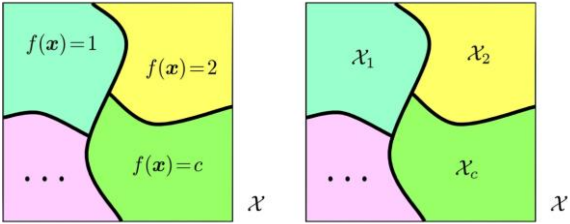
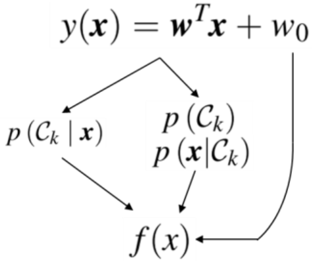
Outline

What is Linear Classification



- Probabilistic Discriminative Models
- Probabilistic Generative Models
- Discriminant Functions



Least Squares for Classification?

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

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

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

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

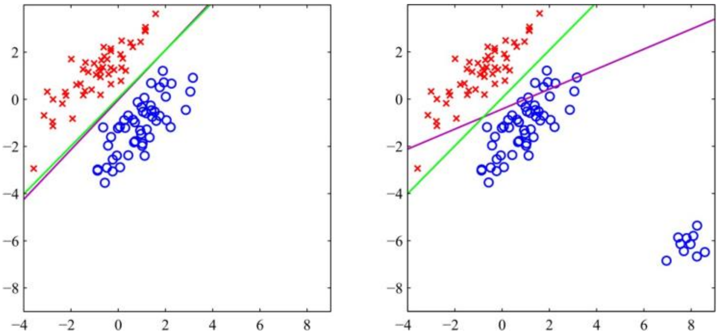
$$\boldsymbol{T} \ - \ n^{\text{th}} \text{ row } \ - \ \mathbf{t}_n^{\mathrm{T}}$$

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

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

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

Least Squares for Classification?



Least squares is highly sensitive to outliers

Probabilistic Discriminative Models

- Logistic regression

$$\begin{array}{c} \boldsymbol{w}^T\boldsymbol{\phi} \\ \big\downarrow \\ p\left(\mathcal{C}_k|\boldsymbol{\phi}\right) \end{array}$$

Probabilistic Discriminative Models

- Logistic regression

$$\boldsymbol{w}^T\boldsymbol{\phi} \xrightarrow{\mathrm{Ber}\left(\sigma\left(\boldsymbol{w}^T\boldsymbol{\phi}\right)\right)} p\left(\mathcal{C}_k|\boldsymbol{\phi}\right)$$

Probabilistic Discriminative Models

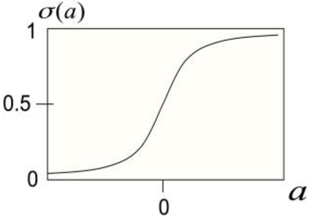
- Logistic regression

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

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

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

$$\begin{tikzcd} \boldsymbol{w}^T \boldsymbol{\phi} \arrow[d, "\mathrm{Ber}\left(\sigma\left(\boldsymbol{w}^T \boldsymbol{\phi}\right)\right)"'] \\ p\left(\mathcal{C}_k \mid \boldsymbol{\phi}\right) \end{tikzcd}$$



Probabilistic Discriminative Models

- Logistic regression

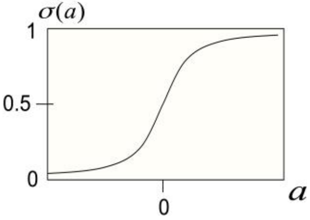
$$p(\mathcal{C}_1 \mid \phi) = \sigma\left(\boldsymbol{w}^T \phi\right)$$

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

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

Why sigmoid function?

$$\begin{array}{c} \boldsymbol{w}^T \phi \\ \Big\downarrow {\scriptstyle \mathrm{Ber}\left(\sigma\left(\boldsymbol{w}^T \phi\right)\right)} \\ p(\mathcal{C}_k | \phi) \end{array}$$



Probabilistic Discriminative Models

- Logistic regression

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

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

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

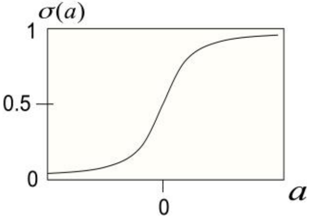
$$\boldsymbol{w}^T\phi \xrightarrow{\mathrm{Ber}\left(\sigma\left(\boldsymbol{w}^T\phi\right)\right)} p(\mathcal{C}_k \mid \phi)$$

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

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

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

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

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

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

Probabilistic Discriminative Models

- Softmax regression

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

Probabilistic Discriminative Models

- Softmax regression

$$p(\mathcal{C}_k|\boldsymbol{\phi}) = y_k(\boldsymbol{\phi}) = \frac{\exp(a_k)}{\sum_j \exp(a_k)} \qquad \text{Multi}\left\{\cdots \frac{\exp(\boldsymbol{w}_k^T\boldsymbol{\phi})}{\sum_j \exp(\boldsymbol{w}_k^T\boldsymbol{\phi})} \cdots\right\}$$

$$\boldsymbol{w}_k^T\boldsymbol{\phi} \downarrow p(\mathcal{C}_k \mid \boldsymbol{x})$$

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

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

- Cross-entropy error function

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

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

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

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

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

Linear?

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

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

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

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

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

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

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

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

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

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

Probabilistic Generative Models

- Maximum likelihood solution for Linear discriminant

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

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

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

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

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

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

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

Probabilistic Generative Models

- Maximum likelihood solution for Linear discriminant

The terms in the log likelihood function that depend on $\mathbf{\Sigma}$ is

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

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

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

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

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

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

Probabilistic Generative Models

- Naïve-Bayes (NB) classifier
- conditional independence

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

- Bernoulli NB classifier

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

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

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

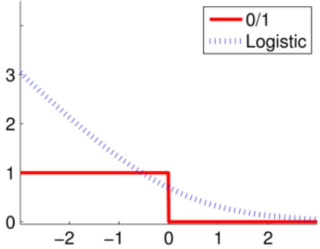
$$a_k = \ln p((\boldsymbol{x} \mid \mathcal{C}_k) p(\mathcal{C}_k))$$

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

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

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

- Log loss

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

equals to

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

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

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

- Log loss

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

equals to

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

Hinge Loss and Support Vector Machines

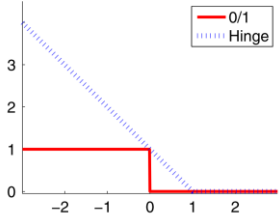
- Hinge Loss

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

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

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

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



- Support Vector Classifier

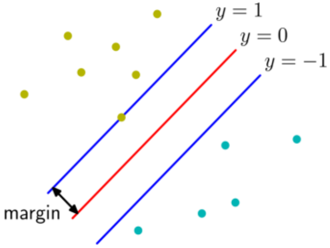
$L_{SVC} = \sum_{n=1}^{N} E_{\text{Hinge}}(t_n, y_n) + \lambda \|\boldsymbol{w}\|^2$

Hinge Loss and Support Vector Machines

- Maximum-Margin View for SVC

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

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

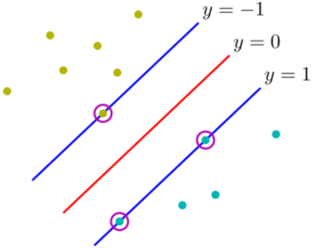


$$\text{s.t. } \min_n\left[t_n\left(\boldsymbol{w}^T\boldsymbol{\phi}\left(\boldsymbol{x}_n\right)+b\right)\right]=1$$

Hinge Loss and Support Vector Machines

- Maximum-Margin View for SVC

$$\arg\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t. } t_n\left(\boldsymbol{w}^T\phi\left(\boldsymbol{x}_n\right)+b\right) \geq 1, n=1,\ldots,N$$



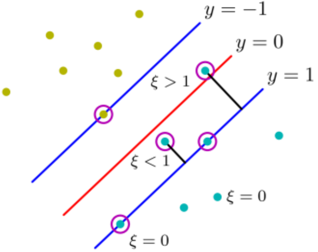
Hinge Loss and Support Vector Machines

- Maximum-Margin View for SVC

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

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

$$\xi_n \geq 0$$

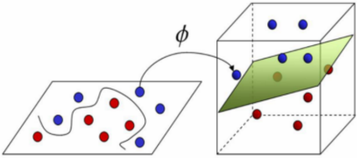


Kernel Trick and Nonlinear Support Vector Machines

- Kernel

$$\boldsymbol{x} \longrightarrow \phi(\boldsymbol{x})$$

$$k\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \phi(\boldsymbol{x})^T \phi\left(\boldsymbol{x}'\right)$$



$$\begin{aligned} k(\boldsymbol{x}, \boldsymbol{z}) &= \left(\boldsymbol{x}^T \boldsymbol{z}\right)^2 = (x_1 z_1 + x_2 z_2)^2 \\ &= x_1^2 z_1^2 + 2 x_1 z_1 x_2 z_2 + x_2^2 z_2^2 \\ &= \left(x_1^2, \sqrt{2} x_1 x_2, x_2^2\right)\left(z_1^2, \sqrt{2} z_1 z_2, z_2^2\right)^T \\ &= \phi(\boldsymbol{x})^T \phi(\boldsymbol{z}) \end{aligned}$$

$$\phi(\boldsymbol{x}) = \left(x_1^2, \sqrt{2} x_1 x_2, x_2^2\right)^T$$

Kernel Trick and Nonlinear Support Vector Machines

- Kernel SVC

dual representation

$$\tilde{L}(\boldsymbol{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m \boldsymbol{\phi}(\boldsymbol{x}_n)\,\boldsymbol{\phi}(\boldsymbol{x}_m)$$

$$a_n \geq 0, \quad n = 1, \ldots, N$$

$$\sum_{n=1}^{N} a_n t_n = 0$$

Kernel Trick and Nonlinear Support Vector Machines

- Kernel SVC

Moreover, we have

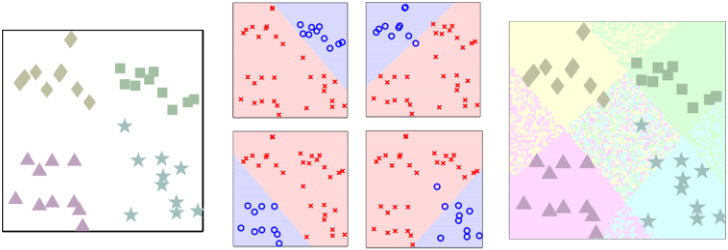
$$y(\boldsymbol{x}) = \sum_{n=1}^{N} a_n t_n \phi(\boldsymbol{x}) \phi(\boldsymbol{x}_n) + b$$

$$b = \frac{1}{N_{\mathcal{S}}} \sum_{n \in \mathcal{S}} \left(t_n - \sum_{m \in \mathcal{S}} a_m t_m \phi(\boldsymbol{x}_n) \phi(\boldsymbol{x}_m) \right)$$

We can replace $\phi(\boldsymbol{x}_n) \phi(\boldsymbol{x}_m)$ by $k(\boldsymbol{x}_n, \boldsymbol{x}_m)$ and $\phi(\boldsymbol{x})\phi(\boldsymbol{x}_n)$ by $k(\boldsymbol{x}, \boldsymbol{x}_n)$ for kernel SVC

Multiclass Classification

- One-versus-all

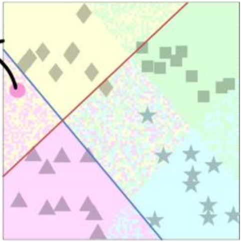


For K classes, we have K classifiers

Multiclass Classification

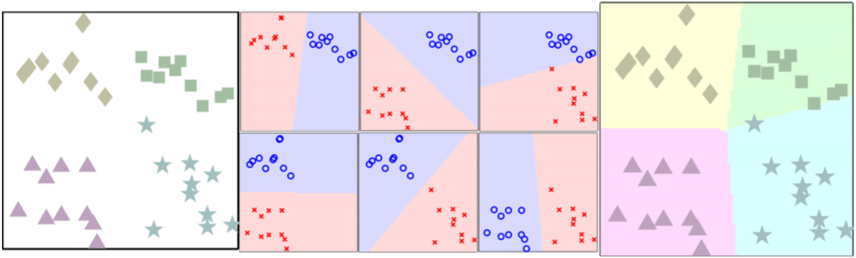
- One-versus-all

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

    $$TP/(TP+FP)$$

  - Recall

    $$TP/(TP+FN)$$

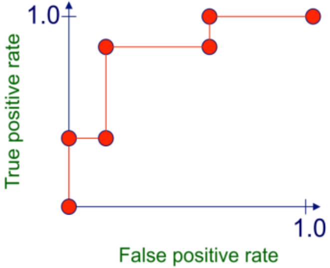
  - F-measure

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

Model Evaluation for Classification

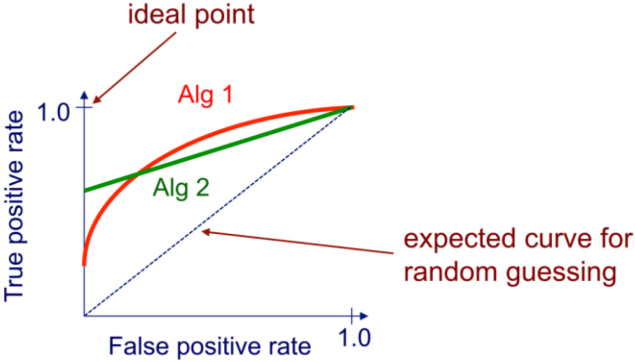
- Performance Matrices
  - ROC curve

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	-



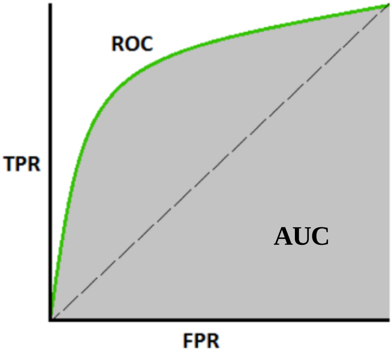
Model Evaluation for Classification

- 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