模式识别笔记

陈鸿峥

2020.01*

目录

*Build 20200105

1 简介

机器学习侧重于处理的算法，而模式识别则包括了数据预处理、实际运算和数据输出的完整过程。

- 模式识别：涵盖的范围广，包括特征提取、特征选择、降维、各种分类器等。
- 机器学习：主要是讲学习，更多关于分类器如何训练模型，而不涉及特征方面的知识。

良好特征的四个特点：

- 可区别性（不同类）
- 可靠性（同类）
- 独立性（特征之间）
- 参数少（复杂性）

一个对象的所有特征参数组成特征向量。同样需要从高维测量空间（样本）中提取特征映射到低维特征空间。

模式识别分为两类

- 结构/句法模式识别
- 统计/神经网络模式识别

2 贝叶斯决策论

2.1 离散变量

处于类别ω_i并具有特征值x，有后验概率[1]

$$\mathbb{P}\left(\omega_i \mid x\right)=\frac{p(x \mid \omega_i)\mathbb{P}\left(\omega_i\right)}{p(x)}$$

即

$$posterior=\frac{likelihood \times prior}{evidence}$$

无论什么情况，当我们观察到特定的x，对于二分类问题有错误率

$$\mathbb{P}\left(error \mid x\right)=\begin{cases}\mathbb{P}\left(\omega_1 \mid x\right) & \text{决策}\omega_2 \\ \mathbb{P}\left(\omega_2 \mid x\right) & \text{决策}\omega_1\end{cases}=\min[\mathbb{P}\left(\omega_1 \mid x\right),\mathbb{P}\left(\omega_2 \mid x\right)]$$

平均错误概率可表示为

$$\mathbb{P}\left(error \mid x\right)=\int_{-\infty}^{\infty}\mathbb{P}\left(error, x\right)\,\mathrm{d}x=\int_{-\infty}^{\infty}\mathbb{P}\left(error \mid x\right)p(x)\,\mathrm{d}x$$

注意$p(x)$是证据，可以看为是固定分布（常量）。

[1]通常用$p(\cdot)$代表概率密度函数（连续变量），用$\mathbb{P}\left(\cdot\right)$代表概率质量函数（离散变量）

定理 1 (贝叶斯决策/最小错误率准则). 若$P(\omega_1 \mid x) > P(\omega_2 \mid x)$，则判定类别为$\omega_1$；否则判为$\omega_2$。依照这种准则可以获得最小错误率，即$P(error \mid x) = \min[P(\omega_1 \mid x), P(\omega_2 \mid x)]$

2.2 连续变量

考虑特征向量$\mathbf{x} \in \mathbb{R}^d$（$\mathbb{R}^d$称为特征空间），令$\{\omega_1, \ldots, \omega_c\}$表示有限的$c$个类别集，$\{\alpha_1, \ldots, \alpha_a\}$表示有限的$a$种可能采取的行为集，损失函数(loss)$\lambda(\alpha_t \mid \omega_j)$描述类别状态为$\omega_j$时采取行动$\alpha_i$的风险。 $p(\mathbf{x} \mid \omega_j)$表示在真实类别为$\omega_j$的条件下$\mathbf{x}$的概率密度函数，$P(\omega_j)$表示类别处于状态$\omega_j$时的先验概率，后验概率$P(\omega_j \mid \mathbf{x})$则通过贝叶斯公式

$$P(\omega_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_j)P(\omega_j)}{p(\mathbf{x})}$$

计算得到，证据变为

$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j)P(\omega_j)$$

与行动α_i相关联的风险(risk)为

$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j)\mathbb{P}\left(\omega_j \mid \mathbf{x}\right)$$

进而得到总损失

$$R = \int R(\alpha(\mathbf{x}) \mid \mathbf{x})p(\mathbf{x})\ \mathrm{d}\mathbf{x}$$

因此得到连续情形下的贝叶斯决策论：

定理 2. 为最小化R，计算条件概率

$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j)\mathbb{P}\left(\omega_j \mid \mathbf{x}\right),\ \forall i = 1, \ldots, a$$

选择α_i使得$R(\alpha_i \mid \mathbf{x})$最小，进而最小化总的风险即称为贝叶斯风险，记为$R^*$

2.3 二类分类

对称损失/0-1损失

$$\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, 2, \ldots, c$$

有条件风险

$$R(\alpha_1 \mid \mathbf{x}) = \lambda_{11}P(\omega_1 \mid \mathbf{x}) + \lambda_{12}P(\omega_2 \mid \mathbf{x})$$
$$R(\alpha_2 \mid \mathbf{x}) = \lambda_{21}P(\omega_1 \mid \mathbf{x}) + \lambda_{22}P(\omega_2 \mid \mathbf{x})$$

可得贝叶斯决策

$$\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

3 极大似然与贝叶斯参数估计

3.1 极大似然估计

假设样本集$\mathcal{D}$中有n个样本$\mathbf{x}_1, \ldots, \mathbf{x}_n$，由于这些样本均独立抽取，故

$$p(\mathcal{D} \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

这里的$\boldsymbol{\theta}$为参数向量。

定义对数似然为

$$\ell(\boldsymbol{\theta}) = \ln p(\mathcal{D} \mid \boldsymbol{\theta})$$

进而

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$$

求解最大似然估计值的必要条件为

$$\nabla_{\boldsymbol{\theta}} \ell = 0$$

而最大后验(maximum a posteriori, MAP)则是使$\ell(\boldsymbol{\theta})p(\boldsymbol{\theta})$取最大值的参数向量$\boldsymbol{\theta}$，注意这里最好**先乘起来再取对数**。

高维(d维)高斯分布

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

若均值和协方差矩阵均未知，则最大似然估计结果为

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$
$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k-\hat{\boldsymbol{\mu}})(\mathbf{x}_k-\hat{\boldsymbol{\mu}})^{\mathrm{T}} \approx \mathbb{E}\left((\mathbf{x}-\hat{\boldsymbol{\mu}})(\mathbf{x}-\hat{\boldsymbol{\mu}})^{\mathrm{T}}\right)$$

注意上述对方差的估计是有偏的估计。

而**样本协方差矩阵**的无差估计如下

$$C = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k-\hat{\mu})(\mathbf{x}_k-\hat{\mu})^{\mathrm{T}}$$

3.2 贝叶斯参数估计

在最大似然估计方法中，将需要估计的参数向量$\boldsymbol{\theta}$看作一个确定而未知的参数，而在贝叶斯方法中，我们将参数向量$\boldsymbol{\theta}$本身看作一个随机变量，已有的训练样本可以使我们把对于$\boldsymbol{\theta}$的初始密度估计转为后验概率密度。

将训练样本依据类别归到c个次样本集$\mathcal{D}_1,\ldots,\mathcal{D}_c$中，结合先验，贝叶斯公式可表成

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i, \mathcal{D}_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j, \mathcal{D}_j)P(\omega_j)}$$

已知训练样本$\mathcal{D}$，这些样本都从固定但未知的概率密度函数$p(\mathbf{x})$中独立抽取，要求根据这些样本估计$p(\mathbf{x} \mid \mathcal{D})$，即贝叶斯学习的核心问题。

得到贝叶斯估计的核心公式

$$p(\mathbf{x} \mid \mathcal{D}) = \int p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}) \,\mathrm{d}\boldsymbol{\theta}$$

根据贝叶斯公式

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta}}$$

再有样本独立性假设

$$p(\mathcal{D} \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

当没有观测样本时，$p(\boldsymbol{\theta} \mid \mathcal{D}^0) = p(\boldsymbol{\theta})$，反复应用上述公式有概率密度函数$p(\boldsymbol{\theta}), p(\boldsymbol{\theta} \mid \mathbf{x}_1), p(\boldsymbol{\theta} \mid \mathbf{x}_1, \mathbf{x}_2)$等，实际上即增量学习(incremental learning)。

3.3 Fisher线性判别

PCA方法寻找的是用来有效表示的主轴方向，而判别分析方法(discriminant analysis)寻找的是用来有效**分类**的方向。

考虑将d维空间中的数据点投影到一条直线上，以最大限度区分各类数据点的投影方向。假设有一组n个d维的样本$\mathbf{x}_1,\ldots,\mathbf{x}_n$分属两个不同类别，其中$n_1$个样本的子集$\mathcal{D}_1$属于$\omega_1$，$n_2$个样本的子集$\mathcal{D}_2$属于$\omega_2$。对$\mathbf{x}$中各个成分做线性组合，得到点积[2]$y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$，进而全部$n$个样本产生$n$个结果$y_1,\ldots,y_n$相应属于$\mathcal{Y}_1$和$\mathcal{Y}_2$。

[2]回忆高中知识，点积相当于做投影，将$\mathbf{x}$往直线$\mathbf{w}$上投影

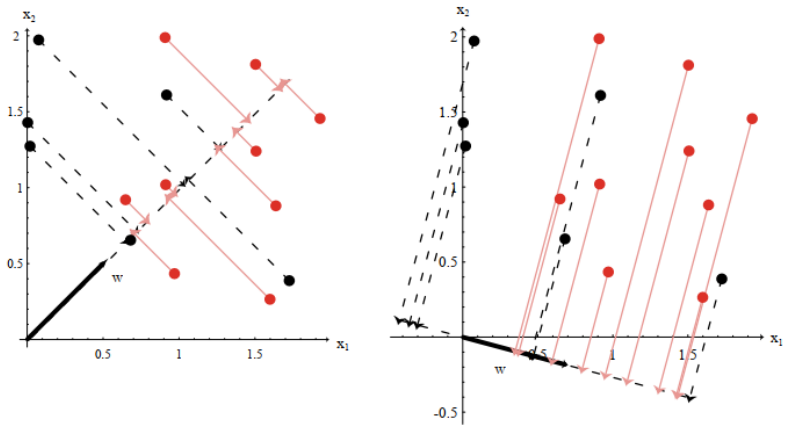

Figure 4.27: Projection of samples onto two different lines. The figure on the right shows greater separation between the red and black projected points.

如何确定最佳的直线方向$\mathbf{w}$得到最佳的分类效果，一个衡量标准即样本均值的差。设原样本第i个类别的均值为$\mathbf{m}_i$，则投影后的样本均值为$\tilde{\mathbf{m}}_i = \mathbf{w}^{\mathrm{T}}\mathbf{m}_i$。样本均值之差为

$$|\tilde{m}_1 - \tilde{m}_2| = |\mathbf{w}^{\mathrm{T}}(\mathbf{m}_1 - \mathbf{m}_2)|$$

可以通过$\mathbf{w}$幅值方法来得到任意大小的均值之差，但这样子没有意义。因此定义类别ω_i的类内散布(scatter)/方差如下

$$\tilde{s}_i^2 = \sum_{y \in \mathcal{Y}_i} (y - \tilde{m})^2$$

这样$1/n(\tilde{s}_1^2 + \tilde{s}_2^2)$即为全部数据总体方差的估计，$\tilde{s}_1^2 + \tilde{s}_2^2$称为投影样本的总类内散布。

故Fisher线性可分性准则要求在投影$y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$下

$$\max_{\mathbf{w}} J(\mathbf{w}) := \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

即均值差尽可能大，同时类内方差尽可能小。

定义类内散布矩阵S_i和总类内散布矩阵S_W如下

$$S_i = \sum_{\mathbf{x} \in \mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^{\mathrm{T}}$$

$$S_W = S_1 + S_2$$

可求得

$$\tilde{s}_i^2 = \mathbf{w}^{\mathrm{T}} S_i \mathbf{w}$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^{\mathrm{T}} S_W \mathbf{w}$$

而投影样本均值之差

$$S_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^{\mathrm{T}}$$

$$(\tilde{\mathbf{m}}_1 - \tilde{\mathbf{m}}_2)^2 = \mathbf{w}^{\mathrm{T}} S_B \mathbf{w}$$

其中S_B称为总类间散布矩阵。 S_W和S_B都是对称且半正定的。

进而原来的准则函数可写成

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}} S_B \mathbf{w}}{\mathbf{w}^{\mathrm{T}} S_W \mathbf{w}}$$

称为广义瑞利商，可证令其最大化

$$S_B \mathbf{w} = \lambda S_W \mathbf{w}$$

一般情况下

$$\arg\max_{\mathbf{w}} J(\mathbf{w}) = S_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

4 非参数技术

4.1 概率密度的估计

向量$\mathbf{x}$落在区域$\mathcal{R}$的概率为

$$P = \int_{\mathcal{R}} p(\mathbf{x}') \, \mathrm{d}\mathbf{x}'$$

即P是概率密度函数$p(\mathbf{x})$平滑后的版本。若假设$p(\mathbf{x})$是连续的，且区域$\mathcal{R}$足够小，以致于在这个区间中p几乎没有变化，则

$$\int_{\mathcal{R}} p(\mathbf{x}') \, \mathrm{d}\mathbf{x}' \approx p(\mathbf{x}) V$$

其中$\mathbf{x}$为一个点，而V为区域$\mathcal{R}$所包含的体积。可以用下述公式作为一个估计

$$p(\mathbf{x}) \approx \frac{k/n}{V}$$

即从n个服从$p(\mathbf{x})$的独立同分布样本落在$\mathcal{R}$中的有k个。

为了估计$\mathbf{x}$的概率密度函数，构造一系列包含$\mathbf{x}$的区域$\mathcal{R}_1, \mathcal{R}_2, \ldots$，第一个区域用1个样本，第二个区域用2个，以此类推。 V_n为区域$\mathcal{R}_n$的体积，k_n为落在区间$\mathcal{R}_n$中的样本个数，而$p_n(\mathbf{x})$表示对$p(\mathbf{x})$的第n次估计：

$$p_n(\mathbf{x}) = \frac{k_n/n}{V_n}$$

若要求$p_n(\mathbf{x})$能够收敛到$p(\mathbf{x})$，则下面3个条件必须满足：

- $\lim_{n\to\infty} V_n = 0$
- $\lim_{n\to\infty} k_n = \infty$
- $\lim_{n\to\infty} k_n/n = 0$

第一个条件保证区域均匀收缩和$p(\cdot)$在点$\mathbf{x}$除连续的情况下，区间平滑了的P/V能够收敛到$p(\mathbf{x})$。第二个

条件保证频率之比能够收敛到概率P。最后一个条件说明虽然最后落在小区域$\mathcal{R}_n$中的样本数目非常大，但是这么多样本在全体样本中所占的比例非常小。

4.2 Parzen窗方法

假设区间$\mathcal{R}_n$是d维超立方体，h_n为一条边长度，体积为

$$V_n = h_n^d$$

定义窗函数

$$\varphi(\mathbf{u}) = \begin{cases} 1 & |\mathbf{u}_j| \le 1/2,\ j = 1, \ldots, d \\ 0 & \text{其他} \end{cases}$$

这样$\varphi(\mathbf{u})$就表示一个中心在原点的单位超立方体。若$\mathbf{x}_i$落在中心点为$\mathbf{x}$的立方体V_n中，那么

$$\varphi((\mathbf{x} - \mathbf{x}_i)/h_n) = 1$$

否则为0，进而可解析表达超立方体样本个数

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

代入估计式有

$$p_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{V_n} \varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right)$$

4.3 k_n近邻估计

最佳的窗函数的选择是个问题，因此一种可行的方案是让体积成为训练样本的函数，而不是硬性规定窗函数为样本个数的某个函数。

比如说可以取$k_n = \sqrt{n}$，有下列迭代过程。

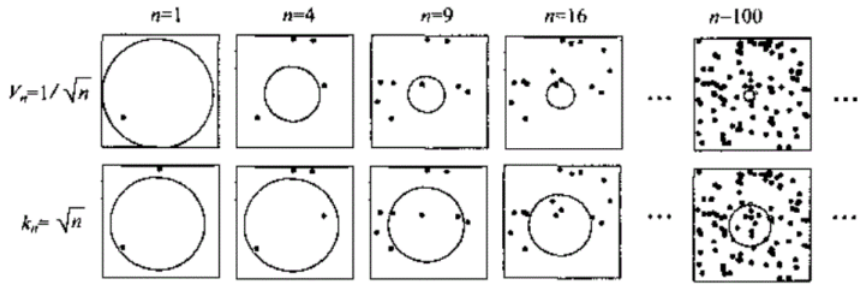

图 4-2　估计某一点处的概率密度函数有两种最基本的方法。这里，我们假设这个点位于图中所示的正方形的中心。第一行表示的方法是从一个以目标样本点为中心的较大的区域开始，根据某个函数，例如 $V_n=1/\sqrt{n}$，逐渐的缩小区域面积。第二种方法如第二行所示。这一方法缩小区域面积的方式是依赖于样本点的。例如，令区域必须包括 $k_n=\sqrt{n}$ 个样本点。这两种情况中的序列都是随机变量，它们一般会收敛，这样就能估计出测试样本点处的真正的概率密度函数

4.4 最近邻规则

定义$\omega_m(\mathbf{x})$为

$$P(\omega_m \mid \mathbf{x}) = \max_i P(\omega_i \mid \mathbf{x})$$

记n个样本的平均误差率为$P_n(e)$，且

$$P = \lim_{n\to\infty} P_n(e)$$

则希望证明

$$P^* \le P \le P^*\left(2 - \frac{c}{c-1}P^*\right)$$

进而推广有k近邻规则(KNN)。

5 多层神经网络

三层神经网络：输入层、隐含层、输出层，也称多层感知器(multilayer perceptron, MLP)

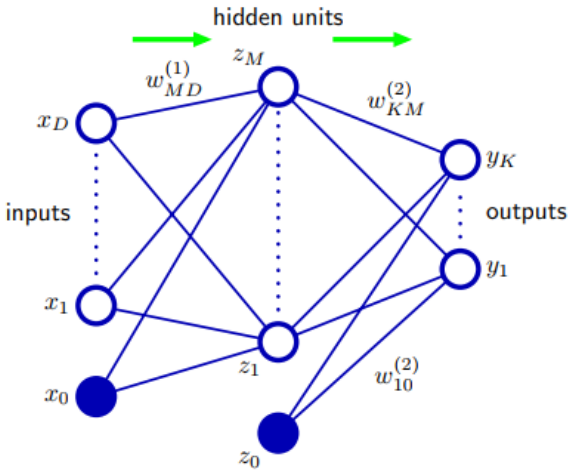

Figure 5.1 Network diagram for the two-layer neural network corresponding to (5.7). The input, hidden, and output variables are represented by nodes, and the weight parameters are represented by links between the nodes, in which the bias parameters are denoted by links coming from additional input and hidden variables x_0 and z_0. Arrows denote the direction of information flow through the network during forward propagation.

前馈运算如下，判别函数$y_k(\mathbf{x}, \mathbf{w})$为每个输出单元产生的信号

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right)$$

任何从输入到输出的连续映射函数都可以用一个三层非线性网络实现，只要有足够的隐单元M、适当的非线性函数和权值。

最小化误差函数

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} \|\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n\|^2$$

6 随机方法

一个系统具有能量E_γ通过下式给出

$$P(\gamma) = \frac{\mathrm{e}^{-E_\gamma/T}}{Z(T)}$$

其中Z是一个归一化常量，

$$Z(T) = \sum_{\gamma'} \mathrm{e}^{-E_{\gamma'}/T}$$