Over-fitting in Linear Regression

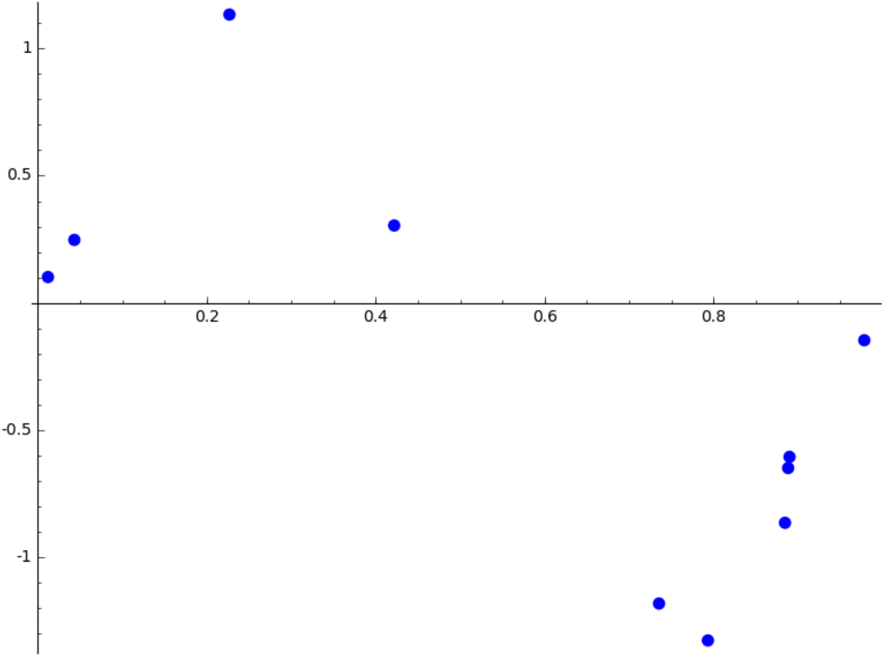
- Input: x; Output: y

  Observations: $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
- Use these observations as training examples.

  Task: Given a new input $\tilde{x}$, predict the output $\tilde{y}$.

- For example, $x \in [0, 1)$, $N = 10$

x	y
0.884644066199	−0.864791215635069
0.793349886821	−1.32738612014193
0.735440841558	−1.18222466237236
0.421871764847	0.304255805886633
0.0118832729931	0.101594120287724
0.226770188973	1.13377458999431
0.978530671629	−0.147028527196347
0.0431076970157	0.247622971933151
0.890003286931	−0.605625802202937
0.888362799625	−0.649537521948140


1
0.5
-0.5
-1
0.2
0.4
0.6
0.8

- It does not look like a line.
- Fit the data using a polynomial

$$y(x, \boldsymbol{w}) = w_1 x^k + w_2 x^{k-1} + \cdots + w_k x + w_{k+1},$$

where $\boldsymbol{w} = [w_1, \ldots, w_k, w_{k+1}]^\top$.

- Introduce the following matrices

$$X = \begin{bmatrix} x_1^k & \cdots & x_1 & 1 \\ x_2^k & \cdots & x_2 & 1 \\ \vdots & \vdots & & \vdots \\ x_N^k & \cdots & x_N & 1 \end{bmatrix}, \quad \boldsymbol{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_k \\ w_{k+1} \end{bmatrix}, \quad \text{and } \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}.$$

  Consider

$$E(\boldsymbol{w}) = \|X\boldsymbol{w} - \boldsymbol{y}\|^2.$$

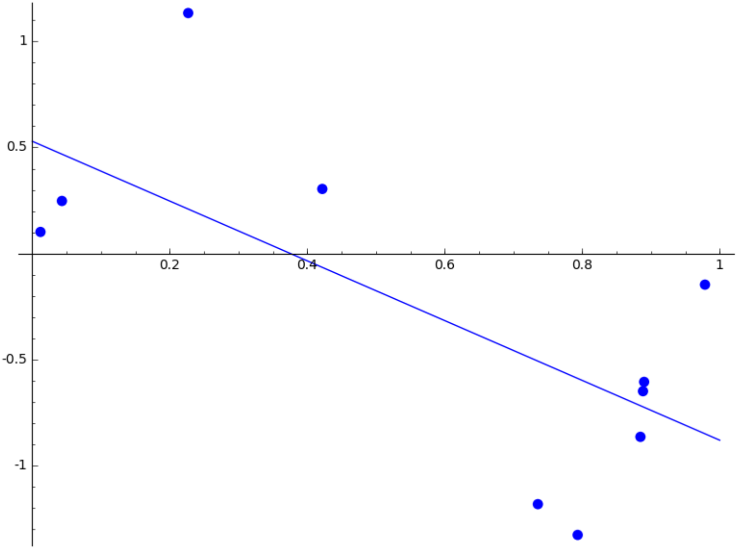
- Polynomial Regression $\rightsquigarrow$ Linear Regression

- Recall $\nabla E(\boldsymbol{w}) = 2X^\top(X\boldsymbol{w} - \boldsymbol{y})$
- We have

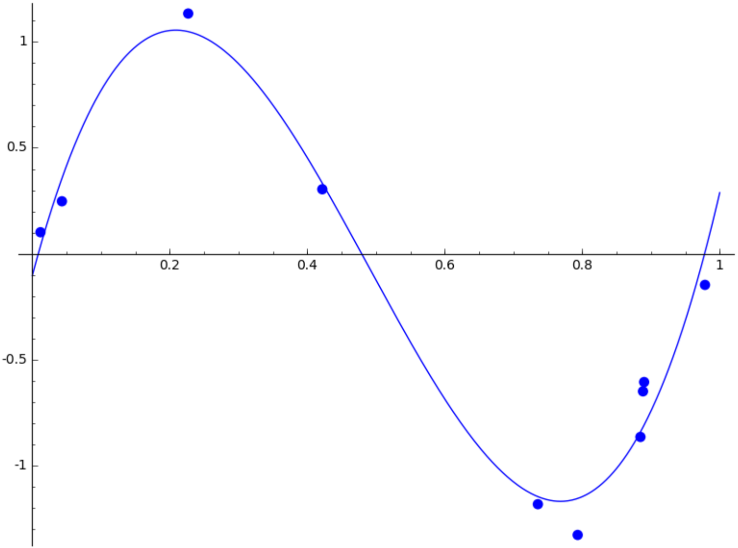
$$\boxed{\boldsymbol{w} = (X^\top X)^{-1} X^\top \boldsymbol{y}}.$$

$\mathbf{w}$	$k = 1$	$k = 3$	$k = 6$	$k = 9$
w_1	-1.41	25.37	43.45	-69519.92
w_2	0.53	-37.18	-210.56	214844.27
w_3		12.21	339.78	-189181.01
w_4		-0.10	-211.33	-61808.88
w_5			34.15	210688.85
w_6			4.49	-141666.70
w_7			0.03	41628.04
w_8				-5191.34
w_9				200.18
w_{10}				-1.61

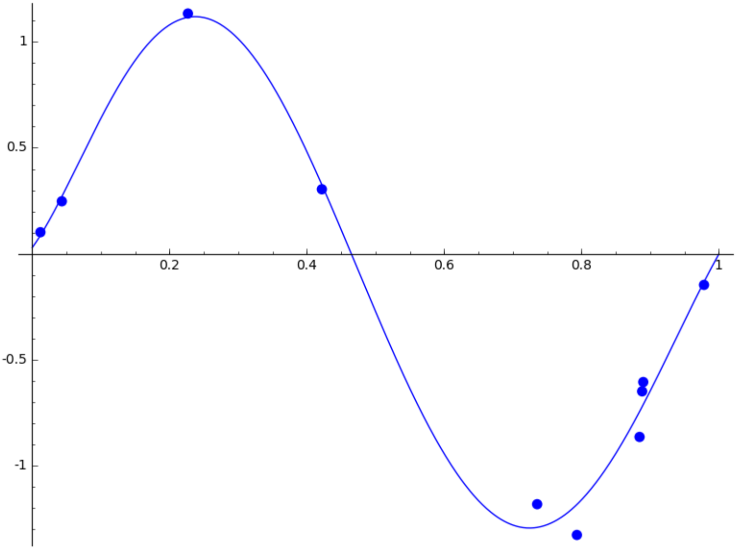
$k = 1$



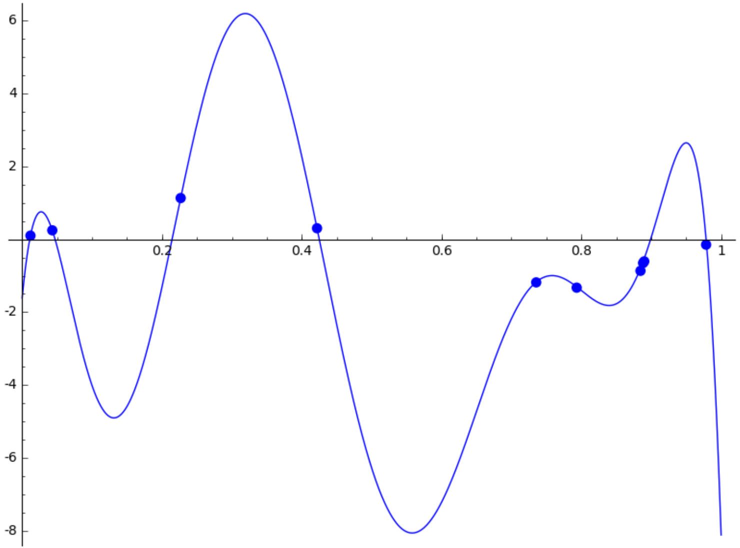
$k = 3$



$k = 6$



$k = 9$



- The case $k = 9$ is over-fitting.
- In order to avoid over-fitting, we can use regularization.
- Ridge regression

$$\boxed{\widetilde{E}(\boldsymbol{w}) = \|X\boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{w}\|^2}.$$

  This can be considered as a result of Bayesian Learning.
- Lasso regression

$$\widetilde{E}(\boldsymbol{w}) = \|X\boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda\sum_{n=1}^{k+1}|w_n|.$$

Bayesian Linear Regression

- Bayesian linear regression avoids the over-fitting problem of maximum likelihood.
- Input: x; Output: y

  Observations: $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
- Basic Assumption:

  Given x, the corresponding value of y has a normal distribution with a mean equal to the value $y_*(x, \boldsymbol{w})$ of the polynomial curve

$$y_*(x, \boldsymbol{w}) = w_1 x^k + w_2 x^{k-1} + \cdots + w_k x + w_{k+1},$$

where $\boldsymbol{w} = [w_1, \ldots, w_k, w_{k+1}]^\top$.

- Write

$$y = y_*(x, \boldsymbol{w}) + \epsilon,$$

where ϵ is a Gaussian noise. Then

$$p(y|x, \boldsymbol{w}, \beta) = \mathcal{N}(y|y_*(x, \boldsymbol{w}), \beta^{-1}),$$

where β is a parameter corresponding to the inverse variance, called the precision.

- Assume that each observation is independent, and that the variance β^{-1} is all the same.
- Then we have

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n | y_*(x_n, \boldsymbol{w}), \beta^{-1}).$$

  This is our probabilistic model.
- It is easy to see

$$\begin{aligned} -\ln p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}, \beta) &= \frac{\beta}{2} \sum_{n=1}^{N} (y_n - y_*(x_n, \boldsymbol{w}))^2 + (\text{constant}) \\ &= \frac{\beta}{2} \|\boldsymbol{y} - X\boldsymbol{w}\|^2 + (\text{constant}). \end{aligned}$$

- Next we need to choose a prior.
- D-dimensional Gaussian distribution:

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \right\}$$

where the D-dimensional vector $\boldsymbol{\mu}$ is the mean, the $D \times D$ matrix Σ is the covariance, and $|\Sigma|$ is the determinant of Σ.

- Choose a prior distribution for $\boldsymbol{w}$:

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1} I).$$

- We have

$$-\ln p(\boldsymbol{w}|\alpha) = \frac{\alpha}{2}\boldsymbol{w}^\top \boldsymbol{w} + (\text{constant}) = \frac{\alpha}{2}\|\boldsymbol{w}\|^2 + (\text{constant}).$$

- Bayes' Theorem gives the posterior

$$p(\boldsymbol{w}|\boldsymbol{x}, \boldsymbol{y}, \alpha, \beta) \propto p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}, \beta)\, p(\boldsymbol{w}|\alpha).$$

  Task: Determine $\boldsymbol{w}$ so that the posterior is maximized.

  This process is called a maximum a posteriori (MAP) estimation.

- Take the negative logarithm of the posterior

$$\begin{aligned} E(\boldsymbol{w}) &= -\ln p(\boldsymbol{w}|\boldsymbol{x}, \boldsymbol{y}, \alpha, \beta) \\ &= -\ln\left[p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}, \beta)\, p(\boldsymbol{w}|\alpha)\right] + \text{(constant)} \\ &= \frac{\beta}{2}\|\boldsymbol{y} - X\boldsymbol{w}\|^2 + \frac{\alpha}{2}\|\boldsymbol{w}\|^2 + \text{(constant)} \end{aligned}$$

- The maximum of the posterior is given by the minimum of

$$\boxed{\tilde{E}(\boldsymbol{w}) = \frac{\beta}{2}\|\boldsymbol{y} - X\boldsymbol{w}\|^2 + \frac{\alpha}{2}\|\boldsymbol{w}\|^2}.$$

Thus maximizing the posterior distribution is equivalent to minimizing the regularized sum-of-square error function.

- We can compute $\boldsymbol{w}$ explicitly:

$$\tilde{E}(\boldsymbol{w}) = \frac{\beta}{2}\|X\boldsymbol{w} - \boldsymbol{y}\|^2 + \frac{\alpha}{2}\|\boldsymbol{w}\|^2,$$

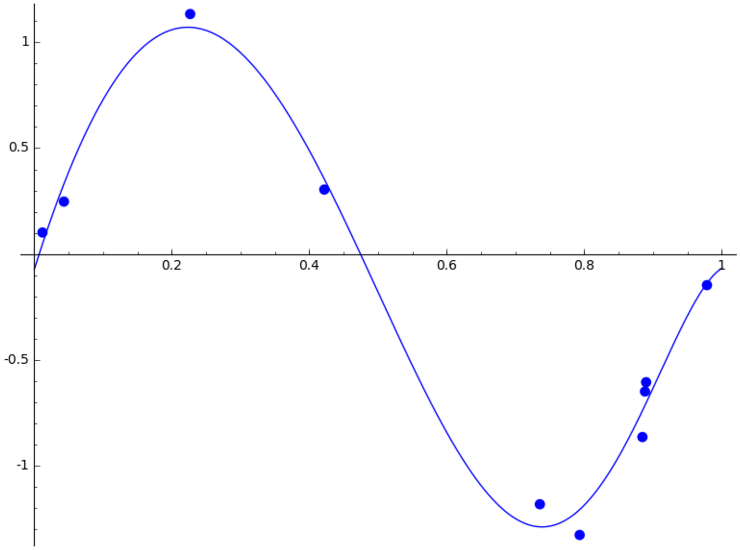
and

$$\nabla\tilde{E}(\boldsymbol{w}) = \beta X^\top(X\boldsymbol{w} - \boldsymbol{y}) + \alpha\boldsymbol{w} = 0.$$

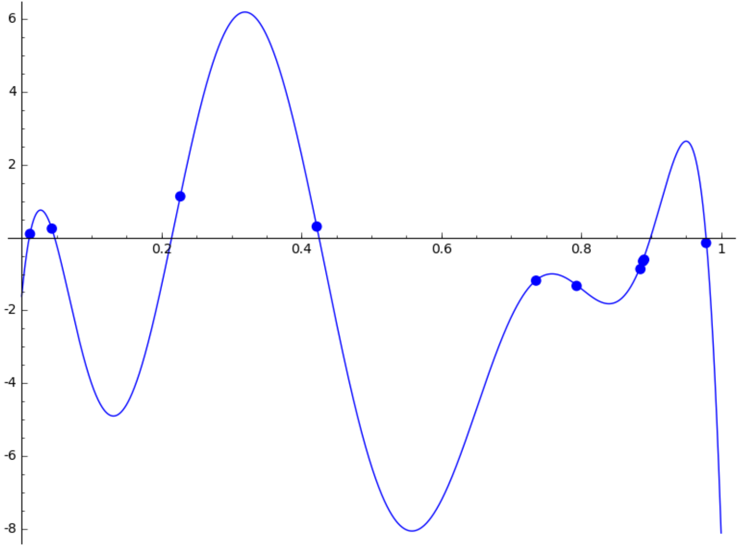
Thus

$$\boxed{\boldsymbol{w} = \beta S X^\top \boldsymbol{y} \quad \text{with} \;\; S^{-1} = \alpha I + \beta X^\top X}.$$

$N = 9$, $\alpha = 0.01$ and $\beta = 1000$



Recall the maximum likelihood gave us



- The posterior can be computed explicitly, since the prior and the likelihood are all Gaussian.
- Indeed, we obtain

$$\boxed{p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, S)},$$

where

$$\boldsymbol{m} = \beta S X^\top \boldsymbol{y} \quad \text{with} \quad S^{-1} = \alpha I + \beta X^\top X.$$