Linear Regression and Classicifation

Wenting Tu

SHUFE, SIME

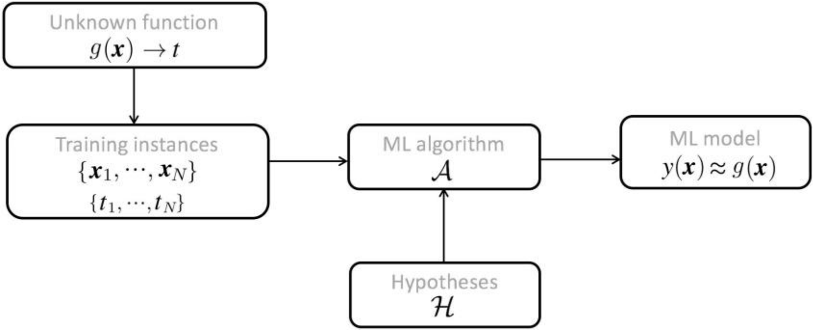
Machine Learning and Deep Lerning

Outline

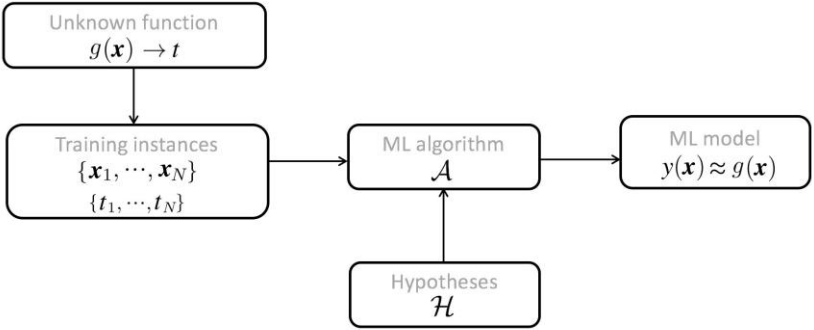
Linear Regression

Linear Classification

Definition


Unknown function
$g(\boldsymbol{x}) \to t$
Training instances
$\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$
$\{t_1, \cdots, t_N\}$
ML algorithm
$\mathcal{A}$
ML model
$y(\boldsymbol{x}) \approx g(\boldsymbol{x})$
Hypotheses
$\mathcal{H}$

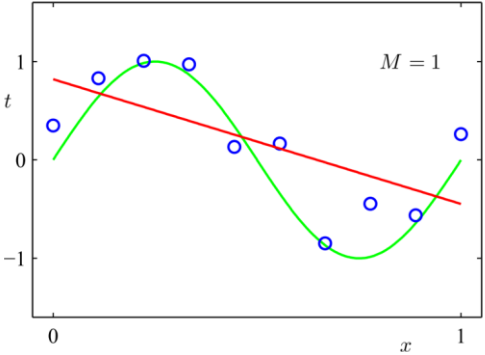
Definition



$$y(x, \boldsymbol{w}) = w_0 + w_1 x$$

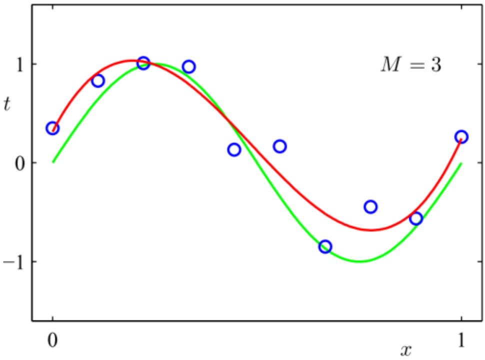
$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_D x_D$$

Illustration



$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \cdots + w_M x^M$$

Illustration



$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \cdots + w_M x^M$$

Linear Basis Function Models

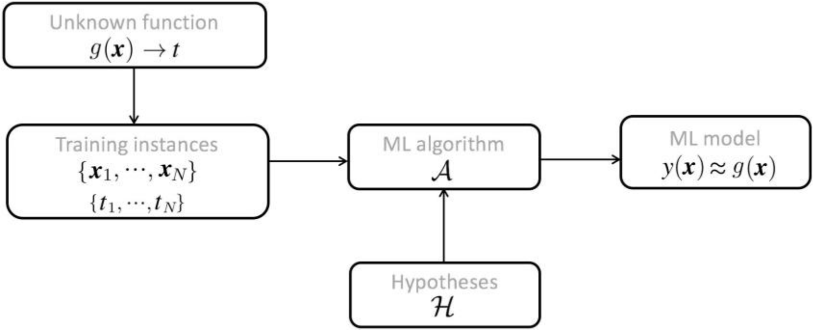
$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\boldsymbol{x})$$

$$y(\boldsymbol{x}, \boldsymbol{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) \quad \phi_0(\boldsymbol{x}) = 1$$

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

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

$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right) \quad \sigma_a = \frac{1}{1+\exp(-a)}$$



Unknown function
$g(\boldsymbol{x}) \to t$
Training instances
$\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$
$\{t_1, \cdots, t_N\}$
ML algorithm
$\mathcal{A}$
ML model
$y(\boldsymbol{x}) \approx g(\boldsymbol{x})$
Hypotheses
$\mathcal{H}$

Least-squares

- Loss function

$L(t, y(\boldsymbol{x})) = \{y(\boldsymbol{x}) - t\}^2$ (squared residuals)

- Empirical risk

$$E_D(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \left\{t_n - \boldsymbol{w}^T \boldsymbol{\phi}\left(\boldsymbol{x}_n\right)\right\}^2$$

Least-squares

- Solution

$$\boldsymbol{w}^{\star} = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^T \mathbf{t}$$

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \cdots & \phi_{M-1}(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \cdots & \phi_{M-1}(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \phi_1(\boldsymbol{x}_N) & \cdots & \phi_{M-1}(\boldsymbol{x}_N) \end{pmatrix}$$

Least-squares and Maximum Likelihood

- Review for maximum-likelihood estimation (MLE)

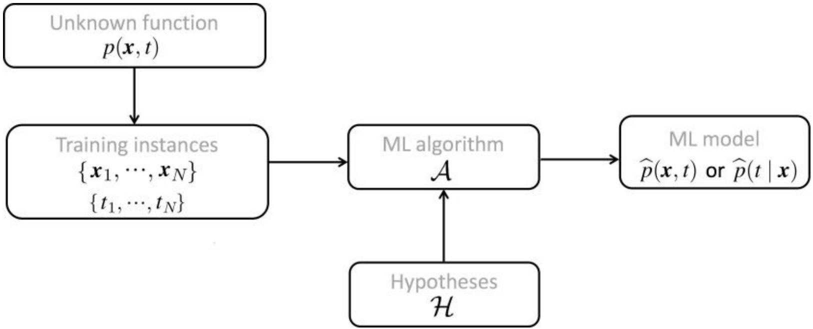
$$q(\boldsymbol{x};\boldsymbol{\theta}) \longrightarrow p(\boldsymbol{x})$$

$$\mathcal{D} = \{\boldsymbol{x}\}_{i=1}^{n}$$

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D} \mid \boldsymbol{\theta}) = \operatorname{argmax} \prod_{i=1}^{n} q(\boldsymbol{x};\boldsymbol{\theta})$$

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} \log L(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} \left[\sum_{i=1}^{n} \log q\left(\boldsymbol{x}_i;\boldsymbol{\theta}\right)\right]$$

$$\widehat{p}(\boldsymbol{x}) = q(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_{\mathrm{ML}})$$



Unknown function
$p(\boldsymbol{x}, t)$
Training instances
$\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$
$\{t_1, \cdots, t_N\}$
ML algorithm
$\mathcal{A}$
ML model
$\widehat{p}(\boldsymbol{x}, t)$ or $\widehat{p}(t \mid \boldsymbol{x})$
Hypotheses
$\mathcal{H}$

Least-squares and Maximum Likelihood

- Relation between LS and MLE

$$t = y(\boldsymbol{x}, \boldsymbol{w}) + \epsilon \quad \epsilon \sim \mathcal{N}\left(0, \beta^{-1}\right)$$

$$p(t \mid \boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}\left(t \mid y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}\right)$$

$$\mathbb{E}[t \mid \boldsymbol{x}] = \int t p(t \mid \boldsymbol{x}) \mathrm{d}t = y(\boldsymbol{x}, \boldsymbol{w})$$

$$p(\mathbf{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \boldsymbol{w}^T \boldsymbol{\phi}\left(\boldsymbol{x}_n\right), \beta^{-1}\right)$$

$$\begin{aligned}\ln p(\mathbf{t} \mid \boldsymbol{w}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}\left(t_n \mid \boldsymbol{w}^T \boldsymbol{\phi}\left(\boldsymbol{x}_n\right), \beta^{-1}\right) \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\boldsymbol{w})\end{aligned}$$

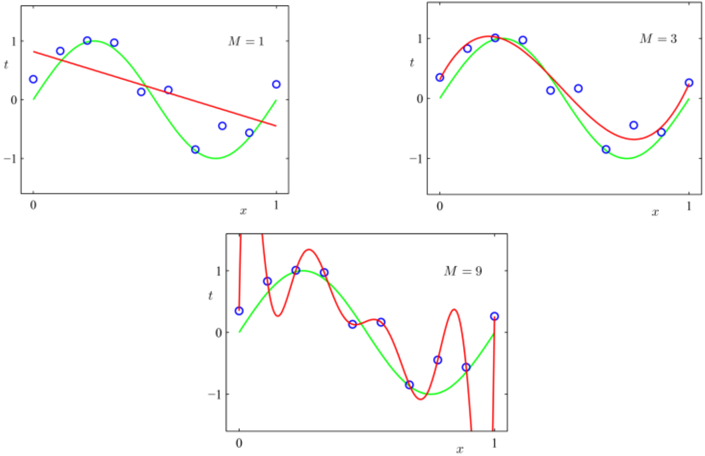
$$\boldsymbol{w}_{MLE}^{\star} = \arg\max_{w} \ln p(\mathbf{t} \mid \boldsymbol{w}, \beta)$$

$$\boldsymbol{w}_{MLE}^{*} = \boldsymbol{w}_{LS}^{*}$$

Underfitting and Overfitting

- Illustration

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \cdots + w_M x^M$$



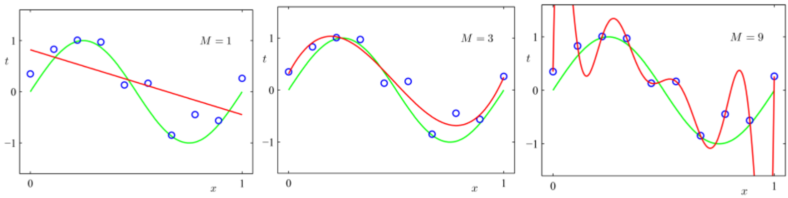
Bias-Variance Decomposition

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}$$

$$(\text{bias})^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})] - h(\boldsymbol{x})\}^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

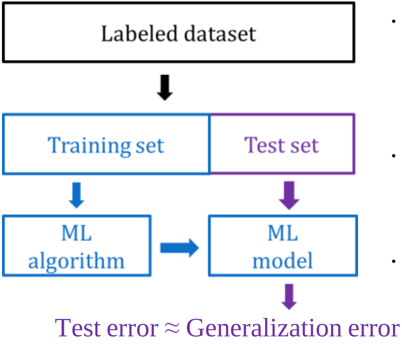
$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}\left[\{y(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})]\}^2\right] p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$\text{noise} = \iint \{h(\boldsymbol{x}) - t\}^2 p(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}\mathrm{d}t$$



Model Evaluation

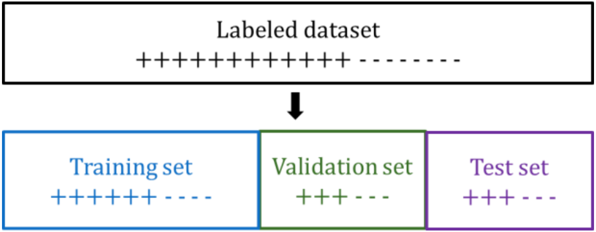
- Generalization ability



- When learning a model, you should pretend that you don't have the test data yet. If the test-set labels influence the learned model in any way, accuracy estimates will be biased.
- Your test set should be large enough to detect meaningful changes in the accuracy of your algorithm, but not necessarily much larger.
- When randomly selecting training or validation sets, we may want to ensure that class proportions are maintained in each selected set.

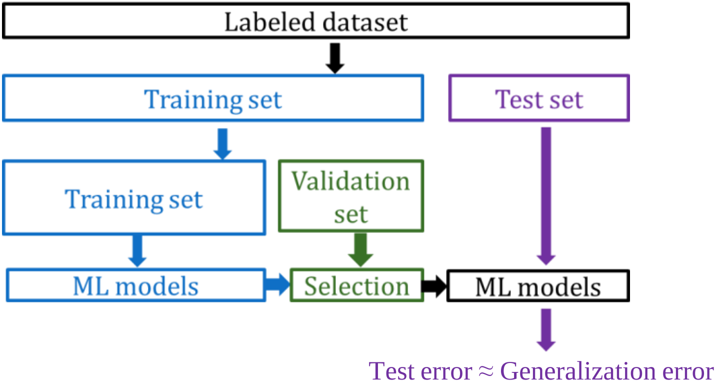
Model Evaluation

- Validation set



Model Evaluation

- Tuning hyperparameters



Model Evaluation

- Cross validation

labeled data set

$$\kappa : \{1, \ldots, N\} \mapsto \{1, \ldots, K\}$$

s_1	s_2	s_3	s_4	s_5

iteration	train on	test on
1	s_2 s_3 s_4 s_5	s_1
2	s_1 s_3 s_4 s_5	s_2
3	s_1 s_2 s_4 s_5	s_3
4	s_1 s_2 s_3 s_5	s_4
5	s_1 s_2 s_3 s_4	s_5

The K results can then be averaged to produce a single estimation.

CV makes efficient use of the available data for testing

Model Evaluation

- Performance Evaluation for Regression

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \left| t_i - f\left(\mathbf{x}_i\right) \right|$$

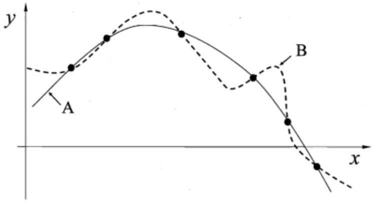
$$MSE = \frac{1}{n}\sum_{i=1}^{n} \left(t_i - f\left(\mathbf{x}_i\right) \right)^2$$

$$RMSE = \sqrt{MSE}$$

Model Selection

- Occam's razor

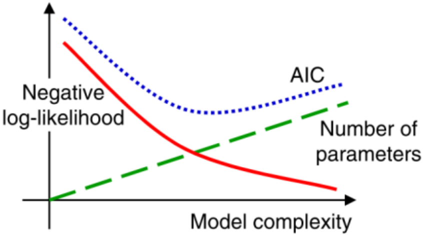
Suppose there exist two explanations for an occurrence. In this case the one that requires the smallest number of assumptions is usually correct.



Model Selection

- Akaike information criterion (AIC)

$$\ln p(\mathcal{D}|\mathbf{w}_{\mathrm{ML}}) - M$$



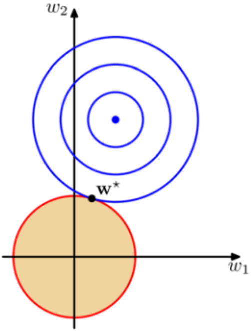
Regularization

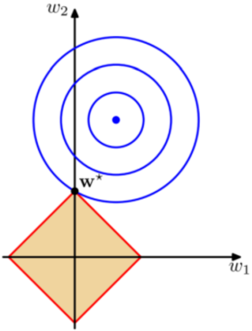
$$E_D(\boldsymbol{w}) + \lambda E_W(\boldsymbol{w})$$

LASSO

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

$$= \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \boldsymbol{w}^T\boldsymbol{\Phi}\left(\boldsymbol{x}_n\right)\right\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$





Ridge Regression

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

$$= \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \boldsymbol{w}^T\phi\left(\boldsymbol{x}_n\right)\right\}^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\boldsymbol{w}^*_{\text{ridge}} = \left(\lambda\boldsymbol{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\mathbf{t}$$

Ridge Regression and Maximum a Posteriori Estimation

- Review for maximum a posteriori (MAP) estimation

Likelihood $p(\mathcal{D}|\boldsymbol{\theta})$

Prior $p(\boldsymbol{\theta})$

Posterior $p(\boldsymbol{\theta}|\mathcal{D})$

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|\mathcal{D})$$

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left(\sum_{i=1}^{n} \log q\left(\boldsymbol{x}_i|\boldsymbol{\theta}\right) + \log p(\boldsymbol{\theta}) \right)$$

Ridge Regression and Maximum a Posteriori Estimation

- Relation between ridge regression and MAP

$$p(\mathbf{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \boldsymbol{w}^T \boldsymbol{\phi}\left(\boldsymbol{x}_n\right), \beta^{-1}\right)$$

$$p(\boldsymbol{w}) = \mathcal{N}\left(\boldsymbol{w} \mid \boldsymbol{m}_0, \boldsymbol{S}_0\right)$$

$$p(\boldsymbol{w} \mid \mathbf{t}) = \mathcal{N}\left(\boldsymbol{w} \mid \boldsymbol{m}_N, \boldsymbol{S}_N\right)$$

$$\boldsymbol{m}_N = \boldsymbol{S}_N\left(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta\boldsymbol{\Phi}^T\mathbf{t}\right), \; \boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}$$

Ridge Regression and Maximum a Posteriori Estimation

- Relation between ridge regression and MAP

$$p(\mathbf{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}\right)$$

$$p(\boldsymbol{w} \mid \alpha) = \mathcal{N}\left(\boldsymbol{w} \mid \mathbf{0}, \alpha^{-1}\boldsymbol{I}\right)$$

$$p(\boldsymbol{w} \mid \mathbf{t}) = \mathcal{N}\left(\boldsymbol{w} \mid \boldsymbol{m}_N, \boldsymbol{S}_N\right)$$

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

$$\begin{aligned}\boldsymbol{w}^*_{MAP} &= \arg\max_w \ln p(\boldsymbol{w}|\mathbf{t}) \\ &= \arg\max_w \left(-\frac{\beta}{2}\sum_{n=1}^{N}\left\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\right\}^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{const}\right)\end{aligned}$$

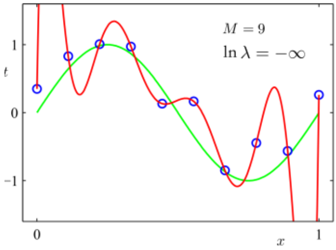
$$\boldsymbol{w}^*_{MAP} = \boldsymbol{w}^\star_{ridge}$$

Ridge Regression

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \sum_{j=0}^{M} w_j x^j\right\}^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}$$

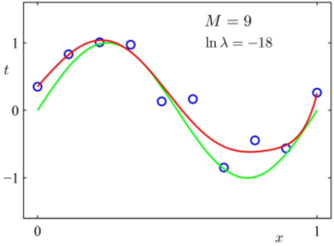
$M = 9$

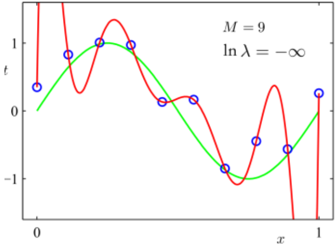
	$\ln\lambda = -\infty$	$\ln\lambda = -18$	$\ln\lambda = 0$
w_0^*	0.35	0.35	0.13
w_1^*	232.37	4.74	-0.05
w_2^*	-5321.83	-0.77	-0.06
w_3^*	48568.31	-31.97	-0.05
w_4^*	-231639.30	-3.89	-0.03
w_5^*	640042.26	55.28	-0.02
w_6^*	-1061800.52	41.32	-0.01
w_7^*	1042400.18	-45.95	-0.00
w_8^*	-557682.99	-91.53	0.00
w_9^*	125201.43	72.68	0.01

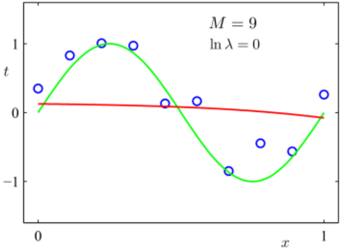


Ridge Regression

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \sum_{j=0}^{M} w_j x^j\right\}^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$







Bayesian Models

- Bayesian estimation

$$\int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathcal{D}) \mathrm{d}\boldsymbol{\theta} \quad \text{(posterior expectation)}$$

- Bayesian inference

$$\widehat{p}_{\text{Bayes}}(\boldsymbol{x}) = \int q(\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) \mathrm{d}\boldsymbol{\theta}$$

$$= \int q(\boldsymbol{x}|\boldsymbol{\theta}) \frac{p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathcal{D})} = \int q(\boldsymbol{x}|\boldsymbol{\theta}) \frac{\prod_{i=1}^{n} q(\boldsymbol{x}_i|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int \prod_{i=1}^{n} q(\boldsymbol{x}_i|\boldsymbol{\theta}') p(\boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}'} \mathrm{d}\boldsymbol{\theta}$$

Bayesian Linear Regression

$$p(t \mid \mathbf{t}, \alpha, \beta) = \int p(t \mid \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \mathbf{t}, \alpha, \beta) \mathrm{d}\boldsymbol{w} \quad \text{(Predictive distribution)}$$

$$p(t \mid \boldsymbol{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}\left(t \mid \boldsymbol{m}_N^T \boldsymbol{\phi}(\boldsymbol{x}), \sigma_N^2(\boldsymbol{x})\right)$$

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