第十章　无监督学习

无监督学习（unsupervised learning）是指从无标签的数据中学习出一些有用的模式。无监督学习算法一般直接从原始数据中学习，不借助于任何人工给出标签或者反馈等指导信息。如果监督学习是建立输入-输出之间的映射关系，无监督学习就是发现隐藏的数据中的有价值信息，包括有效的特征、类别、结构以及概率分布等。

典型的无监督学习问题可以分为以下几类：

聚类 聚类（clustering）是将一组样本根据一定的准则划分到不同的组（也称为集群（cluster））。一个比较通用的准则是组内的样本的相似性要高于组间样本的相似性。常见的聚类算法包括k-means算法、谱聚类等。

密度估计 密度估计（density estimation）是根据一组训练样本来估计样本空间的概率密度。密度估计可以分为参数密度估计（）和非参数密度估计（）。参数密度估计是假设数据服从某个已知概率密度函数形式的分布（比如高斯分布），然后根据训练样本去估计概率密度函数的参数。非参数密度估计是不假设数据服从某个已知分布，只利用训练样本对密度进行估计，可以进行任意形状密度的估计。非参数密度估计的方法有直方图、核密度估计等。

无监督特征学习 无监督特征学习（unsupervised feature learning）是从无标签的训练数据中挖掘有效的特征或表示。无监督特征学习一般用来进行降维、数据可视化或监督学习前期的数据预处理。

特征学习也包含很多的监督学习算法，比如线性判别分析等。

非监督学习的准则非常多，比如最大似然估计、最小重构错误等。以无监督特征学习为例，经常使用的准则为最小化重构错误，同时也经常对特征进行一些约束，比如独立性、非负性或稀释性等。

图 10.1: 主成分分析

10.1 主成分分析

主成份分析（principal component analysis，PCA）一种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。如图10.1所示的两维数据，如果将这些数据投影到一维空间中，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，保留能多的原始数据信息。

图待修改，红色线

假设有一组 d 维的数据 $\mathbf{x}^{(i)} \in \mathbb{R}^d, 1 \leq i \leq N$，我们希望将其投影到一维空间中，投影向量为 $\mathbf{w} \in \mathbb{R}^d$。不失一般性，可以限制 $\mathbf{w}$ 的模为1，即 $\mathbf{w}^{\mathrm{T}}\mathbf{w} = 1$。每个数据点 $\mathbf{x}^{(i)}$ 投影之后的表示为

$$z^{(i)} = \mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)}. \tag{10.1}$$

我们用 $X = [\mathbf{x}^{(1)}\mathbf{x}^{(2)} \cdots \mathbf{x}^{(N)}]$ 表示数据矩阵，$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}^{(i)}$ 为原始数据的中心点，所有数据投影后的方差为

$$\sigma(X; w) = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} - \mathbf{w}^{\mathrm{T}}\bar{\mathbf{x}})^2 \tag{10.2}$$

$$= \frac{1}{N}(\mathbf{w}^{\mathrm{T}}X - \mathbf{w}^{\mathrm{T}}\bar{X})(\mathbf{w}^{\mathrm{T}}X - \mathbf{w}^{\mathrm{T}}\bar{X})^{\mathrm{T}} \tag{10.3}$$

$$= \mathbf{w}^{\mathrm{T}}S\mathbf{w} \tag{10.4}$$

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

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，

$$\sigma(X; w) + \lambda(1 - \mathbf{w}^{\mathrm{T}}\mathbf{w}), \tag{10.5}$$

其中 λ 为拉格朗日乘子。对上式求导并令导数等于 0，可得

$$S\mathbf{w} = \lambda\mathbf{w}. \tag{10.6}$$

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

$$\sigma(X; w) = \mathbf{w}^{\mathrm{T}}S\mathbf{w} = \mathbf{w}^{\mathrm{T}}\lambda\mathbf{w} = \lambda. \tag{10.7}$$

λ 也是投影后数据的方差。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 $W \in R^{d \times d'}$ 将数据投到 d' 维空间，投影矩阵满足 $W^{\mathrm{T}}W = I$，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。

$$SW = W\,\mathbf{diag}(\Lambda), \tag{10.8}$$

其中 $\Lambda = [\lambda_1, \cdots, \lambda_{d'}]$ 为 S 的前 d' 个最大的特征值。

对于 N 个样本组成的数据集，其有效的投影子空间不超过 $N - 1$ 维。

主成分分析是一种无监督学习方法，可以作为监督学习的数据预处理方法，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。提高列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）。

10.2 稀疏编码

稀疏编码（sparse coding）也是一种受哺乳动物视觉系统中*简单细胞*感受野而启发的模型。

在哺乳动物的*初级视觉皮层*（primary visual cortex）中，每个神经元仅对处于其感受野中特定的刺激信号做出响应，比如特定方向的边缘、条纹等特征。

数学小知识 | 完备性

如果 p 个基向量刚好可以 p 维的欧式空间，则这 p 个基向量是完备的。如果 p 个基向量可以 d 维的欧式空间，并且 $p > d$，则这 p 个基向量是过完备的，冗余的。

“过完备”基向量一般是指的是基向量个数远远大于其支撑空间维度。因此这些基向量一般是不具备独立、正交等性质。

带通滤波（bandpass filter）是指容许某个频率范围的信号通过，同时屏蔽其他频段的设备。

局部感受野可以被描述为具有空间局部性、方向性和带通性（即不同尺度下空间结构的敏感性）[Olshausen et al., 1996]。也就是说，外界刺激在视觉神经系统的表示具有很高的稀疏性，即外界信息经过编码后仅有一小部分神经元激活。编码的稀疏性在一定程度上符合生物学的低功耗特性。

在数学上，（线性）编码是指给定一组基向量 $A = [\mathbf{a}_1, \cdots, \mathbf{a}_p]$，将输入样本 $\mathbf{x} \in \mathbb{R}^d$ 表示这些基向量的线性组合

$$\mathbf{x} = \sum_{i=1}^{p} z_i \mathbf{a}_i \tag{10.9}$$

$$= A\mathbf{z}, \tag{10.10}$$

其中基向量的系数 $\mathbf{z} = [z_1, \cdots, z_p]$ 称为输入样本 $\mathbf{x}$ 的编码（encoding），基向量 A 也称为字典（dictionary）。

编码是对 d 维空间中的样本 $\mathbf{x}$ 找到其在 p 维空间中的表示（或投影），其目标通常是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。编码的关键是找到一组“完备”的基向量 A，比如主成分分析等。但是主成分分析得到编码通常是稠密向量，没有稀疏性。

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：

$$L(A, Z) = \sum_{n=1}^{N} \left(\left\| \mathbf{x}^{(n)} - A\mathbf{z}^{(n)} \right\|^2 + \eta \rho(\mathbf{z}^{(n)}) \right), \tag{10.11}$$

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。

对于一个向量 $\mathbf{z} \in \mathbb{R}^p$，其稀疏性定义为非零元素的比例。如果一个向量只有很少的几个非零元素，就说这个向量是稀疏的。

严格的稀疏向量有时比较难以得到，因此如果一个向量只有少数几个远大于零的元素，其它元素都接近于 0，我们也称这个向量为稀疏向量。

稀疏性衡量函数 $\rho(\mathbf{z})$ 是给向量 $\mathbf{z}$ 一个标量分数。$\mathbf{z}$ 越稀疏，$\rho(\mathbf{z})$ 越小。

稀疏性衡量函数有多种选择，最直接的衡量向量 $\mathbf{z}$ 稀疏性的函数是 L_0 范式

$$\rho(\mathbf{z}) = \sum_{i=1}^{p} \mathbf{I}(|z_i| > 0)) \tag{10.12}$$

但 L_0 范式不满足连续可导，因此很难进行优化。在实际中，通常使用 L_1 范式

$$\rho(\mathbf{z}) = \sum_{i=1}^{p} |z_i| \tag{10.13}$$

或对数函数

$$\rho(\mathbf{z}) = \sum_{i=1}^{p} \log(1 + z_i^2) \tag{10.14}$$

或指数函数

$$\rho(\mathbf{z}) = \sum_{i=1}^{p} -\exp(-z_i^2). \tag{10.15}$$

10.2.1 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

$$\min_{\mathbf{x}^{(n)}} \left\| \mathbf{x}^{(n)} - A\mathbf{z}^{(n)} \right\|^2 - \eta \rho(\mathbf{z}^{(n)}), \ \forall n \in [1, N]. \tag{10.16}$$

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

$$\min_{A} \sum_{n=1}^{N} \left(\left\| \mathbf{x}^{(n)} - A\mathbf{z}^{(n)} \right\|^2 \right) + \lambda \frac{1}{2} \|A\|^2, \tag{10.17}$$

其中第二项为正则化项，λ为正则化项系数。

10.2.2 稀疏编码的优点

稀疏编码的每一维都可以看做是一种特征。和*分布式表示*相比，稀疏编码的具有更小的计算量，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。

10.3 自编码器

自编码器（auto-encoder）是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$\mathbf{x}^{(n)} \in \mathbb{R}^d, 1 \leq n \leq N$，没有输出标签。*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。

自编码器的结构可分为两部分：

编码器（encoder）

$$f : \mathbb{R}^d \to \mathbb{R}^p \tag{10.18}$$

和解码器（decoder）

$$g : \mathbb{R}^p \to \mathbb{R}^d. \tag{10.19}$$

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

$$\mathcal{L} = \sum_{n=1}^{N} \|\mathbf{x}^{(n)} - g(f(\mathbf{x}^{(n)}))\|^2 \tag{10.20}$$

$$= \sum_{n=1}^{N} \|\mathbf{x}^{(n)} - f \circ g(\mathbf{x}^{(n)})\|^2 \tag{10.21}$$

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$p \geq d$，一定可以找到一组或多组解使得$f \circ g$为单位函数（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。然而，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类的聚类问题。

最简单的自编码器是三层的神经网络，输入层到隐藏层用来编码，隐藏层到输出层用来解码。层与层之间互相全连接（如图10.2所示）。输入为样本$\mathbf{x}$，中间隐藏层为编码

$$\mathbf{z} = s(W^{(1)}\mathbf{x} + b^{(1)}), \tag{10.22}$$

输出为重构的数据

$$\mathbf{x}' = s(W^{(2)}\mathbf{z} + b^{(2)}), \tag{10.23}$$

其中W, b为网络参数，$s(\cdot)$为激活函数。如果令$W^{(2)}$等于$W^{(1)}$的转置，即$W^{(2)} = W^{(1)^{\mathrm{T}}}$，称为捆绑权重（tied weights）。

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

$$\mathcal{L} = \sum_{n=1}^{N} \|\mathbf{x}^{(n)} - \mathbf{x}'^{(n)})\|^2 + \lambda \|W\|_F^2. \tag{10.24}$$

其中λ为正则化项系数。通过最小化重构错误，可以有效地学习中网络参数。

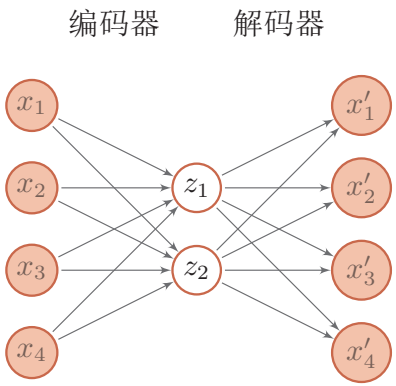


图 10.2: 三层网络结构的自编码器

10.4 稀疏自编码器

自编码器除了可以学习到有效的低维编码之外，还可以高维的稀疏编码。假设中间隐藏层 $\mathbf{z}$ 的维度为 p 大于输入样本 $\mathbf{x}$ 的维度 d，并限制 $\mathbf{z}$ 尽量稀疏，这就是*稀疏自编码器*（sparse auto-encoder）。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。

$$\mathcal{L} = \sum_{n=1}^{N} \|\mathbf{x}^{(n)} - \mathbf{x}'^{(n)})\|^2 + \eta\rho(\mathbf{z}^{(n)})) + \lambda\|W\|^2, \tag{10.25}$$

其中 $\rho(\cdot)$ 为稀疏性度量函数，W 表示自编码器中的参数。

稀疏性度量函数 $\rho(\cdot)$ 除了可以选择公式 (10.13)-(10.15) 的定义外，还可以定义为一组训练样本中每一个神经元激活的频率。

给定 N 个训练样本，隐藏层第 j 个神经元平均活性值为

$$\hat{\rho}_j = \frac{1}{N}\sum_{n=1}^{N} z_j^{(n)}, \tag{10.26}$$

$\hat{\rho}_j$ 可以近似地看作是第 j 个神经元激活的概率。我们希望 $\hat{\rho}_j$ 接近于一个事先给

定的值 ρ^*，比如 0.05，可以通过 KL 距离来衡量 $\hat{\rho}_j$ 和 ρ^* 的差异，即

$$\mathrm{KL}(\rho^*||\hat{\rho}_j) = \sum_{j=1}^{p} \rho^* \log \frac{\rho^*}{\hat{\rho}_j} + (1-\rho^*) \log \frac{1-\rho^*}{1-\hat{\rho}_j}. \tag{10.27}$$

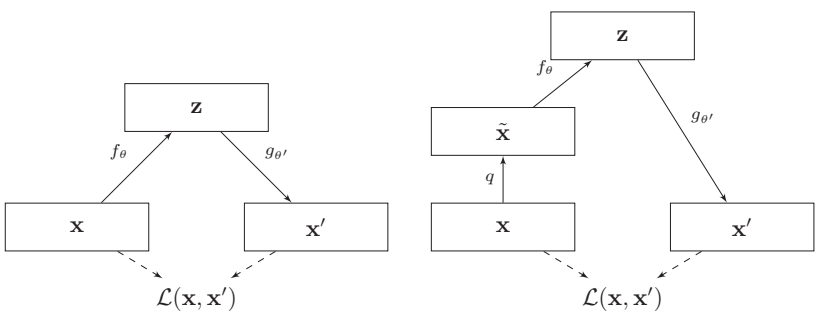
如果 $\hat{\rho}_j = \rho^*$，则 $\mathrm{KL}(\rho^*||\hat{\rho}_j) = 0$。

稀疏性度量函数定义为

$$\rho(\mathbf{z}^{(n)})) = \sum_{j=1}^{p} \mathrm{KL}(\rho^*||\hat{\rho}_j). \tag{10.28}$$

10.5 堆叠自编码器

10.6 降噪自编码器



(a) 自编码器 (b) 降噪自编码器

图 10.3: 自编码器 VS 降噪自编码器。f_θ 为编码器，$g_{\theta'}$ 为解码器。$\mathcal{L}(\mathbf{x}, \mathbf{x}')$ 为重构错误。

?

10.7 自组织特征映射网络

10.8 总结和深入阅读

习题 10-1 分析主成分分析为什么具有数据降噪能力？

习题 10-2 若数据矩阵 $X' = X - \bar{X}$，则对 X' 奇异值分解 $X' = U\Sigma V$，则 U 为主成分分析的投影矩阵。

参考文献

Bruno A Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.