

Linear Regression and Classification

Wenting Tu

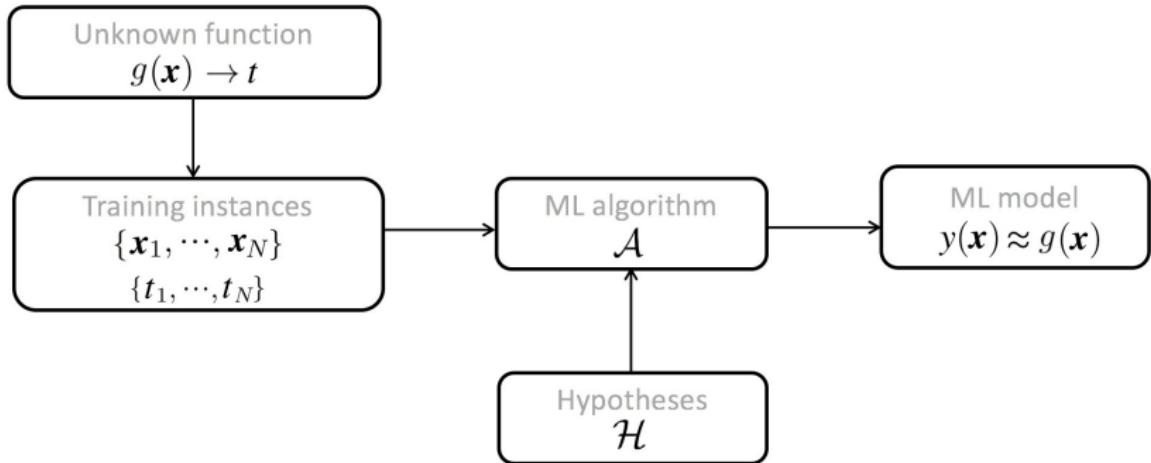
SHUFE, SIME

Outline

Linear Regression

Linear Classification

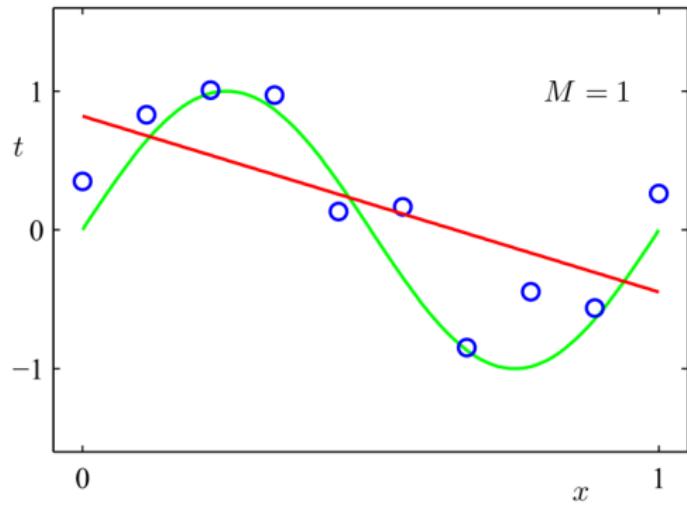
Definition



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

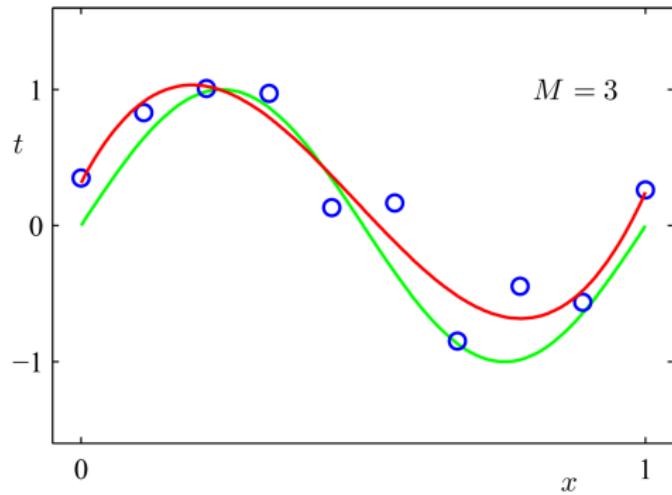
$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x^3 + \dots + w_D x_D$$

Illustration



$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + w_3x^3 + \cdots + w_Mx^M$$

Illustration



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

Linear Basis Function Models

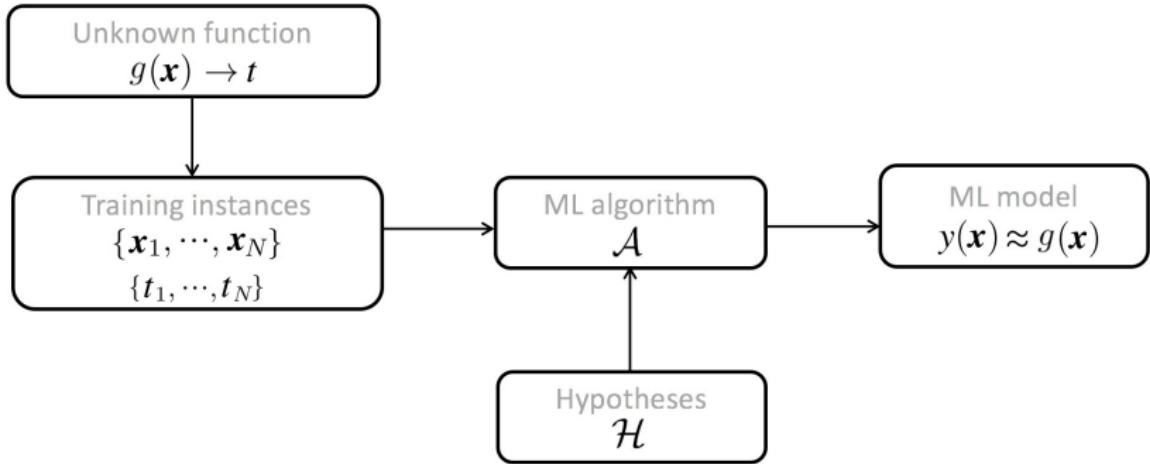
$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

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

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

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

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



Least-squares

- Loss function

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

- Empirical risk

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

Least-squares

- Solution

$$\mathbf{w}^* = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{t}$$

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

Least-squares and Maximum Likelihood

- Review for maximum-likelihood estimation (MLE)

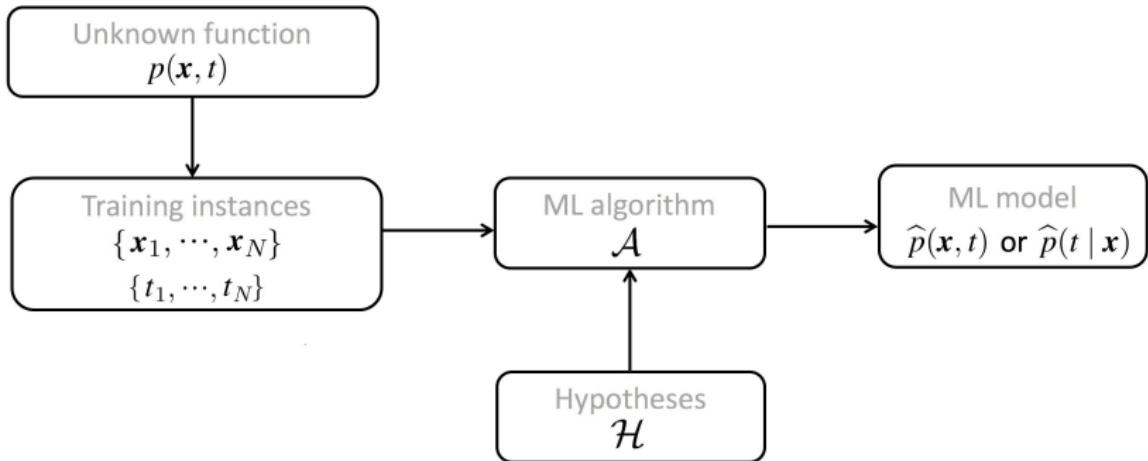
$$q(\mathbf{x}; \boldsymbol{\theta}) \longrightarrow p(\mathbf{x})$$

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

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} p(\mathcal{D} \mid \boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^n q(\mathbf{x}_i; \boldsymbol{\theta})$$

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

$$\hat{p}(\mathbf{x}) = q(\mathbf{x}; \hat{\boldsymbol{\theta}}_{\text{ML}})$$



Least-squares and Maximum Likelihood

- Relation between LS and MLE

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

$$p(t | \mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t | y(\mathbf{x}, \mathbf{w}), \beta^{-1})$$

$$\mathbb{E}[t | \mathbf{x}] = \int t p(t | \mathbf{x}) dt = y(\mathbf{x}, \mathbf{w})$$

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

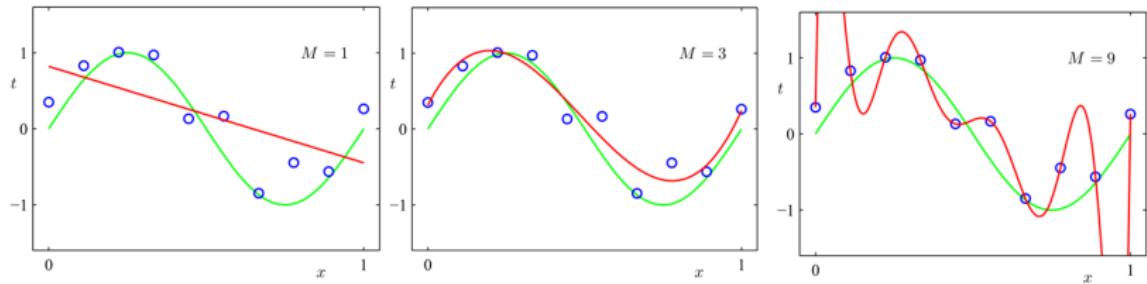
$$\ln p(\mathbf{t} | \mathbf{w}, \beta) = \sum_{n=1}^N \ln \mathcal{N}(t_n | \mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1})$$

$$= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w})$$

$$\mathbf{w}_{MLE}^* = \arg \max_w \ln p(\mathbf{t} | \mathbf{w}, \beta)$$

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

Illustration



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

Bias-Variance Decomposition

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

Bias-Variance Decomposition

Optimal decision $h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int tp(t|\mathbf{x})dt$

Expected squared loss

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

For any given data set \mathcal{D} :

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

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

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

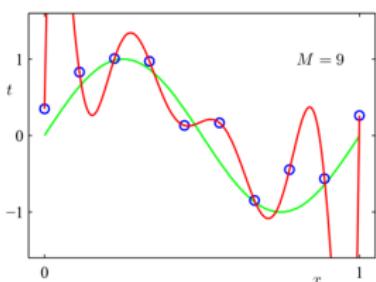
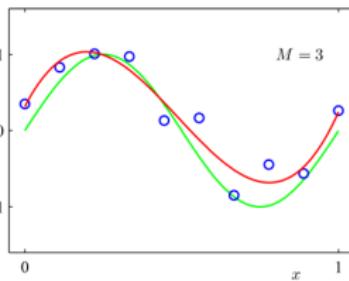
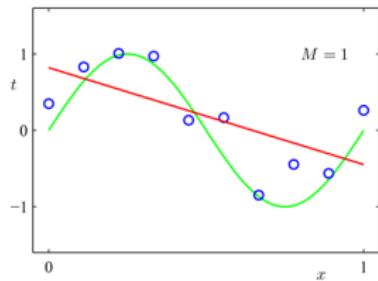
Bias-Variance Decomposition

expected loss = (bias)² + variance + noise

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

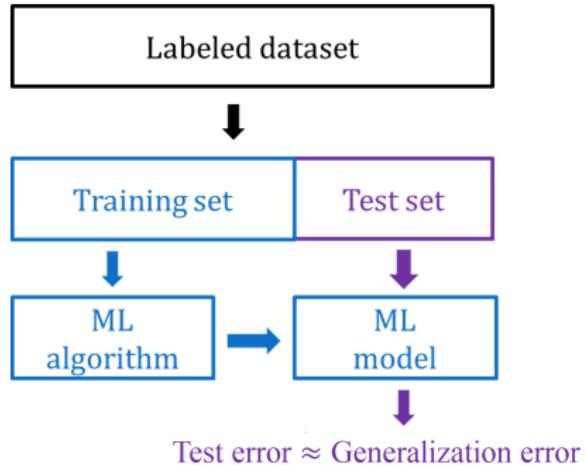
$$\text{variance} = \int \mathbb{E}_{\mathcal{D}} \left[\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2 \right] p(\mathbf{x}) d\mathbf{x}$$

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



Model Evaluation

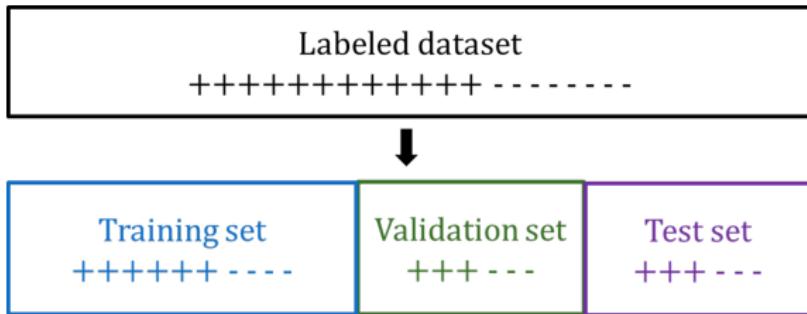
- Generalization Ability



- 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.

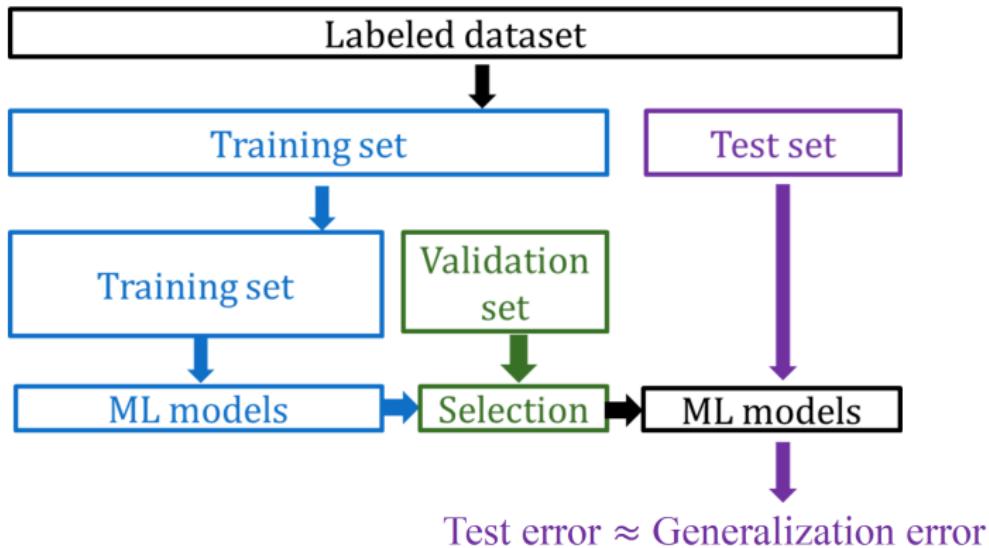
Model Evaluation

- Validation Set



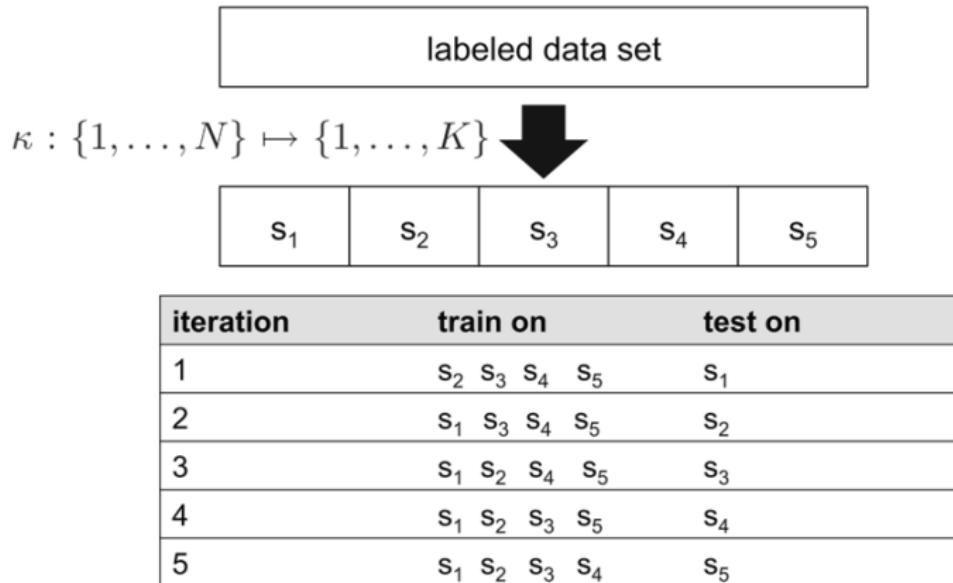
Model Evaluation

- Tuning Hyperparameters



Model Evaluation

- Cross Validation



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

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

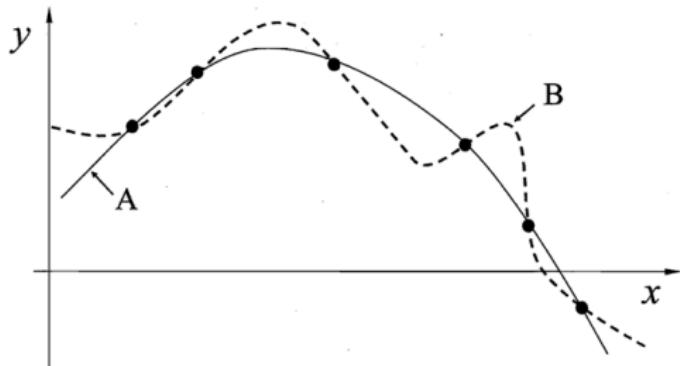
$$MSE = \frac{1}{n} \sum_{i=1}^n (t_i - f(\mathbf{x}_i))^2$$

$$RMSE = \sqrt{MSE}$$

Model Selection

- Occam's razor

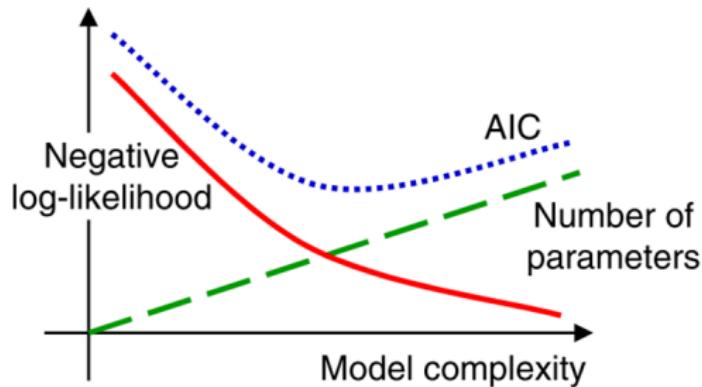
Suppose there exist two explanations for an occurrence. In this case the one that requires the smallest number of assumptions is usually correct.



Model Selection

- Akaike information criterion (AIC)

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



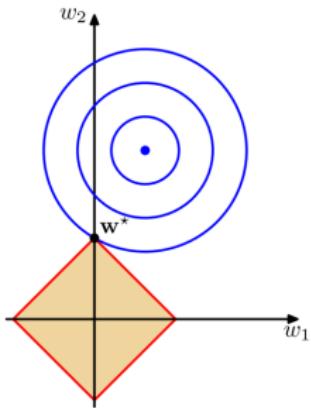
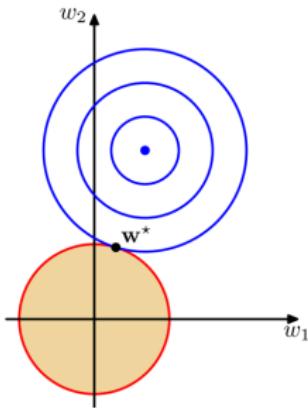
Regularization

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

LASSO

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

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



Ridge Regression

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

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

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

Ridge Regression and Maximum a Posteriori Estimation

- Review for maximum a posteriori (MAP) estimation

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

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

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

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

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

Ridge Regression and Maximum a Posteriori Estimation

- Relation between ridge regression and MAP

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

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

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

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

Ridge Regression and Maximum a Posteriori Estimation

- Relation between ridge regression and MAP

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

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

$$p(\mathbf{w} | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N)$$

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

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

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

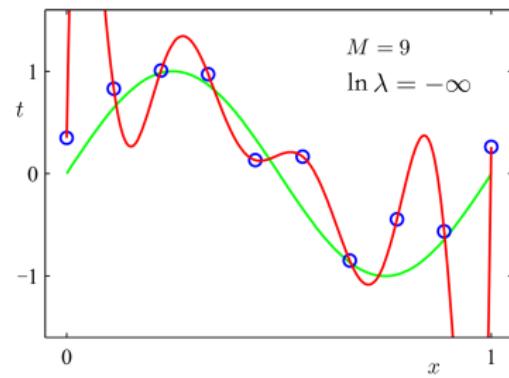
$$\mathbf{w}_{MAP}^* = \mathbf{w}_{ridge}^*$$

Ridge Regression

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

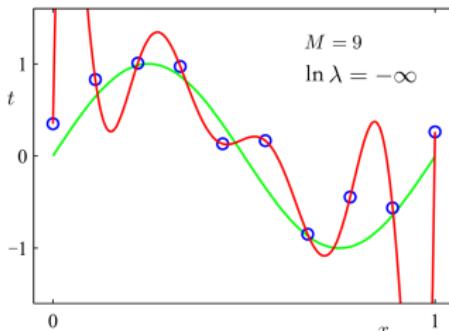
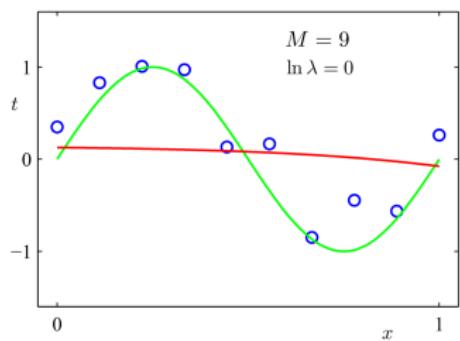
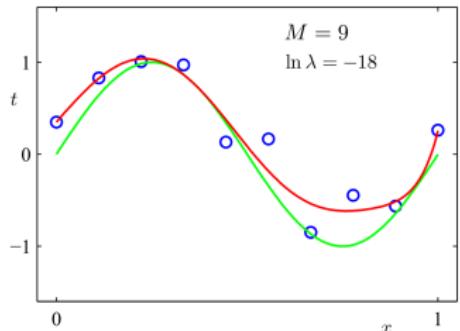
$M = 9$

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
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

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



Bayesian Models

- Bayesian Estimation

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

- Bayesian Estimation

$$\begin{aligned}\hat{p}_{\text{Bayes}}(\mathbf{x}) &= \int q(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} \\ &= \int q(\mathbf{x}|\boldsymbol{\theta}) \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} = \int q(\mathbf{x}|\boldsymbol{\theta}) \frac{\prod_{i=1}^n q(\mathbf{x}_i|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int \prod_{i=1}^n q(\mathbf{x}_i|\boldsymbol{\theta}')p(\boldsymbol{\theta}') d\boldsymbol{\theta}'} d\boldsymbol{\theta}'\end{aligned}$$

Bayesian Linear Regression

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

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

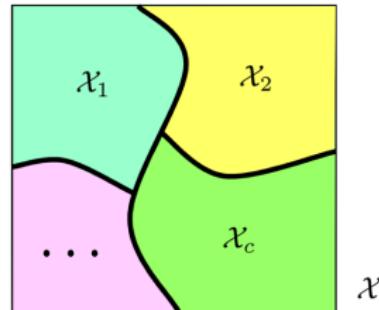
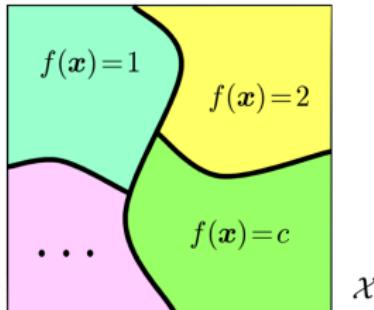
$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})$$

Outline

Linear Regression

Linear Classification

What is Linear classification



- Probabilistic Discriminative Models
- Probabilistic Generative Models
- Discriminant Functions

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

```
graph TD; A[p(Ck | x)] --> B[p(Ck)]; A --> C[p(x | Ck)]; B --> D[f(x)]; C --> D;
```

A flowchart illustrating the calculation of the discriminant function $y(\mathbf{x})$. It starts with two conditional probabilities: $p(\mathcal{C}_k | \mathbf{x})$ and $p(\mathbf{x} | \mathcal{C}_k)$. These probabilities are combined to form the marginal probability $p(\mathcal{C}_k)$. This marginal probability is then used in the discriminant function $f(\mathbf{x})$.

Least squares for classification?

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

$$\mathbf{y}(\mathbf{x}) = \tilde{\mathbf{W}}^T \tilde{\mathbf{x}}, \quad \tilde{\mathbf{w}}_k = (w_{k0}, \mathbf{w}_k^T)^T, \quad \tilde{\mathbf{x}} = (1, \mathbf{x}^T)^T$$

$$\{\mathbf{x}_n, t_n\}, n = 1, \dots, N$$

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

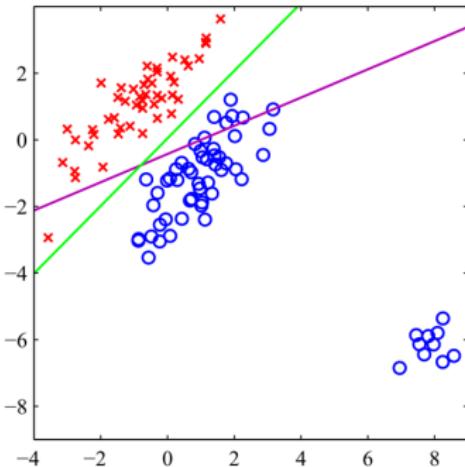
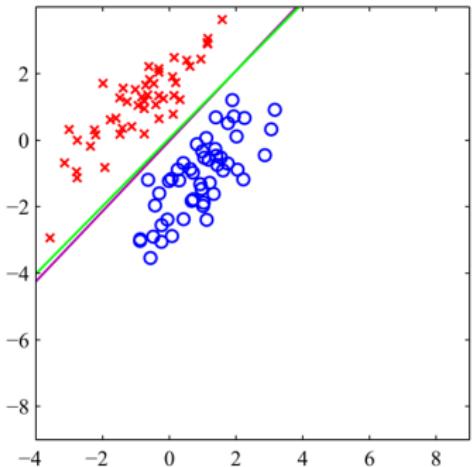
$$\mathbf{T} - n^{\text{th}} \text{ row } - \mathbf{t}_n^T$$

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

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

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

Least squares for classification?



Least squares is highly sensitive to outliers.

Probabilistic Discriminative Models

- Logistic regression

$$p(\mathcal{C}_1 | \phi) = \sigma(\mathbf{w}^T \phi)$$

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

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

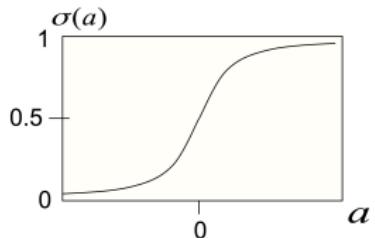
$$\begin{array}{c} \mathbf{w}^T \phi \\ \downarrow \\ \text{Ber}(\sigma(\mathbf{w}^T \phi)) \\ \downarrow \\ p(\mathcal{C}_k | \phi) \end{array}$$

Why sigmoid function?

$$p(\mathcal{C}_1 | \mathbf{x}) = \frac{p(\mathbf{x} | \mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x} | \mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x} | \mathcal{C}_2)p(\mathcal{C}_2)}$$

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

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



Probabilistic Discriminative Models

- Logistic regression

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

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

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

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

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

Probabilistic Discriminative Models

- Softmax regression

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

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

$$\mathbf{w}_k^T \phi$$

$$\text{Multi} \left\{ \cdots \frac{\exp(\mathbf{w}_k^T \phi)}{\sum_j \exp(\mathbf{w}_k^T \phi)} \cdots \right\}$$

$$p(\mathcal{C}_k | \mathbf{x})$$

$$p(\mathbf{T}|\mathbf{w}_1, \dots, \mathbf{w}_K) = \prod_{n=1}^N \prod_{k=1}^K p(\mathcal{C}_k|\phi_n)^{t_{nk}} = \prod_{n=1}^N \prod_{k=1}^K y_{nk}^{t_{nk}}$$

Cross-entropy error function

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

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots, \mathbf{w}_K) = \sum_{n=1}^N (y_{nj} - t_{nj}) \phi_n$$

Probabilistic Generative Models

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



$$p(\mathbf{x}|\mathcal{C}_k)$$

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



$$f(x)$$

Probabilistic Generative Models

- Linear discriminant

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

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

Probabilistic Generative Models

- Linear discriminant

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

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

Linear?

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

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

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

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

Probabilistic Generative Models

- Linear discriminant

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

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

Linear?

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

$$a_k = \ln p(\mathbf{x} | \mathcal{C}_k) p(\mathcal{C}_k)$$

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

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

Probabilistic Generative Models

- Maximum likelihood solution for Linear discriminant

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

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

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

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

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

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 μ_1 is

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

Thus, we obtain

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

Probabilistic Generative Models

- Maximum likelihood solution for Linear discriminant

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

$$\begin{aligned} & -\frac{1}{2} \sum_{n=1}^N t_n \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^N t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^T \Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) \\ & - \frac{1}{2} \sum_{n=1}^N (1 - t_n) \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^N (1 - t_n) (\mathbf{x}_n - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_2) \\ & = -\frac{N}{2} \ln |\Sigma| - \frac{N}{2} \text{Tr} \{ \Sigma^{-1} \mathbf{S} \} \\ \mathbf{S} & = \frac{N_1}{N} \mathbf{S}_1 + \frac{N_2}{N} \mathbf{S}_2 \\ \mathbf{S}_1 & = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \boldsymbol{\mu}_1) (\mathbf{x}_n - \boldsymbol{\mu}_1)^T, \mathbf{S}_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \boldsymbol{\mu}_2) (\mathbf{x}_n - \boldsymbol{\mu}_2)^T \\ \Sigma & = \mathbf{S} \end{aligned}$$

Probabilistic Generative Models

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

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

- Bernoulli NB classifier

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

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

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

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

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

Hinge Loss and Support Vector Machines

- Loss Functions for Classification

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

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

- 0-1 loss

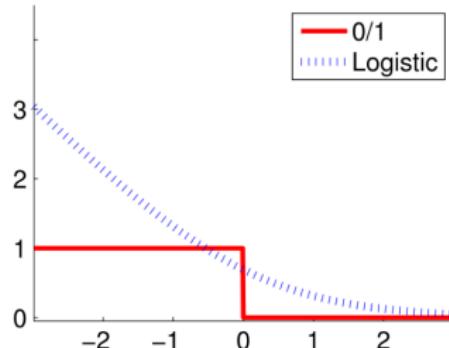
$$E_{0/1}(t_n, y_n) = 1 - \text{sign}\{t_n y(\mathbf{x}_n)\}$$

- Log loss

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

equals to

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



Hinge Loss and Support Vector Machines

- Hinge Loss

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

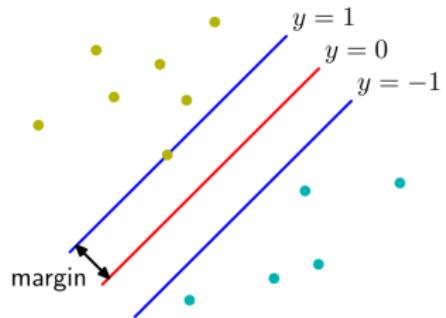
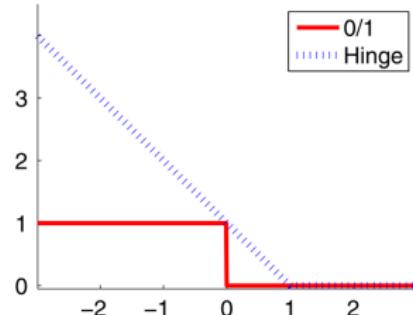
$$y_n > 0 \rightarrow \hat{t}_n = 1, y_n < 0 \rightarrow \hat{t}_n = -1$$

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

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

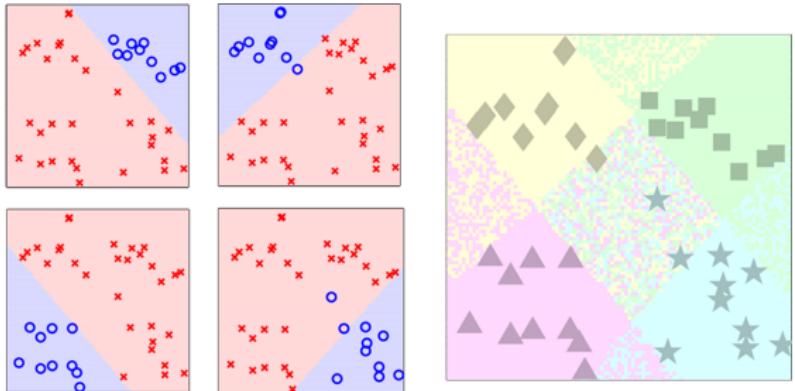
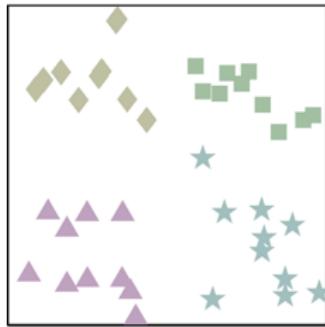
- Support Vector Classifier

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



Multiclass Classification

- One-versus-the-rest

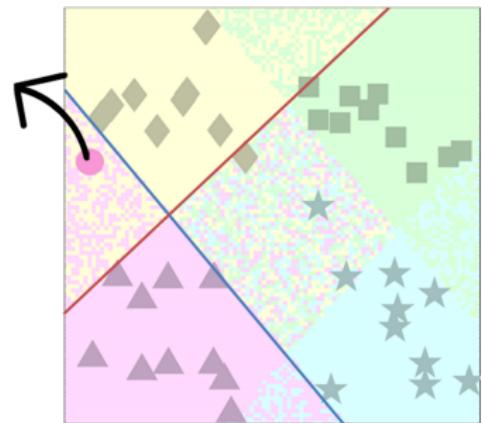


For K classes, we have K classifiers.

Multiclass Classification

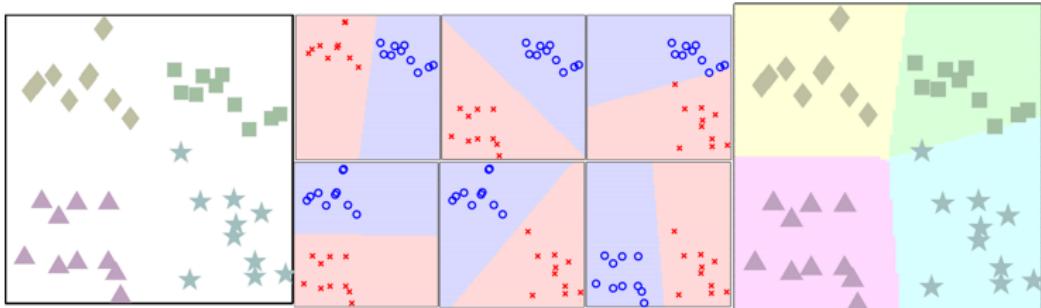
- One-versus-the-rest

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

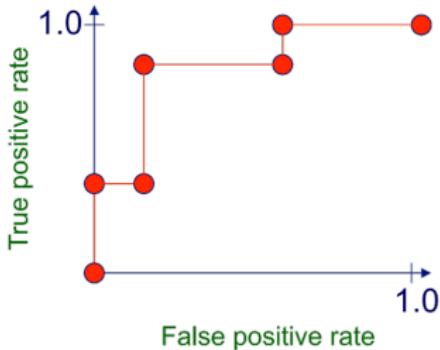
- Precision
 $TP / (TP + FP)$
- Recall
 $TP / (TP + FN)$
- F-measure

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

Model Evaluation for Classification

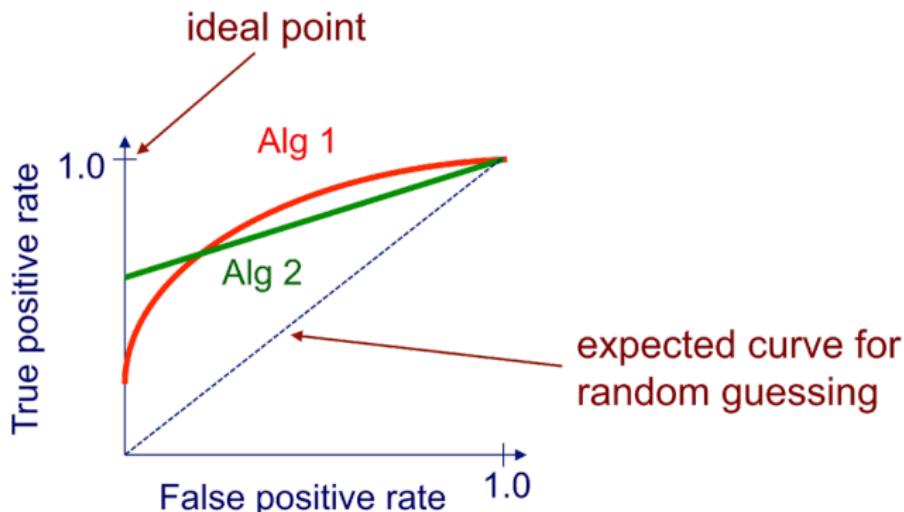
- Performance Matrices
 - ROC curve

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



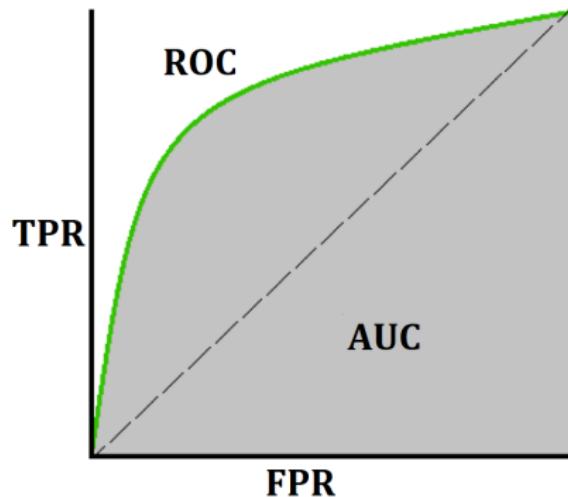
Model Evaluation for Classification

- 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](mailto:tudongtian@163.com)