Bayesian Decision

① $P(H|X) = \frac{P(H)\cdot P(X|H)}{P(X)}$

直接判断.

② X, Ω, A, λ 损失. (A: 动作)

$\Rightarrow$ ☆ $R(\alpha_i|X) = \sum_{j=1}^{c} \lambda(\alpha_i|w_j)\cdot P(w_j|X)$

$\Rightarrow$ Bayes decision rule: 损失小的选.

$\alpha(X) = \arg\min_{\alpha_i \in A} R(\alpha_i|X)$.

分类问题中 α_i 为分类的类别.

二分类: $R(\alpha_1|X) < R(\alpha_2|X)$ $\Leftrightarrow$ $\lambda_{21}\cdot P(w_1|X) + \lambda_{22}\cdot P(w_2|X)$ (判, 真)

$\Leftrightarrow \frac{P(X|w_1)}{P(X|w_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \cdot \frac{P(w_2)}{P(w_1)}$ 前判后真.

多分类: 记得全概率公式 & 条件概率公式

使用0-1损失时: 对0错1.

$R(\alpha_i|X) = 1 - P(w_i|X)$.

$= \sum_{j=1}^{c} \lambda(\alpha_i|w_j)\cdot P(w_j|X)$ 损失小

$\Downarrow$

w_i if $P(w_i|X) > P(w_j|X)$ 谁大选谁.

$P(x|w_i) \sim N(\mu_i, \sigma_i^2)$

高斯分布: $X \sim N(\mu, \sigma^2)$

$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]}$

$E[x] = \int_{-\infty}^{+\infty} x\,p(x) = \mu$

$Var[x] = \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot P(x) = \sigma^2$ ($E(X-EX)^2$)

多元 $P(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp[-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)]$

X、μ 向量, Σ 矩阵.

书26页.

$\Downarrow$

针对Gaussian的判别函数推理 (P28)

$\begin{cases} \sigma I = \Sigma_i \\ \Sigma_i = \Sigma \end{cases} \rightarrow$ linear 带入后二次项无意义.(书上)

Σ_i = arbitary $\rightarrow$ quadratic

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

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

Parametric Estimation

Maximum Likelihood

① $P(w_j) \rightarrow$ count.

② $P(x|w_j) \rightarrow$ 假设服从某一分布

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

$\hat{\theta} = \arg\max_{\theta} P(D|\theta)$.

求解: ①加上ln ($\prod \rightarrow \Sigma$) ② 对θ求导为0 $\Rightarrow$ 必要条件 (Σ还在.)

③检查极值点哪一个最优 {代入 / 二阶导.

应用: Gaussian $\begin{cases} \text{已知}\Sigma, \text{估计}\mu \\ \Sigma, \mu \text{未知} \end{cases}$ P70

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

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

Bayesian Estimation:

求 $P(w_j|x, D^*)$.

$\Rightarrow$ 每个类的后验, 最大的即结果 $\Rightarrow$ 可能求别的后验

此处 $P(x|D) = P(x|w_j, D_j)$.

☆ $P(w_j|x, D) = \frac{P(w_j)\cdot P(x|w_j, D_j)}{\sum_i P(w_i)\cdot P(x|w_i, D_i)}$

乘法公式+条件概率可证

求解: P74 高斯情况, 求 $P(\mu|D)$.

☆ 核心: $P(x|w_j, D_j)$

① $P(x|D) = \int p(x|\theta)p(\theta|D)\,d\theta$.

② $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{}$ $\alpha \Leftarrow \int P(D|\theta)P(\theta)d\theta$.

③ $P(D|\theta) = \prod_{k=1}^{n} P(x_k|\theta)$.

递归 P79. 看题目. 按步骤+化简.

HMM:

Notations:

① $\Omega = \{w_1, \cdots\} \Rightarrow$ state

② $V = \{v_1, \cdots\} \Rightarrow$ signal/symbol

③ $w^T = \{w(1), \cdots\} \Rightarrow$ state sequence (时序)

④ $V^T = \{v(1) \cdots\} \Rightarrow$ symbol sequence

⑤ A、B、π. $\pi_j = P(w(1) = w_j)$

Evaluation: P108 求 $P(V^T|\theta)$

前向 forward. $\Rightarrow$ 记得求和, 从1开始

后向 backward. $\Rightarrow$ 看准下标, 从T开始

F: $\alpha_j(t) = [\sum_{i=1}^{c}\alpha_i(t-1)a_{ij}]\, b_{jv(t)}$ (状态; 前; 后) 结果是 $\alpha_j(1) = \pi_j b_{jv(1)}$ 求和

B: $\beta_j(t) = \sum_{i=1}^{c}\beta_i(t+1)\, a_{ji}\, b_{iv(t+1)}$ (后; 前) $\beta_j(T) = 1$

建议列表做.

☆ Decoding: 求 w^T, $\max P(w^T|\theta, V^T) \Rightarrow \sum_j \pi_j b_{jv(1)} \beta_j(1)$ (P)

① Viterbi

$\delta_j(t) = \max_{w(1)\cdots w(t-1)} P(w(1), \dots, w(t-1), w(t) = w_j, v(1)\cdots v(t)|\theta)$

$\Rightarrow$ 到t的前面state/symbol最高的概率 ($w(t) = w_j$ 时)

$\Rightarrow$ 求解: $\delta_j(1) = \pi_j b_{jv(1)}$, $\psi_j(1) = 0$.

$\delta_j(t) = [\max_i \delta_i(t-1)a_{ij}]\, b_{jv(t)}$

$\psi_j(t) = \arg\max_i \delta_i(t-1)a_{ij} \rightarrow$ 最有可能的w下标

$\Downarrow$ 值是下标. $\Rightarrow$ 有n个(根据t推t-1)由后往前

☆ $w(T) = \arg\max_j \delta_j(T)$ 从后往前. (后处理).

$w(t) = \psi_{w(t+1)}(t+1)$.

小数计算很多 别错了(计算器)

Learning: 求 $\theta = \{A, B, \pi\}$, know V^T.

$\Rightarrow \arg\max_{\theta} P(V^T|\theta)$.

前向-后向算法 (P113).

$\hat{\pi}_i$ / $\hat{a}_{ij}$ / $\hat{b}_{ik}$

Bayesian Belief Network.

P44.

只有父子相关

$\Rightarrow$ 联合: $P(a,b,c) = P(a|Pa(a))\cdot P(b|Pa(b))\cdot P(c|Pa(c))$

边缘: $P(a) = \sum_{b,c} P(a,b,c)$

条件: $P(a|b) = \frac{P(a,b)}{P(b)}$

Non-parametric

P134.

① Density estimation 求 $p(x)$ (pdf):

$\begin{cases}\int\varphi(u)du = 1 \\ \varphi(x) \geq 0\end{cases}$

原理:

$P(x) \approx P = \frac{E[k]}{n} = \frac{k/n}{V}$

1) Parzen Windows P134 ($=h_n^d$ 超立方体)

☆ h_n 为超参 P136 固定 V (参数), 求 k.

$k_n = \sum_{i=1}^{n}\varphi(\frac{x-x_i}{h_n})$ 样本. (边长)

任意pdf. $(\varphi(u) = \begin{cases}1, |u_j| \leq \frac{1}{2} \\ 0\end{cases}$.

待测样本(即求) φ 可用任意pdf

$\Rightarrow P_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi(\frac{x-x_i}{h_n})$, $V_n = h_n^d$

求的相当于时 $P(x|w_k)$ 只用了类k数据.

再用后验决策: $P(w_k|x) = \frac{P(w)\cdot P(x|w)}{P(x)}$ (数数; 别忘了求 $P(x)$; 常数)

2) P143

③ k_n - Nearest-Neighbor

$\Rightarrow$ 人工定 k_n, 在x中心变大 V_n 直到 k_n 个样本, 返回 V_n 算 $P(x|w_j)$.

$P_n(x|w_j)$

$P_n(x) = \frac{k_n/n}{V_n}$ $\rightarrow h_n^d \rightarrow h_n$ 是有 k_n 个样本 时的最小距离

一类一个! =

$\Downarrow$

$P(x|w_j)$, 用 $P(w_j|x)$ 判断!

③ Decision $\Rightarrow$ 直接估计 $P(w|x)$.

1) 最近邻 NN P146 $\Rightarrow$ 选与测试集最近的样本, 二者同类.

2) kNN k近邻. $\Rightarrow$ 选最近k个点, 其中哪个类多选哪个. 最大时部与Bayes同

复杂度 $O(nd)$

各种距离: P153

Evaluation 补:

① 列表算, 每种可能都有一项

② 后处理是求和, t时刻所有状态概率和为序列概率

先写完整公式, 再代入计算! (有分.)

	E	σ:	
0-1	p	p(1-p)	→ Bernoulli(单次实验)
二项	np	np(1-p)	
均匀	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	
正态	μ	6^2	

☆答题先抄算法,正则化,写样本 包括 until

⇒ 还是非参数化来分类.

Linear Discriminant Function

$g(x) = w^T x + w_{i0}$ 本质是超平面.

$= a^T y \rightarrow \binom{1}{x}$ y是样本, $a = \binom{w_{i0}}{w}$

Solution Region 是半空间交集

→ 考试时都normalize好了(也可取负).

⇒ Normalize: $\boxed{a^T y_i > 0}$ for all y_i ⇒ 表示分类正确 无论正负类 (负类y被反转)

定义一个半空间,交集为solution

⇒ 空则线性不可分.

$x \rightarrow (1, x, x^2)$, 以x为自变量

⇒ Method: 1) map 到高维(线性可分) 对于 f(x) 为线性

2) 训练求解 LDF ⇒ Alg. GD/Newton

☆最优η(二阶Taylor) P185.

GD: $a = a - \eta \nabla J(a)$

until $\eta \nabla J(a) < \varepsilon$ (convergence)

Newton: $a = a - H^{-1} \nabla J(a)$ → 二阶导 $\frac{\partial J}{\partial a_i \partial a_j}$ ⇒ $O(d^3)$ 求逆

until convergence 收敛更快

建议列表做

⇒ Criterion Function ☆初始化 k=0

① misclassified number: $\sum_{i=1}^{n} |a^t y_i \le 0$. ($\nabla J = 0$, 不可GD)

⇓

② $J(a) = \sum_{y \in Y} (-a^t y)$, $Y = \{y_i \mid a^t y_i \le 0\}$ ⇒ 可GD ↳ 错分类集合.

$\nabla J(a) = \sum_{y \in Y} -y$ ☆=0 的也算分错了!

③ Batch Perceptron Alg. ⇒ 用一批错误样本来分类

$k = k+1$

$a = a_i + \eta(k) \cdot \sum_{y \in Y_k} y$ → a_k 错分类集合.

until $\|\eta(k) \sum_{y \in Y_k} y\| < \varepsilon$ (convergence) ⇒ 全对则 $\|\eta(k) \sum y\| = 0$

④ Single - Sample Correction ⇒ 有错就立刻更新一次

$k = k+1$; $y^k = y_i$; $a = a + \eta y^k$ → 对 fixed increment Algs, $\eta = 1$

until 遍历一轮后 k 不变 (没出错)

⑤ Others

$J(a) = \sum_{y \in Y} (a^T y)^2$ ⇒ $\nabla J(a) = 2 \sum_{y \in Y} a^T y\, y$

$J(a) = \frac{1}{2} \sum_{y \in Y} \frac{(a^T y - b)^2}{\|y\|^2}$ ⇒ $\nabla J(a) = \sum_{y \in Y} \frac{(a^T y - b) y}{\|y\|^2}$

补⑥: variable - Increment Perceptron with Margin

if $a^t y_i \le b$ margin

$k = k+1$, $y^k = y_i$, $a = a + \eta(k) \cdot y^k$ ⇒ 按Alg来, k定lr,

until n 次、一轮后 k 不变.

记得要Normalize b也会反转!

⑥ Widrow-Hoff / LSE 算法

⇒ 从 $a^t y_i > 0$ 到 $a^t y_i = b_i$

⇓

矩阵表示 $Ya = b$ (b是label)

⇒ $J(a) = \|Ya - b\|_2^2$ ⇒ 使用整个数据集.

MSE ← $= \sum_{i=1}^{n} (a^T y_i - b_i)^2$

$\nabla J(a) = 2 \sum_{i=1}^{n} (a^T y_i - b_i) y_i$

$= 2 (Ya - b)^T Y$

$= 2 Y^T (Ya - b)$

① Closed Form

令 $\nabla J(a) = 0$ ⇒ 要求 $Y^T Y$ 可逆.

⇒ $a = (Y^T Y)^{-1} Y^T b$

② GD ($Y^T Y$ 随意) Widrow-Hoff (LSE)

$k = k+1$

$a = a + \eta(k) (b_k - a^t y^k) y^k$ ⇒ 对每个样本

有轮check.

until $|\eta \cdot (b_k - a^t y^k) y^k| < \theta$.

注意会schedule η.

⇒ 随轮次 ↓.

b也会一起 normalize

⇒ 多分类问题

$a \rightarrow a_i \ (i = 1 \cdots c)$

w_i if $a_i^t y > a_j y$

分类得分最高.

⇓

☆Kesler's Construction (P217)

⇒ 利用 $\boxed{a_i^T y - a_j^T y > 0}$

转化为二分类问题.

构造样本:

每个样本构造 c-1 个

c × d+1 维.

$y_{i1}^k = \begin{bmatrix} y_k \\ -y_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, $y_{i2}^k = \begin{bmatrix} y_k \\ 0 \\ -y_k \\ \vdots \\ 0 \end{bmatrix}$

$a^T y_{i1}^k > 0$

⇒ w_i 分 > w_1 分

⇒ 若 $a^T y_{ij}^k > 0$ (for all j)

⇒ w_i 为分类结果

降维技术 (P94)

① PCA

⇒ 流程:

1) $m = \frac{1}{n} \sum_{k=1}^{n} x_k$

$S = \sum_{k=1}^{n} (x_k - m)(x_k - m)^T$ ↗ 散布矩阵

☆求特征根 / eigen vector, 取前 d' 个大的 value 和对应 d' 个标准化的 vector

3) $W = [V_1, V_2 \cdots V_{d'}]$ ($d \times d'$)

4) $x_k' = W^T (x_k - m)$ 别忘 -m.

☆计算量大,数且,一定按步骤来,别急.

③ LDA / Fisher FDA

流程:

1) $m = \frac{1}{n} \sum_{x \in D} x$, $m_i = \frac{1}{n_i} \sum_{x \in D_i} x$ 类中心

2) $S_B = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^T$ (类间) → 权重

$S_W = \sum_{i=1}^{c} (\sum_{x \in D_i} (x - m_i)(x - m_i)^T)$ (类内)

3) 选 n 个 非0 eigen value of $\boxed{S_W^{-1} S_B}$

对应标准化 vector $[w_1 \cdots w_n]$

☆ $n \ge c-1$, 至多 c-1 个非0 eigen value.

4) $W = [w_1 \cdots w_{c-1}]$

$x' = W^T x$.

MLP NN

forward ⇒ 逐层计算

backward ⇒ 链式法则

$\frac{1}{2} \sum_{k=1}^{c} (t_k - o_k)^2$ = $\frac{1}{2} \|label - output\|^2$ = Loss 用 $\frac{1}{2}$MSE.

① Stochastic training (P239)

⇒ 随机选一个 x^m, 来BP

$w \leftarrow w + \eta \cdot \nabla J(w)$ → x^m 算 until $\|\nabla J(w)\| < \theta$

② Batch BP (P240)

⇒ 全部梯度累计, 再更新一次

$\Delta w = 0$

for $x \in D$

$\Delta w = \Delta w + \nabla J(w)$ → x 算.

$w = w + \Delta w$ until ... ~~两次~~ $\|\nabla J(w)\| < \theta$

概率公式: 导数公式在书上. E(x), D(x)

① $P(A_1 A_2 \cdots A_n) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 A_2) \cdots P(A_n | A_1 \cdots A_{n-1})$

② $P(B) = \sum_{i=1}^{n} P(A_i) \cdot P(B | A_i)$ 可算 P(x).

③ $P(A|B) = \frac{P(AB)}{P(B)}$

统计公式:

① $E[f(x)] = \sum_{i=1}^{c} f(x) \cdot P(x_i) \Big/ \int_{-\infty}^{+\infty} f(x) p(x) dx$

② $Var[x] = E[(x - Ex)^2]$

$= \sum_{i=1}^{c} (x_i - \mu)^2 \cdot P(x_i) \Big/ \int_{-\infty}^{+\infty} (x - \mu)^2 p(x) dx$.

矩阵: 行列式公式在书.

① 求逆 { 待定系数 ⇒ 2×2 快, 再大别; ☆增广矩阵 + 初等行变换 $[A, I] \rightarrow [I, A^{-1}]$ }

② 求 Eigen value & vector

$Ax = \lambda x$ ⇒ $(A - \lambda E) x = 0$

⇒ $|A - \lambda E| = 0$

↓

列λ方程, 解出λ → value

↓

代回原式求 x → vector ↳ 待定系数解方程, x1 随便给一个值1, 之后标准化.

↓

标准化

☆验算!