbf{x}_i), \mathbf{w}) - t_i = \sum_{j=1}^{m} w_j \phi_j(\mathbf{x}_i) - t_i$$

is the residue.

Equivalent to minimizing the empirical risk $\overline{\mathcal{R}}(\mathbf{w})$

Regression loss

Error minimization

- To minimize $E(\mathbf{w})$, set its derivative w.r.t. $\mathbf{w}$ to $\mathbf{0}$
- the quadratic loss is a convex function, which implies that only one (global) minimum is defined
- $E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(y(x_i, \mathbf{w}) - t_i)^2$ is convex itself, being the sum of n convex functions $(y(x_k, \mathbf{w}) - t_k)^2$
- in particular, $E(\mathbf{w})$ quadratic implies that its derivative is linear, hence that it is zero in one point w^*
- The resulting function is $h(x, \mathbf{w}^*) = \phi(\mathbf{x})^T\mathbf{w}^*$

Regression loss

Derivative with respect to $\mathbf{w}$

The derivative w.r.t. $\mathbf{w}$ is indeed a collection of derivatives. A linear system is obtained:

$$\frac{\partial E(\mathbf{w})}{\partial w_k} = 2\sum_{i=1}^{n} r_i(\mathbf{w}) \frac{\partial}{\partial w_k} r_i(\mathbf{w}) = 2\sum_{i=1}^{n} r_i(\mathbf{w})\phi_k(\mathbf{x}_i) = 2\sum_{i=1}^{n} \left(\sum_{j=1}^{m} w_j \phi_j(\mathbf{x}_i) - t_i \right) \phi_k(\mathbf{x}_i)$$

Each of the m equations is linear w.r.t. each coefficient in $\mathbf{w}$. A linear system results, with m equations and m unknowns $w_1, \ldots, w_m$, which, in general and with the exceptions of degenerate cases, has precisely one solution.

Closed form solution

In this case, the solution is defined in closed form by the normal equations for least squares

$$\mathbf{w}^* = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{t}$$

Gradient descent

- The minimum of $E(\mathbf{w})$ can be computed numerically, by means of gradient descent methods
- Initial assignment $\mathbf{w}^{(0)} = (w_1^{(0)}, w_2^{(0)}, \ldots, w_m^{(0)})$, with a corresponding error value

$$E(\mathbf{w}^{(0)}) = \frac{1}{2} \sum_{i=1}^{n} r_i(\mathbf{w}^{(0)})^2$$

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

$$w_k^{(s)} := w_k^{(s-1)} - \eta \frac{\partial E(\mathbf{w})}{\partial w_k}\bigg|_{\mathbf{w}^{(s-1)}} = w_k^{(s-1)} - 2\eta \sum_{i=1}^{n} r_i(\mathbf{w}^{(s-1)}) \phi_k(\mathbf{x}_i)$$

Gradient descent

- In matrix notation:

$$\mathbf{w}^{(s)} := \mathbf{w}^{(s-1)} - \eta \nabla E(\mathbf{w}) \Big|_{\mathbf{w}^{(s-1)}}$$

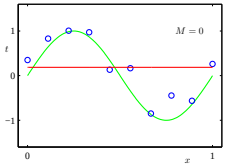
- By definition of $E(\mathbf{w})$:

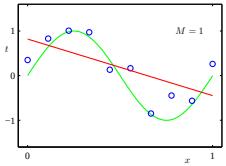
$$\mathbf{w}^{(s)} := \mathbf{w}^{(s-1)} - 2\eta \sum_{i=1}^{n} r_i(\mathbf{w}^{(s-1)}) \phi(\mathbf{x}_i)$$

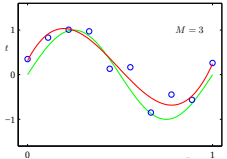
Example: fitting of polynomials

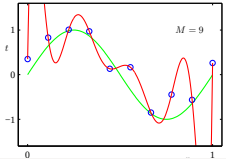
Polynomial degree

- Example of model selection: assigning a value to M determines the model to be used, the choice of M implies the number of coefficients to be estimated
- increasing M allows to better approximate the training set items, decreasing the error
- if $M + 1 = n$ the model allows to obtain a null error (overfitting)









Example: polynomial regression

Overfitting

- The function $h(\phi(\mathbf{x}), \mathbf{w})$ is derived from items in the training set, but should provide good predictions for other items.
- It should provide a suitable generalization to all items in the whole domain.
- If $h(\phi(\mathbf{x}), \mathbf{w})$ is derived as a too much accurate depiction of the training set, it results into an unsuitable generalization to items not in the training set

Example: polynomial regression

Evaluation of the generalization

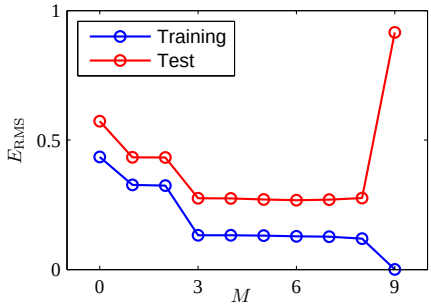
- Training set $\mathcal{T}_{train}$ of 1-dimensional items, generated by uniformly sampling x in $[0, 1,]$ and ε from $\mathcal{N}(0, \sigma^2)$, and computing $t = \sin 2\pi x + \varepsilon$
- Test set $\mathcal{T}_{test}$ of 1-dimensional items, generated in the same way as the training set
- For each M:
  - derives $\mathbf{w}^*$ by minimizing the empirical risk on the training set $\overline{\mathcal{R}}_{\mathcal{T}_{train}}(\mathbf{w})$
  - compute the empirical risk $\overline{\mathcal{R}}_{\mathcal{T}_{test}}(\mathbf{w}^*)$ on the test set: the square root of such value is considered here

$$E_{RMS}(\mathbf{w}^*, \mathcal{T}_{test}) = \sqrt{\overline{\mathcal{R}}_{\mathcal{T}_{test}}(\mathbf{w}^*)} = \sqrt{\frac{1}{|\mathcal{T}_{test}|} \sum_{(\mathbf{x},t)\in\mathcal{T}_{test}} (h(\phi(\mathbf{x}), \mathbf{w}^*) - t)^2}$$

- a lower value of $E_{RMS}(\mathbf{w}^*, \mathcal{T}_{test})$ denotes a good generalization

EXAMPLE: POLYNOMIAL REGRESSION

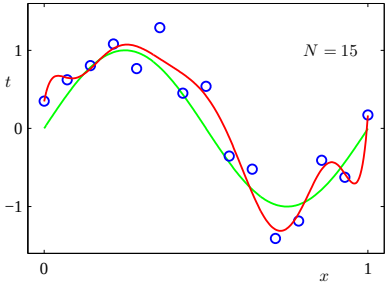
Plot of E_{RMS} w.r.t. M, on the training set and on the test set.

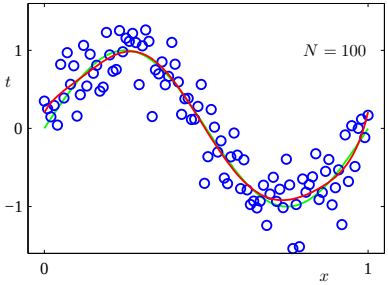


- As M increases, the error on the training set tends to 0.
- On the test set, the error initially decreases, since the higher complexity of the model allows to better represent the characteristics of the data set. Next, the error increases, since the model becomes too dependent from the training set.

Example: polynomial regression

For a given model complexity (such as the degree in our example), overfitting decreases as the dimension of the dataset increases.





The larger the dataset, the higher the acceptable complexity of the model.

How to limit the complexity of the model?

- 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

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

- The resulting overall loss to be minimized is then

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n} r_i(\mathbf{w})^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} = \frac{1}{2}\boldsymbol{r}(\mathbf{w})^T\boldsymbol{r}(\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

  where $\boldsymbol{r}(\mathbf{w})$ is the vector of residues, which can be expressed in terms of $\boldsymbol{\Phi}, \mathbf{w}$ and $\mathbf{t}$ as $\boldsymbol{r}(\mathbf{w}) = \boldsymbol{\Phi}\mathbf{w} - \mathbf{t}$

- this is called ridge regression: its solution can be expressed in closed form as

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

Regularization

- A more general form is obtained by considering the degree of the summed coefficients as a parameter

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n} r_i(\mathbf{w})^2 + \frac{\lambda}{2}\sum_{j=1}^{m} |w_j|^q$$

- The case $q = 1$ is denoted as lasso. Lasso regression has the property of favor sparse models (that is returning parameter vectors with many null values).

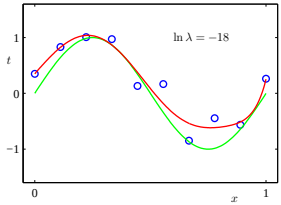
Example: polynomial regression

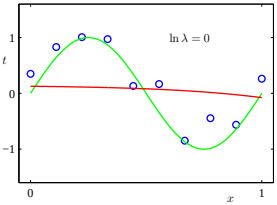
Use of regularization to limit complexity and overfitting.

- inclusion of a penalty term in the error function
- purpose: limiting the possible values of coefficients
- usually: limiting the absolute value of the coefficients

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n} r_i(\mathbf{w})^2 + \frac{\lambda}{2}\sum_{k=0}^{M} w_k^2 = \frac{1}{2}\sum_{i=1}^{n} r_i(\mathbf{w})^2 + \frac{\lambda}{2}||\mathbf{w}||^2$$

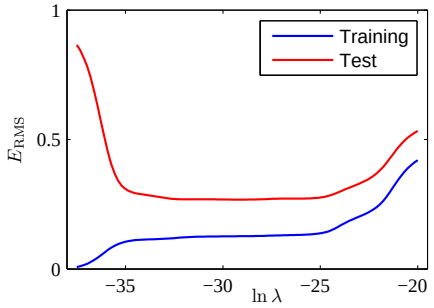
Dependance from the value of the hyperparameter λ.





Example: polynomial regression

Plot of the error w.r.t λ, ridge regression.



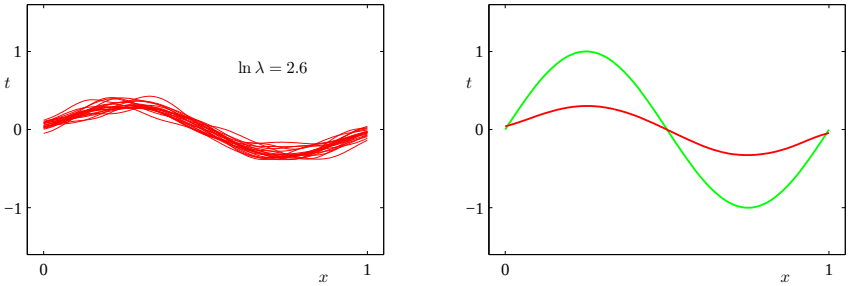
- Small λ: overfitting. Small error on the training set, large error on the test set.
- Large λ: the effect of data values decreases. Large error on both test and training sets.
- Intermediate λ. Intermediate error on training set, small error on test set.

Example: polynomial regression

- 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

$$
\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2}\sum_{i=1}^{n} r_i(\mathbf{w})^2 + \frac{\lambda}{2}\sum_{k=1}^{m} w_k^2 \\
&= \frac{1}{2}(\mathbf{\Phi w} - \mathbf{t})^T(\mathbf{\Phi w} - \mathbf{t}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}
\end{aligned}
$$

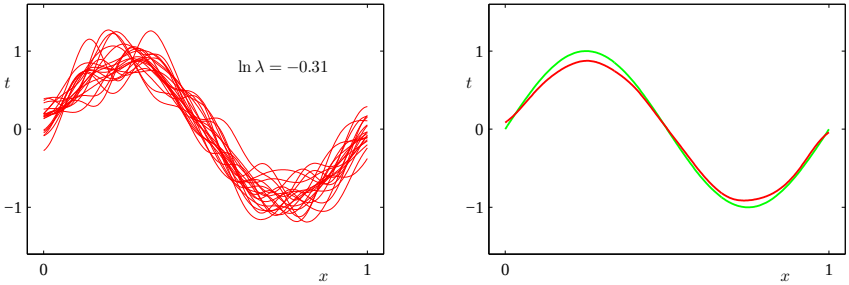
Example: polynomial regression



Left, a possible plot of prediction functions $h_i(\mathbf{x})$ ($i = 1, \ldots, 100$), 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 $f(x) = \sin 2\pi x$.

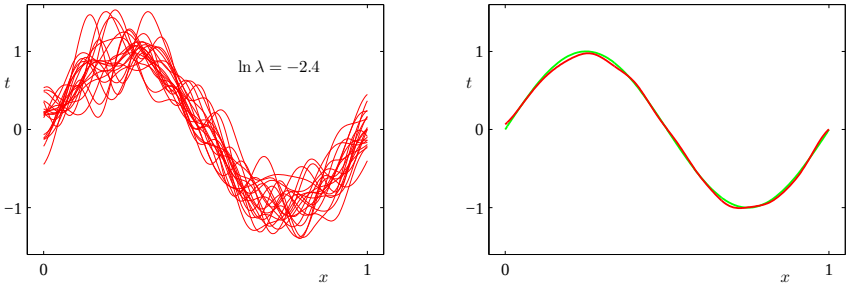
The prediction functions $h_i(\phi(\mathbf{x}))$ do not differ much between them (small variance), but their expectation is a bad approximation of the unknown function (large bias).

Example: polynomial regression



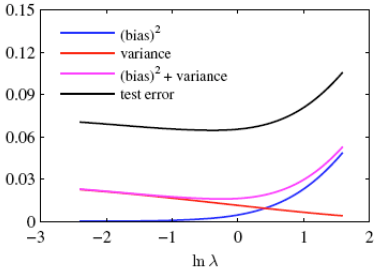
Plot of the prediction functions obtained with $\ln \lambda = -0.31$.

Example: polynomial regression



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

Example: polynomial regression

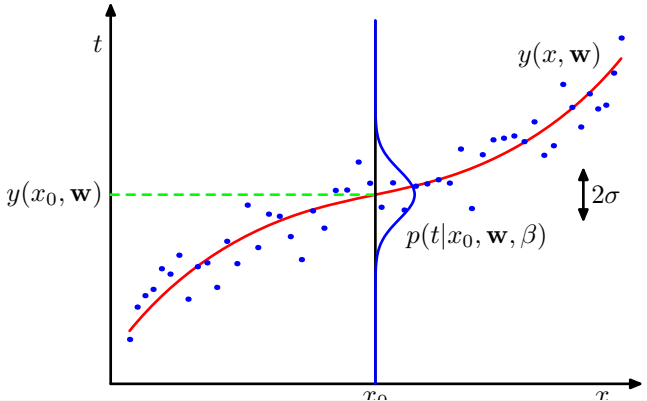


- Plot of (bias)^2, variance and their sum as functions of λ: las λ increases, bias increases and variance decreases. Their sum has a minimum in correspondance to the optimal value of λ.

Probabilistic model for regression

Assume that, given an item $\mathbf{x}$, the corresponding unknown target t is normally distributed around the value returned by the model $\mathbf{w}^T\overline{\mathbf{x}}$, with a given variance $\sigma^2 = \beta^{-1}$:

$$L(\mathbf{w}, \beta | \mathbf{\Phi}, \mathbf{t}) = p(\mathbf{t} | \mathbf{\Phi}, \mathbf{w}, \beta) = \prod_{i=1}^{n} \mathcal{N}(t_i | y(\phi(\mathbf{x}_i), \mathbf{w}), \beta^{-1}) = \prod_{i=1}^{n} \frac{\sqrt{\beta}}{\sqrt{2\pi}} e^{\frac{\beta}{2} r_i(\mathbf{w})^2}$$



Probabilistic model for regression

An estimate of both β_{ML} and the coefficients $\mathbf{w}_{ML}$ can be performed on the basis of the likelihood w.r.t. the assumed normal distribution:

$$L(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^{n} \mathcal{N}(t_i|y(\mathbf{x}_i, \mathbf{w}), \beta^{-1})$$

Parameters $\mathbf{w}$ and β can be estimated as the values which maximize the data likelihood, or its logarithm

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

which results into

$$p(\mathbf{t}|\mathbf{\Phi}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{i=1}^{n} r_i(\mathbf{w})^2 + \frac{n}{2} \log \beta + c$$

Probabilistic model for regression

The maximization w.r.t. $\mathbf{w}$ is performed by determining a maximum w.r.t. $\mathbf{w}$ of the function

$$-\frac{1}{2}\sum_{i=1}^{n}(t_i - y(\mathbf{x}_i, \mathbf{w}))^2$$

this is equivalent to minimizing the least squares sum.
The maximization w.r.t. the precision β is done by setting to 0 the corresponding derivative

$$\frac{\partial l(\mathbf{t}|\mathbf{\Phi}, \mathbf{w}, \beta)}{\partial \beta} = -\frac{1}{2}\sum_{i=1}^{n} r_i(\mathbf{w})^2 + \frac{n}{2\beta}$$

which results into

$$\beta_{ML}^{-1} = \frac{1}{n}\sum_{i=1}^{n} r_i(\mathbf{w})^2$$

Probabilistic model for regression

As a side result, the parameter estimate provides a predictive distribution of t given $\mathbf{x}$, that is the (gaussian) distribution of the target value for a given item $\mathbf{x}$.

$$p(t|\mathbf{x};\mathbf{w},\beta) = \mathcal{N}(t|h(\boldsymbol{\phi}(\mathbf{x}),\mathbf{w}),\beta^{-1}) = \sqrt{\frac{\beta_{ML}}{2\pi}} e^{-\frac{\beta_{ML}}{2}(h(\boldsymbol{\phi}(\mathbf{x}),\mathbf{w}_{ML})-t)^2}$$

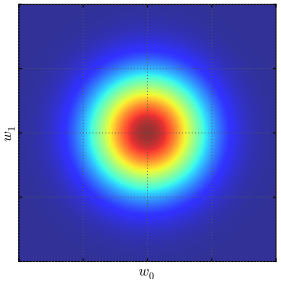
Probabilistic model for regression

- In the maximum likelihood framework parameters are considered as (unknown) values to determine with the best possible precision (frequentist approach).
- 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.
- The bayesian framework looks at parameters as random variables, whose probability distribution has to be derived.

Probabilistic model for regression

Prior distribution of parameters: gaussian with mean $\mathbf{0}$ and diagonal covariance matrix with variance equal to the inverse of hyperparameter α

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{\frac{m}{2}} e^{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}}$$



Why a gaussian prior?

Why a gaussian prior? Because the gaussian distribution is conjugated to itself. This means that the posterior distribution, being proportional to prior times likelihood, is gaussian if the likelihood is gaussian.

$$p(\mathbf{t}|\mathbf{\Phi}, \mathbf{w}, \beta) = \prod_{i=1}^{n} \mathcal{N}(t_i|h(\phi(\mathbf{x}_i), \mathbf{w}), \beta^{-1}) = \prod_{i=1}^{n} e^{-\frac{\beta}{2} r_i(\mathbf{w})^2}$$

Given the prior $p(\mathbf{w}|\alpha)$, the posterior distribution for $\mathbf{w}$ derives from Bayes' rule

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

Why a gaussian prior?

Given the above likelihood, if the prior of $\mathbf{w}$ is a gaussian

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \Sigma_0)$$

than the posterior distribution is itself gaussian

$$p(\mathbf{w}|\mathbf{\Phi}, \mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_p, \Sigma_p)$$

with

$$\Sigma_p = (\Sigma_0^{-1} + \beta \mathbf{\Phi}^T \mathbf{\Phi})^{-1}$$
$$\mathbf{m}_p = \Sigma_p(\Sigma_0^{-1}\mathbf{m}_0 + \beta \mathbf{\Phi}^T \mathbf{t})$$

Why a gaussian prior?

In the case we are considering here, we have

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \prod_{j=1}^{m} \frac{\sqrt{\alpha}}{\sqrt{2\pi}} e^{-\frac{\alpha}{2} w_i^2}$$

The posterior distribution is then a gaussian itself

$$p(\mathbf{w}|\mathbf{t}, \mathbf{\Phi}, \alpha, \sigma) = \mathcal{N}(\mathbf{w}|\mathbf{m}_p, \Sigma_p)$$

with

$$\Sigma_p = (\alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi})^{-1}$$
$$\mathbf{m}_p = \beta\Sigma_p\mathbf{\Phi}^T\mathbf{t}$$

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 $\mathbf{w}$

$$\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\operatorname{argmax}} \ \log p(\mathbf{w}|\Phi, \mathbf{t}, \alpha, \beta) = \underset{\mathbf{w}}{\operatorname{argmax}} \ (\log p(\mathbf{t}|\mathbf{w}, \Phi, \beta) + \log p(\mathbf{w}|\alpha))$$

that is,

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

Maximum a Posteriori

In this case

$$\log p(\mathbf{t}|\mathbf{\Phi}, \mathbf{w}, \beta) = \log \prod_{i=1}^{n} \frac{\sqrt{\beta}}{\sqrt{2\pi}} e^{-\frac{\beta}{2} r_i(\mathbf{w})^2} = \frac{n}{2} \log \beta - \frac{n}{2} \log(2\pi) - \frac{\beta}{2} \sum_{i=1}^{n} r_i(\mathbf{w})^2$$

and

$$\log p(\mathbf{w}|\alpha) = \log \prod_{j=1}^{m} \frac{\sqrt{\alpha}}{\sqrt{2\pi}} e^{-\frac{\alpha}{2} w_i^2} = \frac{m}{2} \log \alpha - \frac{n}{2} \log(2\pi) - \frac{\alpha}{2} \sum_{j=1}^{m} w_j^2$$

The value $\mathbf{w}_{MAP}$ which maximize the probability (mode of the distribution) minimizes

$$-\frac{\beta}{2} \sum_{i=1}^{n} r_i(\mathbf{w})^2 - \frac{\alpha}{2} \sum_{j=1}^{m} w_j^2 + \frac{n}{2} \log \beta + \frac{m}{2} \log \alpha - \frac{n+m}{2} \log(2\pi)$$

this is equivalent to maximizing

$$\frac{\beta}{2} \sum_{i=1}^{n} r_i(\mathbf{w})^2 + \frac{\alpha}{2} \sum_{j=1}^{m} w_j^2 \propto \frac{1}{2} \sum_{i=1}^{n} r_i(\mathbf{w})^2 + \frac{\alpha}{2\beta} \sum_{j=1}^{m} w_j^2$$

This corresponds to a ridge regression with regularization hyperparameter $\lambda = \frac{\alpha}{\beta}$.

Maximum a Posteriori

The same considerations of ML appy here for what concerns deriving the predictive distribution of t given $\mathbf{x}$, which results now

$$p(t|\mathbf{x}; \mathbf{w}_{MAP}, \beta_{MAP}) = \mathcal{N}(t|h(\phi(\mathbf{x}), \mathbf{w}_{MAP}), \beta_{MAP}^{-1}) = \sqrt{\frac{\beta_{MAP}}{2\pi}} e^{-\frac{\beta_{MAP}}{2}(h(\phi(\mathbf{x}), \mathbf{w}_{MAP}) - t)^2}$$

where, as it is easy to see, $\beta_{MAP} = \beta_{ML}$

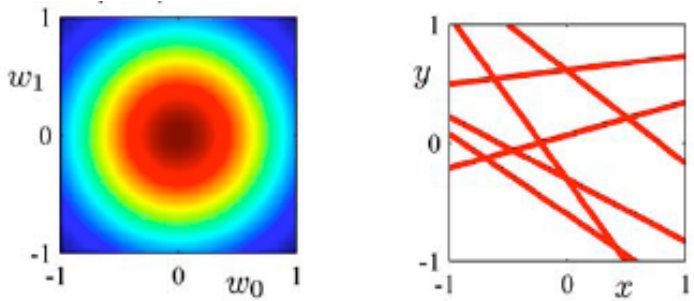
Sequential learning

- 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

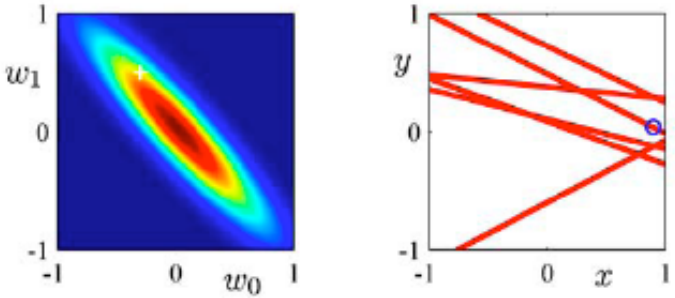
- Input variable x, target variable t, linear regression $y(x, w_0, w_1) = w_0 + w_1 x$.
- Dataset generated by applying function $y = a_0 + a_1 x$ (with $a_0 = -0.3$, $a_1 = 0.5$) to values uniformly sampled in $[-1, 1]$, with added gaussian noise ($\mu = 0$, $\sigma = 0.2$).
- Assume the prior distribution $p(w_0, w_1)$ 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

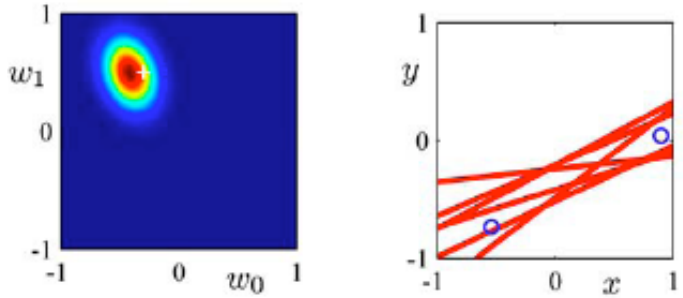
After observing item (x_1, y_1) (circle in right figure).



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

Example

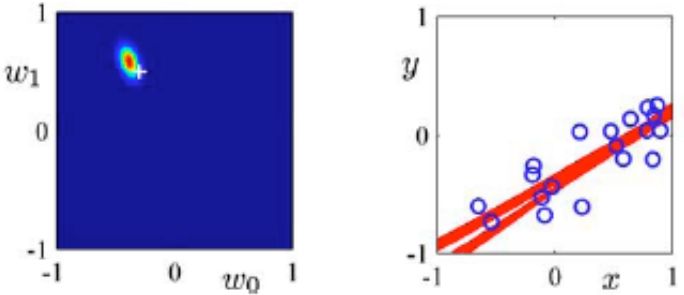
After observing items $(x_1, y_1), (x_2, y_2)$ (circles in right figure).



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 $t = \phi(\mathbf{x})^T \mathbf{w}_{LS}$

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 $\lambda = \frac{\alpha}{\beta}$
- The prediction, for a value $\mathbf{x}$, is a gaussian distribution $p(t|\boldsymbol{\phi}(\mathbf{x})^T\mathbf{w}_{MAP}, \beta)$ 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 $p(\mathbf{w}|\mathbf{t}, \mathbf{\Phi}, \alpha, \beta)$, 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 $p(\mathbf{w}|\mathbf{t}, \mathbf{\Phi}, \alpha, \beta)$, that is

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

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

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

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

Approaches to prediction in linear regression

Fully bayesian

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

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

with mean

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

and variance

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

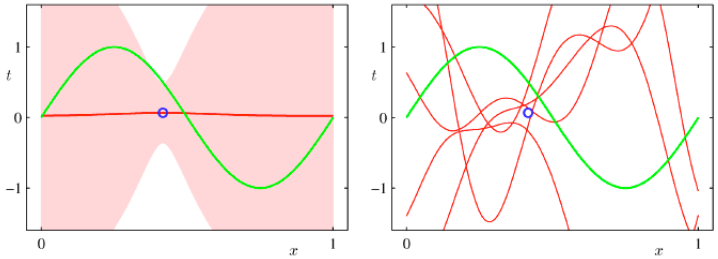
- $\frac{1}{\beta}$ is a measure of the uncertainty intrinsic to observed data (noise)
- $\boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})$ is the uncertainty wrt the values derived for the parameters $\mathbf{w}$
- as the noise distribution and the distribution of $\mathbf{w}$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

Example

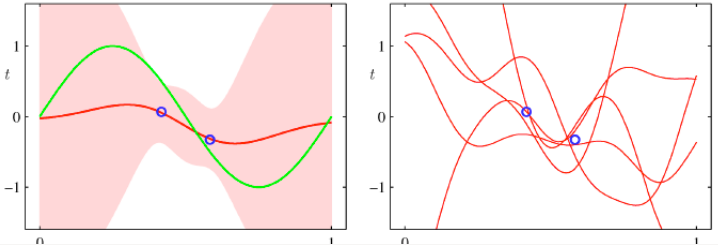
- predictive distribution for $f(\mathbf{x}) = \sin 2\pi x$, applying a model with 9 gaussian base functions and training sets of 1, 2, 4, 25 items, respectively
- left: items in training sets (sampled uniformly, with added gaussian noise); expectation of the predictive distribution (red), as function 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 $f(\mathbf{x}) = \sin 2\pi x$, derived through sampling from the posterior distribution $p(\mathbf{w}|\mathbf{t}, \mathbf{\Phi}, \alpha, \beta)$

Example

$n = 1$

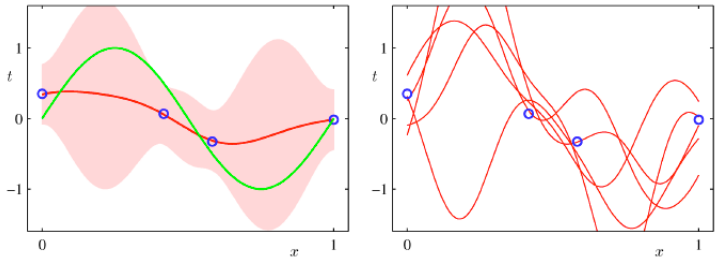


$n = 2$

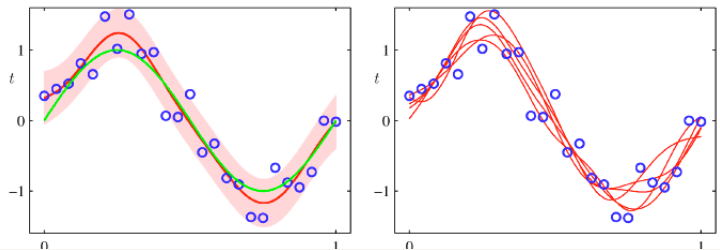


Example

$n = 4$



$n = 25$



Fully bayesian regression and hyperparameter marginalization

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

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

  where, as seen before,
  - $p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}), \beta)$
  - $p(\mathbf{w}|\mathbf{t}, \mathbf{\Phi}, \alpha, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$, with $\mathbf{S}_N = (\alpha \mathbf{I} + \beta \mathbf{\Phi}^T \mathbf{\Phi})^{-1}$ e $\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{\Phi}^T \mathbf{t}$

  this marginalization wrt $\mathbf{w}, \alpha, \beta$ is analytically intractable
- we may consider approximation methods