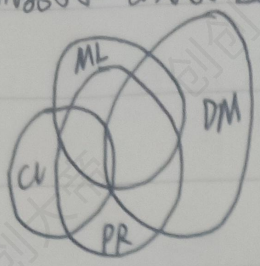
模式识别

绪论

① 模式识别 - 机器学习 - 数据挖掘 - 计算机视觉



CV：模仿人类视觉系统的方法，实现视觉信息高层理解

ML：主要研究通用的处理算法，大部分针对分类

PR：主要研究分类识别方法，面向具体应用

DM：针对各种数据库中的信息挖掘和知识发现

② 模式的两个层次：样本和类别

模式识别核心技术：模式分类

识别对象表示：特征

③ 泛化性能：测试数据集上的分类性能

④ 为什么需要结构方法：1° 表示模式的结构（字符的笔划）

2° 长度/大小不固定的模式

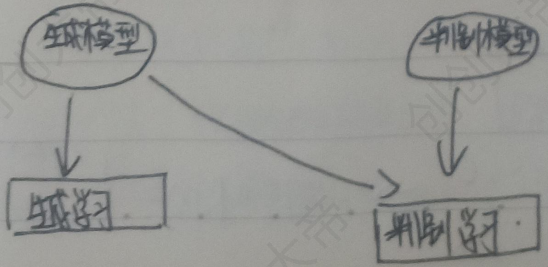
3° 相互关联的多个部件/部件同时分类

⑤ 生成模型：表示各个类别内部结构或特征分布 $P(X|C)$

判别模型：表示不同类别之间的区别，一般为判别函数，边界函数或后验概率

生成学习：从每个类别的结构描述其分布参数，不同类别分别学习

判别学习：从得到判别函数或边界函数的参数，所有类别样本同时学习



第2章 贝叶斯决策理论

先验概率 $P(w_i)$ $\sum_{i=1}^{c} P(w_i)=1$

概率密度函数(条件概率) $P(x|w_i)$

后验概率 $P(w_i|x)=\frac{P(x|w_i)P(w_i)}{P(x)}=\frac{P(x|w_i)P(w_i)}{\sum_{j=1}^{c}P(x|w_j)P(w_j)}$

if $P(x|w_1)\cdot P(w_1) > P(x|w_2)P(w_2)$ 选择 w_1

最小风险决策:

条件风险 $R(\alpha_i|x)=\sum_{j=1}^{c}\lambda(\alpha_i|w_j)P(w_j|x)$

二分类时

$R(\alpha_1|x)=\lambda_{11}P(w_1|x)+\lambda_{12}P(w_2|x)$ $R(\alpha_2|x)=\lambda_{21}P(w_1|x)+\lambda_{22}P(w_2|x)$

$R(\alpha_1|x)<R(\alpha_2|x) \Longleftrightarrow \lambda_{11}P(w_1|x)+\lambda_{12}P(w_2|x)<\lambda_{21}P(w_1|x)+\lambda_{22}P(w_2|x)$

$(\lambda_{21}-\lambda_{11})P(w_1|x)>(\lambda_{12}-\lambda_{22})P(w_2|x)$

$(\lambda_{21}-\lambda_{11})P(x|w_1)P(w_1)>(\lambda_{12}-\lambda_{22})P(x|w_2)P(w_2)$

选择 w_1 $\frac{P(x|w_1)}{P(x|w_2)}>\frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}\frac{P(w_2)}{P(w_1)}$

④ $\lambda(\alpha_i|w_j)=\begin{cases}0 & i=j\\ 1 & i\neq j\end{cases}$ 0-1 loss

$R(\alpha_i|x)=\sum_{j=1}^{c}\lambda(\alpha_i|w_j)P(w_j|x)$

$=\sum_{j\neq i}P(w_j|x)$

$=1-P(w_i|x)$ ⇒ 可以看出, 决策代价越大, 被选择概率越小.

最小误差决策 ⟺ 最大后验概率(MAP)

⑤ 带拒绝的决策

C+1 class: $\lambda(\alpha_i|w_j)=\begin{cases}0 & i=j\\ \lambda_s, & i\neq j\\ \lambda_r, \text{reject} & \end{cases}$ $\lambda_s>\lambda_r$

条件风险 $R(\alpha_i|x)=\sum_{j=1}^{c}\lambda(\alpha_i|w_j)P(w_j|x)$

$=\lambda_s[1-P(w_i|x)]$

$R_i(x)=\begin{cases}\lambda_s[1-P(w_i|x)], & i=1,\cdots,c\\ \lambda_r, \text{reject}\end{cases}$ ⇒ $\arg\min_i R_i(x)=\begin{cases}\arg\max_i P(w_i|x) & \text{if } \max P(w_i|x)>1-\frac{\lambda_r}{\lambda_s}\\ \text{reject} & \text{otherwise}\end{cases}$

⑨ 判别函数：$\arg\max_i g_i(x)$ 最大后验概率决策

⑩ 决策面：$g(x) = g_1(x) - g_2(x)$

$g(x) = P(w_1|x) - P(w_2|x) = 0$.

$P(x|w_1)P(w_1) - P(x|w_2)P(w_2) = \log P(x|w_1) + \log P(w_1) - (\log(x|w_2) + \log P(w_2))$

$= \ln \frac{P(x|w_1)}{P(x|w_2)} + \ln \frac{P(w_1)}{P(w_2)}$

⑪ 概率密度估计方法：1°参数法，$P(x|w_i) = P(x|\theta_i)$

2°非参数法 可以表示任意概率分布，无函数形式 （K-NN）

3°半参数法 近似任意概率分布，有函数形式（GM 高斯混合 EM 期望最大）

⑫ 高斯函数（正态分布）

$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$ （在给定均值和方差的所有分布中，正态分布收敛最大.

⑬ 多元正态分布 → 要考

$P(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\mu)^t \Sigma_i^{-1} (x-\mu)\right]$

均值：$\mu = E[x] = \int x p(x) dx$ $\mu_i = E[x_i]$

协方差矩阵：$\Sigma = E[(x-\mu)(x-\mu)^t] = \int (x-\mu)(x-\mu)^t p(x) dx$.

⑭ 协方差矩阵的本征值和本征向量.

本征值 ⇓ 特征值　　本征向量 ⇓ 特征向量

⑮ PCA：一种降维（特征提取）方法. 其将d维向量投影到低维子空间，使子空间投影的重要误差最小。其选择本征值最大的M个本征向量作为子空间的基。

⑯ 高斯密度下的判别函数：→ 看视频

$g_i(x) = \ln P(x|w_i) + \ln P(w_i)$

$P(x|w_i) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\mu_i)\Sigma_i^{-1}(x-\mu_i)\right]$

$\therefore g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$

情况1：$\Sigma_i = \sigma^2 I$ （方差与类别无关）

$g_i(x) = -\frac{\|x-\mu_i\|^2}{2\sigma^2} + \ln P(w_i)$　展开 $\|x-\mu_i\|^2$ $\Rightarrow$ $g_i(x) = W_i^t x + W_{i0}$

$W_i = \frac{1}{\sigma^2}\mu_i$；$W_{i0} = \frac{-1}{2\sigma^2}\mu_i^t \mu_i + \ln P(w_i)$

DATE / / M T W T F S S

情况2: $\Sigma_i = \Sigma$

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma^{-1}(x-\mu_i) + \ln P(w_i)$$

$$g_i(x) = w_i^t x + w_{i0}$$

$$w_i = \Sigma^{-1}\mu_i \qquad w_{i0} = -\frac{1}{2}\mu_i^t \Sigma^{-1}\mu_i + \ln P(w_i)$$

情况3: Σ_i = 任意

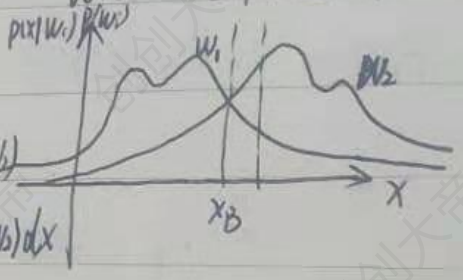
$$g_i(x) = x^t W_i x + w_i^t x + w_{i0}$$

$$W_i = -\frac{1}{2}\Sigma_i^{-1} \qquad w_i = \Sigma_i^{-1}\mu_i \qquad w_{i0} = -\frac{1}{2}\mu_i^t \Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$$

(14) 2类分错错误情况

$$P(error) = P(x \in R_2 | w_1)\cdot P(w_1) + P(x \in R_1 | w_2) P(w_2)$$

$$= \int_{R_2} p(x|w_1) p(w_1) dx + \int_{R_1} p(x|w_2)\cdot p(w_2) dx$$



$$P(correct) = \sum_{i=1}^{c} \int_{R_i} P(x|w_i) P(w_i) dx$$

当决策有的 x_B 时为最小错误率分类.

(15) 用最大后验概率决策情况

$$P(error) = \int_x [1 - \max_i P(w_i|x)] \cdot P(x)\, dx$$

DATE: / / M T W T F S S NOTES

第3章 贝叶斯决策 参数估计

① 贝叶斯决策：1°最小风险决策：$\min R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|w_j) P(w_j|x)$

2°最小错误率（最大后验概率） $\max p(w_i|x)$ MAP

② 条件特征变量：$P(x|w_i) = P(x_1 x_2 \cdots x_d | w_i)$

③ 独立二值特征 $P(x) = P(x_1 x_2 \cdots x_d) = \prod_{i=1}^{d} P(x_i)$

2类：$P(x|w_1) = \prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i}$

$P(x|w_2) = \prod_{i=1}^{d} q_i^{x_i}(1-q_i)^{1-x_i}$

最大似然比 $\frac{P(x|w_1)}{P(x|w_2)} = \prod_{i=1}^{d} \left(\frac{p_i}{q_i}\right)^{x_i} \left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}$

④ 独立二值特征的决策面

$g(x) = \log \frac{P(x|w_1)P(w_1)}{P(x|w_2)P(w_2)} = \sum_{i=1}^{d} \left[x_i \ln \frac{p_i}{q_i} + (1-x_i) \ln \frac{1-p_i}{1-q_i}\right] + \ln \frac{P(w_1)}{P(w_2)}$

$g(x) = \sum_{i=1}^{d} w_i x_i + w_0$

$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)}$ $i = 1, \cdots, d$

$w_0 = \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(w_1)}{P(w_2)}$

例3.1：$P(w_1) = 0.5$ $P(w_2) = 0.5$

$p_i = 0.8$ $q_i = 0.5$ $i = 1, 2, 3$

例3.2：$P(w_1) = 0.5$ $P(w_2) = 0.5$

$p_1 = p_2 = 0.8$ $p_3 = 0.5$ $q_i = 0.5$ $i = 1, 2, 3$

⑤ 复合模式分类

$X = x_1 x_2 \cdots x_n$ $w = w(1)w(2)\cdots w(n)$

$P(w|x) = \frac{P(x|w)P(w)}{P(x)} = \frac{P(x|w)P(w)}{\sum_w P(x|w)P(w)}$ w类别数巨大，$P(x|w)$存在估计困难性

⑥ 朴素贝叶斯：

$P(w_i|e_1, \cdots, e_k) = \frac{P(e_1, \cdots, e_k|w_i)P(w_i)}{P(e_1, \cdots, e_k)}$，$i = 1, \cdots, c$

$P(e_1 = w_{i1}, \cdots, e_k = w_{ik}|w_i) = \prod_{k=1}^{k} P(e_k = w_{ik}|w_i)$

⑦ 分类器设计

1°给定分类器结构形式，从样本中估计参数

2°统计生成模型：密度估计

3°统计判别模型：判别函数

⑧统计生成模型的参数估计

1° ML 最大似然法：作为把参数为确定值，最优估计：似然度最大。

2° 贝叶斯估计：作为把参数为随机变量，估计其分布。

⑨ ML 法：

似然：$P(D|\theta)=\prod_{k=1}^{n} P(x_k|\theta)$

最大似然：$\max_{\theta} P(D|\theta) \Leftrightarrow \nabla_\theta P(D|\theta)=0$

可能有解析解，有可能需要迭代求解。

⑩ Log-Like hood 求解 ML 问题

第一步 变对数 $l(\theta)\equiv \ln P(D|\theta)$ $l(\theta)=\sum_{k=1}^{n}\ln P(x_k|\theta)$

第二步 求 ML 估计 $\hat{\theta}=\arg\max_{\theta} l(\theta)$

$$\nabla_\theta l=\sum_{k=1}^{n}\nabla_\theta \ln P(x_k|\theta)=0$$

$$\frac{\partial l}{\partial \theta_j}=0 \quad j=1,\cdots,p.$$

高斯情况 未知 μ 当 $P(x_k|\mu)=-\frac{1}{2}\ln[(2\pi)^d|\Sigma|]-\frac{1}{2}(x_k-\mu)^t\Sigma^{-1}(x_k-\mu)$

$$\nabla_\theta \ln P(x_k|\mu)=\Sigma^{-1}(x_k-\mu)=0$$

$$\sum_{k=1}^{n}\Sigma^{-1}(x_k-\hat{\mu})=0$$

$$\hat{\mu}=\frac{1}{n}\sum_{k=1}^{n}x_k$$

高斯情况 未知 μ 和 Σ：$\hat{\mu}=\frac{1}{n}\sum_{k=1}^{n}x_k$

$$\hat{\Sigma}=\frac{1}{n}\sum_{k=1}^{n}(x_k-\hat{\mu})(x_k-\hat{\mu})^t$$

⑪ 高维情况贝叶斯估计 ⇒ P16

$$P(\mu|D)=\frac{P(D|\mu)P(\mu)}{\int P(D|\mu)P(\mu)d\mu}=\alpha\prod_{k=1}^{n}P(x_k|\mu)P(\mu)$$ α 为正则化因子。

⑫贝叶斯估计一般情况

1° 后验参数分布：$P(\theta|D)=\frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta}$ $P(D|\theta)=\prod_{k=1}^{n}P(x_k|\theta)$

2° 后验数据分布 $P(x|D)=\int P(x|\theta)\cdot P(\theta|D)d\theta$

⑭ 朴素贝叶斯使用参数极大似然估计

⑮ 最大似然参数估计与贝叶斯估计二者的区别和联系：

最大似然估计和贝叶斯估计最大区别便在于估计的参数不同，最大似然估计是估计的参数是个确定固定但未知的一个常数变量，然后结合真实数据通过最大化似然函数求解固定但未知的常数参数，即其所估计的是将参数视为是有某种由已知此类分布的随机变量，在贝叶斯估计中除了类条件概率密度 $P(X|\omega)$ 符合一定先验分布参数 θ 也符合一定先验分布

同：是参数化估计使得类条件概率密度相对简单，但估计结果的准确性依赖于所假设的概率密度分布形式是否符合潜在的真实数据分布

⑯ 增加特征有什么好处和缺点？

答：增加特征将提高模型的判别性，类别间有差异的特征有助于分类

增加特征也可能导致分类性能下降，因为有模型估计误差。

⑰ 最大似然估计的计算复杂度

1. 参数估计复杂度 μ 为 $O(dN)$ $\hat{\Sigma}$ 为 $O(Nd^2)$ $d \to$ 每个特征维度 $N \to$ 样本数

2. 求逆复杂度，计算逆矩阵此步骤复杂，一般为 $O(d^3)$

⑱ 过拟合：特征维数高、训练样本少导致对模型参数估计不准确

克服方法：降特征降维，参数收缩/平滑。

⑲ 期望-最大法 (EM) → 看作业题 (数据缺失情况下的参数估计)

EM算法保证对数据的对数似然单调递增

⑳ 基于高斯混合的 EM

$$P(X) = \sum_{k=1}^{K} \pi_k P(X|\theta_k) \quad S.t. \sum_{k=1}^{K} \pi_k = 1$$

高斯则为 $P(X|\theta_k) = N(X|\mu_k, \Sigma_k)$

参数估计采用 ML

$$\max LL = \ln \prod_{n=1}^{N} P(X_n) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k P(X_n|\theta_k)$$

$\nabla_{\pi_k} LL = 0 \quad \nabla_{\mu_k} LL = 0 \quad \nabla_{\Sigma_k} LL = 0$

DATE / / M T W T F S S

(20) 隐马尔可夫模型

联合 $P(O,Q)$，$P(O|Q)$

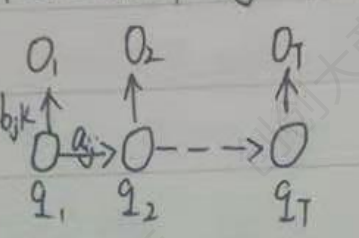
马尔可夫链：$P(q_1 q_2 \cdots q_T) = P(q_1)P(q_2|q_1)P(q_3|q_1,q_2)\cdots P(q_T|q_1,\cdots q_{T-1})$

一阶马尔可夫链：$P(q_t=S_j|q_{t-1}=S_i, q_{t-2}=S_k,\cdots) = P(q_t=S_j|q_{t-1}=S_i)$

$P(q_1,q_2,\cdots q_T) = P(q_1)\cdot P(q_2|q_1)P(q_3|q_2)\cdots P(q_T|q_{T-1})$

状态转移概率 $a_{ij} = P(q_t=S_j|q_{t-1}=S_i)$ $\sum_{j=1}^{N} a_{ij} = 1$

(1)



$O_i \in \{v_1, \cdots, v_M\}$

$q_i \in \{S_1, S_2, \cdots, S_N\}$.

观察序列可能分布：$b_j(k) = P(v_k \text{ at } t | q_t = S_j)$

(21) 隐马尔可夫的计算 → 看作业题

$$P(O|\lambda) = \sum_{\text{all } Q} P(O|Q,\lambda)P(Q|\lambda)$$

$$= \sum_{q_1,\cdots,q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2)\cdots a_{q_{T-1}q_T} b_{q_T}(O_T)$$

条件独立 $P(O|Q,\lambda) = \prod_{t=1}^{T} P(O_t|q_t,\lambda)$

$$= b_{q_1}(O_1) b_{q_2}(O_2)\cdots b_{q_T}(O_T)$$

$P(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3}\cdots a_{q_{T-1}q_T}$

复杂度 $\boxed{2TN^T}$

(22) 隐马尔可夫的三个问题和训练 (双佳)

DATE / / M T W T F S S NOTES

非参数方法

① 基本密度估计 $p(x)=\frac{k}{nV}$

② 非参数概率密度估计:

1° Parzen window:固定局部区域体积,K变化

2° KNN:固定局部样本数,V变化。

③ Parzen窗

$\varphi(u)=\begin{cases}1 & |u_j|\le \frac{1}{2}\\ 0 & \text{otherwise}\end{cases}$ 满足条件 $\varphi(x)\ge 0$ $\int \varphi(u)du=1$

以x为中心,体积为 $V_n=h_n^d$ 的局部区域内样本数

$$k_n=\sum_{i=1}^{n}\varphi\left(\frac{x-x_i}{h_n}\right)$$

概率密度估计 k_n/nV_n。

大h:低可变性,未拟合

小h:高可变性,过拟合

窗宽 h_n 选择经验:一般原则:n越大或密度越大,h_n 越小

$$V_n=V_1/\sqrt{n}$$

④ K近邻估计 → 老师说会考

固定局部区域样本数k,体积V变化 k↑窗口大 更平滑

K-NN分类规则里没有概率密度,但要注意,该规则是从非参数概率密度估计和贝叶斯决策过来的。

最近邻规则的错误率。

K近邻的快速计算 $O(dn)$

近邻搜索三种策略: 1° 部分距离

2° 预结构化

3° 编辑

⑤ 度量学习:距离度量的参数被优化。

⑥ 对于图像平移/形状的变换(平移,缩放,弯曲,旋转),距离将会有很大变化。因此引出了切线距离

Tangent Distance 本质是为了解决经过图像变换后距离度量的问题,其在经过梯度下降算法下求切空

DATE / / M T W T F S S

间中优化求得投影向量和原始图像及其经过变换后图像的空间中最近的点.

① 级数展开估计

测试样本 训练样本

$$\varphi\left(\frac{x-x_i}{h_n}\right)=\sum_{j=1}^{m} a_j \psi_j(x) \chi_j(x_i)$$

$$\sum_{i=1}^{n}\varphi\left(\frac{x-x_i}{h_n}\right)=\sum_{j=1}^{m} a_j \psi_j(x)\sum_{i=1}^{n}\chi_j(x_i)$$

$$P_n(x)=\sum_{j=1}^{m} b_j \psi_j(x) \quad b_j=\frac{a_j}{n}\sum_{i=1}^{n}\chi_j(x_i)$$

b_j 可离线计算. $P_n(x)$ 每加 m 次计算.

线性判别函数

① 不需要有关概率密度函数的确切知识，即可进行判决。因此属于无参数估计方法。

② 模式分类的途径：

1° 估计类条件概率密度函数，利用贝叶斯公式估计后验概率，然后决策。概率密度函数的估计一般较复杂

2° 直接估计后验概率，不需要估计类条件概率密度 K-近邻

3° 直接设计判别函数，直接利用已知类别的判别函数

③ 利用样本设计分类器基本思想：

1° 给定一个判别函数，且已知该函数的参数形式

2° 采用样本来训练判别函数的参数

3° 对于新样本，采用判别函数对其进行决策，并按照一些准则来完成分类。

④ 两类情形 一个作业题，估计任意

1° one vs all 存在不确定区域

2° one vs one 存在不确定区域

对于 i：如果 $g_i(x) > g_j(x)$ x 就将被分类 w_i 类，否则不决策

$x \in w_i$ $g_i(x) = \max_{i=1,2,\cdots,c} g_i(x)$

线性机器将样本空间分为 C 个决策区域 $R_1, \cdots R_c$

⑤ 广义线性判别函数

数据点 x —非线性映射→ 数据点 y

在新的数据点空间中应用线性判别函数方法。

$g(x) = \sum_{i=1}^{d} a_i \boxed{y_i(x)}$ ← 变换函数

令 $a = [a_1, a_2, \cdots a_d]^T$ $y = [y_1, y_2, \cdots, y_d]^T$

$g(x) = a^T y$

a 为广义权向量。 $g(x)$ 对 x 非线性，对 y 线性。当新空间维数足够高时，$g(x)$ 可以通过任意判别函数

⑥ 对线性判别函数采用齐次增广表示：线性齐次空间上增加一个维度，仍可保持原式性质不变，分类

不仅与原来的决策面相同，但分类面将过坐标原点。 $r = \frac{a^T y}{\|a\|}$

⑦ 规范化处理：规范化增广样本，解向量（$a^T y_i > 0$ 的权向量 a）

⑧ 感知准则函数

$a^T y_i > 0$ $i = 1, 2, \cdots n$

在线性可分情形下，a 有无穷个，所以需要引入准则： $J_p(a) = \sum_{y \in Y(a)} (-a^T y)$ 为全部分错样本

$\min_a J_p(a)$

DATE M T W T F S S

$$\frac{\partial J_p(a)}{\partial a} = -\sum_{y\in Y} y$$

$$a_{k+1} = a_k + \eta_k \sum_{y\in Y_k} y$$

⑨ 感知器准则函数存在的问题

1° 每次错分样本点集合可能会改变

2° 什么向保留各坐标原点旋转远的

⑩ 固定增量单样本修正方法，可变增量单样本修正方法。

⑪ 感知器准则收敛性定理证明

⑫ 学习准则：线性准则、平方准则、松弛准则

$J_l(a) = \sum_{y\in Y}(-a^Ty)$ 　 $J_q(a) = \sum_{y\in Y}(a^Ty)^2$ 　 $J_r(a) = \frac{1}{2}\sum_{y\in Y}\frac{(a^Ty-b)^2}{\|y\|^2}$ 归一化 + margin

⇒ 解决异常点影响的 　 Y 为 $a^Ty\le b$ 的

松弛准则好处：① 线性准则为分段线性，梯度不连续

② 平方准则梯度连续，但目标函数过于平滑，收敛速度慢，同时目标函数过于受到最长样本影响

③ 松弛准则避免了上述缺点。

$$\frac{\partial J_r(a)}{\partial a} = \sum_{y\in Y}\frac{a^Ty-b}{\|y\|^2}y$$

$$a_{k+1} = a_k - \eta_k \sum_{y\in Y}\frac{a^Ty-b}{\|y\|^2}y$$

对于单样本更新：$a_{k+1} = a_k - \eta_k \frac{a_k^Ty^k-b}{\|y^k\|^2}y^k$

$$= a_k - \eta_k \frac{a_k^Ty^k-b}{\|y^k\|}\cdot\frac{y^k}{\|y^k\|}$$

$\frac{a_k^Ty^k-b}{\|y^k\|}$ ↓ 点到超平面距离

$\frac{y^k}{\|y^k\|}$ ⇒ 单位向量方向

$0<\eta_k<1$	靠近超平面
$\eta_k=1$	到达超平面
$\eta_k=2$	超过超平面

⑬ 松弛方法收敛性证明

⑭ 最小平方误差(MSE)准则函数

$a^Ty_i = b_i > 0$ 　 b_i 为任意给定的正常数

$e = Ya - b$ 　 Y 为矩阵 a 为向量，b 为向量

$$J_s(a) = \|e\|^2 = \|Ya-b\|^2 = \sum_{i=1}^{n}(a^Ty_i - b_i)^2$$

$$\frac{\partial J_s(a)}{\partial a} = \sum_{i=1}^{n}2(a^Ty_i-b_i)y_i = 2Y^T(Ya-b)$$ 令偏导为0 ⇒ $a = Y^+b$

时 Y^+ 计算多次时，可以采用梯度下降法 $a_{k+1} = a_k + \eta_k Y^T(b - Ya_k)$

其收敛条件是 $Y^T(b-Ya)=0$.

$a_{k+1} = a_k + \eta_k (b_k - a_k^T y^k) y^k$ 考虑单个样本对误差的贡献.

⑮ Widrow-Hoff方法. η_k 随着 k 增加而逐渐减小. $\eta_k = \eta_1/k$.

⑯ Widrow 相对于感知器准则，最小平方准则方法可能并不收敛于可分超平面，即使该平面是存在的.

MSE方法寻找的为最小化样本至超平面的距离平方和.

⑰ Ho-kashyap

$J_s(a,b) = \|Ya-b\|^2$ a,b均未知，进行优化，b可理解为margin，解决在 Widrow-Hoff 方法得到最优解不一定在可分超平面上.

若 $e_k = Ya_k - b_k$ 全为0，此时，b_k 将不再更新. 如果 e_k 有一部分元素小于0，则可以证明该问题不是线性可分的.

⑱ 分类线性判别函数 { 1° MSE的梯度下降
2° 感知器准则扩展方法——多样本修正法
3° Kelser构造.

DATE

人工神经网络第1讲

① 转移函数应是非线性，否则网络计算能力并不比单层网络强

② 前馈网络与反馈网络比较

1° 前馈型网络"不存储记忆"，输出仅为当前输入的加权和再加激活函数

2° 反馈型网络要将以前的输出作为这次的输入

3° 反馈型类似于"人类的短期记忆"，网络的输出状态会取决于以前输入

③ 无监督学习本质是通过对样本所隐含的统计特性

④ 梯度下降法

$$\Delta W_{ij} = -\eta \frac{\partial E}{\partial W_{ij}}$$

⑤ 线性单元训练：δ规则

$$E(W) = \frac{1}{2}\sum_{k\in D}(t_j^k - z_j^k)^2 = \frac{1}{2}\sum_{k\in D}(t_j^k - \sum W_{ij}x_i^k)^2$$

$$\Delta W_{ij} = -\eta\frac{\partial E}{\partial W_{ij}} = \eta\sum_{k\in D}(t_j^k - z_j^k)x_i^k = \eta\,\delta_j^k x_i^k \qquad \delta_j = t_j^k - z_j^k$$

⑥ 考虑转移函数，转移函数是非线性函数，且处处可微分。

单个样本的求导：$\Delta W_{ij} = \eta\,\delta_j^k x_i^k$ $\quad \delta_j^k = \frac{-\partial E}{\partial net_j^k} = f'(net_j^k)[t_j^k - z_j^k]$

解释：边的起始结点的单输出乘以边的指向结点"经导数缩放之后的误差"

⑦ Sigmod函数 $f(s) = \frac{1}{1+e^{-s}}$

$$f'(s) = \frac{e^{-s}}{(1+e^{-s})^2} = \frac{1}{1+e^{-s}}\left(1 - \frac{1}{1+e^{-s}}\right) = y(1-y)$$

$$\delta_j^k = y_j^k(1-y_j^k)[t_j^k - y_j^k]$$

⑧ BP算法基本原理：利用输出层的误差来估计输出层的前一层的误差，再用这个误差估计前一层的误差，如此一层一层地传递下去，从而获得所有其他各层的误差估计

⑨ BP算法权重更新推导（看作业题）

DATE / / M T W T F S S NOTES

人工神经网络第2讲

① 对于多分类问题，采用 one-hot 编码

② 交叉熵准则：$E_{ce}(W) = -\sum_{n} t_j^k \ln(z_j^k)$

③ 激励函数：1°非线性

2°有界连续可导

3°最好单调.

④ 隐含层数设定：对于线性问题，隐含层的存在决定了网络的表达能力，决定了决策面的复杂程度.

⑤ 结点个数：太少，不能建立复杂的判别界面

太多，容易过拟合，失去推广能力.

一般设置较大，允许有冗余结点可能 1°正则化神经网络

2°修剪过连接的神经网络

3°dropout 技术.

⑥ 初始权重：初始权重 $\Delta W_{ij} \neq 0$，否则不会学习了

⑦ 正则化技术 防止网络出现 overfitting 的一种有效方法

⑧ 学习率.

⑨ 附加冲量项.

⑩ 训练停止准则：没有固定准则，过度训练会 overfitting，训练不可以令出现识别精度不足

⑪ BP 算法中存在的问题：

1°完全对准训练：网络的麻痹现象，梯度消失，局部极小

2°训练时间过长，尤其对复杂问题，需要很长时间训练，可能选取不恰当的

⑫ 网络的麻痹现象：由于在计算权重修正量时，误差 δ 正比于 $f'(net)$，当 $f(net) \to 0$ 时，$\delta \to 0$，从而 $\Delta W_{ij} \to 0$，从而其相当于调节过程停滞.

⑬ 梯度消失：由于 $\delta_n^k = f'(net_n^k) \sum W_{ij} [\delta 0]_j^k$，由于 $f'(net_n^k)$ 通常小于 1，所以在多层神经网络前向神经网络中，随着当前向层数增多，容易出现梯度消失

DATE / / M T W T F S S

⑭ 径向基函数网络：其可以对任意连续的非线性函数进行逼近，收敛速度比通常的基于神经网络的

⑮ 径向基网络结构简化，采用聚类技术对数据进行聚类，每个隐含层结点代表一个聚类中心，

经过聚类处理，还应防止overfitting，增强网络的泛化能力，提高精度

⑯ Hopfield网络：从初始状态按能量减小的方向进行变化，直到达到稳定状态，稳定状态即为网络的输

⑰ BM，一种随机Hopfield网络，BM一部分神经元与外部相连，另一部分不与外部相连

⇓ 缺：网络结构复杂，训练代价大，局部不稳定

⑱ RBM 层内结点不相连，信息双向流动

⑲ 自组织映射

DATE / / M T W T F S S NOTES

人工神经网络第3讲

①卷积神经网络降低网络权重数量：局部连接 → 图像空间相关性

权值共享 → 采用滤波器，每个滤波器可以看成是学习一种特征

②卷积神经网络网络训练采用仿传播算法

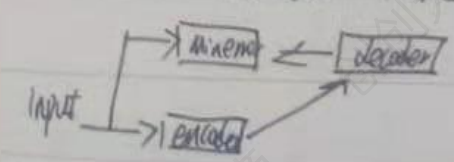
③卷积神经网络核心思想：1°使用卷积后的特征是因为图像具有一种"静态性"的属性，这也就意味着在一个图像区域有用的特征极有可能在另一个区域同样适用

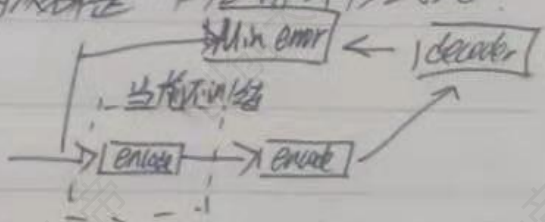
2°局部感受野、权值共享以及时间或空间下采样这三种结构化思想结合起来获得了某种程度的位移、尺度、形变不变性。

④自编码器：给定一个神经网络，假设其输出与输入是相同的，通过训练该网络，可以得到各层中的权重。

⑤自编码器是一种尽可能重构其输入信号的神经网络

⑥自编码器使用逐层训练思想，每次都是一个三层的神经网络





⑦ Autoencoder 可用作特征提取器使用，只应用编码器，最后一层编码器的输出作为该样本的新特征。

⑧Autoencoder 实质是一个特征学习，层级变换，特征的语义性，抽象性，结构性表达明显。

对分类器，可以用Autoencoder学习，学习得到权值为初始权重，采用带有标签数据去进行再次学习，去进行微调。

⑨

DATE / / M T W T F S

特征提取与特征选择

①SIFT 尺度不变特征转换：1°检测极值点空间的极值点

2. 准确的关键点定位

3°特征方向估计

4°关键点描述计算.

②特征选择：

③降维数削减 作用：1°低维数据可视化

2°避免维度诅咒

3°加速学习算法

4°更少存储的.

④PCA ⇒有作业题 计算流程

⑤CCA 典型相关性分析：探索两个协变量之间的关联关系，这两个协变量来自一个相同个体

1.数据降维：用少量的线性组合来解释两组变量之间的相关作用

2.数据解释：寻找特征值，这些特征值对解释两个变量之间相互作用十分关键.

⑥LDA 线性判别分析：不同于PCA方差最大化理论，LDA算法的思想是将数据投影到低维空间使得同一类数据尽可能紧凑，不同类的数据尽可能分散

⑦LDA = 白化 + PCA

⑧LDA的约束或假设：1°每个类是单模态高斯分布

2°每一个类的协方差与其他都不相同

3°降维后的维数不能超过C-1

4°类别分开问题在

⑨独立成分分析ICA. ~~ICA算法~~ 与PCA不同，ICA追求的是各个变量相互独立，而非仅不相关，因此ICA

~~需要~~利用数据分布的高阶统计信息而非仅二阶信息

~~很多~~应用中，ICA提取的特征优于PCA.

DATE / / M T W T F S S NOTES

⑩ ICA会恢复出幅值，但对应关系消失。

⑪ 非线性打度，流形，保持邻居关系

解决方法：Kernel PCA

使用流形基于方法（LLE. Isomap . . . .）

⑫ Kernel PCA：对于样本空间中的矩阵X，我们先用一个非线性映射把X中的所有样本用映射到一个高维甚至是无穷维的空间，然后在这个高维空间进行PCA降维

⑬ LLE：显式利用"局部线性"的假设，流形时子的局部区域具有欧式空间的性质，用数学在LLE中，先假设某点x_i坐标可以由它周围的一些点的坐标线性组合求出。

保持局部邻域几何结构

和重对样本集自的几何变换仍具有不变性。

主要步骤：1°寻找每个样本点的k个近邻点。

2°由每个样本点的近邻点计算出该样本点的局部重建权值矩阵

3°由该样本点的局部重建权值矩阵和其近邻点计算出样本点的全局坐标

⑭ ~~Isomap~~ MDS降维算法：将高维空间中的点投影到低维空间，保持点间此之间的相似度尽可能不变。

算法流程：1°计算原始空间中数据点矩阵

2°计算内积矩阵B

3°对矩阵B进行特征值分解，获得特征值矩阵Λ和特征向量矩阵V

4°取特征值矩阵最大的前d个及其对应的特征向量 $Z=V_z\Lambda_z^{\frac{1}{2}}$

⑮ Isomap：基于MDP算法，不同之处在于Isomap用图中两点的最短路径距离替代了MDS中原来空间的欧氏距离，这样能更好的利用流形的数据。

算法流程 (1) 设置每个点最近邻点数以，构建连通图的邻接矩阵

(2) 通过图的最短路径构建原始空间中的距离矩阵

(3) 计算内积矩阵B

DATE / / M T W T F S S

(14) 对矩阵B进行特征值分解，获得特征值矩阵Λ和特征向量矩阵V

(15) 目对特征值矩阵最大的前d项及其对应特征向量 $Z = V_d A_d^{\frac{1}{2}}$.

(16) LPP局部保持投影：非线性降维方法的线性化。相当间有某种非线性关系的样本点，在降维后仍然保持这种非线性关系

(17) 特征选择：降维 VS 特征选择

降维：1°仍有原始特征被使用
2°转化特征是原始特征线性组合.

特征选择：原始特征的若干个子集的选择.

(18) 特征选择：不能一个个特征看，不能忽略特征相关性

- 过滤法：按照发散性或者相关性对各个特征进行评分，设定阈值或者待选择阈值个数，选择特征
- Wrapper：包装法，根据目标函数（通常是预测效果评分），每次选择若干特征，或者排除若干特征
- Embedded：嵌入法，用某些算法和模型进行训练，得到各个特征的权值系数，根据系数从大到小选择特征

DATE / / M T W T F S S NOTES

模型选择

① 机器学习：泛化　过拟合　维数灾难　理论上的保证　相关并不意味因果

随着样本维数增加，正确泛化难度指数级增加。

② 集成学习：若干单个分类器集成。{单一学习器错误低于0.5；每个分类器应当各不相同。

③ 层次泛化：分层结构，第一层的输出作为第二层的输入

④ Bagging：训练一组基分类器，每个基分类器通过有放回地随机训练样本集来训练，后通过投票进行统计。

⑤ 随机子空间：对于每一个分类器，从D个特征中选择d个子特征来构建一个训练集合同时学习个分类器

⑥ Adaboost：从弱学习算法出发，反复学习得到一系列弱分类器。然后加权组合这些弱分类器，构成一个强分类器

1° 提高前一轮弱分类器分类错误样本的权重，降低之前被正确分类样本权重

2° 加大分类误差率较小的分类器的权重

DATE / / M T W

数据聚类

① 聚类方法分类：基于距离的聚类方法，基于密度的聚类方法，基于连通性的聚类方法

② $\nabla_{\theta_i} f_{lk}(\theta|x_k) = P(w_i|x_k,\theta)\nabla_{\theta_i} \ln L P(x_k|w_i,\theta_i))$

单个样本 x_k 对"待确似然函数关于 θ_i 的梯度"之贡献等于"x_k 属于第 i 个成分的后验概率"乘以"x_k 属于第 i 个成分密度 $P(x|w_i,\theta_i)$ 的对数关于 θ_i 的梯度"

③ 基于混合密度的聚类

④ 谱聚类：其本质是将聚类问题正则化为顶点划分的最优问题

拉普拉斯矩阵：度矩阵减去邻接矩阵

子图相似度：连接两个子图所有边的权重之和

最小二分切割：切开之后两个子图之间的相似性最小

归一化最小二分切割：采用子图的边数或者体积来对切割进行归一化

⑤ 如果图 G 具有 K 个连通子图，若每个连通子图为一个聚类，那么采用其拉普拉斯矩阵的零特征值对应的特征向量可以分别这些子图分别开来

⑥ 考察拉普拉斯矩阵最小的特征值对应的特征向量，并由这些特征向量组成新的特征空间

⑦ 谱聚类算法核心步骤：1° 利用点对间相似性，构建亲和度矩阵

2° 构建拉普拉斯矩阵

3° 求解拉普拉斯矩阵最小的特征值对应的特征向量

4° 由这些特征向量构成点的新特征，采用 K-means 等聚类方法完成最后的聚类

⑧ 未标准化的谱聚类 Shi算法，Ng算法 核心是将原始数据点 X_i 转换为特征空间的数据点 Y_i 在新的空间对原始数据进行描述

⑨

DATE / / M T W T F S S NOTES

支撑向量机与核方法

① VC维

② 硬边界-SVM

$$\underset{w,b}{\operatorname{argmax}} \underset{x_i \in D}{\operatorname{argmin}} \frac{|b + x_i \cdot w|}{\sqrt{\sum_{i=1}^{d} w_i^2}} \quad s.t.\ \forall x_i \in D : y_i(x_i \cdot w + b) \geq 0$$

$$\Downarrow$$

$$\underset{w,b}{\operatorname{argmin}} \sum_{i=1}^{d} w_i^2 \quad s.t.\ \forall x_i \in D\ y_i(x_i \cdot w + b) \geq 1$$

③ soft-SVM

$$\{\vec{w}, b^*\} = \min_{w,b} \sum_{i=1}^{d} w_i^2 + C \sum_{j=1}^{N} \xi_j$$

$$y_1(\vec{w} \cdot \vec{x_1} + b) \geq 1 - \xi_1,\ \xi_1 \geq 0$$

$$y_N(\vec{w} \cdot \vec{x_N} + b) \geq 1 - \xi_N,\ \xi_N \geq 0$$

引入hinge Loss

$$\underset{f \in H}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(x_i)) + \lambda \|f\|_H^2$$

④ 硬 $\max W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1} \alpha_i \alpha_j y_i y_j x_i^T x_j \quad s.t.\ \alpha_i \geq 0 \quad \sum_{i=1}^{n} \alpha_i y_i = 0.$

软 $\max W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \quad s.t.\ C \geq \alpha_i \geq 0 \quad \sum_{i=1}^{n} \alpha_i y_i = 0$

⑤

决策树算法

① 熵 $H(X) = -\sum_{x \in X} P(x) \log_2 P(x)$

② 条件熵 $H(Y|X) = \sum_{i=1}^{n} P_i H(Y|X=x_i)$

信息增益：得知特征X的信息而使类别的信息不确定性减少的程度

$$g(D|A) = H(D) - H(D|A)$$

③ 信息增益算法 矢考

输入：训练数据集D和特征A；

输出：特征A对训练数据集D的信息增益 $g(D,A)$

(1) 计算数据集D的经验熵 $H(D)$

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|}$$

(2) 计算特征A对数据集D的经验条件熵 $H(D|A)$

$$H(D|A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{K} \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|}$$

(3) 计算信息增益 $g(D,A) = H(D) - H(D|A)$

④ ID3采用信息增益作为度量，其偏向于具有大量值的属性，在训练集中某个属性所取的不同值的样例非常多甚至都有可能每个都不一样，则ID3会选择它作为分裂属性

⑤ C4.5采用信息增益率

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

⑥ 分类与回归树 CART 采用二叉树，基尼指数

Gini $Gini(p) = \sum_{k=1}^{K} P_k(1-P_k) = 1 - \sum_{k=1}^{K} P_k^2$

$$Gini(D,A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

⑦ 决策树对连续值，有分支

⑧ 随机森林：RF使用了两次随机抽取，1° 对训练样本的随机抽取 (bagging)

2° 对变量的随机抽取 (从中对特征值随机选取)

DATE / / M T W T F S S

⑨ 随机森林中的两个随机性 1° 防止过拟合，随机有放回抽样
2° 对特征空间抽取.

⑩ 影响RF分类性能因素 1° 分类强度，每棵树的分类强度越大，RF分类性能越好
2° 树之间相关度 越小越好