无监督学习

无监督特征学习

从无标注的数据中自动学习有效的数据表示，从而能够帮助后续的机器学习模型更快速地达到更好的性能。

主成分分析、稀疏编码、自编码器等。

主成份分析（Principal Component Analysis， PCA）

最常用的数据降维方法，使得在转换后的空间中数据的方差最大。选择数据方差最大的方向进行投影，才能最大化数据的差异性，保留更多的原始数据信息。用来去除噪声并减少特征之间的相关性，但是它并不能保证投影后数据的类别可分性更好。提高两类可分性的方法一般为监督学习方法，比如线性判别分析（Linear Discriminant Analysis， LDA）。

每个样本点 $\mathbf{x}^{(n)}$ 投影之后的表示为

$$z^{(n)} = \mathbf{w}^{\mathrm{T}}\mathbf{x}^{(n)}. \tag{9.1}$$

我们用矩阵 $X = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(N)}]$ 表示输入样本，$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}^{(n)}$ 为原始样本的中心点，所有样本投影后的方差为

$$\sigma(X;\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(n)} - \mathbf{w}^{\mathrm{T}}\bar{\mathbf{x}})^2 \tag{9.2}$$

$$= \frac{1}{N}(\mathbf{w}^{\mathrm{T}}X - \mathbf{w}^{\mathrm{T}}\bar{X})(\mathbf{w}^{\mathrm{T}}X - \mathbf{w}^{\mathrm{T}}\bar{X})^{\mathrm{T}} \tag{9.3}$$

$$= \mathbf{w}^{\mathrm{T}}S\mathbf{w}, \tag{9.4}$$

其中 $\bar{X} = \bar{\mathbf{x}}\mathbf{1}_d^{\mathrm{T}}$ 为 d 列 $\bar{\mathbf{x}}$ 组成的矩阵，$S = \frac{1}{N}(X - \bar{X})(X - \bar{X})^{\mathrm{T}}$ 是原始样本的协方差矩阵。

最大化投影方差 $\sigma(X;w)$ 并满足 $\mathbf{w}^{\mathrm{T}}\mathbf{w} = 1$，利用拉格朗日方法转换为无约束优化问题，

$$\max_{\mathbf{w}} \mathbf{w}^{\mathrm{T}}S\mathbf{w} + \lambda(1 - \mathbf{w}^{\mathrm{T}}\mathbf{w}), \tag{9.5}$$

其中 λ 为拉格朗日乘子。对上式求导并令导数等于 0，可得

$$S\mathbf{w} = \lambda\mathbf{w}. \tag{9.6}$$

从上式可知，$\mathbf{w}$ 是协方差矩阵 S 的特征向量，λ 为特征值。同时

$$\sigma(X;w) = \mathbf{w}^{\mathrm{T}}S\mathbf{w} = \mathbf{w}^{\mathrm{T}}\lambda\mathbf{w} = \lambda. \tag{9.7}$$

λ 也是投影后样本的方差。因此，主成分分析可以转换成一个矩阵特征值分解问题，投影向量 $\mathbf{w}$ 为矩阵 S 的最大特征对应的特征向量。

如果要通过投影矩阵 $W \in R^{d\times d'}$ 将样本投到 d' 维空间，投影矩阵满足 $W^{\mathrm{T}}W = \mathbf{I}$，只需要将 S 的特征值从大到小排列，保留前 d' 个特征向量，其对应的特征向量即使最优的投影矩阵。

$$SW = W\,\mathrm{diag}(\Lambda), \tag{9.8}$$

稀疏编码（Sparse Coding）

编码的各个维度都是统计独立的，并且可以重构出输入样本。

为了得到稀疏的编码，我们需要找到一组“超完备”的基向量（即 p>d）来进行编码。在超完备基向量之间往往会存在一些冗余性，因此对于一个输入样本，会存在很多有效的编码。如果加上稀疏性限制，就可以减少解空间的大小，得到“唯一”的稀疏编码。

给定一组 N 个输入向量 $\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)}$，其稀疏编码的目标函数定义为：

$$L(A, Z) = \sum_{n=1}^{N} \left(\left\| \mathbf{x}^{(n)} - A\mathbf{z}^{(n)} \right\|^2 + \eta \rho(\mathbf{z}^{(n)}) \right), \tag{9.11}$$

其中 $\rho(\cdot)$ 是一个稀疏性衡量函数，η 是一个超参数，用来控制稀疏性的强度。

训练方法

给定一组 N 个输入向量 $\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)}$，需要同时学习基向量 A 以及每个输入样本对应的稀疏编码 $\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(N)}$。

稀疏编码的训练过程一般用交替优化的方法进行。

1) 固定基向量 A，对每个输入 $\mathbf{x}^{(n)}$，计算其对应的最优编码

$$\min_{\mathbf{x}^{(n)}} \left\| \mathbf{x}^{(n)} - A\mathbf{z}^{(n)} \right\|^2 - \eta \rho(\mathbf{z}^{(n)}), \ \forall n \in [1, N]. \tag{9.16}$$

2) 固定上一步得到的编码 $\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(N)}$，计算其最优的基向量

$$\min_{A} \sum_{n=1}^{N} \left(\left\| \mathbf{x}^{(n)} - A\mathbf{z}^{(n)} \right\|^2 \right) + \lambda \frac{1}{2} \|A\|^2, \tag{9.17}$$

特点：稀疏编码的每一维都可以看作是一种特征。和基于稠密向量的分布式表示相比，稀疏编码具有更小的计算量和更好的可解释性等优点。

计算量 稀疏性带来的最大好处就是可以极大地降低计算量。

可解释性 因为稀疏编码只有少数的非零元素，相当于将一个输入样本表示为少数几个相关的特征。这样我们可以更好地描述其特征，并易于理解。

特征选择 稀疏性带来的另外一个好处是可以实现特征的自动选择，只选择和输入样本相关的最少特征，从而可以更好地表示输入样本，降低噪声并减轻过拟合。

自编码器（Auto-Encoder， AE）

通过无监督的方式来学习一组数据的有效编码（或表示）。

假设有一组 d 维的样本 $\mathbf{x}^{(n)} \in \mathbb{R}^d, 1 \leq n \leq N$，自编码器将这组数据映射到特征空间得到每个样本的编码 $\mathbf{z}^{(n)} \in \mathbb{R}^p, 1 \leq n \leq N$，并且希望这组编码可以重构出原来的样本。

自编码器的学习目标是最小化重构错误（reconstruction errors）。

如果特征空间的维度 p 小于原始空间的维度 d，自编码器相当于是一种降维或特征抽取方法。如果 p≥d，一定可以找到一组或多组解使得 f∘g 为单位函数

$$\mathcal{L} = \sum_{n=1}^{N} \|\mathbf{x}^{(n)} - g(f(\mathbf{x}^{(n)}))\|^2 \tag{9.20}$$

$$= \sum_{n=1}^{N} \|\mathbf{x}^{(n)} - f \circ g(\mathbf{x}^{(n)})\|^2. \tag{9.21}$$

（Identity Function），并使得重构错误为 0。但是，这样的解并没有太多的意义。但是如果再加上一些附加的约束，就可以得到一些有意义的解，比如编码的稀疏性、取值范围，f 和 g 的具体形式等。如果我们让编码只能取 k 个不同的值(k<N)，那么自编码器就可以转换为一个 k 类的聚类（Clustering）问题。

给定一组样本$\mathbf{x}^{(n)} \in [0,1]^d, 1 \leq n \leq N$，其重构错误为

$$\mathcal{L} = \sum_{n=1}^{N} \|\mathbf{x}^{(n)} - \mathbf{x}'^{(n)})\|^2 + \lambda \|W\|_F^2. \tag{9.24}$$

使用自编码器是为了得到有效的数据表示，因此在训练结束后，一般去掉解码器，只保留编码器。编码器的输出可以直接作为后续机器学习模型的输入。

稀疏自编码器（Sparse Auto-Encoder）

自编码器除了可以学习低维编码之外，也学习高维的稀疏编码。假设中间隐藏层 z 的维度为 p 大于输入样本 x 的维度 d，并让 z 尽量稀疏，这就是稀疏自编码器。和稀疏编码一样，稀疏自编码器的优点是有很高的可解释性，并同时进行了隐式的特征选择。

通过给自编码器中隐藏层单元$\mathbf{z}$加上稀疏性限制，自编码器可以学习到数据中一些有用的结构。

$$\mathcal{L} = \sum_{n=1}^{N} \|\mathbf{x}^{(n)} - \mathbf{x}'^{(n)})\|^2 + \eta \rho(\mathbf{z}^{(n)})) + \lambda \|W\|^2, \tag{9.25}$$

堆叠自编码器（Stacked Auto-Encoder， SAE）

使用逐层堆叠的方式来训练一个深层的自编码器，逐层训练（layer-wise training）来学习网络参数。

降噪自编码器（Denoising Autoencoder）

能够从部分损坏的数据中得到有效的数据表示，并能够恢复出完整的原始信息。通过引入噪声来增加编码鲁棒性的自编码器，并提高模型的泛化能力。

对于一个向量 x，我们首先根据一个比例 μ 随机将 x 的一些维度的值设置为 0，得到一个被损坏的向量$\tilde{\mathbf{x}}$。然后将被损坏的向量$\tilde{\mathbf{x}}$输入给自编码器得到编码 z，并重构出原始的无损输入 x。

概率密度估计（Probabilistic Density Estimation）

参数密度估计和非参数密度估计。

参数密度估计（Parametric Density Estimation）

根据先验知识假设随机变量服从某种分布，然后通过训练样本来估计分布的参数。

令$D = \{x^{(n)}\}_{i=1}^{N}$为从某个未知分布中独立抽取的N个训练样本，假设这些样本服从一个概率分布函数$p(x|\theta)$，其对数似然函数为

$$\log p(D|\theta) = \sum_{n=1}^{N} \log p(x^{(n)}|\theta)$$

$$\theta^{ML} = \arg\max_{\theta} \sum_{n=1}^{N} \log p(x^{(n)}|\theta)$$

正态分布

假设样本$\mathbf{x} \in \mathbb{R}^d$服从正态分布

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big), \tag{9.31}$$

其中$\boldsymbol{\mu}$和Σ分别为正态分布的均值和方差。其对数似然函数为

$$\log p(\mathcal{D}|\boldsymbol{\mu}, \Sigma) = -\frac{N}{2}\log\Big((2\pi)^2|\Sigma|\Big) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}). \tag{9.32}$$

分别上式关于$\boldsymbol{\mu}, \Sigma$的偏导数，并令其等于0。可得，

$$\boldsymbol{\mu}^{ML} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}^{(n)}, \tag{9.33}$$

$$\Sigma^{ML} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}. \tag{9.34}$$

多项分布

假设样本服从K个状态的多项分布，令onehot向量$\mathbf{x} \in [0,1]^K$来表示第k个状态，即$x_k = 1$，其余$x_{i,i\neq k} = 0$。样本$\mathbf{x}$的概率密度函数为

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K}\mu_k^{x_k}, \tag{9.35}$$

其中μ_k为第k个状态的概率，并满足$\sum_{k=1}^{K}\mu_k = 1$。

数据集$\mathcal{D} = \{\mathbf{x}^{(n)}\}_{n=1}^{N}$的对数似然函数为

$$\log p(\mathcal{D}|\boldsymbol{\mu}) = \sum_{n=1}^{N}\sum_{k=1}^{K}x_k^{(n)}\log(\mu_k). \tag{9.36}$$

多项分布的参数估计为约束优化问题。引入拉格朗日乘子λ，将原问题转换为无约束优化问题。

$$\max_{\boldsymbol{\mu},\lambda} \quad \sum_{n=1}^{N}\sum_{k=1}^{K}x_k^{(n)}\log(\mu_k) + \lambda\Big(\sum_{k=1}^{K}\mu_k - 1\Big). \tag{9.37}$$

问题：（1）模型选择问题：即如何选择数据分布的密度函数。实际数据的分布往往是非常复杂的，而不是简单的正态分布或多项分布。

（2）不可观测变量问题：即我们用来训练的样本只包含部分的可观测变量，还有一些非常关键的变量是无法观测的，这导致我们很难准确估计数据的真实分布。

（3）维度灾难问题：即高维数据的参数估计十分困难。随着维度的增加，估计参数所需要的样本数量指数增加。在样本不足时会出现过拟合。

非参数密度估计（Nonparametric Density Estimation）

不假设数据服从某种分布，通过将样本空间划分为不同的区域并估计每个区域的概率来近似数据的概率密度函数。

对于高维空间中的一个随机向量$\mathbf{x}$，假设其服从一个未知分布$p(\mathbf{x})$，则$\mathbf{x}$落入空间中的小区域$\mathcal{R}$的概率为

$$P = \int_{\mathcal{R}} p(\mathbf{x})d\mathbf{x}. \tag{9.39}$$

给定N个训练样本$\mathcal{D} = \{\mathbf{x}^{(n)}\}_{n=1}^{N}$，落入区域$\mathcal{R}$的样本数量$K$服从二项分布

$$P_K = \binom{N}{K} P^K (1-P)^{1-K}, \tag{9.40}$$

其中K/N的期望为$\mathbb{E}[K/N] = P$，方差为$\text{var}(K/N) = P(1-P)/N$。当$N$非常大时，我们可以近似认为

$$P \approx \frac{K}{N}. \tag{9.41}$$

假设区域$\mathcal{R}$足够小，其内部的概率密度是相同的，则有

$$P \approx p(\mathbf{x})V, \tag{9.42}$$

其中V为区域$\mathcal{R}$的体积。结合上述两个公式，得到

$$p(\mathbf{x}) \approx \frac{K}{NV}. \tag{9.43}$$

（1）固定区域大小V，统计落入不同区域的数量，这种方式包括直方图方法和核方法两种。（2）改变区域大小以使得落入每个区域的样本数量为K，这种方式称为K近邻方法。

直方图方法（Histogram Method）

非常直观的估计连续变量密度函数的方法，可以表示为一种柱状图。

直方图方法的关键问题是如何选取一个合适的区间宽度Δ。如果Δ太小，那么落入每个区间的样本数量会比较少，其估计的区间密度也具有很大的随机性。如果Δ太大，其估计的密度函数变得十分平滑，很难反映出真实的数据分布。

$$p_m = \frac{K_m}{N\Delta_m}, \qquad 1 \le m \le M$$

直方图通常用来处理低维变量，可以非常快速地对数据的分布进行可视化，但其缺点是很难扩展到高维变量。需要的样本数量会随着维度d的增加而指数增

长，从而导致维度灾难（Curse of Dimensionality）问题。

核方法

假设$\mathcal{R}$为d维空间中的一个以点$\mathbf{x}$为中心的“超立方体”，并定义核函数

$$\phi\Big(\frac{\mathbf{z}-\mathbf{x}}{h}\Big)=\begin{cases}1 & \text{if } |z_i-x_i|<\frac{h}{2}, 1\leq i\leq d\\ 0 & \text{else}\end{cases} \tag{9.45}$$

来表示一个样本$\mathbf{z}$是否落入该超立方体中，其中h为超立方体的边长，也称为核函数的宽度。

给定N个训练样本$\mathcal{D}=\{\mathbf{x}^{(n)}\}_{n=1}^{N}$，落入区域$\mathcal{R}$的样本数量$K$为

$$K=\sum_{n=1}^{N}\phi\Big(\frac{\mathbf{x}^{(n)}-\mathbf{x}}{h}\Big), \tag{9.46}$$

则点$\mathbf{x}$的密度估计为

$$p(\mathbf{x})=\frac{K}{Nh^d}=\frac{1}{Nh^d}\sum_{n=1}^{N}\phi\Big(\frac{\mathbf{x}^{(n)}-\mathbf{x}}{h}\Big), \tag{9.47}$$

其中h^d表示区域$\mathcal{R}$的体积。

除了超立方体的核函数之外，我们还可以选择更加平滑的核函数，比如高斯核函数，

$$\phi\Big(\frac{\mathbf{z}-\mathbf{x}}{h}\Big)=\frac{1}{(2\pi)^{1/2}h}\exp\Big(-\frac{\|\mathbf{z}-\mathbf{x}\|^2}{2h^2}\Big), \tag{9.48}$$

其中h^2可以看做是高斯核函数的方差。这样点$\mathbf{x}$的密度估计为

$$p(\mathbf{x})=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi)^{1/2}h}\exp\Big(-\frac{\|\mathbf{z}-\mathbf{x}\|^2}{2h^2}\Big). \tag{9.49}$$

K 近邻方法

核密度估计方法中的核宽度是固定的，因此同一个宽度可能对高密度的区域过大，而对低密度区域过小。设置一种可变宽度的区域，并使得落入每个区域中样本数量为固定的 K。

如果 K 太小，无法有效地估计密度函数，而 K 太大也会使得局部的密度不准确，并且增加计算开销。

K 近邻方法也经常用于分类问题，称为 K 近邻分类器。当 K=1 也称为最近邻分类器。最近邻分类器的一个性质是，当 N→∞时，其分类错误率不超过最优分类器错误率的两倍。

无监督学习缺少有效的客观评价方法，导致很难衡量一个无监督学习方法的好坏。