Linear Regression and Classicifation

Wenting Tu

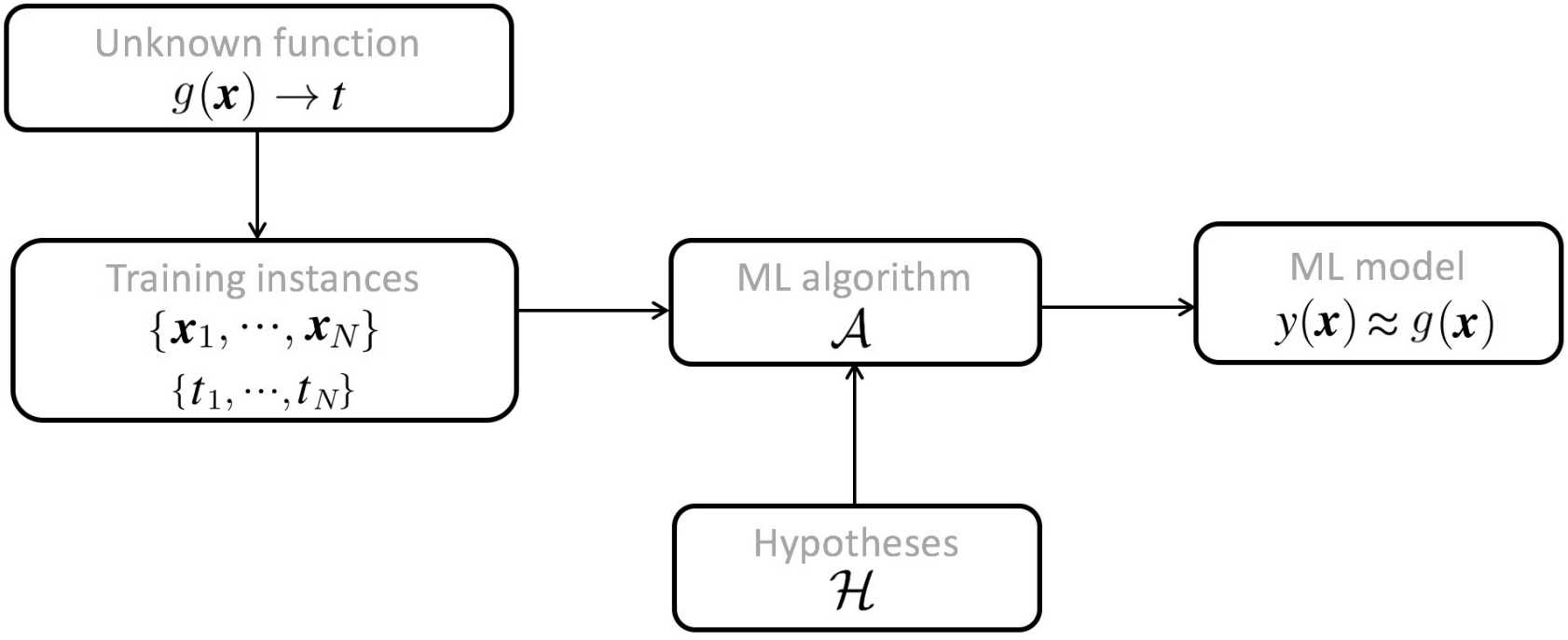
SHUFE, SIME

Machine Learning and Deep Lerning

Course No. 1638

Outline

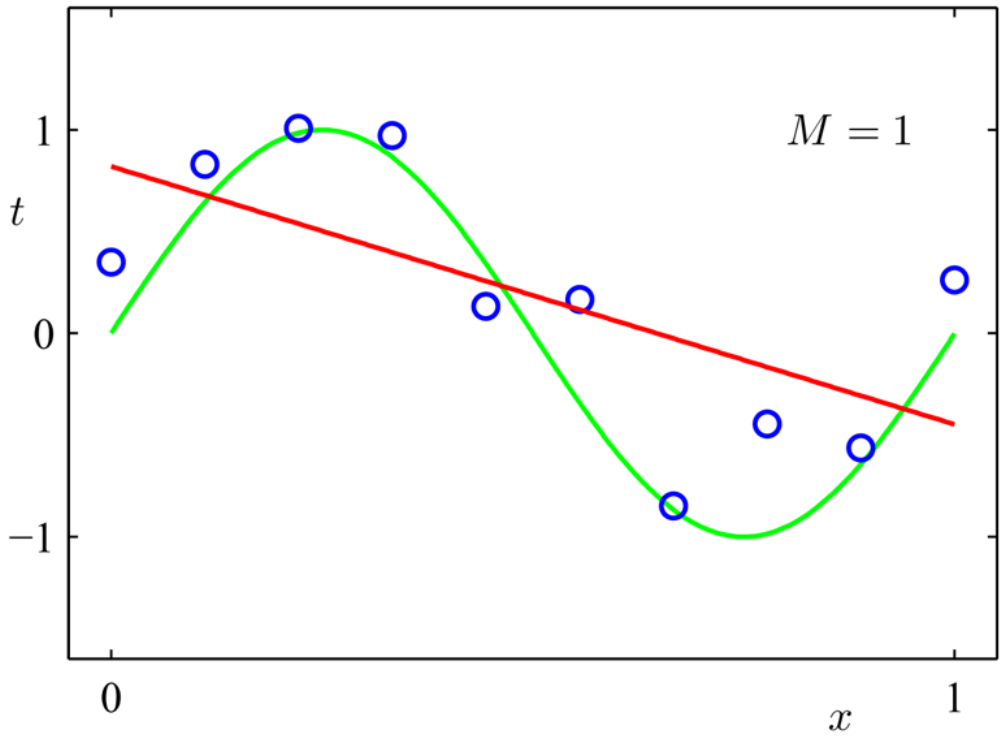
Definition



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

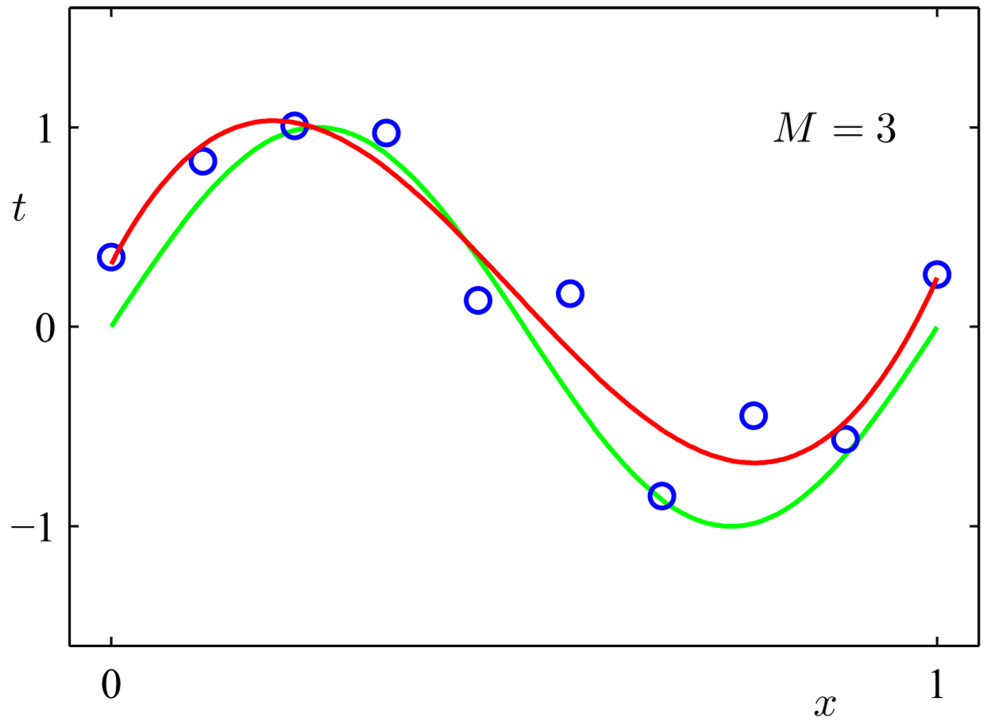
$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x^3 + \cdots + w_D x_D$$

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

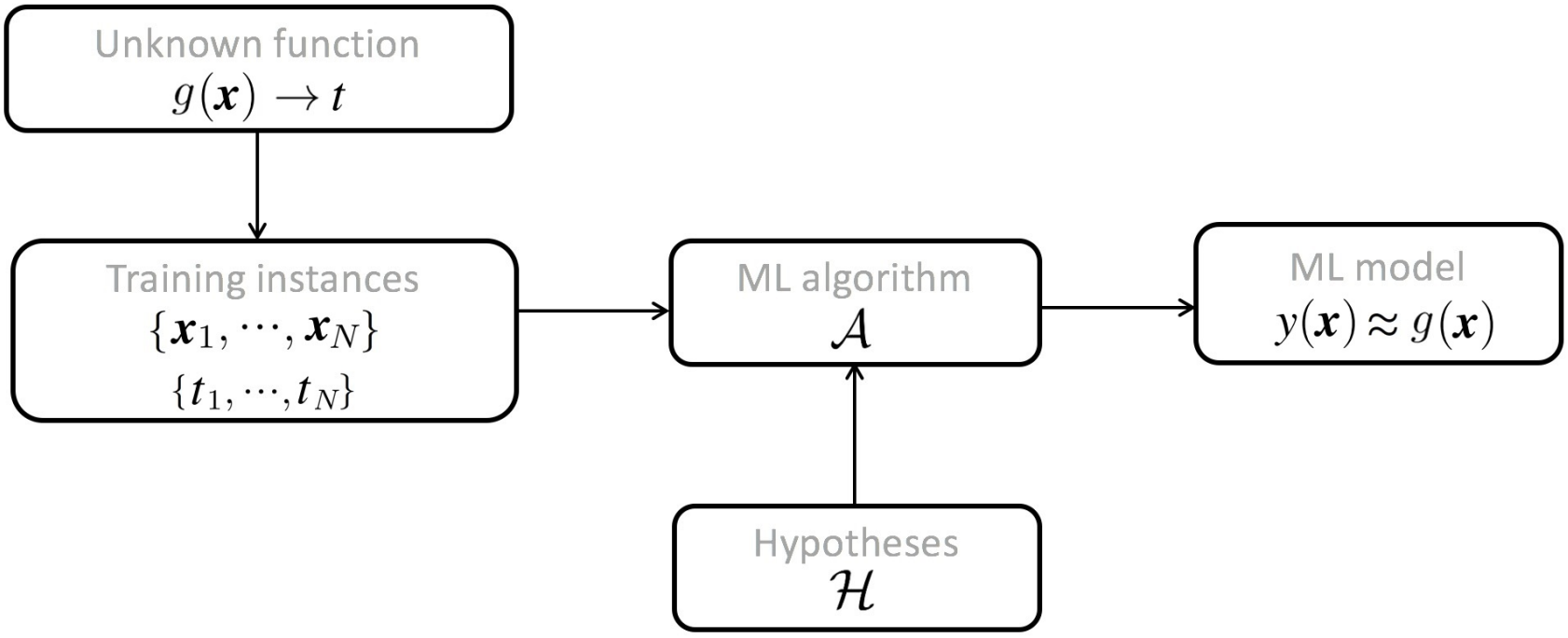
$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\boldsymbol{x})$$

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

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

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

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


Unknown function
$g(\boldsymbol{x}) \rightarrow t$
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)

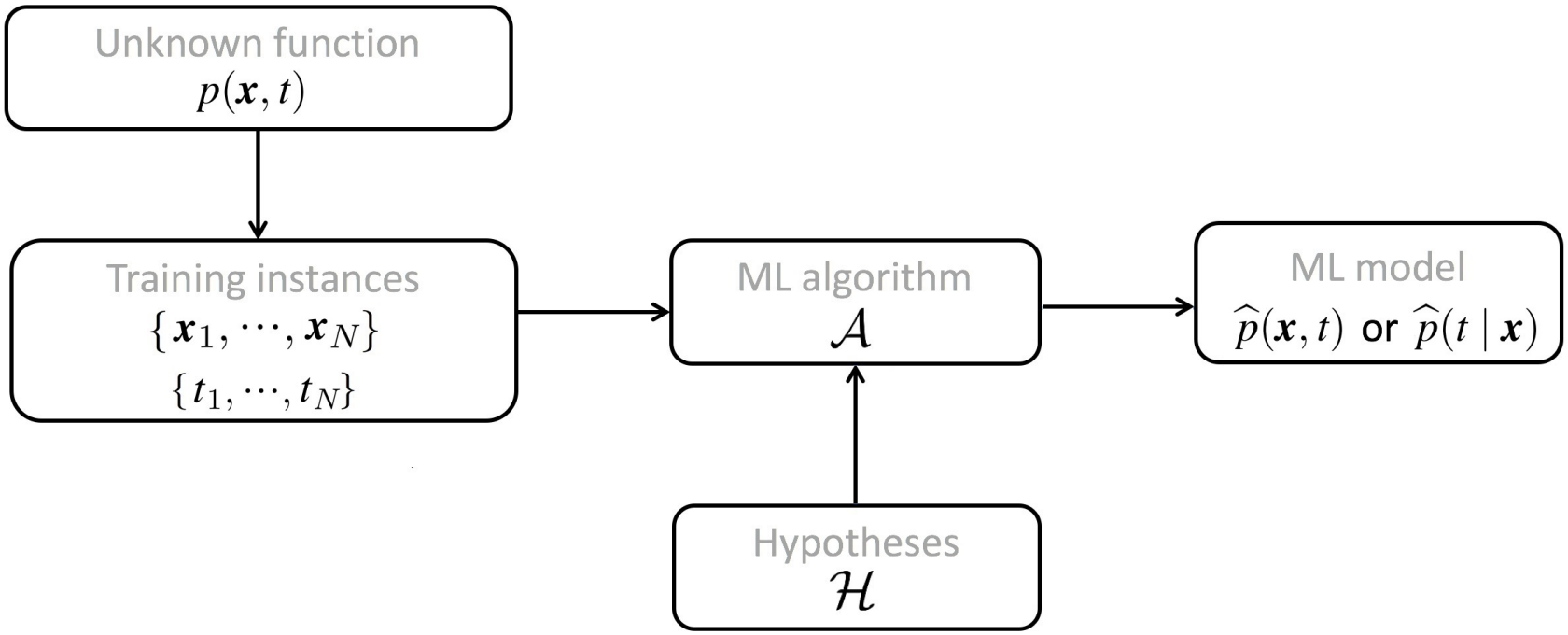
  $q(\boldsymbol{x};\boldsymbol{\theta}) \longrightarrow p(\boldsymbol{x})$

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

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

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

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


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

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

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

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

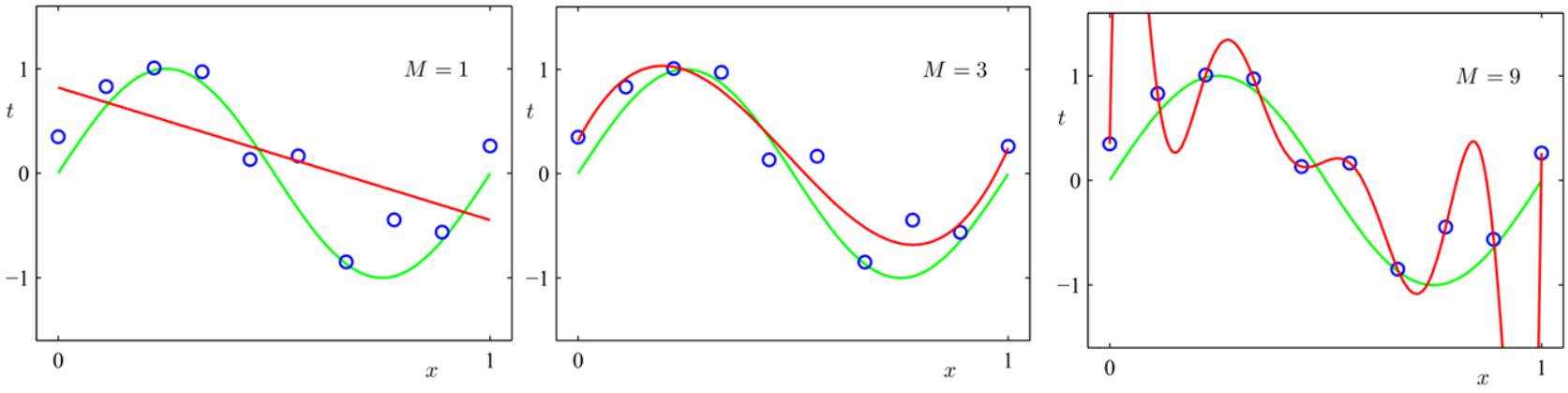
$$p(\mathbf{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \boldsymbol{w}^T \boldsymbol{\phi}\left(\boldsymbol{x}_n\right), \beta^{-1}\right)$$

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

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

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

Illustration



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

Bias-Variance Decomposition

$$\{y(\mathbf{x}) - t\}^2 = \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2$$

$$= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})\mathrm{d}\mathbf{x} + \int \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x}, t)\mathrm{d}\mathbf{x}\mathrm{d}t$$

Bias-Variance Decomposition

Optimal decision $h(\boldsymbol{x}) = \mathbb{E}[t|\boldsymbol{x}] = \int t p(t|\boldsymbol{x}) \mathrm{d}t$

Expected squared loss

$$\mathbb{E}[L] = \int \{y(\boldsymbol{x}) - h(\boldsymbol{x})\}^2 p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \iint \{h(\boldsymbol{x}) - t\}^2 p(\boldsymbol{x}, t) \mathrm{d}\boldsymbol{x} \mathrm{d}t$$

For any given data set $\mathcal{D}$:

$$\begin{aligned}\{y(\boldsymbol{x}; \mathcal{D}) - h(\boldsymbol{x})\}^2 &= \{y(\boldsymbol{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x}; \mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x}; \mathcal{D})] - h(\boldsymbol{x})\}^2 \\ &= \{y(\boldsymbol{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x}; \mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x}; \mathcal{D})] - h(\boldsymbol{x})\}^2 \\ &\quad + 2\{y(\boldsymbol{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x}; \mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x}; \mathcal{D})] - h(\boldsymbol{x})\}\end{aligned}$$

Take the expectation of this expression with respect to $\mathcal{D}$:

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}\left[\{y(\boldsymbol{x}; \mathcal{D}) - h(\boldsymbol{x})\}^2\right] \\ &= \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x}; \mathcal{D})] - h(\boldsymbol{x})\}^2}_{\text{(bias)2}} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\boldsymbol{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x}; \mathcal{D})]\}^2\right]}_{\text{variance}}\end{aligned}$$

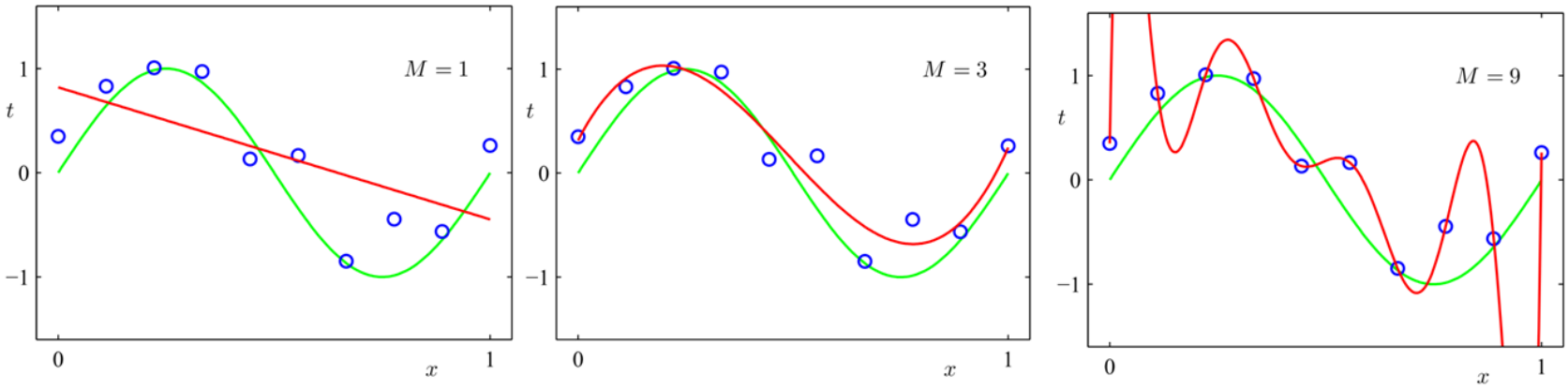
Bias-Variance Decomposition

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

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

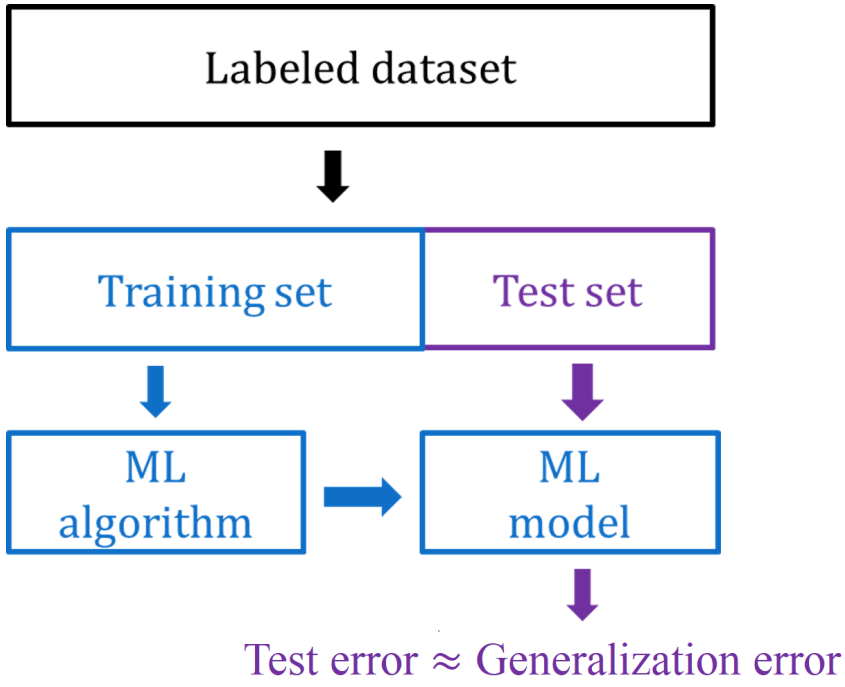
$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}\left[\{y(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})]\}^2\right] p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

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



Model Evaluation

- Generalization Ability



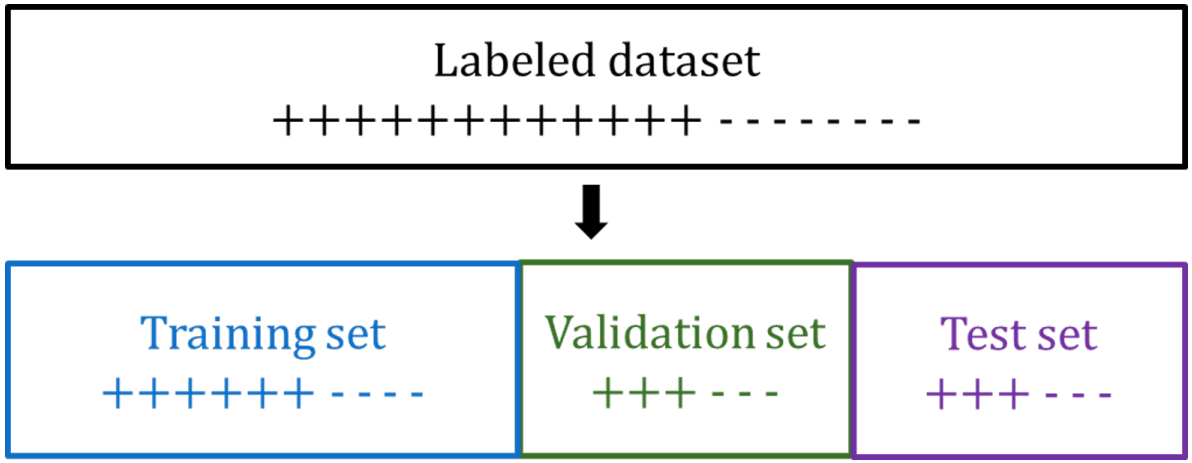
· When learning a model, you should pretend that you don't have the test data yet. If the test-set labels influence the learned model in any way, accuracy estimates will be biased.

· 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.

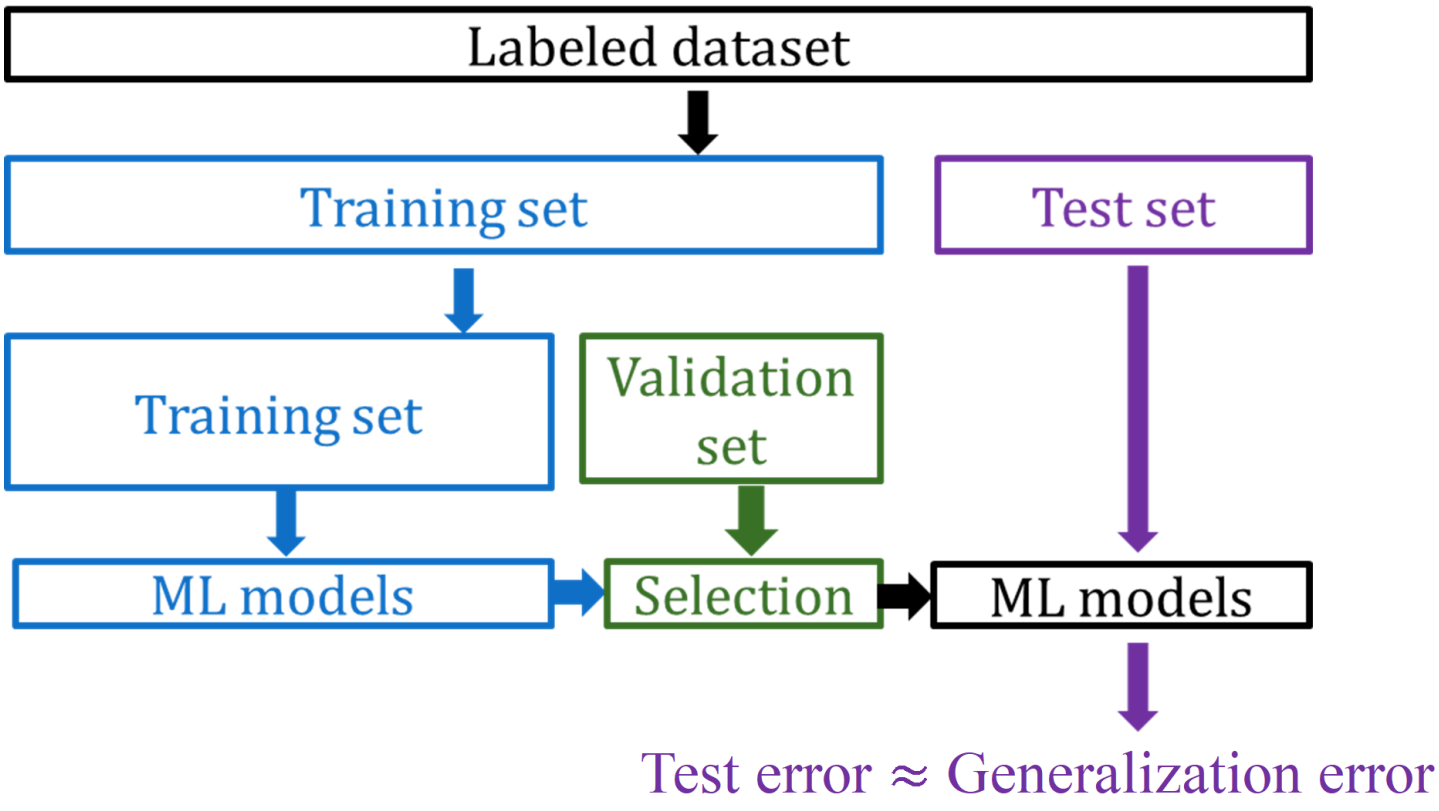
Model Evaluation

- Validation Set



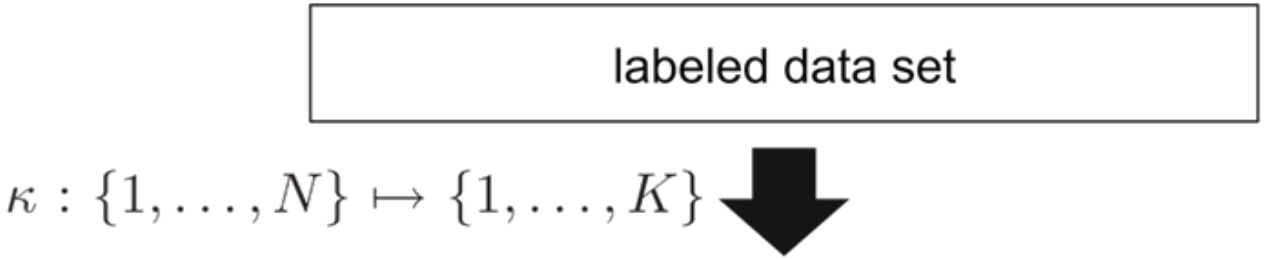
Model Evaluation

- Tuning Hyperparameters



Model Evaluation

- Cross Validation



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

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

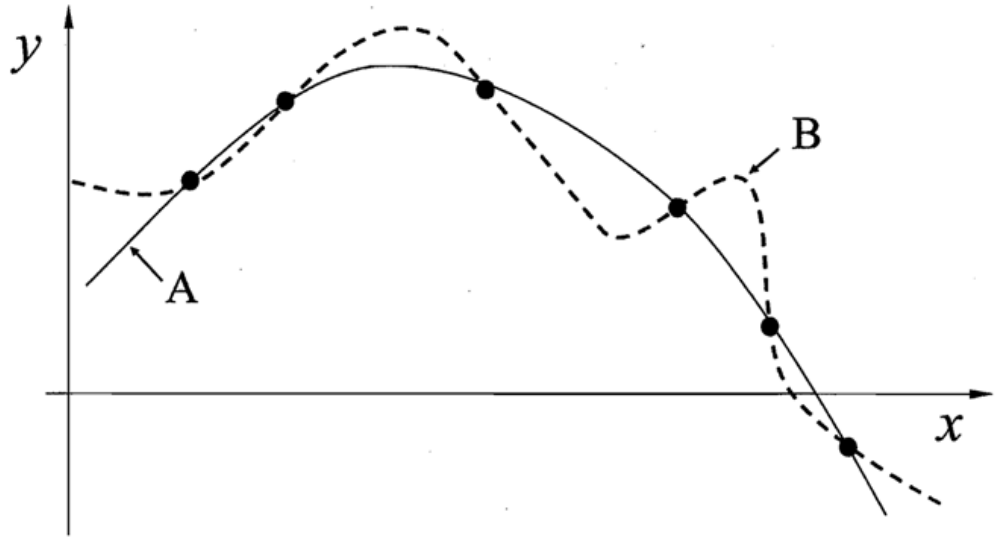
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(t_i - f(\mathbf{x}_i))^2$$

$$RMSE = \sqrt{MSE}$$

Model Selection

- Occam's razor

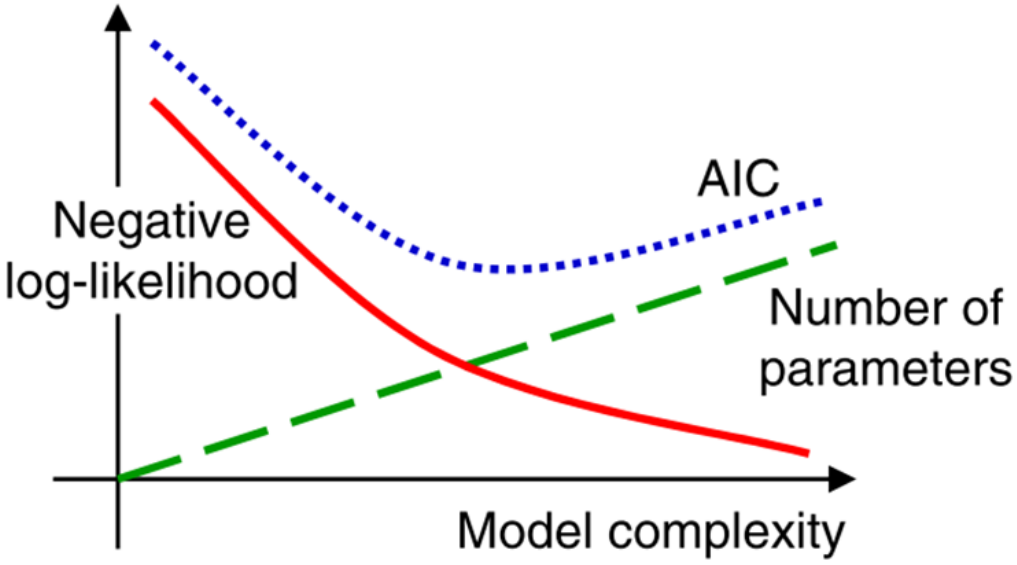
  Suppose there exist two explanations for an occurrence. In this case the one that requires the smallest number of assumptions is usually correct.



Model Selection

- Akaike information criterion (AIC)

  $\ln p\left(\mathcal{D}|\boldsymbol{w}_{ML}\right) - M$



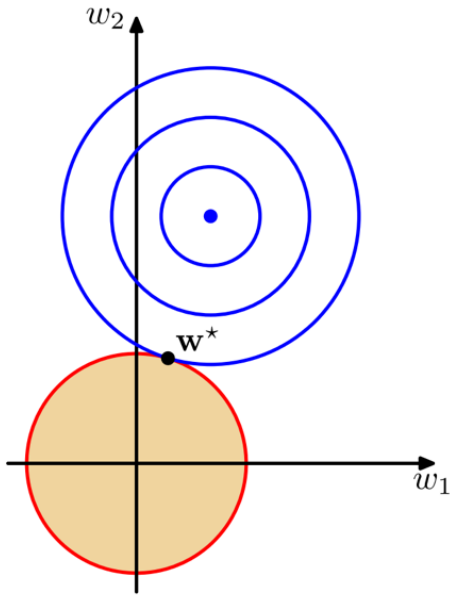
Regularization

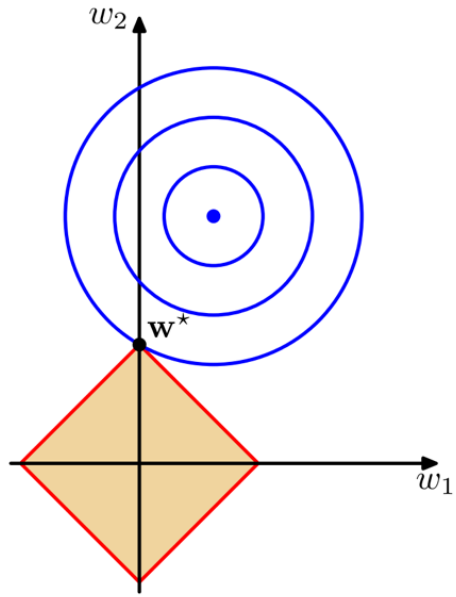
$$E_D(\boldsymbol{w}) + \lambda E_W(\boldsymbol{w})$$

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

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

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

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

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

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

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

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

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

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

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

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

$$\boldsymbol{w}_{MAP}^* = \arg\max_{w} \ln p(\boldsymbol{w}|\mathbf{t})$$

$$= \arg\max_{w} \left(-\frac{\beta}{2} \sum_{n=1}^{N} \left\{ t_n - \boldsymbol{w}^T \boldsymbol{\phi}\left(\boldsymbol{x}_n\right) \right\}^2 - \frac{\alpha}{2} \boldsymbol{w}^T \boldsymbol{w} + \text{const} \right)$$

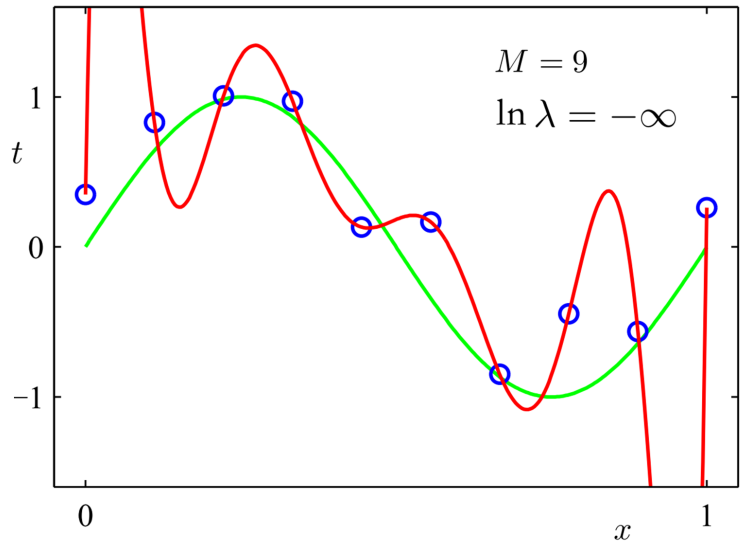
$$\boldsymbol{w}_{MAP}^* = \boldsymbol{w}_{ridge}^\star$$

Ridge Regression

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

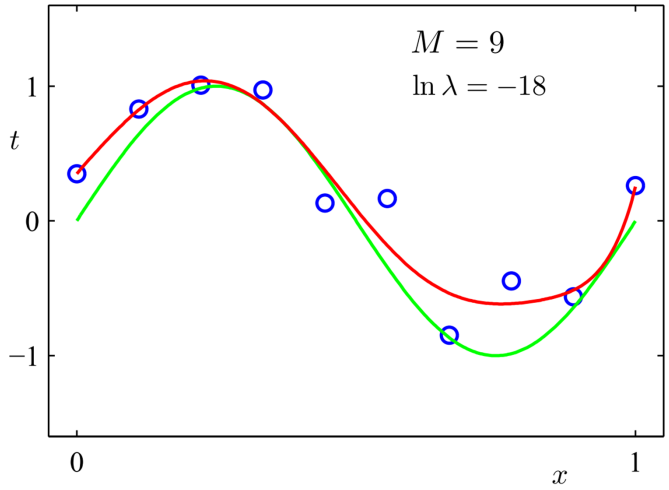
$M = 9$

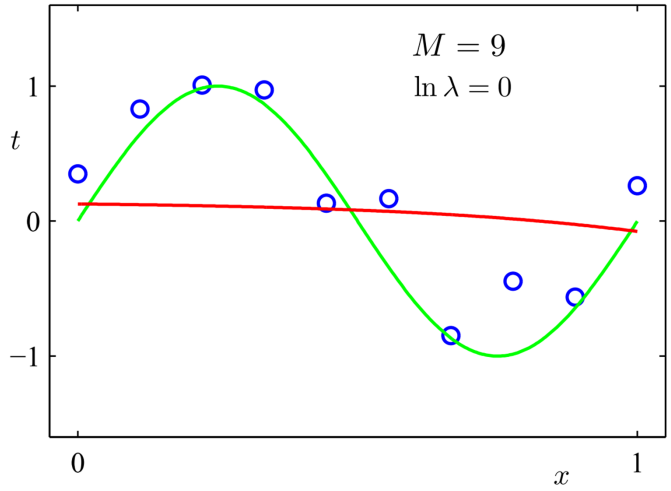
	$\ln\lambda = -\infty$	$\ln\lambda = -18$	$\ln\lambda = 0$
w_0^*	0.35	0.35	0.13
w_1^*	232.37	4.74	-0.05
w_2^*	-5321.83	-0.77	-0.06
w_3^*	48568.31	-31.97	-0.05
w_4^*	-231639.30	-3.89	-0.03
w_5^*	640042.26	55.28	-0.02
w_6^*	-1061800.52	41.32	-0.01
w_7^*	1042400.18	-45.95	-0.00
w_8^*	-557682.99	-91.53	0.00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

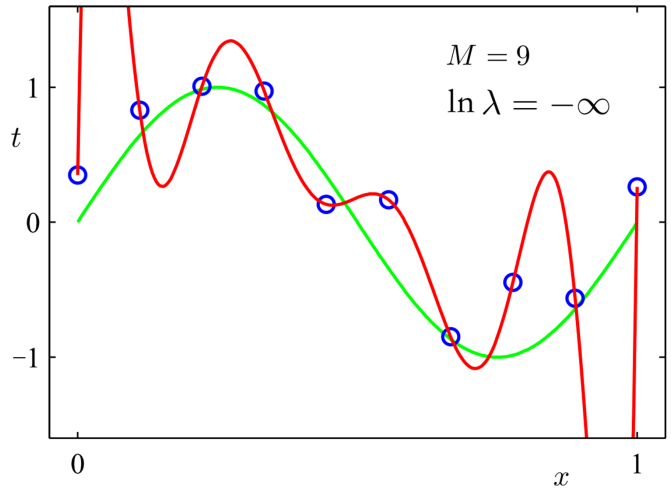


Ridge Regression

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







Bayesian Models

- Bayesian Estimation

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

- Bayesian Estimation

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

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

Bayesian Linear Regression

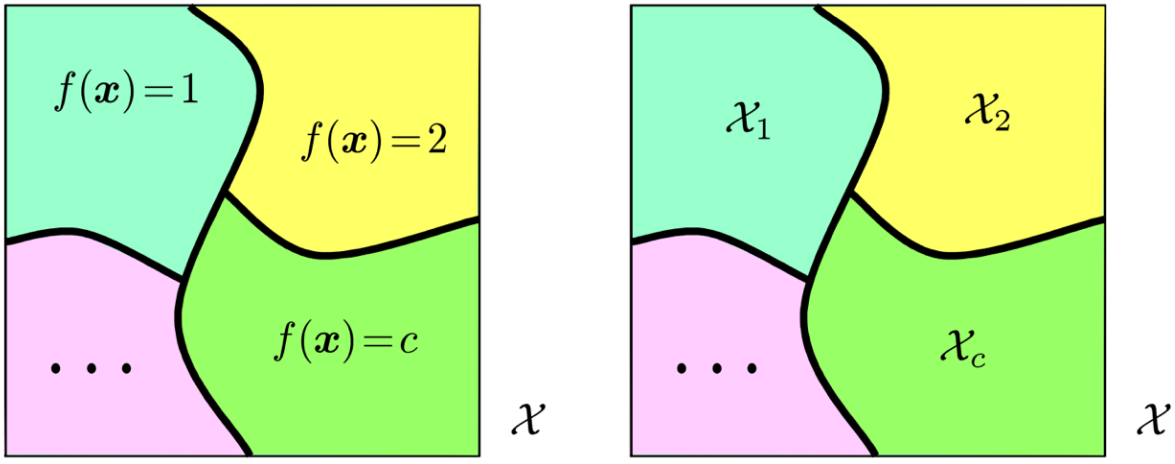
$$p(t \mid \mathbf{t}, \alpha, \beta) = \int p(t \mid \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \mathbf{t}, \alpha, \beta) \mathrm{d}\boldsymbol{w} \quad \text{(Predictive distribution)}$$

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

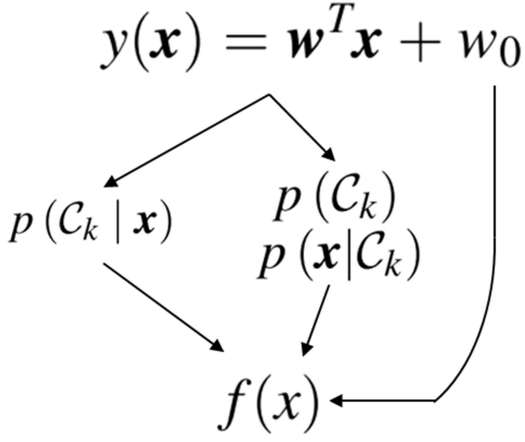
$$\sigma_N^2(\boldsymbol{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\phi}(\boldsymbol{x})$$

Outline

What is Linear classification



- Probabilistic Discriminative Models
- Probabilistic Generative Models
- Discriminant Functions



Least squares for classification?

$$y_k(\boldsymbol{x}) = \boldsymbol{w}_k^T \boldsymbol{x} + w_{k0}$$

$$\boldsymbol{y}(\boldsymbol{x}) = \tilde{\boldsymbol{W}}^T \tilde{\boldsymbol{x}},\ \tilde{\boldsymbol{w}}_k = \left(w_{k0}, \boldsymbol{w}_k^T\right)^T,\ \tilde{\boldsymbol{x}} = \left(1, \boldsymbol{x}^T\right)^T$$

$$\left\{\boldsymbol{x}_n, \boldsymbol{t}_n\right\}, n = 1, \ldots, N$$

$$\tilde{\boldsymbol{X}} \ - \ n^{\text{th}} \text{ row } \ - \ \tilde{\mathbf{x}}_n^{\mathrm{T}}$$

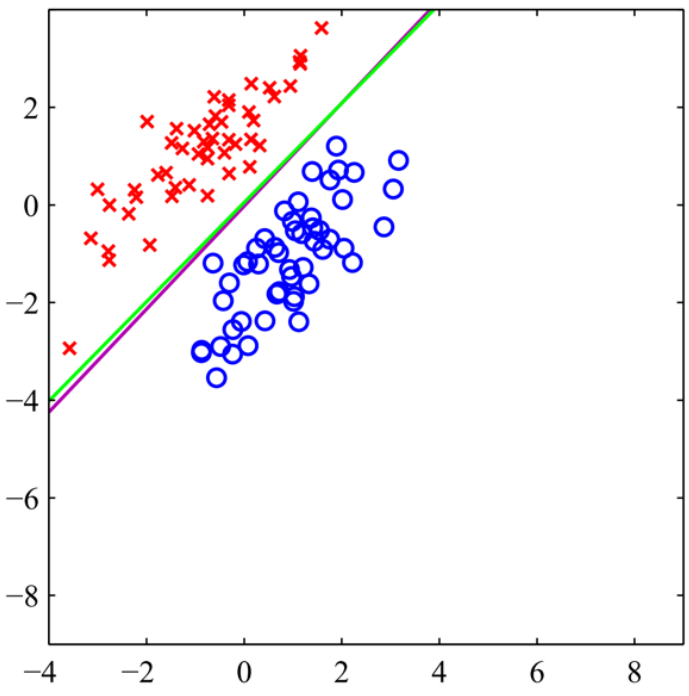
$$\boldsymbol{T} \ - \ n^{\text{th}} \text{ row } \ - \ \mathbf{t}_n^{\mathrm{T}}$$

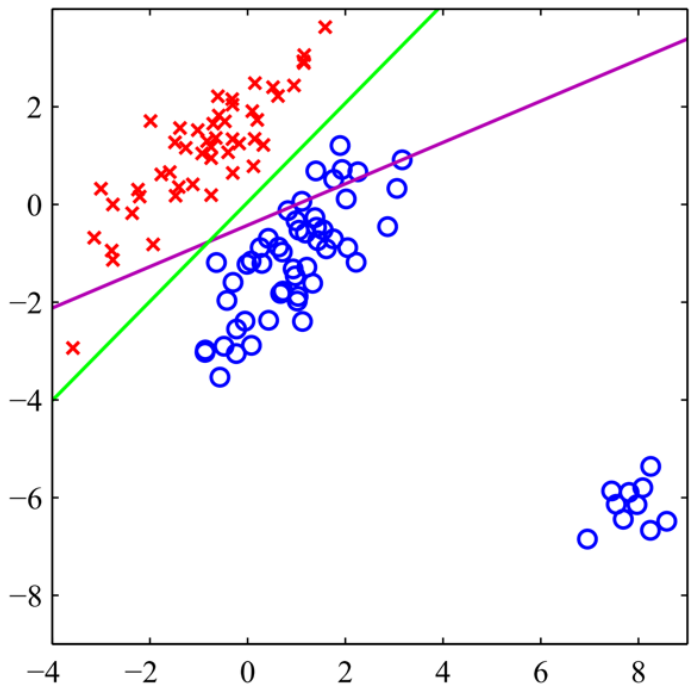
$$E_D(\tilde{\boldsymbol{W}}) = \frac{1}{2} \operatorname{Tr}\left\{(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{W}} - \boldsymbol{T})^T (\tilde{\boldsymbol{X}}\tilde{\boldsymbol{W}} - \boldsymbol{T})\right\}$$

$$\tilde{\boldsymbol{W}} = \left(\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}}\right)^{-1} \tilde{\boldsymbol{X}}^T \boldsymbol{T} = \tilde{\boldsymbol{X}}^{\dagger} \boldsymbol{T}$$

$$y(\boldsymbol{x}) = \tilde{\boldsymbol{W}}^T \tilde{\boldsymbol{x}} = \boldsymbol{T}^T \left(\tilde{\boldsymbol{X}}^{\dagger}\right)^T \tilde{\boldsymbol{x}}$$

Least squares for classification?





Least squares is highly sensitive to outliers.

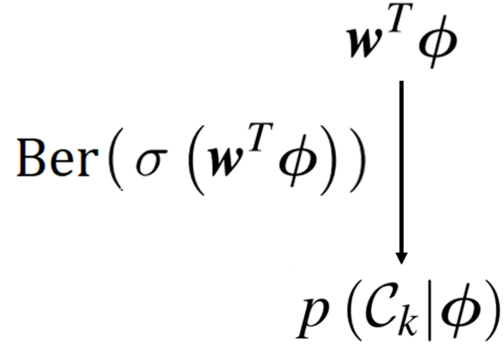
Probabilistic Discriminative Models

- Logistic regression

$$p(\mathcal{C}_1 \mid \boldsymbol{\phi}) = \sigma\left(\boldsymbol{w}^T \boldsymbol{\phi}\right)$$

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

$$p(\mathcal{C}_2 \mid \boldsymbol{\phi}) = 1 - p(\mathcal{C}_1 \mid \boldsymbol{\phi})$$

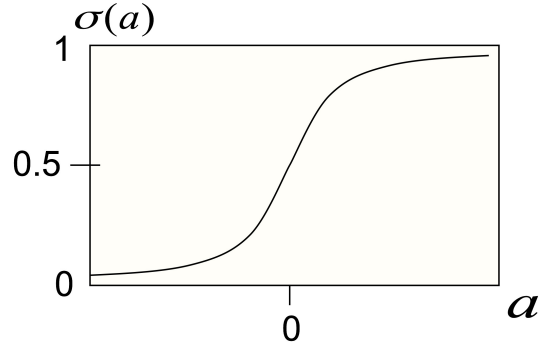


Why sigmoid function?

$$p(\mathcal{C}_1 \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \mathcal{C}_1)\, p(\mathcal{C}_1)}{p(\boldsymbol{x} \mid \mathcal{C}_1)\, p(\mathcal{C}_1) + p(\boldsymbol{x} \mid \mathcal{C}_2)\, p(\mathcal{C}_2)}$$

$$p(\mathcal{C}_1 \mid \boldsymbol{x}) = \frac{1}{1 + \exp(-a)} = \sigma(a)$$

$$a = \ln \frac{p(\boldsymbol{x} \mid \mathcal{C}_1)\, p(\mathcal{C}_1)}{p(\boldsymbol{x} \mid \mathcal{C}_2)\, p(\mathcal{C}_2)} = \ln \frac{p(\mathcal{C}_1 \mid \boldsymbol{x})}{p(\mathcal{C}_2 \mid \boldsymbol{x})}$$



Probabilistic Discriminative Models

- Logistic regression

$$p(\mathbf{t} \mid \boldsymbol{w}) = \prod_{n=1}^{N} \{p\left(\mathcal{C}_1 \mid \boldsymbol{\phi}_n\right)\}^{t_n} \{1 - p\left(\mathcal{C}_1 \mid \boldsymbol{\phi}_n\right)\}^{1-t_n}$$

$$y_n = p\left(\mathcal{C}_1 \mid \boldsymbol{\phi}_n\right)$$

$$E(\boldsymbol{w}) = -\ln p(\mathbf{t} \mid \boldsymbol{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln (1 - y_n)\}$$

$$\nabla E(\boldsymbol{w}) = \sum_{n=1}^{N} (y_n - t_n)\, \boldsymbol{\phi}_n$$

$$\boldsymbol{w}^{(\tau+1)} = \boldsymbol{w}^{(\tau)} - \eta \nabla E_n$$

Probabilistic Discriminative Models

- Softmax regression

$$p\left(\mathcal{C}_k|\boldsymbol{\phi}\right) = y_k(\boldsymbol{\phi}) = \frac{\exp\left(a_k\right)}{\sum_j \exp\left(a_k\right)}$$

$$\boldsymbol{w}_k^T\boldsymbol{\phi} \;\Big|\; \text{Multi}\left\{\cdots \frac{\exp\left(\boldsymbol{w}_k^T\boldsymbol{\phi}\right)}{\sum_j \exp\left(\boldsymbol{w}_k^T\boldsymbol{\phi}\right)} \cdots\right\} \;\Big|\; p\left(\mathcal{C}_k \mid \boldsymbol{x}\right)$$

$$a_k = \boldsymbol{w}_k^T\boldsymbol{\phi}$$

$$p\left(\boldsymbol{T}|\boldsymbol{w}_1,\ldots,\boldsymbol{w}_K\right) = \prod_{n=1}^{N}\prod_{k=1}^{K} p\left(\mathcal{C}_k|\boldsymbol{\phi}_n\right)^{t_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

Cross-entropy error function

$$E\left(\boldsymbol{w}_1,\ldots,\boldsymbol{w}_K\right) = -\ln p\left(\boldsymbol{T}|\boldsymbol{w}_1,\ldots,\boldsymbol{w}_K\right) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln y_{nk}$$

$$\nabla_{\boldsymbol{w}_j} E\left(\boldsymbol{w}_1,\ldots,\boldsymbol{w}_K\right) = \sum_{n=1}^{N}\left(y_{nj} - t_{nj}\right)\boldsymbol{\phi}_n$$

Probabilistic Generative Models

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0$$

$$\downarrow$$

$$p\left(\boldsymbol{x}|\mathcal{C}_k\right)$$

$$p\left(\mathcal{C}_k\right)$$

$$\downarrow$$

$$f(x)$$

Probabilistic Generative Models

- Linear discriminant

  $$p(\boldsymbol{x} \mid \mathcal{C}_k) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}$$

  (assuming that features are continuous and all classes share the same covariance matrix)

Probabilistic Generative Models

- Linear discriminant

$$p(\boldsymbol{x} \mid \mathcal{C}_k) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}$$

(assuming that features are continuous and all classes share the same covariance matrix)

Linear?

$$p(\mathcal{C}_1 \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \mathcal{C}_1)\, p(\mathcal{C}_1)}{p(\boldsymbol{x} \mid \mathcal{C}_1)\, p(\mathcal{C}_1) + p(\boldsymbol{x} \mid \mathcal{C}_2)\, p(\mathcal{C}_2)} = \frac{1}{1+\exp(-a)} = \sigma(a)$$

$$a = \ln \frac{p(\boldsymbol{x} \mid \mathcal{C}_1)\, p(\mathcal{C}_1)}{p(\boldsymbol{x} \mid \mathcal{C}_2)\, p(\mathcal{C}_2)}$$

$$p(\mathcal{C}_1 \mid \boldsymbol{x}) = \sigma\left(\boldsymbol{w}^T \boldsymbol{x} + w_0\right)$$

$$\boldsymbol{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad w_0 = -\frac{1}{2}\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 + \ln \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

Probabilistic Generative Models

- Linear discriminant

$$p(\boldsymbol{x} \mid \mathcal{C}_k) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}$$

(assuming that features are continuous and all classes share the same covariance matrix)

Linear?

$$p(\mathcal{C}_k \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \mathcal{C}_k)\, p(\mathcal{C}_k)}{\sum_j p(\boldsymbol{x} \mid \mathcal{C}_j)\, p(\mathcal{C}_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

$$a_k = \ln p\left((\boldsymbol{x} \mid \mathcal{C}_k)\, p(\mathcal{C}_k)\right)$$

$$a_k(\boldsymbol{x}) = \boldsymbol{w}_k^T \boldsymbol{x} + w_{k0}$$

$$\boldsymbol{w}_k = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k \quad w_{k0} = -\frac{1}{2}\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln p(\mathcal{C}_k)$$

Probabilistic Generative Models

- Maximum likelihood solution for Linear discriminant

$\{\boldsymbol{x}_n,\ t_n\}_{n=1}^{N}, t_n = 1 \longleftrightarrow \mathcal{C}_1,\ t_n = 0 \longleftrightarrow \mathcal{C}_2$

$p(\mathcal{C}_1) = \pi,\ p(\mathcal{C}_2) = 1 - \pi$

$p(\boldsymbol{x}_n, \mathcal{C}_1) = p(\mathcal{C}_1)\, p(\boldsymbol{x}_n \mid \mathcal{C}_1) = \pi \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma})$

$p(\boldsymbol{x}_n, \mathcal{C}_2) = p(\mathcal{C}_2)\, p(\boldsymbol{x}_n \mid \mathcal{C}_2) = (1 - \pi) \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$

$$p(\mathbf{t}, \boldsymbol{X} \mid \pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \left[\pi \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma})\right]^{t_n} \left[(1 - \pi) \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma})\right]^{1 - t_n}$$

The terms in the log likelihood function that depend on π is

$$\sum_{n=1}^{N} \{t_n \ln \pi + (1 - t_n) \ln(1 - \pi)\}$$

Thus, we obtain

$$\pi = \frac{1}{N} \sum_{n=1}^{N} t_n = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$$

Probabilistic Generative Models

- Maximum likelihood solution for Linear discriminant
  The terms in the log likelihood function that depend on $\boldsymbol{\mu}_1$ is

$$\sum_{n=1}^{N} t_n \ln \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}\right) = -\frac{1}{2} \sum_{n=1}^{N} t_n \left(\mathbf{x}_n - \boldsymbol{\mu}_1\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_n - \boldsymbol{\mu}_1\right)$$

  Thus, we obtain

$$\boldsymbol{\mu}_1 = \frac{1}{N_1} \sum_{n=1}^{N} t_n \boldsymbol{x}_n, \ \boldsymbol{\mu}_2 = \frac{1}{N_2} \sum_{n=1}^{N} \left(1 - t_n\right) \boldsymbol{x}_n$$

Probabilistic Generative Models

- Maximum likelihood solution for Linear discriminant
  The terms in the log likelihood function that depend on $\boldsymbol{\Sigma}$ is

$$-\frac{1}{2}\sum_{n=1}^{N} t_n \ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} t_n \left(\mathbf{x}_n - \boldsymbol{\mu}_1\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_n - \boldsymbol{\mu}_1\right)$$

$$-\frac{1}{2}\sum_{n=1}^{N} (1-t_n) \ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} (1-t_n) \left(\mathbf{x}_n - \boldsymbol{\mu}_2\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_n - \boldsymbol{\mu}_2\right.$$

$$= -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{N}{2}\mathrm{Tr}\left\{\boldsymbol{\Sigma}^{-1}\mathbf{S}\right\}$$

$$\boldsymbol{S} = \frac{N_1}{N}\boldsymbol{S}_1 + \frac{N_2}{N}\boldsymbol{S}_2$$

$$\boldsymbol{S}_1 = \frac{1}{N_1}\sum_{n\in\mathcal{C}_1} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_1\right)\left(\boldsymbol{x}_n - \boldsymbol{\mu}_1\right)^T, \boldsymbol{S}_2 = \frac{1}{N_2}\sum_{n\in\mathcal{C}_2} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_2\right)\left(\boldsymbol{x}_n - \boldsymbol{\mu}_2\right)^T$$

$$\boldsymbol{\Sigma} = \boldsymbol{S}$$

Probabilistic Generative Models

- Naïve-Bayes (NB) classifier
- Conditional Independence Assumption

  $x_i \perp x_{\{j \neq i\}} \mid t$
- Bernoulli NB classifier

  $x_i \in \{0, 1\}$ & $p(x_i \mid \mathcal{C}_k) \sim \text{Ber}(\mu_{ki})$

$$p(\boldsymbol{x} \mid \mathcal{C}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i} (1 - \mu_{ki})^{1 - x_i}$$

$$p(\mathcal{C}_k \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \mathcal{C}_k) p(\mathcal{C}_k)}{\sum_j p(\boldsymbol{x} \mid \mathcal{C}_j) p(\mathcal{C}_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

$$a_k = \ln p((\boldsymbol{x} \mid \mathcal{C}_k) p(\mathcal{C}_k))$$

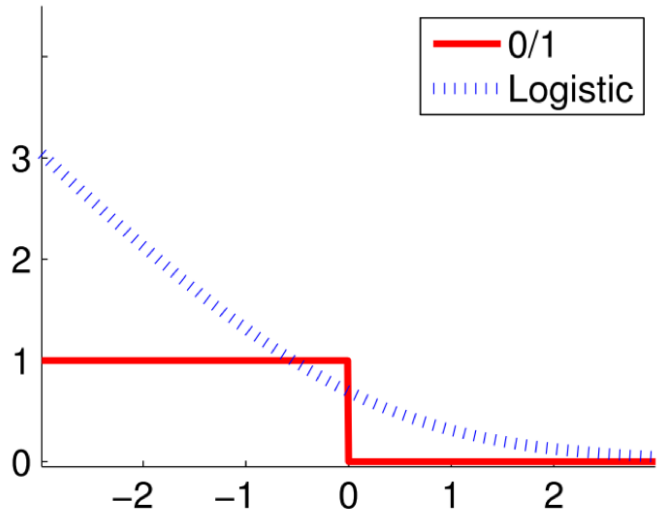
$$a_k(\boldsymbol{x}) = \sum_{i=1}^{D} \{x_i \ln \mu_{ki} + (1 - x_i) \ln (1 - \mu_{ki})\} + \ln p(\mathcal{C}_k)$$

Hinge Loss and Support Vector Machines

- Loss Functions for Classification

$t_n \in \{-1, 1\}$

$y_n > 0 \leftrightarrow \widehat{t}_n = 1,\ y_n < 0 \leftrightarrow \widehat{t}_n = -1$



· 0-1 loss

$$E_{o/1}\left(t_n, y_n\right) = 1 - \text{sign}\left\{t_n y\left(\boldsymbol{x}_n\right)\right\}$$

· Log loss

$$E_{log}\left(t_n, y_n\right) = \ln\left\{1 + \exp\left(-y_n t_n\right)\right\}$$

equals to

$$E_{\text{cross-ent}}\left(t_n, y_n\right) = \left\{t_n \ln y_n + \left(1 - t_n\right) \ln\left(1 - y_n\right)\right\} \left(t_n \in \{0, 1\}\right)$$

Hinge Loss and Support Vector Machines

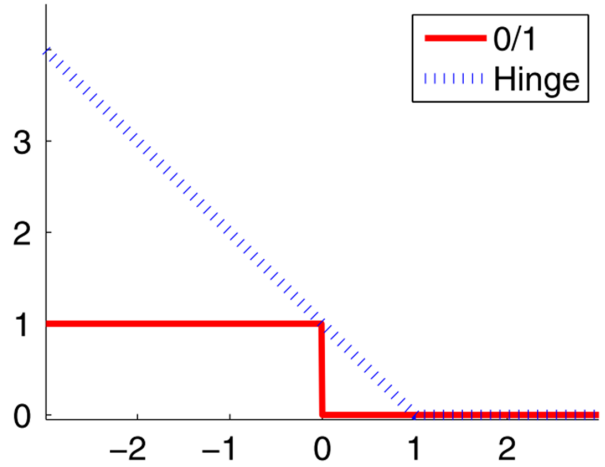
- Hinge Loss

$t_n \in \{-1, 1\}$

$y_n > 0 \rightarrow \hat{t}_n = 1,\ y_n < 0 \rightarrow \hat{t}_n = -1$

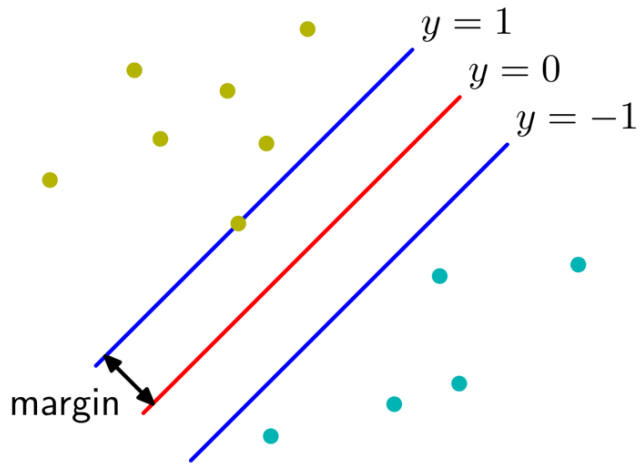
$E_{\text{Hinge}}(t_n, y_n) = [1 - y_n t_n]_+$

$[\cdot]_+$ denotes the positive part



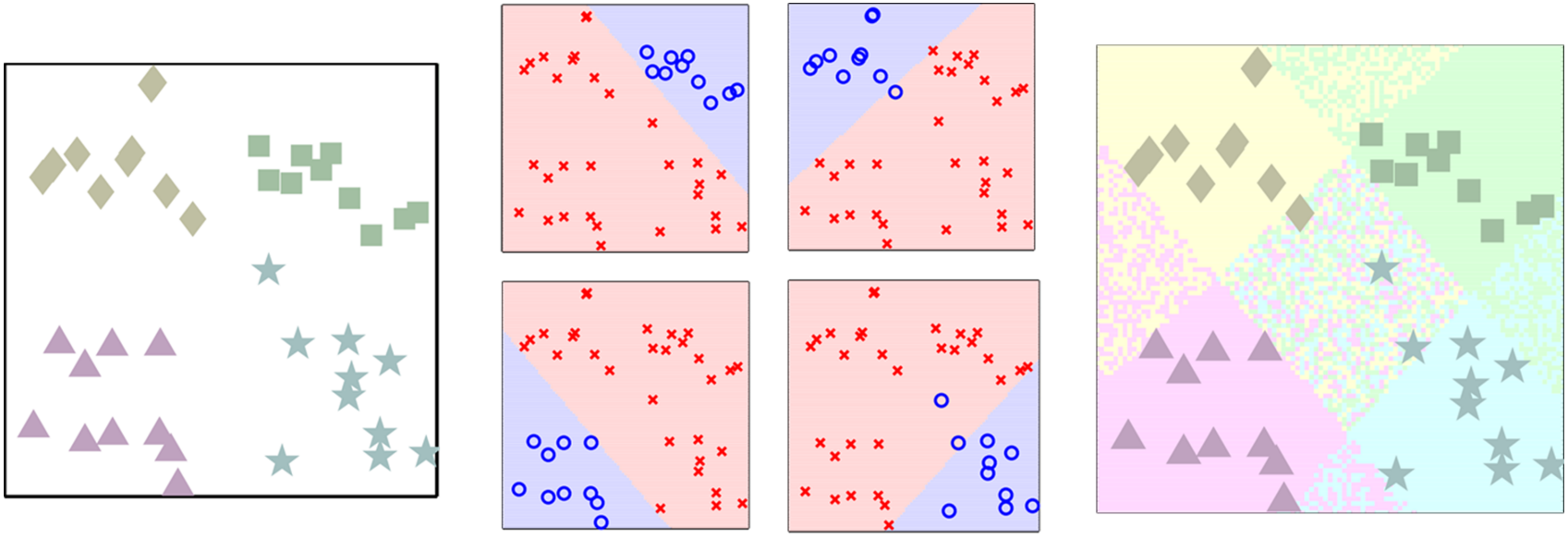
- Support Vector Classifier

$L_{SVC} = \sum_{n=1}^{N} E_{\text{Hinge}}(t_n, y_n) + \lambda \|\boldsymbol{w}\|^2$



Multiclass Classification

- One-versus-the-rest

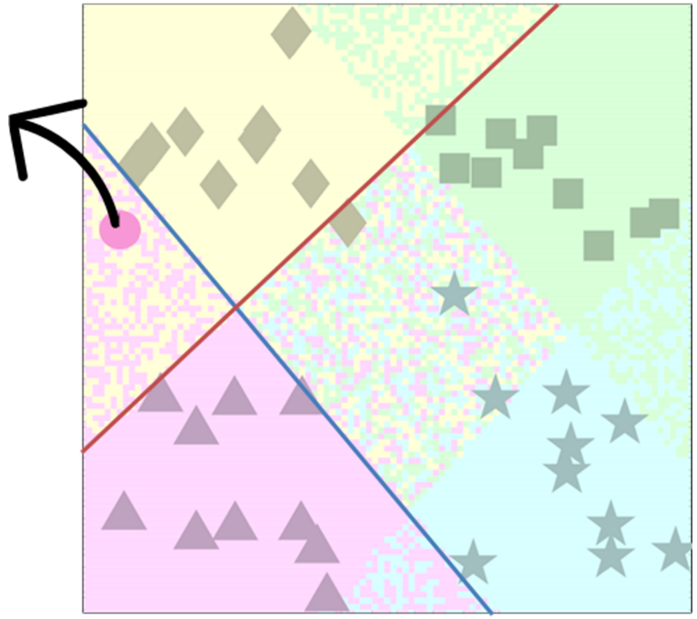


For K classes, we have K classifiers.

Multiclass Classification

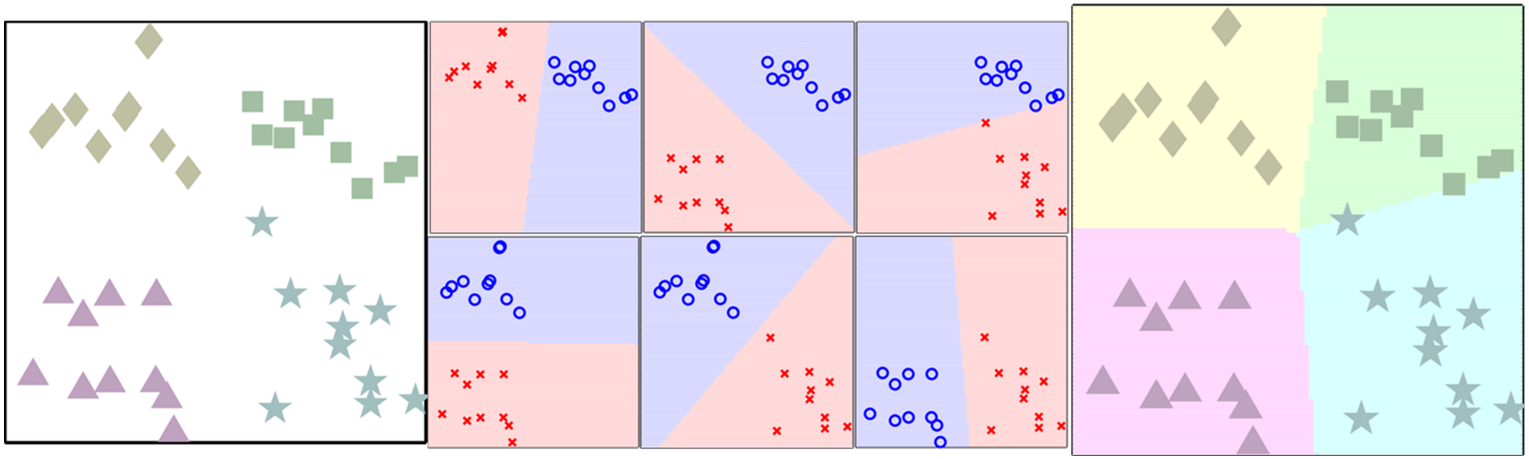
- One-versus-the-rest

How to choose the one that makes the strongest prediction?



Multiclass Classification

- One-versus-one



Model Evaluation for Classification

- Performance Matrices
  - · Confusion matrix

		Actual	
		Class +	Class -
Predicted	Class +	TP	FP
	Class -	FN	TN

  - · Accuracy

$$\frac{TP + TN}{TP + FP + FN + TN}$$

  - · Error rate

$$\frac{FP + FN}{TP + FP + FN + TN}$$

Model Evaluation for Classification

- Performance Matrices
  - · Confusion matrix

		Actual	
		Class +	Class -
Predicted	Class +	TP	FP
	Class -	FN	TN

  - · Precision

$$\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$$

  - · Recall

$$\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$$

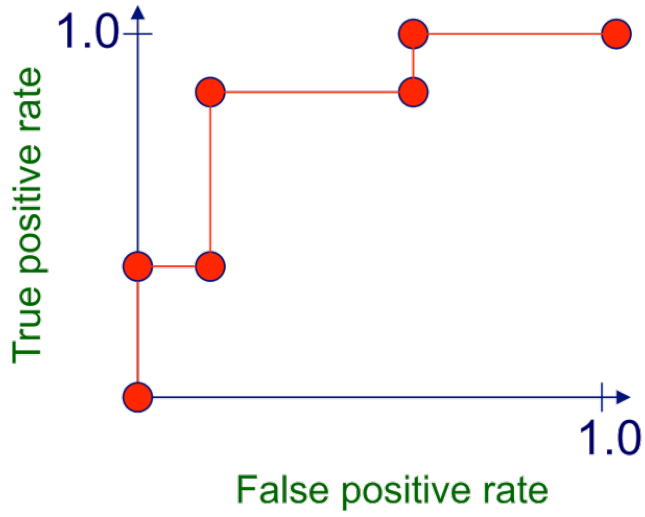
  - · F-measure

$$F_\beta = (1+\beta^2)\cdot\frac{\text{precision}\cdot\text{recall}}{(\beta^2\cdot\text{precision})+\text{recall}}$$

Model Evaluation for Classification

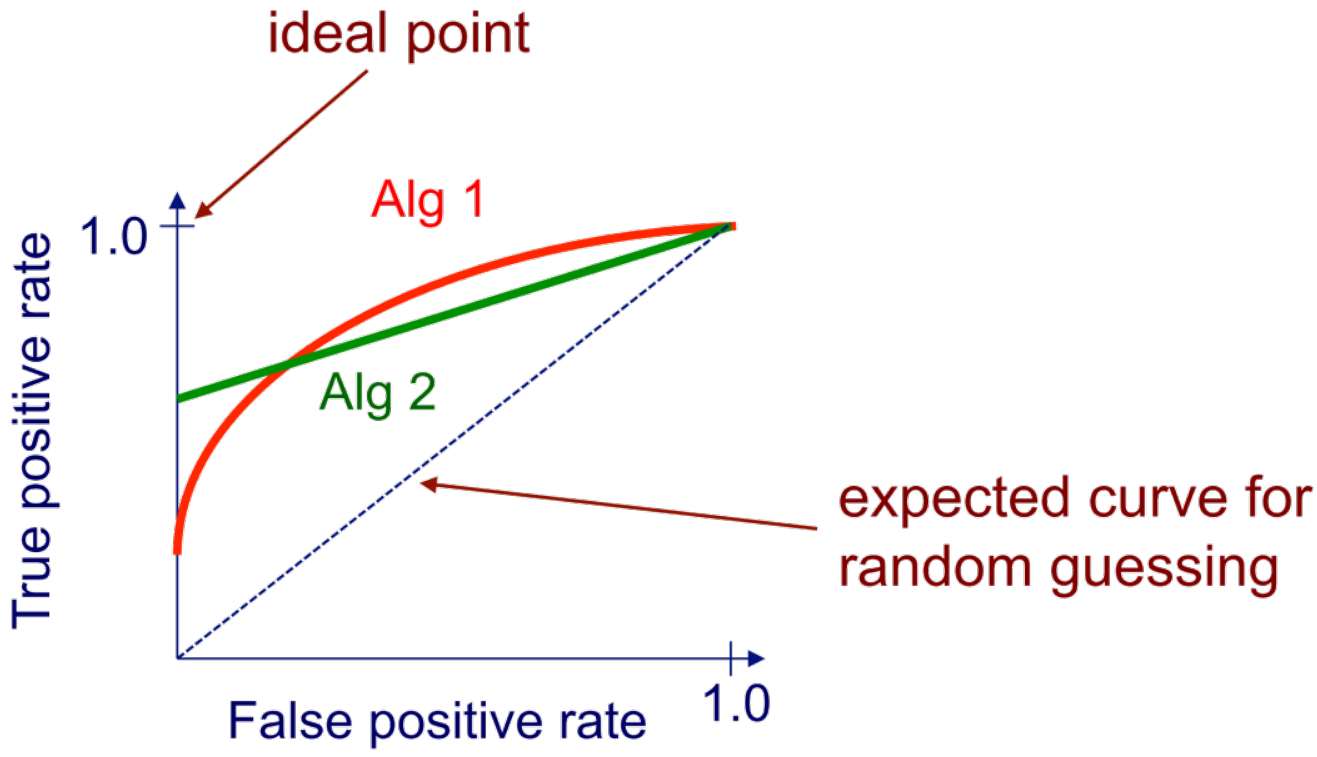
- Performance Matrices
  - · ROC curve

instance	confidence positive		correct class
Ex 9	.99		+
Ex 7	.98	TPR= 2/5, FPR= 0/5	+
Ex 1	.72	TPR= 2/5, FPR= 1/5	-
Ex 2	.70		+
Ex 6	.65	TPR= 4/5, FPR= 1/5	+
Ex 10	.51		-
Ex 3	.39	TPR= 4/5, FPR= 3/5	-
Ex 5	.24	TPR= 5/5, FPR= 3/5	+
Ex 4	.11		-
Ex 8	.01	TPR= 5/5, FPR= 5/5	-



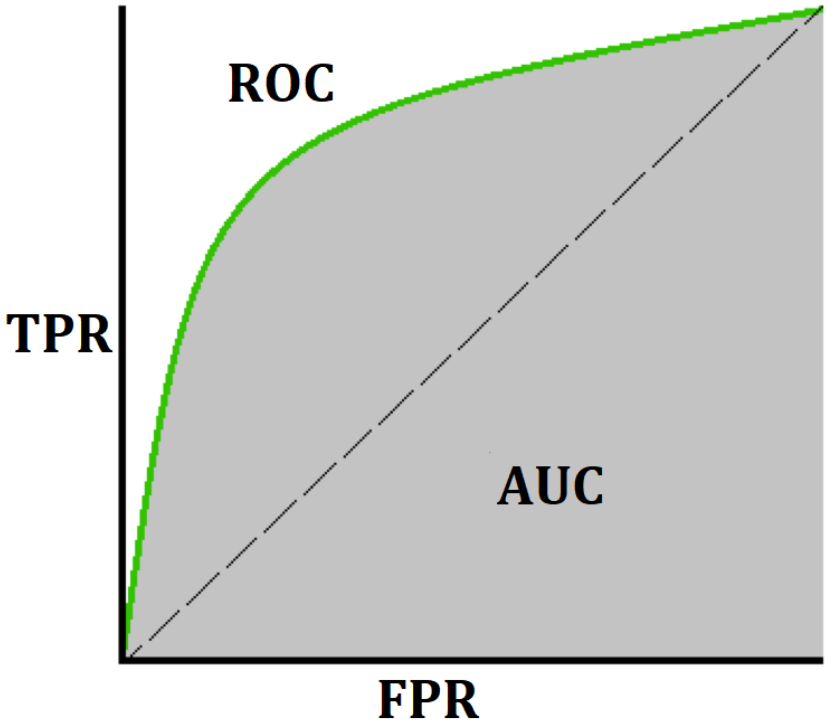
Model Evaluation for Classification

- Performance Matrices
  - · ROC curve



Model Evaluation for Classification

- Performance Matrices
  - AUC



Thanks

Some images and slides are from the internet.
If related to copyright, please contact me.

tu.wenting@mail.shufe.edu.cn