贝叶斯决策

决策理论

- 贝叶斯公式： $P(\omega_i|\vec{x}) = \frac{p(\vec{x}|\omega_i)P(\omega_i)}{\sum p(\vec{x}|\omega_j)P(\omega_j)}$
- 联合概率分布和条件概率分布：

$$P(B|A)P(A) = P(B,A)$$

- 最小错误率方法：

$$\arg\max_i P(\omega_i|\vec{x}) = \arg\max_i P(\vec{x}|\omega_i)P(\omega_i)$$

- 最小错误率分析： $P(e) = P(\omega_2)P_2(e) + P(\omega_1)P_1(e)$, $P_2(e)$是把2分类错误成1的概率
- 多类正确率分析： $P(c) = \sum P_i(c)P(\omega_j)$ 对应最高的曲线加先验的面积
- False postive: 将负样本判成正样本。False negative：将正样本判成负样本
- ROC曲线：横轴 False Postive，纵轴 True Positive
- 最小风险贝叶斯决策： $\lambda(\alpha_i,\omega_j)$是在$\omega_j$状态上采用$\alpha_i$决策的代价。(选对了可以认为是0)。 决策$\arg\min\sum_j \lambda(\alpha_i,\omega_j)p(\vec{x}|\omega_j)p(\omega_j)$
- 限定一类错误率，求第二类错误率最小值：拉格朗日待定系数法 $\gamma = P_1(e) + \lambda(P_2(e) - \epsilon_0) = 1 - P_1(c) + \lambda P_2(e) - \lambda\epsilon_0$。对$\lambda$求偏导，一维情况得到决策面满足$\lambda = \frac{p(t|\omega_1)}{p(t|\omega_2)}$，决策面限定错误率达到指标上限。(图解：看残余面积)。
- 最小最大决策： 风险已知， 先验概率未知。 $R = \int_{\mathcal{R}_1} \lambda_{11}p(\omega_1)P(\vec{x}|\omega_1) + \lambda_{12}p(\omega_2)P(\vec{x}|\omega_2) + \int_{\mathcal{R}_2} \lambda_{21}p(\omega_1)P(\vec{x}|\omega_1) + \lambda_{22}p(\omega_2)P(\vec{x}|\omega_2) = a + p(\omega_1)b$风险不可确定，以$b=0$作为最坏情况的最优解（防止$P(\omega_2)$产生变化
- 多类决策：可以进行逐一比较。或者算最大的$P(\omega_i|\vec{x}) > P(\omega_j|\vec{x}) \Rightarrow p(\vec{x}|\omega_i)p(\omega_i) > p(\vec{x}|\omega_j)p(\omega_j) \Rightarrow p(\vec{x}|\omega_i)p(\vec{x}|\omega_j) > p(\omega_j)p(\omega_i) \Rightarrow \ln p(\vec{x}|\omega_i) + \ln p(\omega_i) > \ln p(\vec{x}|\omega_j) \ln p(\omega_j)$（对数似然）
- 决策面以及两类情况的决策面：$g(\vec{x}) \begin{cases} > 0 \Rightarrow \in \omega_1 \\ < 0 \Rightarrow \in \omega_2 \end{cases}$ $g(\vec{x}) =$0是决策面

朴素贝叶斯决策和正态分布决策

- 朴素贝叶斯：认为所有变量独立：$P(\vec{x}|\omega) = \prod_i p(x_i|\omega)$
- 正态分布基本性质：$P(\vec{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(\vec{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\vec{x}-\mu)$
- 性质：等密度点是超椭球面$(\vec{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\vec{x}-\mu) = const$，马氏距离，不相关性$E_{12} = E_1E_2 ==$ 独立性$P_{12} = P_1P_2$(仅在正态分布下)
- 边缘分布和条件分布，线性变化正态性。$\vec{y} = \mathbf{A}\vec{x} \rightarrow N(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^{\mathbf{A}^{\mathrm{T}}})$
- $\Sigma_i = \sigma^2\mathbf{I}$（协方差相等均对角），分类面为圆形（先验不相等）或中位（先验相等）
- $\Sigma_i = \Sigma$（协方差相等），分类面为椭圆形（先验不相等）或直线（先验相等，相对中位线发生偏转）
- 其他可能性：二次曲线。计算错误率：公式，近似计算上界，实验估计

概率密度函数的估计

参数和非参数估计

- 参数空间，点估计，区间估计，统计量等概念。MLE方法$\max\sum_i \ln(p(\vec{x}_i|\theta))$（有的时候不需要求导，
- 正态分布的MLE估计方法：$\mu = \frac{1}{N}\sum_i \vec{x}_i, \Sigma = \frac{1}{N}\sum(\vec{x}-\mu)(\vec{x}-\mu)^{\mathrm{T}}$
- 可识别性：参数空间内每两个不同参数均有不同的概率密度函数，连续的往往是可以识别的，离散的往往不能识别
- 贝叶斯参数估计： 将参数认为成随机变量， 并有先验。$P(\theta|\vec{x}) = \frac{P(\vec{x}|\theta)p(\theta)}{p(\vec{x})}$
- 正态分布MAP： $\vec{x} \sim N(\mu,\sigma^2), p(\mu) \sim N(\mu_0,\sigma_0^2)$ （均值有不确定性） $P(\mu|\vec{x}) = \frac{1}{p(\vec{x})}\prod\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}\frac{\vec{x}_k-\mu}{\sigma}^2\frac{1}{\sqrt{2\pi}\sigma_0}\exp(-\frac{1}{2}\frac{\vec{m}u-\mu_0}{\sigma_0}^2) \sim N(\mu_N,\sigma_N), \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}m_N + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$。

结果仍然是正态分布。$N=0 \rightarrow$仅能靠先验，没有实验信息。$N \rightarrow \infty$实验信息足够多，先验信息没有用。$\sigma_0 \rightarrow 0$先验太强，实验信息被忽视。$\sigma_0 \rightarrow \infty$先验信息太弱，被忽视。
当先验可靠时， 利用更多的信息。 目的是$\max P(\vec{x}|\theta)p(\theta)$。同等情况下MLE更简单

- 最小二乘等价于对认为误差是正态分布的MLE方法，此方法上的加参数先验为正态分布的MAP方法就是L-2正则化
- Parzen窗法：窗口大小的选择，正态分布Parzen窗
- 三类误差：贝叶斯误差：特征一旦选定就无法改变的误差。模型误差：模型不标准或者错误。估计误差：参数估计的时候产生的误差。维数问题：普遍x100个样本一维。依靠PCA，独立性等方式降低维数灾难
- 过拟合： MLE的普遍特性， 考虑增加样本或者引入MAP方案， 通过简单模型 （参数化模型， 对角矩阵， 共享参数）降低参数
- 错误率估计：先验未知$\epsilon = \frac{k}{N}, E(k) = N\epsilon, Var(k) = N\epsilon(1-\epsilon), Var(\hat{\epsilon}) = \frac{\epsilon(1-\epsilon)}{N}$。先验已知：$E(\hat{\epsilon}) = \epsilon, Var(\hat{\epsilon}) = \frac{1}{N}\sum P(\omega_i)\epsilon_i(1-\epsilon_i)$。已知先验方差更低。交叉验证和留一法等

EM, GMM 算法

- GMM算法：$P(X|\Theta) = \sum\alpha_i p_i(X|\theta_i), \sum\alpha_i = 1, p_i(X|\theta_i) \sim N(\mu_i,\sigma_i)$log-MLE : $\sum_i \ln\sum_j N(x_i|\mu_j,\Sigma_j)P(\omega_j)$
- 已知概率是多类高斯混合而成的，每一类的比例（先验）未知，每一类高斯的参数未知
- $P(\omega_k|x_i,\mu_k,\Sigma_k) = \frac{N(x_i|\mu_k,\Sigma_k)P(\omega_k)}{\sum_j N(x_i|\mu_j,\Sigma_j)P(\omega_j)}$
- $\hat{P}(\omega_k|x_i,\mu_k,\Sigma_k) = \frac{1}{N}\sum_i P(\omega_k|x_i,\mu_k,\Sigma_k)$
- 对μ_k求偏导为零得到$\hat{\mu}_k = \frac{\sum_i P(\omega_k|x_i,\mu_k,\Sigma_k)x_i}{\sum_i P(\omega_k|x_i,\mu_k,\Sigma_k)}$
- 对Σ_k：$\hat{\Sigma}_k = \frac{\sum_i P(\omega_k|x_i,\mu_k,\Sigma_k)(x_i-\mu_k)(x_i-\mu_k)^{\mathrm{T}}}{\sum_i P(\omega_k|x_i,\mu_k,\Sigma_k)}$
- 解决策略：E步利用上一步的东西计算$P(\omega_k|x_i,\mu_k,\Sigma_k)$，M步计算参数迭代
- EM 算法：解决数据缺失或隐变量的问题X显变量，Y隐变量。$P(X,Y|\theta) = p(Y|X,\theta)p(X|\theta)$
- $L(\theta) = \ln p(X|\theta) = \ln(\sum_Y p(X,Y|\theta))$，设$Y \sim q(Y)$

线性判别函数

- 线性判别，广义线性判别，增广向量等
- 感知准则函数：假设样本线性可分，$\begin{cases} a^{\mathrm{T}}y > 0 \text{对一切} y_i \in \omega_1 \\ a^{\mathrm{T}}y < 0 \text{对一切} y_i \in \omega_2 \end{cases}$

采用$y_i' = \begin{cases} y_i \\ -y_i \end{cases}$ 得到$\forall i, a^{\mathrm{T}}y_i' > 0$，错分样本集合设为$Y^k$

- 目标函数$J_P(a) = \sum_{y\in Y^k} -a^Ty \geq 0$，希望优化到0。梯度下降得到$ak+1 = a_k + \rho\sum_{y\in Y^k} y$
- 1-bit SGD, $\rho = 1$ 可以达到最小值（可以硬证明）

SVM

- 优化问题：$\min\frac{1}{2}w^{\mathrm{T}}w$, subject to $d_i(w^{\mathrm{T}}x_i + b) > 1$，凸优化问题加拉格朗日乘子转对偶方法优化
- $J = \frac{1}{2}w^{\mathrm{T}}w - \sum\alpha_i[d_i(w^{\mathrm{T}}x_i + b) - 1], \alpha_i > 0$ 转优化$J(\vec{W},\vec{\alpha})$。
- 鞍点是最值点：$J(\vec{W}',\alpha) \leq J(\vec{W}',\alpha') \leq J(\vec{W},\alpha')$，左侧得到$\sum(\alpha_i' - \alpha)g_i(\vec{w}') \leq 0 \Rightarrow \sum g_i(\vec{w}') > 0$说明是可行点，取$\alpha_i = 0$ 得到$\alpha_i'g_i(\vec{w}') = 0$
- 右侧：$f(\vec{w}') \leq f(\vec{w}) - \alpha'g(\vec{w}) \leq f(\vec{w})$得到最优点（强对偶条件）
- 对偶求鞍点：$w = \sum\alpha_i d_i x_i, \sum alpha_i d_i = 0$
- 对偶方法： $\max Q(\alpha) = \sum\alpha_i - \frac{1}{2}\sum\sum\alpha_i\alpha_j d_i d_j H_{ij}, H_{ij} = x_i^{\mathrm{T}}x_j$，之后可以更换为核函数。
- 线性不可分松弛：$\min\frac{1}{2}w^{\mathrm{T}}w + C\sum\xi$。松弛$d_i(w^{\mathrm{T}}x_i + b) > 1 - \xi_i$，C越大，要求性能越好，间隔越小，C越小要求间隔越大。
- 非线性核函数方法，多项式$K(x,y) = (xy+1)^p$，RBF：$\exp(-\|x-y\|^2/2\sigma^2)$等、
- 采用L-1正则化和L-2正则化的函数的SVM， L-1稀疏学习，L-2防过拟合
- 产生式模型（优点：可以提供大量信息量）和判别式模型（优点：简单）

神经网络

- 损失函数：TSSE：$\frac{1}{2}(d_i - y_i)^2$，可以进一步补充为对整个batch的求导
- BP算法，矩阵求导，反向传播，链式法则等
- CNN，RNN，LSTM的结构（有空补上）。序列长度：一个sample的长度（sample之间没有关系）

决策树

- 交叉熵：$E = -\sum p_i \ln p_i$
- Gini 不纯度：$1 - \sum P(w_n)^2$
- 误差不纯度：$1 - \max_j P(\omega_j)$
- 信息熵的增量：$l - p(l)l_l - p(r)l_r$（左右两支加权计算）

近邻法和距离

- 最近邻法错误率在1$\tilde{2}$倍贝叶斯错误率中间
- 压缩近邻法：首先用近邻法测试所有测试样本，如果测错了放进评测集，通过这种方式扩充评测集
- 距离度量：1. s-Minkowski: $D(x,y) = [\sum|x_i - y_i|^s]^{1/s}$。$s = 2$为欧几里得距离。Chebyshev 距离：$\max_i|x_i - y_i|$，马氏距离$(x-y)^{\mathrm{T}}Q(x-y)$
- 距离的正定性，三角性，Holder不等式等（不会）
- KL 散度作为概率PDF之间距离，切距离：做流形变换之后点到直线的距离（单边切距离，双边切距离）

特征提取和特征选择

PCA, KL变换等线性方法

- 目的：防止维数灾难，方便可视化和理解
- Fisher 方法：找一个方向使得投影结束后类间方差最大，类间方差最小。类间方差$(m_1-m_2)(m_1-m_2)^{\mathrm{T}}$。类内方差$\sum_j\sum_i(x_{ij}-m_j)(x_{ij}-m_j)^{\mathrm{T}}$
- 投影之后方差变化做比$J(w) = \frac{w^{\mathrm{T}}S_bw}{w^{\mathrm{T}}S_ww}$最大转拉格朗日乘子：$L = w^{\mathrm{T}}S_bw - \lambda(w^{\mathrm{T}}S_ww - c)$（控制一个类内方差不变）得到$w$是$S_w^{-1}S_b$的最大特征值对应的特征向量。
- Fisher投影之后的分类问题：按照两类重心的重点、或者按照加权重心或者$\frac{m_1+m_2}{2} + \frac{\ln P(w_1)-\ln P(w_2)}{N_1+N_2-2}$先验。也可以采用Bayes决策手段
- 其他的Fisher手段，局部方法，非线性方法等，选取其他判据方案等（如矩阵迹）
- 多维度Fisher方法：选取最大的几个向量
- PCA： 最小化剩余方差$\min E(x-\hat{x})^{\mathrm{T}}(x-\hat{x}) = E[\sum_{i=d+1}^{\infty}c^2]$
- PCA：找S（对称正交矩阵）的从大到小的特征值，对应的特征向量组成变换矩阵。

特征选择方案

- 向前向后方法，顺序前进法，顺序后退法（找收益最大的和损失最小的）
- Relief方案，遗传算法等，存在问题：特征选择过学习

MDS, LLE, ISOMap，遗传算法

- MDS方案：给定两点之间的距离， 找p维的x使得$d_{rs}^2 = (x_r-x_s)^{\mathrm{T}}(x_r-x_s) = x_r^{\mathrm{T}}x_r + x_s^{\mathrm{T}}x_s - 2x_r^{\mathrm{T}}x_s$可以构造$B = XX^{\mathrm{T}}, n\times n$，进行$p$维PCA提出 $B = V\Lambda V^{\mathrm{T}}, X = V_1\Lambda^{\frac{1}{2}}$，展开成$P$维空间
- ISOMAP 方案：先通过kNN或者ϵ圆构建一个整个图的支撑树，之后采用Dijkstra算法找到图中任意两点的最近距离构建距离矩阵，之后使用MDS方法确定坐标
- LLE（局部线性化）：先通过kNN或者ϵ圆找到每一个点的近邻然后把每一个面元铺平
- GA：初始化种群，交叉和编码方式等

Ensemble

- Adaboost：采用一组弱学习器进行学习。方案：1. 初始化$w_1(i) = \frac{1}{N}$. 2. 归一化：$p_l(i) = \frac{w_l(i)}{\sum_i w_l(i)}$ 3. $\epsilon_l = \sum_{mis} p_l(i)$ 4. $a_l = \frac{1}{2}\ln\frac{1-\epsilon_l}{\epsilon_l}$ 5. 调整：$w_{l+1}(i) = w_l(i)\exp(\pm a_l)$正：错分样本加权重。负：正确样本减权重。5. 训练结束之后按a_l对L个分类器加权投票。
- 随机森林：随机抽取若干特征构造树， 按照准确率进行投票，投票悖论等

聚类分析，非监督学习方法

- GMM聚类：对样本进行GMM算法，每一个成分分成一类
- kmeans方案：对分割进行调整规则：$\frac{N_j}{N_j+1}\|y-m_j\|^2 < \frac{N_k}{N_k-1}\|y-m_k\|^2$将$y$从$k$类移到$j$类，复杂度$\mathcal{O}(n)$
- 分类个数的选取：选择肘点（二阶差分最大（一阶差分变化最大））
- 核函数方法，选择核函数代替欧几里得距离
- 多级聚类方案： 将两个最近/最远/平均距离最近的两类每次聚类，直到只剩下一类。形成聚类树。$\mathcal{O}(n^3)$复杂度
- 谱聚类：首先构造相似距离矩阵D（越远的距离越大，可以使用RBF核或者其他的ϵ圆或者kNN方案。将行相加构造对角矩阵D。$L = D - W, L_{rw} = D^{-1}L$构造非归一化和归一化的拉普拉斯矩阵。计算两个矩阵的前k个向量组成n行k列的矩阵，对此矩阵进行聚类。复杂度$\mathcal{O}(n^2)$

各种情况下的EM算法

一般情况下的EM算法

- X是显变量，Z是隐变量（一般来说是离散的）。联合分布$p(X,Z|\theta)$确定。目标是最大化$p(X|\theta)=\sum_Z p(X,Z|\theta)$
- 强行引入Z的分布$q(Z)$，立即得到$\ln(p(X|\theta))=\mathcal{L}(q,\theta)+KL(q\|p)$
- $\mathcal{L}(q,\theta)=\sum_Z q(Z)\ln\frac{p(X,Z|\theta)}{q(Z)}$
- $KL(q\|p)=-\sum_Z q(Z)\ln\frac{p(Z|X,\theta)}{q(Z)}$
- 以上根据$\ln P(X,Z|\theta)=\ln P(Z|X,\theta)+\ln P(X|\theta)$，$KL(q\|p)$是KL散度，当且仅当$q(Z)=p(Z|X,\theta)$（估计完全正确）时等于0否则大于零，注意$\sum_Z q(z)\ln p(X|\theta)=\ln(p(X|\theta))$（$\ln p(X|\theta)$与$Z$无关
- $\mathcal{L}(q,\theta)\le\ln(p(X|\theta))$，得到优化下界
- E-step：使用q最大化$\mathcal{L}(q,\theta)$，由于$\ln(p(X|\theta))$不变，故只能KL散度为0，$q(Z)=p(Z|X,\theta)$时$\mathcal{L}$达到最大化。
- M-step：使用θ优化$\mathcal{L}$，这将使得$p(Z|X,\theta)$导致非零的KL散度，给E-step留空间。
- 当总共有n个样本时，$P(Z|X,\theta)=\frac{P(X,Z|\theta)}{\sum_Z P(X,Z|\theta)}=\frac{\prod p(x_n,z_n|\theta)}{\sum_Z\prod p(x_n,z_n|\theta)}=\prod p(z_n|x_n,\theta)$。其中用到了交换求和求积顺序，同时需要求条件概率。

基于上述解释的EM算法

- E-step：计算$p(Z|X,\theta)$
- M-step：计算$\theta=\arg\max_\theta Q(\theta,\theta^{old})=\sum_Z p(Z|X,\theta^{old})\ln p(X,Z|\theta)$。这里忽略了上面的分母因为和$\theta$无关

EM算法用于有先验的贝叶斯参数估计，GEM算法

- $\ln P(\theta|X)=\ln P(X|\theta)+\ln P(\theta)-\ln P(X)$进行优化
- GEM 算法：对M-step进行梯度下降而不是一次到最优点

GMM算法的EM表示

- E-step: $\gamma(z_{nk})=\frac{\pi_k\mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_j\pi_j\mathcal{N}(x_n|\mu_j,\Sigma_j)}$
- M-step: $\begin{cases}\pi_k=\frac{1}{N}\sum\gamma(z_{nk}),N_k=\sum\gamma(z_{nk})\\ \mu_k=\frac{1}{N_k}\sum\gamma(z_{nk})x_n\\ \Sigma_k=\frac{1}{N_k}\sum\gamma(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^{\mathrm{T}}\end{cases}$
- 对数似然：$\sum\ln\sum\pi_k\mathcal{N}(x_n|\mu_k,\Sigma_k)$】
- 隐变量是二值的Z,Z_{nk}表征第n个样本是否在k的聚类中。$P(X,Z|\theta)=\prod_n\prod_k\pi_k^{z_{nk}}\mathcal{N}(x_n|\theta)^{z_{nk}}$
- $p(Z|X,\theta)=const\prod_n\prod_k\pi_k^{z_{nk}}\mathcal{N}(x_n|\theta)^{z_{nk}}$具有$X$在第k个聚类中的概率，可以用条件公式贝叶斯等
- M-step 必然得到GMM算法的结果

使用琴生不等式推出的EM算法

- $\ln P(x|\theta)=\ln\sum_Z P(x,z|\theta)=\ln\sum_z q(z)\frac{p(x,z|\theta)}{q(z)}\ge\sum_z q(z)\ln p(x,z|\theta)-\sum_z q(z)\ln q(z)=F(q,\theta)$
- E-step 仍然是对$F(q,\theta)$进行优化，得到结论仍是$q(Z)=p(Z|X,\theta)$，原因是此时处理之后$F(q,\theta)=\mathcal{L}(\theta)$。我们优化的是$q$
- M-step 优化$Q(\theta,\theta^{old})=\sum_z p(Z|X,\theta^{old})\ln p(x,z|\theta)$和前面的相同

GMM对应一个batch的样本

- $P(Z|X,\theta)=\prod p(z_n|x_n,\theta)$，混合模型下$p(x,z|\theta)=\sum alpha_{z_i}p_{z_i}(x|\theta)$。注意这里的表示方式和上面的二值表示不同，这里表示的是下标参数，因此$\sum_Z$需要展开成那个很复杂的式子。设$y_i=k$表示的是i样本在k类中，直接得到M步为$\sum_{y_1=1}^M\cdots\sum_{y_1=1}^M\sum_{i=1}^N ln(\alpha_{z_i}p_{z_i}(x|\theta_{z_i}))$。
- 这样做没有上面那么做好，二值的可以表示的更简洁。求解过程中注意区分$\theta,\theta^{old},p(z|x,\theta_old)$中不含我们需要的东西