

PRML

def 机器学习

$E \uparrow \Rightarrow P(T) \uparrow$ 更多的 data 导向系统性能提升

监督与无监督, 强化学习, 推荐系统

归纳学习

↓
泛化能力

§ 线性模型

$$w^* = \arg \min_w L(w)$$

$$w = w - \eta \nabla_w L(w)$$

GD, SGD, normal equation (闭式解)

线性回归

线性分类

$$\frac{k(k-1)}{2}$$

one vs. one / one vs. rest

$k-1$ 线性鉴别

Cross Entropy 梯度算法

感知问题

perception

Logistic Regression

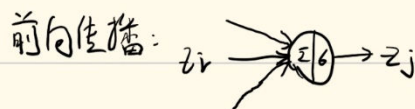
§ 神经网络

curse of dimensionality

动机: 线性模型参数 \uparrow , 可样本需求 \uparrow 指数增长

仿生学出发点 brain (脑回路用, 单一算法假设)

[考题] XOR 问题 线性不可分 \rightarrow 感知器不收敛 \rightarrow 组合 bool 函数 (析取)

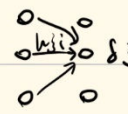


$$z_j = \phi\left(\sum_i w_{ji} z_i\right)$$

$$= \phi(w^T \cdot z)$$

* 反向传播: $\frac{\partial L}{\partial w_{ji}} = z_i \cdot \delta_j$

$s_i = \sum_j w_{ji} \cdot \delta_j$

$z_i \rightarrow$ 

DL $= \sum_j h'(a_i) w_{ji} \cdot \delta_j$ 对误差的修正

随机梯度下降 E_n 单样本损失 $\frac{\partial E_n}{\partial w_{ji}}$

$DE = \sum_n \frac{\partial E_n}{\partial w_{ji}}$ Batch

$DE = DE_n$ SGD

$DE = \sum_{n \in \text{batch}} \frac{\partial E_n}{\partial w_{ji}}$ mini-batch gradient descent

正则化 weight-decay + early-stopping

样本增强 (data augmentation)

预处理 结构 \rightarrow 神经网络 (CNN)

贝叶斯学习

先验与似然估计 点估计 \rightarrow 精度 (方差)

线性 / 判别式模型

计算学习理论

偏差-方差分解

down \uparrow overfitting (泛化误差)

误差 = 噪声 + 偏差 + 方差

low complexity (underfitting)

模型复杂度 有限复杂度 pol

VC维 \rightarrow 能正确分类样本的数量 ex-线性: 4个样本(xor) $VC=4$

§ 模型评估与选择

1. 经验误差 $E = \frac{a}{m} \rightarrow \begin{matrix} a \rightarrow \text{误分样本} \\ m \rightarrow \text{全部样本} \end{matrix}$ $acc = 1 - E = 1 - \frac{a}{m}$

误差: 输出与标签的差异

→ 训练集误差

{ 过拟合: $P \neq D \Rightarrow$ 无法解决 (避免)

{ 欠拟合: complexity \uparrow

2. 评估方法 期望: 最小化误差 \rightarrow 训练集测试误差近似估计

训练样本 $\begin{cases} \text{i.i.d} \\ \text{真实分布抽样} \end{cases}$

划分方法 ① 留出法 (hold-out) - 随机, 分属样本

② 交叉验证 (cross-validation)

③ 留一法 (leave-one-out) $\downarrow k=N$ 计算量 \uparrow

④ 自助法 (bootstrapping) m 次有放回随机抽样 \rightarrow Dataset

$$\lim_{m \rightarrow \infty} (1 - \frac{1}{m})^m = \frac{1}{e} \approx 0.368 \rightarrow \text{test set}$$

(out of bag estimate)

parameter tuning

参数配置

划分 trainset / validation set / test set

训练

样本集, 用于模型选择与评估

训练定标估计

① 网格搜索 (grid search) 与 随机搜索 (random search)

性能度量 ① 回归任务 $E(f, D) = \frac{1}{2m} \sum_{i=1}^m (f(x_i) - y_i)^2$ \downarrow 推广

mean square error (MSE) or $E(f, D) = \int_{x \sim D} (f(x) - y_i)^2 p(x) dx$

② 分类任务 错误率 (error rate) \times 准确率 (accuracy)

$$\mathbb{I}(i, j) = 1 \Leftrightarrow i = j$$

$$E(f, D) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(f(x_i) \neq y_i)$$

$$acc(f, D) = 1 - E(f, D)$$

$$E(f, D) = \int_{x \sim D} \mathbb{I}(f(x_i) \neq y_i) p(x) dx$$

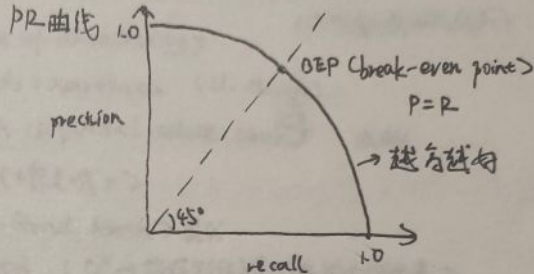
\downarrow 推广

② 分类任务：混淆矩阵

		predict	
		pos	neg
ground true	pos	TP	FN
	neg	FP	TN

精确率/查准率 (precision) $P = \frac{TP}{TP+FP}$

召回率/查全率 (recall) $R = \frac{TP}{TP+FN}$



F1 score

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P+R}$$

↓ 重要性加权

$$F_\beta = \frac{(1+\beta^2)PR}{(\beta^2 P) + R} \quad \beta > 0$$

对各组实验评价进行综合 ① macro-P / macro-R / macro-F1 (宏)

(指标平均) $\text{macro-P} = \frac{1}{n} \sum_{i=1}^n P, \text{macro-R} = \frac{1}{n} \sum_{i=1}^n R$
 $\text{macro-F}_1 = \frac{2 \times \text{macro-P} \times \text{macro-R}}{\text{macro-P} + \text{macro-R}}$

② micro-P / micro-R / micro-F1 (微)

(结果平均) $\text{micro-P} = \frac{\sum TP}{\sum TP + \sum FP}, \text{micro-R} = \frac{\sum TP}{\sum TP + \sum FN}$

ROC曲线 (receiver operating characteristic) 接收者工作特性

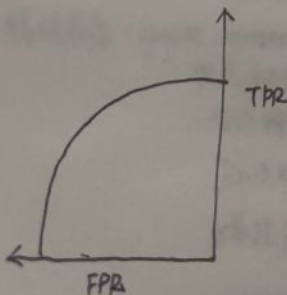
$$\text{TPR} = \frac{TP}{TP+FN} \Leftrightarrow \text{recall}$$

$$\text{FPR} = \frac{FP}{TN+FP}$$

ROC曲线越靠近左上角，分类性能越好

AUC (Area Under ROC Curve) 指标

如何绘制 ROC? 根据排序与变化点阈值截断



特征选择

OK: 奥卡姆剃刀: 简单的就是最好的

没有免费午餐: 没有一种机器学习算法适用于所有情况。

1. 子集搜索 相关特征 无关特征 relevant / irrelevant features
(与当前任务相关/无关)

def 维数灾难 (curse of dimensionality)

冗余特征 $d_3 = \lambda d_1 + \lambda_2 d_2$ (d_1, d_2, d_3)

① 穷举前向子集搜索 (sequential subset search) 依次

前向/后向/双向 (增加/减少)

② 分支定界 (Branch-and-Bound Search) 复杂

2. 子集评估 ① 信息增益 + 类别划分 ($D^v \rightarrow$ 根据特征子集产生的划分结果)

$$\text{Gain}(D) = \text{Ent}(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Ent}(D^v)$$

$$\text{Ent}(D) = - \sum_{n=1}^N p_n \log_2 p_n$$

3. 特征选择算法 ① Filter

特征选择独立于分类器选择 \rightarrow 相关性统计量

同类最近邻 (near-hit) (击中)

异类最近邻 (near-miss) (脱靶)

$$\text{对特征 } j \quad S_j^i = \sum_{i=1}^n \text{diff}(x_i^j, x_i^j, n_i)^2 + \text{diff}(x_i^j, x_i^j, n_m)^2$$

单特征评价: 平方 \rightarrow 阈值 or 近邻

② Wrapper

综合考量特征选择与分类器表现 (依据) \rightarrow 鲁棒性

Wrapper 通常好于 filter. 计算量大

LDS VELORS WRAPPER

4. 稀疏表示 (sparse representation)

① L1 正则嵌入

合并特征选择与分类器训练, 自动特征选择

$$J(w) = \sum_{i=1}^m (y^{(i)} - y(x^{(i)}, w))^2 + \lambda \|w\|_1$$

稀疏结果, 非零表示 \downarrow

② 字典学习 (dictionary learning) or (sparse coding)

$$D = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{pmatrix}$$

将样本力以个化稀疏表示

$$\min_{B, d_i} \sum_{i=1}^m \|x_i - B d_i\|_2^2 + \lambda \sum_{i=1}^m \|d_i\|_1$$

\downarrow \downarrow
重归损失 稀疏性

B 为字典, d_i 是对 x_i 的稀疏表示

找到一个字典, 使得样本转换为稀疏表示, 从而优化结构与模型

iteration ① 找到 d_i

⇔ EM 算法

② 根据 d_i 更新 B

高维与度量学习

dimensionality reduction and metric learning

升维 trade-off \Rightarrow curse of dimensionality / over fitting

$\frac{n}{d} > 0$ projection 复杂度可约化

给定 $X \in \mathbb{R}^d$, 找到线性变换 $y = A^T x \in \mathbb{R}^{d'}$, 且 $d' < d$

1. 主成分分析 (PCA)

核心: 表示

principle component analysis

$\overbrace{x \rightarrow y \rightarrow x}^{\text{子空间}}$
重建损失

$$J_{d'} = \sum_{i=1}^n \left\| \sum_{k=1}^{d'} y_{ik} e_k - x_i \right\|^2 \rightarrow \text{重建损失评价}$$

$e_1, \dots, e_{d'}$ 为子空间的基向量

$$\text{最小化重建损失 } J = \frac{1}{N} \sum_{n=1}^N \|x_n - \tilde{x}_n\|^2$$

$$\text{找一组正交基 } \{u_i\} \Rightarrow u_i^T \cdot u_j = \delta_{ij} \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

$$x_n = \sum_{i=1}^p (x_n^T u_i) \cdot u_i \Rightarrow \text{分量投影+重构}$$

$$\tilde{x}_n = \sum_{i=1}^M z_{ni} u_i + \sum_{i=M+1}^p b_{ni} u_i$$

二维情形

$$\min J = u_2^T \cdot S u_2$$

$$S = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})(x_n - \bar{x})^T$$

$$\text{s.t. } u_2^T \cdot u_2 = 1$$

$$L = u_2^T \cdot S u_2 + \lambda_2 (1 - u_2^T u_2)$$

$$\frac{\partial L}{\partial u_2} = 2S u_2 - 2\lambda_2 u_2 = 0$$

$$S u_2 = \lambda_2 u_2 \rightarrow \text{特征值 (最小的)} \Rightarrow \lambda_2 = u_2^T S u_2$$

$$\max J = u_1^T S u_1 \quad (\text{最大})$$

$$\text{s.t. } u_1^T \cdot u_1 = 1$$

$$\Rightarrow \lambda_1 = u_1^T S u_1 \rightarrow \text{最大值}$$

方法: 求 S 矩阵特征值, 从小到大排列, 取前 k 个大

$$\text{① 样本均值中心化} \quad \text{② } \Sigma = \frac{1}{N} X X^T$$

$$\text{③ 求特征值, 取最大} \quad \text{④ 对应 } \lambda \text{ 特征向量 } \text{为投影方向}$$

2. 自动编码器 (Autoencoder) 最小化重构损失
自关联 $E(w) = \frac{1}{2} \sum_{n=1}^N \|y(x_n, w) - x_n\|^2 \Leftrightarrow \text{PCA}$ 前两层 Autoencoder 做深

3. 线性判别分析 (LDA) 核心: 判别

找到投影 $y = W^T x$, 使 D_1, D_2 可分

① 最大化类间间距 $\max J = W^T (m_2 - m_1)$

s.t. $W^T W = I$

$\Rightarrow W \propto m_2 - m_1$ 效果不好!

② Fisher $\max J = W^T S_B W$

s.t. $W^T S_W W = I$

$\Rightarrow W \propto S_W^{-1} (m_2 - m_1)$

$$J = \frac{W^T S_B W}{W^T S_W W}$$

4. 流形学习 (manifold learning) 核

↓
保持邻域关系

① 等度量特征映射 (Isometric Feature Mapping) 核心: 保持距离

保持内在几何 测地线距离度量 (geodesic distance)

近邻: 欧氏距离 远点: 近邻累加短跳

a. 找近邻

b. 建立稀疏图 G , 计算近邻距离 (并加入图中) (ϵ -Isomap / k -Isomap)

c. 由 $d_a(i, j)$ 计算 $d_m(i, j)$ 测地线距离

d. 多维缩放, 保持内在几何

$$E = \|\tau(DG) - \tau(CD)\|_{L^2}^2, \tau(D) = -\frac{HSH}{2}$$

② 局部线性嵌入 (Locally Linear Embedding) 核心: 保持关系

数据充足 \rightarrow 局部线性, 由邻居重构聚类中心

$$\text{最小化重构损失 } E(W) = \sum_i \|x_i - \sum_j w_{ij} x_j\|^2$$

\rightarrow 邻居重构的系数, 要求线性保持

$$\text{s.t. } \sum_j w_{ij} = 1$$

保持不变性 $x_i \rightarrow y_i$

$$\min \Phi(Y) = \sum_i \|y_i - \sum_j \underbrace{w_{ij}}_{\text{不变}} y_j\|^2$$

③ 随机邻居嵌入 (stochastic neighbor embedding)

用概率表示点之间的相关关系, 并进行保持。

j 为 i 邻居的概率:

$$p_{ji} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma^2)} \quad [\text{原始}]$$

↓ 投影

$$q_{ji} = \frac{\exp(-\|z_i - z_j\|^2 / 2\sigma^2)}{\sum_{k \neq i} \exp(-\|z_i - z_k\|^2 / 2\sigma^2)} \quad [\text{投影}]$$

目标: 使投影后 P, Q 分布尽可能靠近

$$\min KL(P \| Q)$$

④ 度量学习 (Metric Learning)

学习合适的距离度量 \Rightarrow 学习 Metric

$$\text{马氏距离: } \text{dist}_{\text{mah}}(x_i, x_j) = (x_j - x_i)^T \underbrace{M}_{\text{学习度量矩阵}} (x_j - x_i) = \|x_i - x_j\|_M^2$$

$$\min_M \sum_{(x_i, x_j) \in M} \|x_i - x_j\|_M^2 \longrightarrow \text{相似样本度量} \downarrow$$

$$\text{s.t. } \sum_{(x_i, x_j) \in C} \|x_i - x_j\|_M^2 \geq 1 \longrightarrow \text{不相似} \uparrow$$

$$M \succeq 0 \longrightarrow \text{半正定对称}$$

概率图模型 (probabilistic graphical models)

前置知识

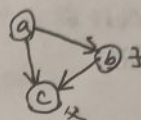
$$\begin{cases} p(x) = \sum_y p(x, y) \\ p(x, y) = p(x|y) \cdot p(y) = p(y|x) \cdot p(x) \\ p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)} \\ \sum_y p(y|x) = 1 \end{cases}$$

图 { 节点 (nodes): 随机变量
连接 (links): 变量间的概率关系

{ 有向: 因果
无向: 关联

1. 贝叶斯网络 (Bayesian Network)

$\forall a, b, c, p(a, b, c) = p(c|a, b) \cdot p(a, b)$
 $= p(c|a, b) \cdot p(b|a) \cdot p(a)$
 ↓ 拓展



$p(x_1, \dots, x_k) = p(x_k | x_1, \dots, x_{k-1}) \dots p(x_2 | x_1) \cdot p(x_1)$ 全连接网络

概率图: $p(x) = \prod_{k=1}^K p(x_k | pa_k)$ $\rightarrow x_k$ 的父

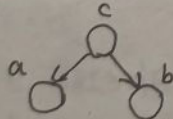
有向无环图 directed acyclic graphs DAGs

2. 条件独立性 (Conditional Independence)

$p(a|b, c) = p(a|c)$ or $p(a, b|c) = p(a|c) p(b|c)$

↪ 给定 c 下, a 与 b 条件独立 $\Rightarrow a \perp b | c$

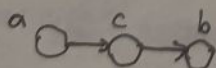
① 尾到尾 (tail to tail)



c 被观测: $p(a, b|c) = \frac{p(a, b, c)}{p(c)}$
 $= p(a|c) \cdot p(b|c)$

$p(a, b, c) = p(c) p(a|c) p(b|c)$ 此时 $a \perp b | c$

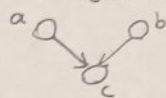
② 头到尾 (head to tail)



c 被观测: $p(a, b|c) = \frac{p(a, b, c)}{p(c)}$
 $= \frac{p(a) p(c|a)}{p(c)} \cdot p(b|c)$
 $= p(a|c) p(b|c)$

此时 $a \perp b | c$

③ 头到头 (head to head)



c 未被检测到时

$$p(a, b, c) = p(a)p(b)p(c|ab)$$

$$p(a, b, c) = p(a)p(b)p(c|ab)$$

$$\sum_c p(a, b, c) = p(a, b) = \sum_c p(a)p(b)p(c|ab) = p(a)p(b)$$

$$p(ab) = p(a)p(b)$$

故 $a \perp b | \phi$

只有当 (头到头) 结点 or 并头继续被检测到时, $a \rightarrow b$ 被阻隔 (blocked)

[条件独立阻断信息传播]

④ D 划分

考虑 a, b, c 三个不相交随机结点集合

$a \rightarrow b$ 的所有路径 { 头到尾, 尾到尾结点在 C 中, 将 a 到 b 在 C 中划分
头到头结点及其后继不在 C 中 } 被独立阻断路径

3. 马尔可夫随机场 (Markov Random Fields)

无向图

$a \rightarrow b$ 的所有路径 在 C 中结点被阻隔

$$A \perp B | C$$

团块: 中所有结点相连

最大团块

$$p(x_i, x_j | x_{-i, -j}) =$$

$$= p(x_i | x_{-i, -j}) \cdot p(x_j | x_{-i, -j})$$

联合分布: 给定数据 \rightarrow 最大团块

$$p(x) = \frac{1}{Z} \prod_c \psi_c(x_c)$$

$$\psi_c(x_c) = \exp[-E_c(x_c)]$$

能量函数

4. 有向图 \rightarrow 无向图

均匀. 高斯分布

双变信信

5. 推理

$$p(x) = \frac{1}{Z} \prod_c \psi_c(x_c)$$

$$p(x_n) = \sum_{x_1} \dots \sum_{x_{n-1}} \sum_{x_{n+1}} \dots \sum_{x_N} p(x)$$

$$= \frac{1}{Z} \left[\sum_{x_{n-1}} \psi_{n-1, n}(x_{n-1}, x_n) \dots \left[\sum_{x_2} \psi_{2, 3}(x_2, x_3) \left[\sum_{x_1} \psi_{1, 2}(x_1, x_2) \right] \right] \right]$$