No.

Date.

chap 1 机器学习概述

模型
学习准则
优化算法
} 机器学习算法三要素 ⇒ 统计推断问题

{ 频率学派：θ固定常数
贝叶斯：θ随机变量，且存在先验分布

2.1 基本术语

No.

Date. . .

2.1 基本概念

特征向量 $\vec{x} = [x_1, x_2, \cdots x_d]^T$

训练集 $D = \{(\vec{x}^{(1)}, y^{(1)}), \cdots, (\vec{x}^{(m)}, y^{(m)})\}$

标签 y

决策函数预测标签的值 $y = f(\vec{x}, \theta)$

标签的条件概率 $p(y|\vec{x}) = f_y(\vec{x}, \theta)$

学习算法 A 找到一组参数 θ^*，使得 $f(\vec{x}, \theta^*)$ 近似真实关系

模型 $f(\vec{x}, \theta)$

$\mathbb{I}(\cdot)$ 指示函数

2.2 机器学习的三个基本要素

⇒ 2.2.1 模型

$\mathcal{X}$ 输入空间 $(\vec{x}, y) \in \mathcal{X} \times \mathcal{Y}$

$\mathcal{Y}$ 输出空间

$p_r(y|\vec{x})$ 真实条件概率分布

$\mathcal{F}$ 假设空间，通常为参数化的函数族

No.

Date. . .

⇒⇒⇒ 2.2.1.1 线性模型

$$f(\vec{x}, \theta) = \vec{w}^T \vec{x} + b$$

⇒⇒⇒ 2.2.1.2 非线性模型

可以写成多个非线性基函数 $\phi(\vec{x})$ 的线性组合.

$$f(\vec{x}, \theta) = \vec{w}^T \phi(\vec{x}) + b$$

$$\phi(\vec{x}) = [\phi_1(\vec{x}), \cdots, \phi_K(\vec{x})]^T$$

⇒⇒ 2.2.2 学习准则

$$|f(\vec{x}, \theta^*) - y| < \varepsilon$$

$$|f_y(\vec{x}, \theta^*) - p_r(y|\vec{x})| < \varepsilon$$

$R(\theta)$ 期望风险

$$R(\theta) = \mathbb{E}_{(\vec{x},y)\sim p_r(\vec{x},y)}[L(y, f(\vec{x}, \theta))]$$

$p_r(\vec{x}, y)$ 真实的数据分布

$L(y, f(\vec{x}, \theta))$ 损失函数

No.

Date. . .

习习习 2.2.2.1 损失函数

0-1 损失函数

$$L(y, f(\vec{x},\theta)) = \begin{cases} 0 & \text{if } y = f(\vec{x},\theta) \\ 1 & \text{if } y \neq f(\vec{x},\theta) \end{cases} = \mathbb{I}(y \neq f(\vec{x},\theta))$$

平方损失函数

$$L(y, f(\vec{x},\theta)) = \frac{1}{2}(y - f(\vec{x},\theta))^2$$

交叉熵损失函数

~~$L(\vec{y}, f(\vec{x},\theta)) =$~~

假设 $y \in \{1, \cdots, C\}$ 个类别，$f(\vec{x},\theta) \in [0,1]^C$

$p(y=c|\vec{x},\theta) = f_c(\vec{x},\theta)$ 称为样本的条件概率分布

交叉熵

$$L(\vec{y}, f(\vec{x},\theta)) = -\sum_{c=1}^{C} y_c \log f_c(\vec{x},\theta)$$

例如：$\vec{y} = [0,0,1]^T$，$f(\vec{x},\theta) = [0.3, 0.3, 0.4]^T$

$$L(\theta) = -(0\times\log 0.3 + 0\times\log 0.3 + 1\times\log 0.4)$$

$$= -\log 0.4$$

对 one-hot 向量

$$L(y, f(\vec{x},\theta)) = -\log f_y(\vec{x},\theta)$$

No.

Date.

多 $f_y(\vec{x},\theta)$ 可以看作 y 的似然函数

负对数似然损失函数

Hinge 损失函数

两分类问题 $\begin{cases} -1 \\ +1 \end{cases}$

$$L(y, f(\vec{x},\theta)) = \max(0, 1-yf(\vec{x},\theta))$$

$$= [1-yf(\vec{x},\theta)]_+$$

⇒⇒⇒ 2.2.2.2 风险最小化准则

$$R_D^{emp}(\theta) = \frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, f(x^{(n)},\theta))$$

$$\theta^* = \arg\min_{\theta} R_D^{emp}(\theta)$$

ERM 经验风险最小化

过拟合

No.

Date. . .

SRM 结构风险最小化

$$\theta^* = \arg\min_{\theta} R_D^{struct}(\theta)$$

$$= \arg\min_{\theta} R_D^{emp}(\theta) + \frac{1}{2}\lambda\|\theta\|^2$$

$$= \arg\min_{\theta} \frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, f(x^{(n)};\theta)) + \frac{1}{2}\lambda\|\theta\|^2$$

$\|\theta\|$ 是 l_2 范数的正则化项

λ 正则化系数

l_1 正则化引入了稀疏性

凸优化

⇒ 2.2.3 优化算法

参数

超参数:

⇒⇒ 2.2.3.1 梯度下降法

$$\theta_{t+1} = \theta_t - \alpha\frac{\partial R_D(\theta)}{\partial\theta}$$

$$= \theta_t - \alpha\cdot\frac{1}{N}\sum_{n=1}^{N}\frac{\partial L(y^{(n)}, f(x^{(n)};\theta))}{\partial\theta}$$

α 学习率

No.

Date. . .

⇒⇒⇒ 2.2.3.2 提前停止

验证集 从训练集中划分出来

验证集上错误率不再下降

⇒⇒⇒ 2.2.3.3 随机梯度下降法

SGD

$$\theta \leftarrow \theta - \alpha \frac{\partial L(\theta; (x^{(n)}, y^{(n)}))}{\partial \theta}$$

小批量梯度下降 Mini-Batch Gradient Descent

$$\theta_{t+1} \leftarrow \theta_t - \alpha \cdot \frac{1}{K} \sum \frac{\partial L(y, f(\vec{x}, \theta))}{\partial \theta}$$

No.

Date. . .

2.3 机器学习的简单示例：线性回归.

LR 自变量数量为1时称为简单回归

自变量 ≥1 时… 多元回归.

模型 $f(\vec{x};\vec{w},b) = \vec{w}^T\vec{x} + b$

求导为 $f(\vec{x};\vec{\hat{w}}) = \vec{\hat{w}}^T\vec{\hat{x}}$

$\vec{\hat{w}} = \vec{w} \oplus b = \begin{bmatrix} \vec{w} \\ b \end{bmatrix}$ 增广权重向量

$\hat{x} = \vec{x} \oplus 1 = \begin{bmatrix} \vec{x} \\ 1 \end{bmatrix}$ 增广特征向量.

⇒ 2.3.1 参数学习

4种参数估计方法

- ERM
- SRM
- 最大似然估计
- 最大后验估计

No.

Date.

2.3.1.1 经验风险最小化

ERM: $R_{emp}(w) = \sum_{n=1}^{N} L(y^{(n)}, f(\vec{x}^{(n)}, \vec{w}))$

$= \frac{1}{2} \sum_{n=1}^{N} (y^{(n)} - \vec{w}^T \vec{x}^{(n)})^2$

这里是个转置.

$= \frac{1}{2} \| \vec{y} - X^T \vec{w} \|^2$

在这里 $\vec{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{bmatrix}$ $\vec{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_N \\ b \end{bmatrix}$

$X = [\vec{x}^{(1)}, \dots, \vec{x}^{(N)}] = \begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(N)} \\ \vdots & & \vdots \\ x_d^{(1)} & & x_d^{(N)} \\ 1 & & 1 \end{bmatrix}$

$\frac{\partial R(\vec{w})}{\partial \vec{w}} = \frac{1}{2} \frac{\partial \| \vec{y} - X^T \vec{w} \|^2}{\partial \vec{w}}$

$= \frac{1}{2} \cdot 2 (\vec{y} - X^T \vec{w}) \cdot \frac{\partial (\vec{y} - X^T \vec{w})}{\partial \vec{w}}$

存疑

$= (\vec{y} - X^T \vec{w}) \cdot (-X)$

$= -X (\vec{y} - X^T \vec{w})$

No.

Date. . .

令 $\frac{\partial R(\vec{w})}{\partial \vec{w}} = 0$,

$-X(\vec{y} - X^T\vec{w}) = 0$

~~$X^{-1}X(\vec{y} - X^T\vec{w}) = 0$~~

$X\vec{y} = X^T\vec{w}$

$X^TX\vec{y} = \vec{w}$ $\quad \vec{w} = (X^T)^{-1} \times \vec{y}$ ~~$= (X^T)^{-1}(X^{-1}) \times X\vec{y}$~~

LSE 最小二乘法估计

$\vec{w} \leftarrow \vec{w} - \alpha X(y - X^T\vec{w})$

⇒⇒ 2.3.1.2 SRM

岭回归 给 XX^T 对角线元素都加一个 λI, 仅得 $(XX^T + \lambda I)$ 秩为 0

~~本质~~ 本质上是

$R(\vec{w}) = \frac{1}{2}\|\vec{y} - X^T\vec{w}\|^2 + \frac{1}{2}\lambda\|\vec{w}\|^2$

⇒⇒ 2.3.1.3 最大似然估计

假设 y 是一个随机变量, 服从均值为 $f(\vec{x}, \vec{w}) = \vec{w}^T\vec{x}$ 为中心, 方差为 σ^2 的高斯分布

$p(y|\vec{x}, \vec{w}, \sigma) = N(y|\vec{w}^T\vec{x}, \sigma^2)$

$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - \vec{w}^T\vec{x})^2}{2\sigma^2}\right)$

No.

Date.

似然函数

$$p(\vec{y} \mid X, \vec{w}, \sigma) = \prod_{n=1}^{N} p(y^{(n)} \mid \vec{x}^{(n)}, \vec{w}, \sigma)$$

$$= \prod_{n=1}^{N} \mathcal{N}(y^{(n)} \mid \vec{w}^T \vec{x}^{(n)}, \sigma^2)$$

对数似然函数

$$\log p(\vec{y} \mid X, \vec{w}, \sigma) = \sum \log \mathcal{N}(y^{(n)} \mid \vec{w}^T \vec{x}^{(n)}, \sigma^2)$$

MLE 最大似然估计

$$\Rightarrow \frac{\partial \log p(\vec{y} \mid X, \vec{w}, \sigma)}{\partial \vec{w}} = 0 \quad \Rightarrow \vec{w}^{ML} = (XX^T)^{-1} X\vec{y}$$

等等 ∂

$=$

$\Rightarrow\Rightarrow$ 2.3.1.中最大后验估计

假设 $\vec{w}$ 为一个随机向量.

$p(\vec{w} \mid v)$ 为各向同性的高斯分布

$p(\vec{w} \mid v) = \mathcal{N}(\vec{w} \mid 0, v^2 I)$ ~ $p(H)$

v^2 为每一维上的方差

$\vec{w}$ 的后验概率分布为

$$p(\vec{w}|X,\vec{y},v,\sigma) = \frac{p(\vec{w},\vec{y}|X,v,\sigma)}{\sum_{\vec{w}} p(\vec{w},\vec{y}|X,v,\sigma)}$$

$$\propto p(\vec{y}|X,\vec{w},\sigma)\, p(\vec{w}|v)$$

$p(\vec{y}|X,\vec{w},\sigma)$：$\vec{w}$ 的似然函数　　$p(\vec{w}|v)$：$\vec{w}$ 的先验

MAP

$$\vec{w}^{MAP} = \arg\max_{\vec{w}} p(\vec{y}|X,\vec{w},\sigma)\, p(\vec{w}|v)$$

$$\log p(\vec{w}|X,\vec{y},v,\sigma) \propto \log p(\vec{y}|X,\vec{w},\sigma) + \log p(\vec{w}|v)$$

$$\propto$$

⇒

No.

Date.

2.4 偏差与方差分解

* $P_r(x,y)$ 真实分布 $f(x)$ 模型

$R(f) = E_{(x,y)\sim P_r(x,y)}[(y-f(x))^2]$ 期望错误

$f^*(x) = E_{y\sim P_r(y|x)}[y]$ 最优化模型

$\varepsilon = E_{(x,y)\sim P_r(x,y)}[(y-f^*(x))^2]$ 模型损失

$f(x)$ 期望错误分解

$$R(f) = E_{(x,y)\sim P_r(x,y)}[(y-f^*(x)+f^*(x)-f(x))^2]$$

$$= E_{x\sim P_r(x)}[(f(x)-f^*(x))^2] + \varepsilon \quad (2.63)$$

当前模型与真实模型之差距

D上的期望差距 $f_D(x)$ 与 $f_D^*(x)$

$$E_D[(f_D(x)-f^*(x))^2]$$

$$= E_D[(f_D(x)-E_D[f_D(x)]+E_D[f_D(x)]-f^*(x))^2] \quad (2.64)$$

$$= \underbrace{(E_D[f_D(x)]-f^*(x))^2}_{\text{偏差}^2} + \underbrace{E_D[(f_D(x)-E_D[f_D(x)])^2]}_{\text{方差}} \quad (2.65)$$

No.

Date. . .

期望/错误可以分解为

$$R(f) = (bias)^2 + variance + \varepsilon$$

$$(bias)^2 = E_{\vec{x}}\left[(E_D[f_D(\vec{x})] - f^*(\vec{x}))^2\right]$$ 最优模型与真实的偏差. (2.67)

$$variance = E_{\vec{x}}\left[E_D\left[(f_D(\vec{x}) - E_D[f_D(\vec{x})])^2\right]\right]$$ 模型与模型的偏差. (2.68)

当样本较多时. 方差. 较小, 我们可以选择能力强的来减小偏差.

多条模型.

No.

Date.

2.1 机器学习算法的类型

监督学习 { Regression
Classification
结构化学习 输入是序列，树或图。

$\phi(x, y)$ 联合特征向量

结构化学习：

$$y = \arg\max_{y \in Gen(x)} f(\phi(x, y), \theta)$$

无监督学习 UL

聚类，密度估计，特征学习，降维。

强化学习 RL

	监督学习	UL	RL
训练样本	$\{(x^{(i)}, y^{(i)})\}$	$\{x^{(i)}\}$	轨迹 τ，奖励 G_t
优化目标	$y=f(x)$ 或 $p(y\|x)$	$p(x)$ 或 $p(x\|\theta)$	期望总回报 $E_\tau[G_t]$
算法	期望风险最小化 最大似然估计	最大似然估计 最小重构错误	策略评估 策略改进

No.

Date.

2.6 数据的特征表示

图像特征

文本特征　Bag-of-Words, Bow 模型

向量中的每一维 x_i 代表词表 V 中的第 i 个词是否在 x 中出现

每个连续词构成一个基本单元，二元特征 BoW

表示学习

=> 2.6.1 传统的特征学习

=> => 2.6.1.1 特征选择 FS

子集搜索

贪心策略：
- 前向搜索　添加该维最优特征
- 反向搜索　删除最无用特征

过滤式　每次增加最有信息量的特征

信息增益

包裹式　每次增加对后续机器学习模型最有用的特征

~~包裹式~~

归一化　解决特征

No.

Date.

⇒⇒ 2.6.1.2 特征抽取 FE

$$\vec{x}' = P\vec{x}$$

$\vec{x}$ 原始特征向量 $\vec{x} \in R^d$

$\vec{x}'$ 新特征向量 $\vec{x}' \in R^k$

P 映射矩阵 $P \in R^{k \times d}$

监督方法：抽取对一个特定的预测任务最有用的特征

LDA

非监督方法：去掉冗余信息和噪声.

PAC

	监督学习	无监督学习
FS	子集搜索 L正则化 决策树	子集搜索
FE	LDA	PAC. 奇异成分分析 流形学习. 自编码器

No.

Date. 2.7 评价指标

$T = (\vec{x}^{(1)}, y^{(1)}), (\vec{x}^{(2)}, y^{(2)}), \cdots, (\vec{x}^{(N)}, y^{(N)})$ 测试集

$Y = \hat{y}^{(1)}, \cdots, \hat{y}^{(N)}$ 预测结果.

准确率 $Acc = \frac{1}{N} \sum \mathbb{I}(y^{(N)} = \hat{y}^{(N)})$

错误率 $\varepsilon = 1 - Acc$

查准率和查全率

TP	FN	真的
FP	TN	假的
认为对的	认为错的	

查准率 Precision

$$P_c = \frac{TP_c}{TP_c + FP_c}$$ 真正对的 / 所有认为对的

望闻中真警察的比例

查全率 Recall

$$R_c = \frac{TP_c}{TP_c + FN_c}$$ 真对 / 真实

~~警队中抓中真警~~ 真警察出逃比例在局

No.

Date. . .

F1值. 查全率和查准率的调和平均.

$$F_1 = \frac{2 \times P_c \times R_c}{P_c + R_c}$$

宏平均和微平均.

macro

micro

宏平均 每一类性能指标的平均值.

$$P_{macro} = \frac{1}{C} \sum P_c$$

R_{macro} =

$F1_{macro}$ =

AUC

ROC 曲线

PR 曲线

Top N

交叉验证

No.

Date.

2.8 理论学习理论

⇒ 2.8.1 PAC 学习理论.

可能近似正确 { 可能 / 近似正确

同等条件下. 模型越复杂. 泛化误差越大.

模型选择. 模型越复杂越耗算力

⇒ 2.8.2 没有免费午餐定理 NFL

基于迭代的最优化算法. 不存在某种算法对所有问题都有效.

⇒ 2.8.3 正则化.

⇒ 2.8.4 奥卡姆剃刀

最小描述长度.

$$\max_f \log p(f|D) = \max_f \log p(D|f) + \log p(f)$$

$$= \min_f -\log p(D|f) - \log p(f)$$

⇒ 2.8.5 归纳偏置. Inductive Bias

机器学习的假设 贝叶斯中的先验 priors