Linear Regression and Classifcation

Wenting Tu

SHUFE, SIME

Machine Learning and Deep Lerning

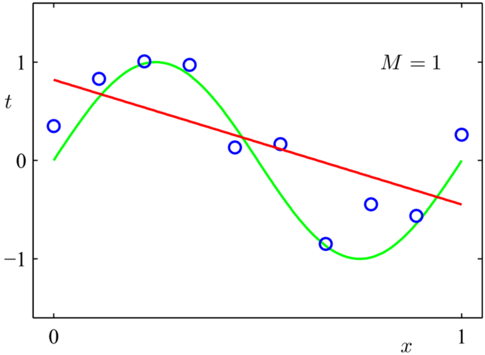
Course No. 1638

Outline

Linear Regression

Linear Classification

Illustration



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

Linear basis function models

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

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

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

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

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

Least-squares

- Loss function

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

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

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

Least-squares and Maximum Likelihood

- Review for maximum-likelihood estimation (MLE)

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

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

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

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

Least-squares and Maximum Likelihood

- Relation between LS and MLE

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

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

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

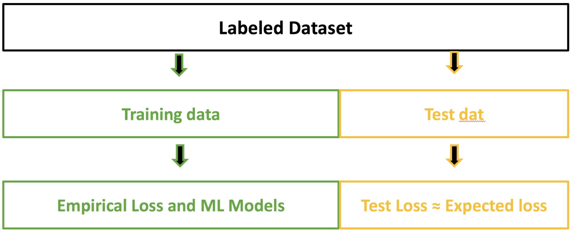
$$p(\mathbf{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \boldsymbol{w}^T \boldsymbol{\phi}\left(\boldsymbol{x}_n\right), \beta^{-1}\right)$$

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

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

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

Model Evaluation



- 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.

Ridge regression

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

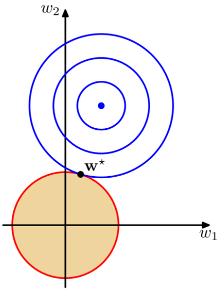
$$= \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \boldsymbol{w}^T\phi\left(\boldsymbol{x}_n\right)\right\}^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}$$

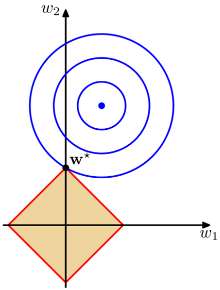
$$\boldsymbol{w}^*_{\text{ridge}} = \left(\lambda\boldsymbol{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\mathbf{t}$$

LASSO

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

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





RR and Maximum a Posteriori Estimation (MAP)

- Review for maximum a posteriori estimation (MAP)

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

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

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

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

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

RR and Maximum a Posteriori Estimation (MAP)

- Relation between RR and MAP

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

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

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

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

RR and Maximum a Posteriori Estimation (MAP)

- Relation between RR and MAP

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

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

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

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

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

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

Bayesian Models

- Bayesian Estimation

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

- Bayesian Estimation

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

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

Bayesian Linear Regression

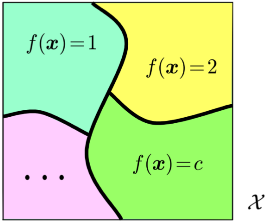
$$p(t \mid \mathbf{t}, \alpha, \beta) = \int p(t \mid \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \mathbf{t}, \alpha, \beta) \mathrm{d}\boldsymbol{w} \qquad \text{(Predictive distribution)}$$

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

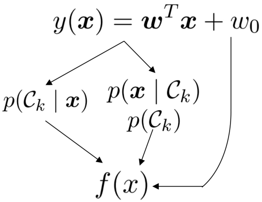
$$\sigma_N^2(\boldsymbol{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\phi}(\boldsymbol{x})$$

Outline

What is Linear classification



- Probabilistic Discriminative Models
- Probabilistic Generative Models
- Discriminant Functions



Least squares for classification?

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

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

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

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

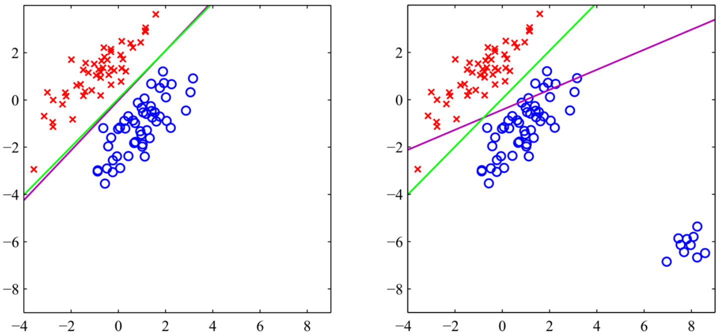
$$\boldsymbol{T}\ -\ n^{\text{th}}\ \text{row}\ -\ \mathbf{t}_n^{\mathrm{T}}$$

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

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

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

Least squares for classification?



least squares is highly sensitive to outliers

Probabilistic Discriminative Models

- Logistic regression

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

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

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

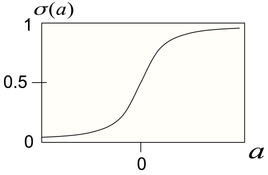
$$\boldsymbol{w}^T\boldsymbol{\phi} \xrightarrow{p(\mathcal{C}_k \mid \boldsymbol{x}) \sim \mathrm{Ber}(\sigma(\boldsymbol{w}^T\boldsymbol{\phi}))} p(\mathcal{C}_k \mid \boldsymbol{x})$$

Why sigmoid function?

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

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

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



Probabilistic Discriminative Models

- Logistic regression

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

$$y_n = p(\mathcal{C}_1 \mid \phi_n)$$

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

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

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

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

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

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

Probabilistic Generative Models

- Linear discriminant

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

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

Linear?

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

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

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

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

Probabilistic Generative Models

- Linear discriminant

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

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

Linear?

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

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

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

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

Probabilistic Generative Models

- Maximum likelihood solution for Linear discriminant

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

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

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

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

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

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

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

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

Probabilistic Generative Models

- Naïve-Bayes (NB) classifier

- Naïve-Bayes (NB) assumption

  $$x_i \perp x_{\{j \neq i\}} \mid t$$

- Bernoulli NB classifier

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

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

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

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

Hinge Loss and Support Vector Machines

- Loss Functions for Classification

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

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

· 0-1 loss

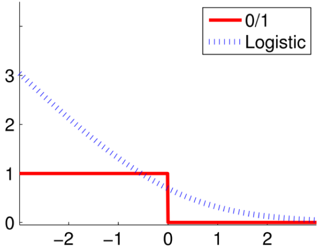
$E_{o/1}\left(t_n, y_n\right) = 1 - \mathrm{sign}\left\{t_n y\left(\boldsymbol{x}_n\right)\right\}$

· 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

- Loss Functions for Classification

  · Hinge Loss

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

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

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

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

- Support Vector Classifier

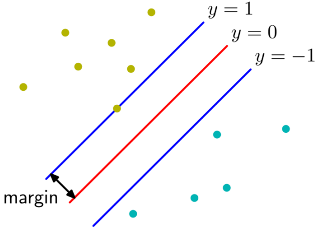
  $L_{SVC} = \sum_{n=1}^{N} E_{\text{Hinge}}\left(t_n, y_n\right) + \lambda \|\boldsymbol{w}\|^2$

Hinge Loss and Support Vector Machines

- Maximum-Margin View for SVC

$$\arg\max_{\boldsymbol{w},b} \left\{ \frac{1}{\|\boldsymbol{w}\|} \min_n \left[t_n \left(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n) + b \right) \right] \right\}$$

$$\text{s.t. } t_n \left(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n) + b \right) \geq 0, \; n = 1, \ldots, N$$



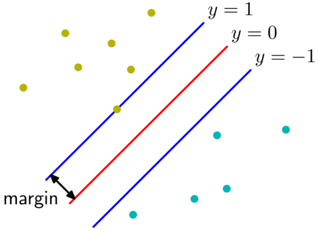
$$\text{s.t. } \min_n \left[t_n \left(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n) + b \right) \right] = 1$$

Hinge Loss and Support Vector Machines

- Maximum-Margin View for SVC

$$\arg\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$\text{s.t. } t_n\left(\boldsymbol{w}^T\phi\left(\boldsymbol{x}_n\right)+b\right) \geq 1, n=1,\ldots,N$$



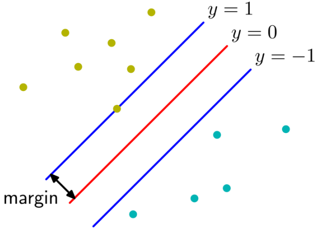
Hinge Loss and Support Vector Machines

- Maximum-Margin View for SVC

$$\underset{\boldsymbol{w},b,\boldsymbol{\xi}}{\arg\min}\; C\sum_{n=1}^{N}\xi_n + \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$\text{s.t. } t_n\left(\boldsymbol{w}^T\boldsymbol{\phi}\left(\boldsymbol{x}_n\right)+b\right) \geq 1-\xi_n, n=1,\ldots,N$$

$$\xi_n \geq 0$$



Model Evaluation for Classification

- Performance Matrices
  - · Confusion matrix

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

  - · Accuracy

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

  - · Error rate

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

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

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

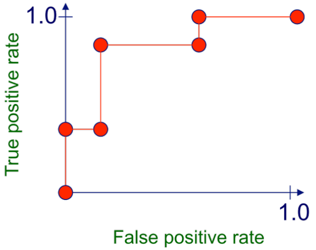
  - · F-measure

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

Model Evaluation for Classification

- Performance Matrices
  - ROC

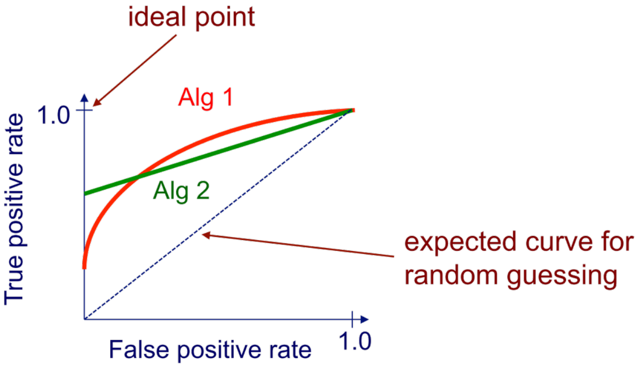
instance	confidence positive		correct class
Ex 9	.99		+
Ex 7	.98	TPR= 2/5, FPR= 0/5	+
Ex 1	.72	TPR= 2/5, FPR= 1/5	-
Ex 2	.70		+
Ex 6	.65	TPR= 4/5, FPR= 1/5	+
Ex 10	.51		-
Ex 3	.39	TPR= 4/5, FPR= 3/5	-
Ex 5	.24	TPR= 5/5, FPR= 3/5	+
Ex 4	.11		-
Ex 8	.01	TPR= 5/5, FPR= 5/5	-



Model Evaluation for Classification

- Performance Matrices

  · ROC



Model Evaluation for Classification

- Performance Matrices

  · AUC