机器学习
1. 生成式 vs. 判别式
2. 过拟合问题

夏睿
计算机学院
南京理工大学
rxia@njust.edu.cn

概要

- 判别式 vs. 生成式
  - 模型假设
  - 决策
  - 学习
- 过度拟合
  - ML - MAP
  - 正则

假设 – 学习- 决策

- 判别式模型
  - 直接建模预测函数

$$y = f(x)$$

例子:
感知机, SVMs

  - 对条件分布建模

$$p(y|x)$$

例子:
逻辑回归

- 产生式模型(对联合分布建模)

$$p(x, y) = p(y)p(x|y)$$

例子:
Naïve Bayes, GMM

假设 – 学习 – 决策

- 判别式模型
  - 建模预测函数

$$\theta^* = \arg\max_\theta J(\theta)$$

优化一些损失函数，如最小均方（LMS），交叉熵（CE），最大余量等

  - 建模后分布

$$\theta^* = \arg\max_\theta \sum_i \log p(y^{(i)}|x^{(i)})$$

ML, MAP (用于后验分布)
⇔在某些情况下等同于某些

- 产生式模型(建模联合/边缘分布)

$$\theta^* = \arg\max_\theta \sum_i \log p(x^{(i)}, y^{(i)})$$

ML, MAP, 贝叶斯推理(用于联合或边缘分布)

假设 – 学习 – 决策

- 判别式模型
  - 条件分布 $\arg\max_y p(y|x)$
  - 预测函数 $y = f(x)$
- 产生式模型
  - 贝叶斯公式 $p(y|x) = \frac{p(x,y)}{p(x)}$

$$\arg\max_y p(y|x) = \arg\max_y p(x,y) = \arg\max_y p(x|y)p(y)$$

分类问题的生成式模型

- 建模联合分布

$$p(x, y = c_j) = p(c_j)p(x|c_j)$$

类-条件概率

类先验概率

- 不同类别条件分布
  - 多项式分布

$$p(x, c_j|\theta) = p(c_j|\theta)p(x|c_j;\theta) = p_j \prod_{t=1}^{M} \theta_{t,j}^{N(\omega_t,x)}$$

  - 高斯分布

$$p(x, c_j|\theta) = p(c_j|\theta)p(x|c_j;\theta) = p_j \mathcal{N}(x|\mu_j, \Sigma_j)$$

生成式例子：朴素贝叶斯

- 模型假设

朴素贝叶斯假设

$$p(x|y=c_j)=\prod_{i=1}^{V}p(t_i|c_j)^{N(t_i,x)}=\prod_{i=1}^{V}\theta_{i|j}^{N(t_i,x)}$$

联合分布

$$p(x,y=c_j)=p(c_j)p(x|c_j)=\pi_j\prod_{i=1}^{V}\theta_{i|j}^{N(t_i,x)}$$

模型参数

生成式例子：朴素贝叶斯

- 学习

最大似然
(联合分布)

$$\max_{\pi,\theta} L(\pi,\theta) = \log \prod_{k=1}^{N} p(x_k, y_k)$$

$$s.t. \begin{cases} \sum_{j=1}^{C} \pi_j = 1 \\ \sum_{i=1}^{V} \theta_{i|j} = 1, j = 1, \dots, C \end{cases}$$

拉格朗日乘数

$$J = L(\pi,\theta) + \alpha(1 - \sum_{j=1}^{C} \pi_j) + \sum_{j=1}^{C} \beta_j (1 - \sum_{i=1}^{M} \theta_{i|j})$$

生成式例子：朴素贝叶斯

- 决策

朴素贝叶斯模型 $p(x, y = c_j) = \pi_j \prod_{i=1}^{V} \theta_{i|j}^{N(t_i,x)}$

贝叶斯公式 $p(y = c_j|x) = \frac{p(x, y = c_j)}{p(x)} = \frac{p(x, y = c_j)}{\Sigma_j p(x, y = c_j)}$

决策规则 $C^* = \arg\max_j p(y = c_j|x) = \arg\max_j p(x, y = c_j)$

判别式例子：Logisitc回归

- 模型假设

$$P(y=1|x;\theta)=h_\theta(x)=\frac{1}{1+exp-\theta^T x}$$

$$P(y=0|x;\theta)=1-h_\theta(x)$$

直接建模后验概率

- 决策

$$c_j^*=\arg\max_j p(y'=c_j|x')$$

判别式例子：逻辑回归

- 学习

$$J_c(\theta) = \sum_{i=1}^{n} \log p(y^{(i)}|x^{(i)})$$
$$= \sum_{i=1}^{n} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)$$

最大似然（条件分布） ⟺ 等价于最小交叉熵准则

朴素贝叶斯- 逻辑回归

- 多类别逻辑回归

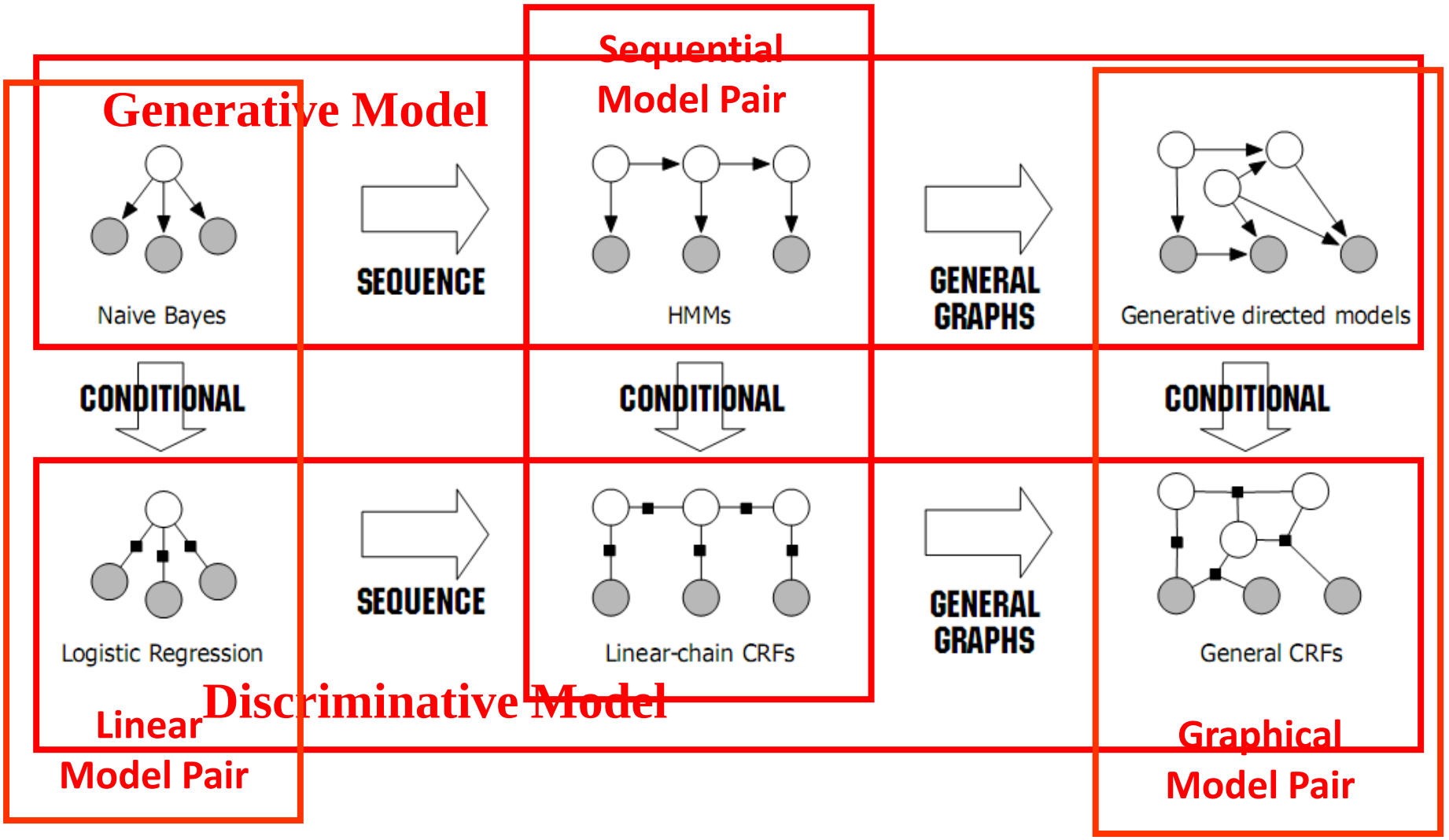
$$p(c_k|x) = y_k(x) = \frac{\exp(w_k^T x)}{\sum_{j=1}^{C} \exp(w_j^T x)}$$

- 朴素贝叶斯

$$p(x|c_k) = \frac{p(c_k) \prod_{t=1}^{M} p(\omega_t|c_k)^{N(\omega_t,x)}}{\sum_{j=1}^{C} \mathrm{p}(c_j) \prod_{t=1}^{M} p(\omega_t|c_j)^{N(\omega_t,x)}}$$

朴素贝叶斯和多类别逻辑回归是一种生成-判别模型对!

生成-判别模型对

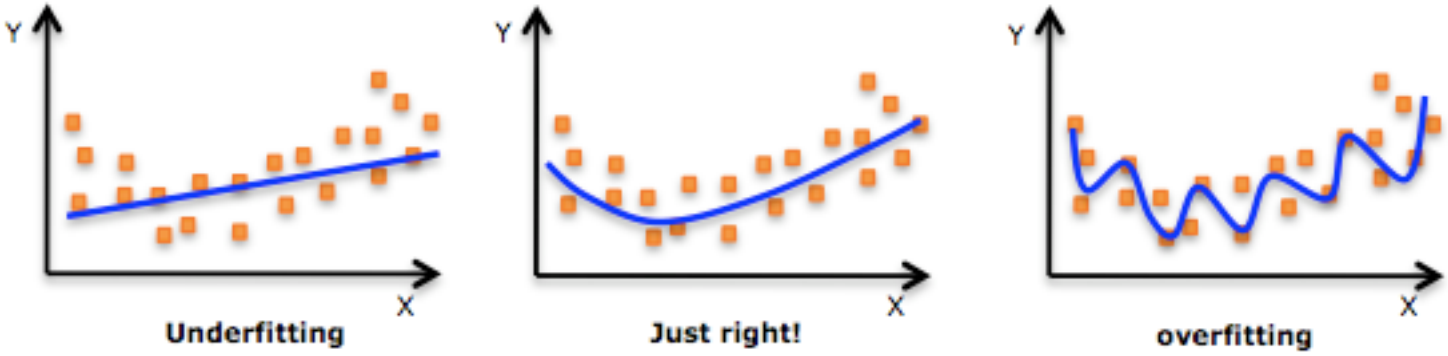

Generative Model
Sequential
Model Pair
Naive Bayes
SEQUENCE
HMMs
GENERAL
GRAPHS
Generative directed models
CONDITIONAL
CONDITIONAL
CONDITIONAL
Logistic Regression
SEQUENCE
Linear-chain CRFs
GENERAL
GRAPHS
General CRFs
Discriminative Model
Linear
Model Pair
Graphical
Model Pair

概要

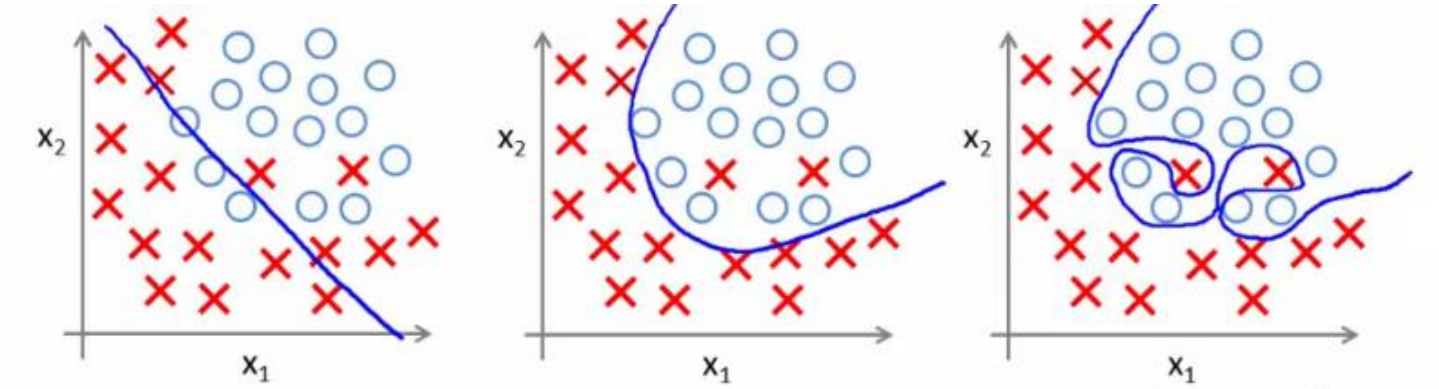
- 生成式 vs. 判别式
  - 模型假设
  - 决策
  - 学习
- 过拟合
  - ML - MAP
  - 正则

过拟合

回归


Y
X
Underfitting
Y
X
Just right!
Y
X
overfitting

分类


x_2
x_1
$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$
(g = sigmoid function)
UNDERFITTING
(high bias)
x_2
x_1
$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$
x_2
x_1
$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \ldots)$
OVERFITTING
(high variance)

ML - MAP

- 最大似然(ML)

$$\theta_{ML}^{*} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} p(X|\theta)$$
$$= \operatorname{argmax}_{\theta} \sum_{x \epsilon X} \log p(x|\theta)$$

likelihood

- 最大后验(MAP)

$$\theta_{MAp}^{*} = \arg\max_{\theta} p(\theta|X) = arg\max_{\theta} \frac{p(X|\theta)p(\theta)}{p(X)}$$
$$= \arg\max_{\theta} p(X|\theta)\, p(\theta)$$
$$= \arg\max_{\theta} \sum_{x \in X} \log p(x|\theta) + \log p(\theta)$$

likelihood • prior

正则项

- ML - MAP

$$\theta_{ML}^{*} = \arg\max_{\theta} \sum_{x \in X} \log p(x|\theta)$$

$$\theta_{MAp}^{*} = \arg\max_{\theta} \sum_{x \in X} \log p(x|\theta) + \log p(\theta)$$

正则项

- 损失函数加正则化

$$\theta^{*} = \arg\max_{\theta} J(\theta)$$

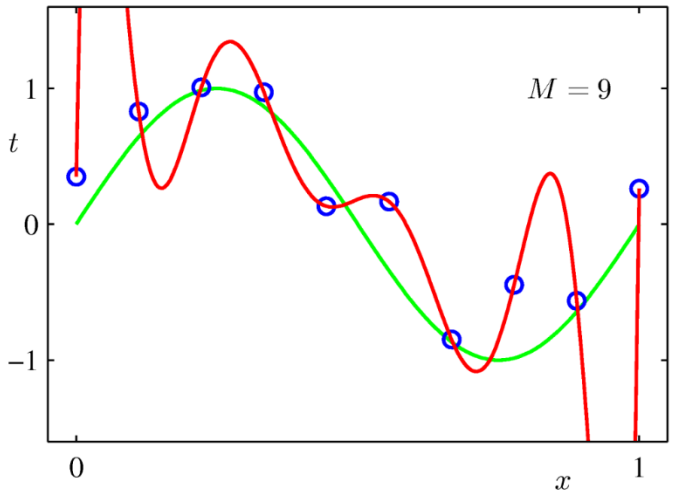
$$\theta^{*} = \arg\max_{\theta} \hat{J}(\theta) = \arg\max_{\theta} J(\theta) + \lambda R(\theta)$$

正则项

示例: 多项式曲线拟合

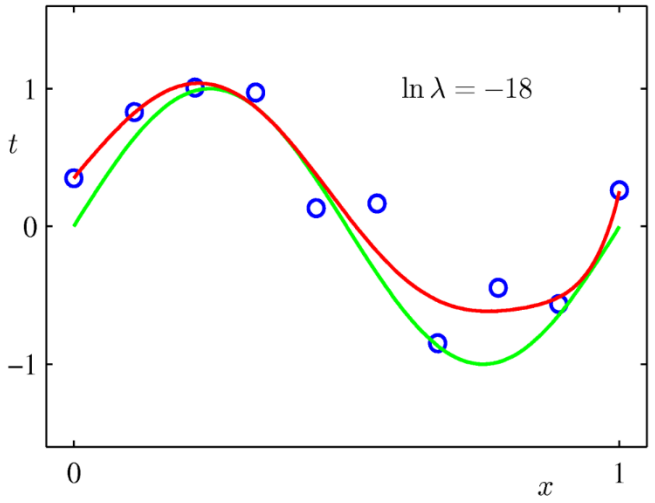
- ML ➔ MAP (PRML Equation 1.62, 1.67)

$$\ln p(t|x,w,\beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\boldsymbol{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$



ML

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\boldsymbol{w}) - t_n\}^2 + \frac{\alpha}{2}w^T w$$



MAP

示例：Logistic回归

- ML

$$J_c(\theta) = \sum_{i=1}^{n} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)$$

$$\text{where } h_\theta(x) = \frac{1}{1 + \exp -\theta^T x}$$

- MAP

$$\hat{J}_c(\theta) = \sum_{i=1}^{n} y^{(i)} \log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right) + \frac{1}{2}||\theta||^2$$

示例：伯努利实验

- Bernoulli 分布

$$Bern(x|\mu) = \mu^x(1-\mu)^{1-x}, x\epsilon\{0,1\}$$

- Log-likelihood

$$\log L(\mu|X) = \log \prod_{i=1}^{N} p(x_i|\mu) = \sum_{i=1}^{N} \log p(x_i|\mu)$$

$$= m_1 \log p(1|\mu) + m_0 \log p(0|\mu)$$

$$= m_1 \log \mu + m_0 \log(1-\mu)$$

- ML solution

$$\frac{\partial logL}{\partial \mu} = \frac{m_1}{\mu} - \frac{m_0}{1-\mu} = 0 \Leftrightarrow \hat{\mu}_{ML} = \frac{m_1}{N}$$

示例：伯努利实验

- 预分配

$$p(\mu|\alpha,\beta)=\frac{1}{B(\alpha,\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}\triangleq Beta(\mu|\alpha,\beta)$$

$$\text{where}\quad B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\qquad \Gamma(x)=\int_0^\infty u^{x-1}e^{-u}du$$

- MAP solution

$$\frac{\partial}{\partial\mu}\log L(\mu|X)+\log p(\mu)=\frac{m_1}{\mu}-\frac{m_0}{1-\mu}+\frac{\alpha-1}{\mu}-\frac{\beta-1}{1-\mu}=0$$

$$\Leftrightarrow \hat{\mu}_{MAP}=\frac{m_1+\alpha-1}{N+\alpha+\beta-2}$$

回顾

- 生成式 vs. 判别式
  - 模型假设
  - 决策
  - 学习
- 过拟合
  - ML - MAP
  - 正则

欢迎提问！