S Ch2. 统计判别

1. 通过对被识别对象进行多次观察&测量，抽取特征构成特征向量，并将其作为判别规则的输入
⇒ 对样本进行分类.

2. 在现实生活中，只有通过大量观察，其结果才能具有规律性.

3. 基于上述情况，我们抽取出的特征向量实际是随机向量.

一. 利用 Bayes 判别进行分类

欲确定样本 x 属于 ω_1 类 or ω_2 类，则要看它属于 ω_1 的概率大还是属于 ω_2 的概率大.

即: if: $P(\omega_1|x) > P(\omega_2|x)$ ① 则 $x \in \omega_1$

反之，$x \in \omega_2$

而由 Bayes 公式，后验概率 $P(\omega_i|x) = \frac{P(x|\omega_i)P(\omega_i)}{P(x)}$

$= \frac{P(x|\omega_i)P(\omega_i)}{\sum_{i=1}^{2} P(x|\omega_i)P(\omega_i)}$ ②

将 ② 代入 ①，有: 若 $P(x|\omega_1)\cdot P(\omega_1) > P(x|\omega_2)\cdot P(\omega_2)$ ③
则 $x \in \omega_1$；反之，$x \in \omega_2$

再将 ③ 变形:

若 $l_{12}(x) = \frac{P(x|\omega_1)}{P(x|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)} = \theta_{21}$ ← 似然阈值

似然比

则 $x \in \omega_1$；反之 $x \in \omega_2$.

二. Bayes 最小风险判别

Bayes 决策过程实际上是一个总风险的优化过程

我们选择条件风险最小化来使得期望损失最小化.

① M类分类问题的条件平均风险

$$r_j(x) = \sum_{i=1}^{M} L_{ij} P(w_i|x) \quad (*)$$

L_{ij} 是本应属于 w_i 类却判成了 w_j 类的somethings价

$$L_{ij} = \begin{cases} 0, & i=j \\ 1, & i \neq j \end{cases}$$

若对每一个模式 x, 都计算出全部类别的平均风险值 $r_1(x), r_2(x), \cdots r_M(x)$, 并且将 x 指定为具有最小风险的那一类, 则这种分类器称之为最小平均风险类器, 它基于: $r_j(x) = \frac{1}{p(x)} \sum_{i=1}^{M} L_{ij} p(x|w_i) \cdot P(w_i)$

(由 * 经 Bayes 公式而来), 可简化为:

$$r_j(x) = \sum_{i=1}^{M} L_{ij} p(x|w_i) \cdot P(w_i)$$

这种分类器是按平均条件风险作为标准判别

三. Normal Distribution 下的 Bayes 分类器

① M种模式下的判别函数

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

将概率密度代入 $\ln P(w_i) - \frac{1}{2}\ln|C_i| - \frac{1}{2}(x-m_i)^T C_i^{-1} (x-m_i)$

其中 $m_i = E_i(x)$ $C_i = E_i[(x-m_i)(x-m_i)^T]$ $i=1,2,\cdots M$

$d(x)$ 可进一步写为 $d(x)=\vec{W}^T\vec{x}$

$\vec{x}=(x_1,x_2,\cdots,x_n,1)$ 叫增广模式向量

$\vec{w}=(w_1,w_2,\cdots,w_n,w_{n+1})$ 叫增广权向量

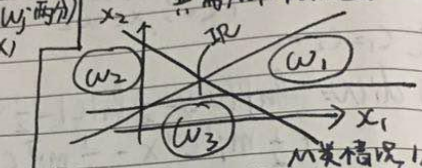
① 两类情况：

$d(x)=\vec{W}^T\vec{x}$ $\begin{cases}<0, & x\in\omega_1\\ >0, & x\in\omega_2\end{cases}$

② M类情况1：$(\omega_i/\overline{\omega_i}$ 两分法)

用 $d(x)$ 将 $\in\omega_i$ 和 $\notin\omega_i$ 的分开 $d_i(x)=\begin{cases}>0, & x\in\omega_i\\ <0, & x\notin\omega_i\end{cases}$ $i=1,2,\cdots,M$

共需M个判别函数



M类情况1条件过宽

缺点：若对某一区域，$d_i(x)>0$ 的条件超过1个，则分类失效。该区域称为不确定区域(IR)

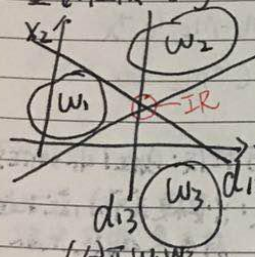
③ M类情况2(ω_i/ω_j 两分法)

采用两对划分，每个 $d(x)$ 只能分开两个界面

$d_{ij}(x)=\vec{W}_{ij}^T\vec{x}$

若 $d_{ij}(x)>0$，$\forall i\neq j$，则 $x\in\omega_i$

重要性质：$d_{ij}=-d_{ji}$



d_{23} (分开 ω_2,ω_3 不管 ω_1)

d_{12} (分开1,2 不管 ω_3)

d_{13} (分开 ω_1,ω_3 不管 ω_2)

共需 $C_M^2=\frac{M(M-1)}{2}$ 个判别函数

若 $\forall d_{ij}(x)$，找不到 $\forall i\neq j$，$d_{ij}(x)>0$

则称为 IR.

④ M类情况3(情况2的特例)：没有IR的情况

把M类问题分成 M-1 个两类问题，共需M个判别函数

把情况2中 $d_{ij}(x)$ 报成 $d_{ij}(x)=d_i(x)-d_j(x)=(w_i-w_j)^T x$ 则 $d_{ij}(x)>0$ 相当于 $d_i(x)>d_j(x)$，$\forall j\neq i$

<2> 总结

若模式可用线性判别函数分开，则称之为线性可分，反之为线性不可分

GDA中 我们 ...
X离散，服从多项分布. 以一个 ... $dim(x) = 5 ...$
$x = (1\ 0\ 0\ \cdots\ 1\ \cdots\ 0)$ 建模，则有 2^{5w} 个不同值. ...
考虑接对多项分布 $P(X|y)$ 建模，则有 ... 则 $p(x_1, \cdots x_{5w}|y) = p(x_1|y) \cdot p(x_2|x_1 \ldots$
在NB中我们假设X的每一维都是独立的. 则 $p(x_1, \ldots$
$\cdots p(x_{5w}|y, x_1, \cdots x_{49999}) = \prod_{j=1}^{n} p(x_j|y)$. 则 对似然 ...
$$l = \sum_{i=1}^{m}[y^{(i)} \log \phi_y + (1-y^{(i)}) \log(1-\phi_y) + \sum_{j=1}^{n}(x_j^{(i)} \log \phi_{j|y^{(i)}} + (1-x_j^{(i)}) \log(1-\phi_{j|y^{(i)}}))]$$

五. Bayes 模型

得 $\phi_{j|y=0} = \sum_{i=1}^{m} 1\{x_j^{(i)} = k \wedge y^{(i)} = 0\} / \sum_{i=1}^{m} 1\{y^{(i)} = 0\}$
即所有 y=0 的样本中有单词 x_j 的邮件数除以 y=0 样...
同理得 $\phi_{j|x=1}$ (后续见开头)

background: 估计. (参数?) 通过训练样本做出
预测 (未知样本的输出?)

Bayes 决策分类: $choose = \begin{cases} 1, & if\ P(C=1|x) > P(C=0|x) \\ 0, & else \end{cases}$

Bayes 公式: $P(C|x) = \dfrac{P(x|C) \cdot P(C)}{P(x)} \propto P(x|C)P(C)$

loss: λ_{ik}: 本类属于 k, 却分到 i 的风险损失

α_i: 分到 i

采取 α_i 的风险 $R(\alpha_i|x) = \sum_{k=1}^{K} \lambda_{ik} P(C_k|x)$

Choose α_i, if $R(\alpha_i|x) = \min_k R(\alpha_k|x)$

0-1 Loss: $\lambda_{ik} = \begin{cases} 0, & i=k \\ 1, & i \neq k \end{cases}$

则 $R(\alpha_i|x) = \sum_{k=1}^{K} \lambda_{ik} P(C_k|x) = \sum_{k \neq i} P(C_k|x) = 1 - P(C_i|x)$

Choose α_i if $P(C_i|x) = \max P(C|x)$

分类的一个方法是用判别函数 $g(x)$

Choose C_i if $g_i(x) = \max_k g_k(x)$

例, $g_i(x)$ 可取 $-R(C_i|x)$
$P(C_i|x)$
$P(x|C_i)P(C_i)$

对于2类判别问题，我们可令 $g(x)=g_1(x)-g_2(x)$

即 h，$\text{otherwise}=\begin{cases}c_1, & \text{if } g(x)>0\\ c_2, & \text{else}\end{cases}$

Bayesian Networks 贝叶斯网：其实就是一个图 ← Node：随机变量；Edge：2个变量间的关系

Causal graph & Diagnostic inference

已知原因推结果　　已知结果推原因

例：T.雨　喷k　P(R)=a4　P(S)=a2

已知草湿求喷水概率：

$P(S|W)$

$=\cdots$

$=0.35$

因此，这是一个 diagnostic inference

$P(W|R,S)=0.95$

$P(W|R,\bar{S})=0.9$

$P(W|\bar{R},S)=0.9$

$P(W|\bar{R},\bar{S})=0.1$

下雨和喷水是草湿的原因，这是一个 causal graph

explaining away：

$0.21=P(S|R,W)$

$<P(S|W)=0.35$

在多因一果的条件，若已知其中某个原因导致的概率，对其他原因导致的概率会下降

$P(C,S,R,W,T)=P(C)P(S|C)P(R|C)P(W|S,R)P(T|R)$

因果网络分类其实就是 Diagnostic inference

一般地 $P(x_1\cdots x_d)=\prod_{i=1}^{d}P(x_i|\text{parents}(x_i))$

Naive Bayes' classifier

$P(x_1|C)$　$P(x_2|C)$　$P(x_3|C)$

$P(\vec{x}|C)=\prod_{j=1}^{d}P(x_j|C)$，$x_j$ 相互独立

$P(C|\vec{x})=\frac{P(\vec{x}|C)P(C)}{P(\vec{x})}$

NB分类器不考虑输入变量 x_j 之间可能存在的依赖性

最大似然估计 (MLE)：寻找 θ，使得 $P(X|\theta)$ 中的样本 $X^{(i)}$ 最有可能出现

对于样本 $X=\{x^{(i)}\}_{i=1}^{N}$，通过样本 X 求出 θ，即为估计的参数 θ（即已知出现了样本 X，参数 θ 为多少时 X 出现的可能最大？此时的 θ 就被我们当成真实的 θ）

似然函数 $L(\theta|X)=P(X|\theta)=\prod_{i=1}^{N}P(x^{(i)}|\theta)$

$\mathcal{L}(\theta|X)=\log L(\theta|X)=\sum_{i=1}^{N}\log P(x^{(i)}|\theta)$

$\hat{\theta}=\arg\max_{\theta}\mathcal{L}(\theta|X)$

贝叶斯估计.

MLE 认为 θ 是固定值, 而 Bayes 认为 θ 有先验分布 $p(\theta)$

给定训练集 $p(\theta|\chi) = \frac{p(\chi|\theta)p(\theta)}{p(\chi)} = \frac{p(\chi|\theta)p(\theta)}{\int p(\chi|\theta')p(\theta')d\theta'}$

Full Bayesian approach: Estimation of density at x

$p(x|\chi) = \int p(x|\theta,\chi)p(\theta|\chi)d\theta = \int p(x|\theta)p(\theta|\chi)d\theta$

prediction:

$y = \int f(x|\theta)p(\theta|\chi)d\theta$

积分难算, 转为:

Maximum a posteriori (MAP): 最大后验估计

$\theta_{map} = \arg\max_\theta p(\theta|\chi)$ (使最后验概率最大的估计)

MLE: 最大似然估计:

$\theta_{ML} = \arg\max_\theta p(\chi|\theta)$

Bayes' 贝叶斯估计:

$\theta_{Bayes} = E[\theta|\chi] = \int \theta p(\theta|\chi)d\theta$

[例] 设 $x^{(i)} \sim N(\mu, \sigma^2)$ 而 $\mu \sim N(\mu_0, \sigma_0^2)$

则似然 $p(\chi|\mu) = \prod_{t=1}^{N} p(x^{(t)}|\mu) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{-\frac{\sum_t (x^{(t)}-\mu)^2}{2\sigma^2}}$

$p(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}$

可证 $p(\mu|\chi) \propto p(\chi|\mu)p(\mu)$

正态指数: $p(\mu|\chi) = -\frac{1}{2\sigma^2}(\mu-\mu_N)^2$

$= -\frac{1}{2\sigma^2}(\mu^2 - 2\mu_N\mu + \mu_N^2)$

$= -\frac{1}{2\sigma^2}\mu^2 + \frac{\mu_N}{\sigma^2}\mu + Const$

$= -\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\right) + \mu\left(\frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2}\sum x^{(i)}\right) + (Const)$

$\therefore \mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML} \to \frac{1}{N}\sum_{t=1}^{N} x^{(t)}$

$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$

支持向量机 (SVM)

① 硬边界 SVM

定义 hyper plane

再定义两条线

设这两个点分

则: $W^T x$

$\therefore W$

要最大化 margin

约束条件: W

转为优化:

这是一个二

拉格朗日

$Lp(W, W_0$

$= \frac{1}{2} W^T W$

$\frac{\partial L_p}{\partial W} = 0$

$\frac{\partial L_p}{\partial W_0} = 0$

$Ld(\{\alpha_i$