15. 表示学习

良好的信息表示可以简化后续学习任务/信息处理，表示的选择通常取决于后续的学习任务。

例如，监督学习中接近顶层的隐藏层，其表示应能够更加容易的完成训练任务。

表示学习提供了进行无监督学习和半监督学习(unsupervised and semi-supervised learning)的一种方法。

深度学习训练中，高质量标注数据比较珍贵，我们通常有大量的未标注数据和相对较少的标注数据。

假设：**未标注数据可以学习出良好的表示** -> 无/半监督学习的可应用性

15.1 贪心逐层无监督预训练(Greedy Layer-Wise Unsupervised Pretraining)

贪心算法(greedy algorithm)：通过每一步都做出**局部最优**选择，**期望**最终获得**全局最优解**。

局部最优、不可回退、不一定能得到全局最优解

表示的**可迁移性**：无监督学习获取的表示，有助于开展具有相同输入域的监督学习

监督学习：可利用预训练得到的顶层特征，训练一个简单分类器；或对预训练得到的网络进行监督微调

应用：

贪心学习过程可为多层联合训练过程寻找好的初始值，有助于训练**多层神经网络**

15.1.1 (贪心)无监督预训练为何有效?

！ Idea

1. 初始参数的选择可以显著影响正则化效果

2. 学习输入分布有助于学习从输入到输出的映射

应用场景

初始表示较差时

例：one-hot向量：高维、稀疏、彼此正交的向量，所有不同的独热向量L2距离均相等，难以捕获词的相似性

通过无监督预训练(如Word2Vec和GloVe)，模型能够学习到词嵌入(word-embeddings)，将单词映射到一个连续的向量空间，通过L2距离、余弦相似度(cosine similarity)等表示语义相似度

标注样本数量有限时

目标函数较复杂时

改进优化分类器、降低测试集误差

神经网络训练的**非确定性**：

梯度接近零的点：局部最优

早停(预防过拟合)、但未达到最优解

梯度过大但难以下降(Hessian矩阵的病态条件)

无监督预训练获得的参数：收敛到更小的参数空间，减小方差，更加稳定，降低过拟合风险

> 相关文献：Erhan et al: Why Does Unsupervised Pre-training Help Deep Learning?

不足

双阶段训练(无监督+有监督)：更加复杂、耗时
超参数设置复杂、反馈延迟
无法灵活调整正则化强度
缺乏对成熟监督学习方法的竞争力
！此处，可通过中等数据集MNIST对比监督学习和无监督学习的效果(结合d2l)。！注意，PR前删掉data文件夹，这是跑MNIST时自动下载的

15.2 迁移学习(Transfer learning)和领域自适应(Domain adaption)

利用一个情景（例如，分布 P1）中已经学到的内容，去改善另一个情景（比如分布 P2）中的泛化情况。
迁移学习：从一个任务（P1）中获得知识，并将其应用于另一个不同但相关的任务（P2）
决定迁移成功的因素：联合学习三种因素（**表示A，表示B和AB之间的关系**）
领域自适应：P1,P2的输入->输出映射关系是相同的，但输入的分布略有不同。
概念漂移(concept drift)：数据分布随时间发生变化
输入语义共享：共享底层和任务相关上层的学习框架
计算机视觉：边缘、视觉形状、几何变化、光照变化
输出语义共享：共享神经网络的上层（输出附近）和进行任务特定的预处理
语音识别
迁移学习的特殊形式：
一次学习（one-shot learning）：仅使用一个标注样本来完成新任务的学习
预训练的表示空间已经学习了**不变性特征**（即与类别无关的变化因素），从而可以有效区分类别
零次学习（zero-shot learning）：在没有任何标注样本的情况下进行迁移学习

15.3 半监督解释因果关系

什么原因能够使一个表示比另一个表示更好？ -- **表示能够更好地区分原因**。
单一特征 -> 观测数据的潜在成因
不同特征(特征空间中不同的方向) -> 不同成因

表示学习与半监督学习^[1]的关系

> 核心观点：如果能够通过**无监督学习**获得建模数据的**潜在结构**，尤其是找到与标签紧密相关的**潜在因素**，那么在这种情况下，半监督学习可以取得成功。

半监督学习何时失败?
例：p(x)均匀分布，此时仅通过对p(x)的无监督学习，不能为p(y|x)的学习提供足够的信息。
半监督学习何时成功?
例：p(x)混合分布^[2]，且每个y对应一个混合分量，针对p(x)的无监督学习可以识别出混合分量的结构；进而，如果每个类都有一个标注样本，模型就可以精确地学习p(y∣x)。

即：**当y与x的生成过程高度相关时，通过建模p(x)的无监督表示可以为p(y|x)提供有用的信息**

半监督学习的现实应用

事实：现实世界中大多数观测数据是由许多潜在成因h_i共同作用产生的,但无监督学习器并不知道具体是哪些h_i；

思路：暴力求解，试图学习一种能够捕获**所有潜在生成因子**的表示

问题：成本过高、现实数据难以逐一捕捉 > 人们不会察觉到环境中和他们所在进行的任务并不立刻相关的变化。[https://link.springer.com/content/pdf/10.3758/bf03208840.pdf](Simons & Levin,1998)

研究前沿：**如何确定在特定任务中，哪些潜在成因是需要被编码的(即：最为关键)?**

相关研究：

基于均方误差训练的自编码器

GAN(生成对抗网络)

... （最新研究！！）

15.4 分布式表示

分布式表示

分布式表示：使用高维向量，通过多个维度的组合编码信息

应用：词向量(Word2Vec)、特征提取等

特点：具有丰富的**相似性空间**；不同的输入共享同一组维度，不同维度的激活状态组合成复杂的模式，能够捕捉数据中的细微差别。(**共享表示**)

优势：表达能力更强，对噪声更鲁棒；"当一个明显复杂的结构可以用较少参数紧致地表示"时，适合**分布式表示**

为什么分布式表示具有较强的泛化能力?

非分布式表示

非分布式表示(符号表示)：每个输入信息与单独的符号或特征关联，彼此**没有共享或组合的表示方式。**

应用：独热编码(One-hot向量，只有1位激活的n维二元向量)

特点：参数足够时，容易拟合训练数据；但**只能通过平滑先验来局部泛化**(即：难以学习复杂函数，只能在与训练数据接近的区域内泛化)

平滑先验的局限性：在高维数据和复杂函数的学习中，存在严重的维数灾难

[1] 半监督学习：使用少量的标注数据（有标签的数据）和大量的未标注数据（无标签的数据）来训练模型。

可以给些半监督学习失败/成功的例子。 [2] $$p(x) = \sum_{i=1}^{K} \pi_i p_i(x)$$

备用：无监督学习

无监督学习（p(x)）可以识别出混合分量的结构，这是因为在混合分布中，数据（x）的生成过程是由多个不同的分量（component）共同作用的，而这些分量可能对应着不同类别或不同的潜在特征。通过无监督学习（p(x)），模型可以识别出这些分布的不同部分，从而捕捉到混合分布中每个分量的结构。具体原因如下：

1. **混合分布的特性**:

混合分布由多个分量分布构成，比如高斯混合模型（GMM），其概率密度函数是多个高斯分布的加权和： $ p(x) = \sum_{k=1}^{K} \pi_k p_k(x) $ 其中，$ p_k(x) $ 是第 $ k $ 个分量分布，$ \pi_k $ 是对应的混合权重。每个分量可能代表某个类或潜在的结构。
在混合分布中，数据 $ x $ 是从不同的分量中抽样得到的，因此不同的 $ x $ 值实际上代表了多个分量分布的样本。通过无监督学习（例如聚类或密度估计方法），模型可以尝试找到每个分量的结构特征和分布形状。

2. **无监督学习如何识别混合分量**:

无监督学习的目标是通过观察数据 $ x $ 的分布，建模其潜在的生成过程。在混合分布的情况下，每个 $ y $ 对应一个混合分量，即 $ p(x) $ 是不同分量的组合。模型通过学习数据的总体分布，能够将数据自动归类到不同的分量，这实际上是在无监督地识别出每个分量的结构。
比如，GMM 会通过最大期望（EM）算法学习到数据中每个混合分量的参数（均值、方差等），从而捕捉到数据中不同类别或特征的潜在结构。

3. **识别出的结构有助于分类**:

每个分量对应着不同的 $ y $ 值或类别。例如，在图像分类中，不同的混合分量可能对应着不同物体的类别。通过识别 $ p(x) $ 的混合分量，模型可以将数据划分为不同的类或特征空间。这样一来，即使只有少量的标注样本，通过这些已识别的结构，模型可以更容易地学习 $ p(y|x) $。

4. **数据的稀疏性和聚集性**:

混合分布通常具有稀疏和聚集性，数据点会围绕着各个分量的中心聚集。无监督学习可以利用这种稀疏和聚集性，识别出不同类别或特征的划分。因为混合分量通常有显著的统计差异，模型能够通过数据的这些差异识别出不同分量的结构。

总结:

无监督学习 $ p(x) $ 可以识别出混合分量的结构，是因为混合分布的生成过程包含多个分量，这些分量代表了数据的不同类别或潜在特征。通过建模 $ p(x) $，无监督学习能够自动识别并分离出这些潜在的分量，从而为后续的分类或预测任务（如学习 $ p(y|x) $）提供有用的结构信息。