***COS 598D*：机器学习中克服难解性问题。**
***Sanjeev Arora*，普林斯顿大学。*2015*年春季。**

讲座1：机器学习和机器学习理论的快速调查。

学习的各种含义。通常的假设：数据由某个分布的iid样本组成。 （哲学上的旁白：DeFinetti的定理，可交换性。）本讲座中的一个运行示例将是线性分类器。

- 无监督与有监督。

  无监督：无标签数据。 通常需要某种模型来解释数据生成的方式（“故事”），然后恢复模型参数。 示例：k--‐m*ea**ns*和其他形式的聚类，对数线性模型，贝叶斯网络等。
  通常是NP--‐hard。

  有监督：训练数据由人类标记（标签可以是二进制或[1..k]）。算法需要预测未来数据的标签。 示例：决策树，支持向量机，k--‐NN。

  泛化界限：将训练数据的性能与未见数据（即整个分布）的性能联系起来。

  粗略想法：假设存在一个可用M位表示且在完整分布上的误差最多为\epsilon的分类器。 那么，如果训练集的大小至少为f(M + K,\epsilon)，那么任何可用K位描述且在训练集上的误差最多为\epsilon的分类器，在整个分布上的误差最多为2\epsilon。 此外，如果任何分类器在整个分布上的误差最多为\epsilon/2，则在样本上的误差最多为\epsilon。 （因此，该方法是完整且可靠的：如果存在一个好的分类器，可以通过检查一小组训练点找到它，反之，每个在样本上足够好的分类器也足够好用于整个分布。）

  证明概要：切尔诺夫界限。 只有2^M个可以用M位表示的分类器。 如果任何M位分类器在分布的2\epsilon比例的点上有错误，则它在训练集上只有\epsilon比例的错误的概率小于2^{--M}。因此，没有坏的分类器可以在训练点上表现良好。

  有一个更一般的理论来计算训练点的数量，涉及到VC维度。请参考在线资源。

  经典的哲学原则奥卡姆剃刀与此有关。

  训练的例子：感知器算法用于线性分类器。 （可以将其视为确定特征权重的一种方法。）完全非贝叶斯描述。

也可以使用边际的概念将其转化为凸规划问题。

- 判别式与生成式。

  判别式：只知道P(label| data)。 (示例1：label = 线性阈值对应于SVM。请注意，这是确定性的。示例2：逻辑回归。
  平滑版本的SVM。判别式学习器的示例：决策树，SVM，核SVM，深度网络，逻辑回归等。SVM和逻辑回归可以通过凸优化来解决。

  生成式：知道P(label, data)的表达式。 通过贝叶斯规则和计算P(label, data)/P(data)来估计P(label|data)。 有关示例，请参阅Cynthia Rudin关于朴素贝叶斯的垃圾邮件分类的讲义。

  还请参阅Mitchell的书中的章节，该章节显示逻辑回归对应于具有独立同分布高斯坐标的朴素贝叶斯估计器。

  为了更深入地理解，可以参考克里斯·曼宁的讲义中的实例。

  https://web.stanford.edu/class/cs124/lec/Maximum_Entropy_Classifiers.pdf

- （顺便说一句：逻辑回归很棒，对吧？如果我们将逻辑回归单元堆叠在一起，会得到深度神经网络。训练这些网络是一个非平凡的任务，我们只知道一些启发式算法。我们不知道这种深度网络分类器的一个好的贝叶斯解释。）

- 每种方法都有其优势。判别式方法需要较少的假设。
  生成式方法更容易适应半监督设置。

  请参阅

  Tom Mitchell的书中关于生成式与判别式的相关章节。

  关于判别式与生成式的比较：Ng和Jordan在NIPS 2001中的一篇论文中比较了逻辑回归和朴素贝叶斯。

  生成式还是判别式？ 通过Bishop和Lasserre在2007年的贝叶斯统计学中获得最佳效果。

- 正则化。用于避免过拟合的技术。 为此，最好使用一个较简单的解决方案，并在目标函数中添加一个正则化项可以帮助解决这个问题。（与泛化理论相关；一个粗略的类比是限制自己

只使用较少位数描述的解决方案。 这只是一个粗略的
直觉）

我们将重点关注无监督学习。

最大似然和最大熵原理。

在许多可能拟合数据的分布中进行选择时，选择具有最大熵的分布。

例子：我们有一个骰子，投掷后产生期望值为4.7。
产生5的机会是多少？解决方案：设
p_i= 产生i的概率。然后i p_i的平均值为4.7。计算使得熵最大的p_i的值，同时
满足这个平均值。

例子2：如果我们只知道分布的均值，与之一致的最大熵分布是指数分布
。（如果变量是n维的，分布是对数线性的。）

例子3：如果我们只知道分布的均值和协方差
那么与之一致的最大熵分布就是高斯分布。

最大似然方法

找到能最大化观测数据的参数向量Theta。

（顺便说一句：令人惊讶的是，香农在1948年除了信息论之外还发明了自然语言处理
他描述了语言的n-gram模型，并建议用他的熵度量来衡量它们
我们可以将其视为最大似然。
http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf)

例子：来自逻辑回归的最大对数似然表达式
http://ufldl.stanford.edu/wiki/index.php/Softmax_Regression

Recall that in logistic regression, we had a training set $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$ of m labeled examples, where the input features are $x^{(i)} \in \Re^{n+1}$. (In this set of notes, we will use the notational convention of letting the feature vectors x be $n + 1$ dimensional, with $x_0 = 1$ corresponding to the intercept term.) With logistic regression, we were in the binary classification setting, so the labels were $y^{(i)} \in \{0, 1\}$. Our hypothesis took the form:

$$h_\theta(x) = \frac{1}{1 + \exp(-\theta^T x)},$$

and the model parameters θ were trained to minimize the cost function

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))\right]$$

这个表达式是凸的，所以梯度下降方法可以很快找到最佳拟合。 事实上，这种计算的简便性是逻辑回归受欢迎的原因之一--﹣--﹣--﹣这种受欢迎程度可以追溯到计算机出现之前的时代，当时人们使用滑尺等工具。

不幸的是，当你计算大多数其他设置的对数似然时，表达式变得非凸。 这种非凸优化通常是NP--﹣hard的（对于许多设置已经被证明）。

简单的例子：具有相同半径的球形高斯混合。最大化对数似然等同于k--﹣means聚类。

（来自Arora--﹣Kannan：学习分离的非球形高斯混合；https://www.cs.princeton.edu/~arora/pubs/gaussians.pdf)

试图克服这种难解性是本课程的主要目标。

4. Max-likelihood estimation. Now we describe an algorithm for max-likelihood fit of a mixture of k spherical Gaussians of equal radius to (possibly) unstructured data. First we derive a combinatorial characterization of the optimum solution in terms of the *k-median* (*sum of squares*, *Steiner version*) problem. In this problem, we are given M points $x_1, x_2, \ldots, x_M \in \Re^n$ in $\Re^n$ and an integer k. The goal is to identify k points $p_1, p_2, \ldots, p_k$ that minimize the function

$$\sum_{i=1}^{M} |x - p_{c(j)}|^2, \tag{26}$$

where $p_{c(j)}$ is the point among $p_1, \ldots, p_k$ that is closest to j and $|\cdot|$ denotes Euclidean distance.

THEOREM 13. *The mixture of k spherical Gaussians that minimizes the log-likelihood of the sample is exactly the solution to the above version of k-median.*

PROOF. Recall the density function of a spherical Gaussian of variance σ (and radius $\sigma\sqrt{n}$) is

$$\frac{1}{(2\pi\sigma)^{n/2}} \exp\left(-\frac{|x-p|^2}{2\sigma^2}\right).$$

Let $x_1, x_2, \ldots, x_M \in \Re^n$ be the points. Let $p_1, p_2, \ldots, p_k$ denote the centers of the Gaussians in the max-likelihood solution. For each data point x_j let $p_{c(j)}$ denote the closest center. Then the mixing weights of the optimum mixture $w_1, w_2, \ldots, w_k$ are determined by considering, for each i, the fraction of points whose closest center is p_i.

The log-likelihood expression is obtained by adding terms for the individual points to obtain

$$-\left[\text{Constant} + \frac{Mn}{2}\log\sigma + \sum_j \frac{|x_j - p_{c(j)}|^2}{2\sigma^2}\right].$$

The optimum value $\hat{\sigma}$ is obtained by differentiation,

$$\hat{\sigma}^2 = \frac{2}{Mn}\sum_j |x_j - p_{c(j)}|^2, \tag{27}$$

which simplifies the log-likelihood expression to

$$\text{Constant} + \frac{Mn}{2}\log\hat{\sigma} + \frac{Mn}{4}.$$

Thus the goal is to minimize $\hat{\sigma}$, which from (27) involves minimizing the familiar objective function from the sum-of-squares version of the k-median problem. □

普林斯顿大学。 Spr'15 cos 598D：克服机器学习中的难解性

讲座2：线性化模型和SVD

讲师：*Sanjeev Arora* 记录员：

1 关于上次的总结思考和今天的过渡话题

从理论角度来看，监督学习的主导模型是SVM或核SVM。我们上次讨论了SVM；只是一个线性分类器。

核函数是从数据点 x到高维空间中的点 $\phi(x)$ 的映射。例子：将 $(x_1, x_2, \ldots\ldots\ldots\ldots, x_n)$ 映射为 n^3维向量 $(x_{i_1}x_{i_2}x_{i_3}$)。可以将 $\phi(x)$ 的坐标看作特征。

使核支持向量机相对实用的两个事实是：(i) 标准的支持向量机拟合算法只需要能够计算数据点对的内积。 因此，如果核函数满足给定x, y时计算 $\langle\phi(x), \phi(y)\rangle$容易，它们也适用于核支持向量机。 这对于所有流行的核函数都成立。 因此，在上面的例子中，没有必要使用显式的 n^3维表示。 (ii) 算法的运行时间和样本复杂度与$1/\lambda$成比例，其中 λ是0示例和1示例之间的间隔。 因此，即使核函数隐式地映射到非常高维的空间，运行时间也可以很小。

核支持向量机的理论证明是边际假设：对于每个分类任务，都存在一个合理的核函数，该函数也具有较大的边际（因此可以高效拟合）

核支持向量机的诱惑在于它们承诺将特征学习与分类任务结合起来。 不幸的是，在实践中，人们需要使用更明确的方法来学习特征，这为数据提供了一种新的表示，然后可以在其上拟合一个分类器。

这就引出了今天的主题，即特征学习。 与分类一样，这可以在生成和非生成的环境中完成。

示例1在 k均值问题中，给定点 x_1，x_2，...，$\in \Re^d$，并且试图找到 k个点（称为均值） c_1，c_2，...，c_k以使目标函数最小化。 ，$\in$ <d并且试图找到k个点（称为均值）c1，c2，...，ck以使目标函数最小化。

$$\sum_i |x_i - p_i|^2,$$

其中 p_i是最接近 x_i的均值。

在学习这些均值之后，每个点都被标记为从1到 k，对应于它最接近的均值。

这个的生成模拟可以是 k个高斯混合模型。 每个数据点可以用一个 k元组来标记，描述它从每个高斯模型生成的概率。

非生成特征学习问题（如k-means）通常是NP难的，而生成版本似乎更容易（平均情况）。 所以我对生成设置很感兴趣。

2 线性代数++

与特征学习最直接相关的数学问题与线性代数的扩展有关。 回想一下，你的大一线性代数课包括以下三个主要方法。 (a) 解线性方程。 $Ax = b$。 (b) 计算秩，我们也可以将其视为矩阵分解：给定一个 $n \times m$矩阵 M，将其重写为 $M = AB$，其中 A为 $n \times r$， B为 $r \times m$，且 r尽可能小。 (c) 特征值/特征向量和奇异值/奇异向量。 例如，每个对称矩阵 M都可以重写为

$$\sum_i \lambda_i u_i u_i^T,$$

其中 u_i是特征向量， λ_i是特征值。 这被称为谱分解（如果矩阵不对称，则类似的表达被称为奇异值分解）。

线性代数$++$是我对上述问题的名称，其中包含以下任意子集的扩展：（i）要求解的某些坐标为非负数。 （ii）要求解具有指定数量或非零模式。 （iii）在存在某种噪声的情况下求解。

例子2如果我们需要解决 $Ax = b$的问题，同时要求 x为非负数，则相当于线性规划问题，该问题在1979年才被发现可以在多项式时间内解决。

如果我们需要解决 $Ax = b$的问题，同时要求 x只有 k个非零元素，则这是稀疏恢复问题，该问题是NP难的。

如果我们需要在存在坐标噪声的情况下解决 $M = AB$，我们可能正在寻找期望的 A，B，使得我们最小化

$$\sum_{ij} |M_{ij} - (AB)_{ij}|^2.$$

通过将SVD截断到前 r个项来最小化这个值。 我们将最佳秩为$_k$的近似值 M表示为 Mk.

你在大一线性代数课上从未见过这些扩展，因为它们是
NP-hard问题。但它们在机器学习领域中无处不在。

3个线性化模型

特征的概念通常在某种线性化的设置中出现：主题模型、稀疏编码、
稀疏恢复。 我们将在后面的讲座中看到这些内容。

讲座的其余部分涵盖了稀疏恢复（Moitra的讲座笔记）；SVD（我在COS 521的讲座笔记）以及使用SVD进行聚类（Hopcroft-Kannan书）。

COS 598C: 检测重叠社区，以及学习深度网络和字典的理论框架

讲师：Sanjeev Arora
记录员：Max Simchowitz

2015年4月8日

今天我们介绍一些关于可证明学习深度神经网络和字典的思想，这两个模型非常重要（且相关）。它们的共同点是一种用于检测网络中重叠社区的简单算法。虽然社区检测通常被认为是发现大型社交网络等结构的一种方法，但在这里我们将其作为一种通用的算法工具来理解潜在变量模型的结构。学习深度神经网络和字典的算法首先通过识别变量之间的相关性，并使用图形表示这些成对的相关性。然后它使用社区发现来揭示潜在的连接结构。

1 在网络中检测重叠社区

社区检测在已知社区是不相交的情况下已经得到了很好的研究。我们在之前的讲座中讨论了随机块模型。具体情况是我们给定了$G = (V, E)$，其中顶点V被划分为两个集合S和S^c，并且在S和S^c内部的边以概率p绘制，在S和S^c之间的边以概率q绘制，使得p-q=$\Omega(1)$。然后，只要$\min(S, S^c)$=$\Omega($ $\sqrt{\text{通过}}$使用SVD或半定规划[6]，我们可以轻松地恢复 S和 S^c。

然而，当社区重叠时，这个问题似乎无法通过SVD解决。回顾一下符号表示，让 G=(V, E) 是我们的图，让我们将用户分配给（可能多个）社区 $C_1, \ldots, C_m$。在最简单的设置中 - 例如在字典学习问题中出现的设置 - 当且仅当存在一个社区 C_j使得（v_1, v_2） $\in E$时，才有 $v_1 \in C_j$和 $v_2 \in C_j$。在这种情况下，我们可以将 C_j识别为 G的子图，所有这些子图都是完全图，而 G恰好是这些完全图的并集。如果图是任意完全图的并集，我不知道如何找到这些完全图的算法。

幸运的是，在字典学习的环境中， G的结构不是敌对确定的。相反，我们假设顶点 $v \in V$在社区中分布相对均匀，并且每个 v不属于太多的社区（比如，没有中心节点）。我们将 G的生成过程形式化如下：

定义 1.1（与重叠社区对应的种植问题）。设 G =(V, E)，其中 $|V| = N$。假设存在 m个社区 $C_1, \ldots$，每个顶点都以均匀随机的方式分配到 k个社区。最后，如果u, v属于同一个

社区，则$Pr[(u,v)$ 是一条边 $] = p \geq 0.9$。 如果它们没有任何重叠的社区，则没有边。

备注。如果 $p=1$，则 $C_1, \ldots,$ ，C_m是团，我们回到了字典学习的聚类设置。

我们生成过程的好处是它允许局部搜索启发式算法，如[2]中所述。 首先，设置 $T = k$ N/m，这大致是每个社区的预期大小。 现在，如果(u,v)在一个社区 C_i中，那么它们之间共享的边的预期数量约为 pT，所以根据Chernoff不等式，至少有 $.9pT$个顶点与 u和 v高概率连接。

另一方面，假设(u,v)不在同一个社区中。 然后，虽然它们不一定通过一条边连接，但可能存在顶点 w，使得(u,w)在一个社区 C_i中，而(u,v)在另一个社区 C_j中，从而产生从 u和 v到 w的边。

现在，有多少这样的虚假边？ 也就是说，在忽略共享社区结构的情况下，边在任意两个顶点之间出现的概率是多少？嗯，有
$\binom{N}{2}$种选择顶点对的方式，并且图中的边数不会超过一个社区中边的数量之和，该社区集中在p $\binom{T}{2}$使用标准的切尔诺夫论证。 对所有 m个社区取并集，我们可以看到两个顶点之间存在边的概率不会超过

$$p_0 := pm\binom{T}{2}\binom{N}{2}^{-1} \tag{1.1}$$

因此， w和 u之间以及 w和 v之间的虚假边的数量大约集中在 $p^2{}_0N$。 因此，为了区分仅共享虚假边和由于共同社区成员资格而共享非平凡边的 u和 v之间的区别，我们希望确保

$$p_0^2 N \ll pT \tag{1.2}$$

这相当于对要求进行约束

$$\frac{N \cdot T^4 m^2}{N^4} \ll T \iff m \ll (N/T)^{3/2} \iff m \ll (m/k)^{3/2} \tag{1.3}$$

大多数连接 u和 v之间的边是因为它们在同一个社区中。
因此，我们可以通过考虑共同边的数量来贪婪地将顶点分配给社区。

1.1 符号

给定一个向量 $x \in \mathbb{R}^n$，我们将用 $x(\ i)$表示它的第i个元素。 我们将两个向量$x, y \in \mathbb{R}^n$的内积表示为 $\langle x, y \rangle$，或者 $x^T y$互换使用。 给定一个矩阵$A \in \mathbb{R}^{n \times m}$，我们用 $A\ i$表示它的第$_i$列。

2个神经网络

在详细阐述[2]中描述的社区发现算法的应用之前，我们将简要介绍神经网络的世界：这可能是当代机器学习中最流行的工具之一。 从一个非常基本的角度来看，神经网络模仿了物理大脑的结构。 模拟大脑的一种抽象方法是将生物神经元网络视为大型图，其中顶点是神经元，边是突触（或其他形式的连接）。 这样一个神经网络的状态由每个神经元被激活的电位（以及其他因素如电流）描述，而突触决定了电位从一个神经元节点传递到下一个神经元节点的程度。

受常见的实现实践和理论可行性的启发，我们将研究分解为 L层的人工神经网络。 因此，我们可以用一个 L元组的向量 $x^{(1)}, \ldots, x^{(L)}$来描述网络在给定时间的状态，其中向量 $x^{(l)} \in R^{N_l}$的条目记录了对应层中神经元的电位。 例如， $x^{(2)(1)}$是第二层中第一个神经元的电位。我们将 $x(1)$称为顶层， $x^{(L)}$称为底层。

神经网络的迷人之处在于潜在向量 $x^{(l)}$在不同层之间的关系。 在生物神经组织中，电位和化学信号不断交换。 在我们的设定中，我们假设在离散的面试中，$t= 1, \ldots T,$ 自然会绘制顶层的潜在向量 $x^{(1)}\ t$。 然后，每个连续层中的潜在向量 $x^{(l)}$由一个确定函数的噪声反对给出，该函数是由潜在向量 $x^{(l-1)}$在层 $x^{(1)}\ t$ 中的确定函数的噪声反对给出。

我们将这种潜在向量的传递建模为

$$x^{(l+1)} = h(A^{(l)}x^{(l)}) \tag{2.4}$$

其中 h是一个（通常是非线性的）函数，它以逐个元素的方式对每个元素进行操作，并且 $A^{(l)} \in \mathbb{R}^{N^{(l)} \times N^{(l+1)}}$是一个指定一个层中的潜在向量如何传递到下一层的矩阵。 等价地，我们可以将 $A^{(l)}$看作是一个二分图 $G^{(l)}$的邻接矩阵，其边表示神经元之间的连接。 在接下来的内容中，我们将在 $G^{(l)}$的顶点和 $x^{(l)}$的条目之间进行交换，这两者在语义上对应于第 l 层的神经元。

为了简化符号并便于阐述，这些笔记的大部分内容将集中在只有两层的学习网络上：一层由稀疏向量 x编码，对观察者隐藏，并从适当的生成过程中绘制，另一层由密集向量 y 编码，可以被观察到。 我们将使用 G和 A来表示 x和 y之间的连接图以及其邻接矩阵。

3 字典学习、神经网络和社区发现

我们可以想象，即使对于任意的非线性函数h，两层问题也相当困难。 因此，从考虑 h只是恒等函数的简单情况开始是有意义的；也就是说 $Ax = y$。 这个问题被称为字典学习，邻接矩阵 A被称为字典。

在字典学习问题中，我们给出了观察到的电位样本 y_1，...，y_N以及我们的目标是重构 A和隐藏样本 x_1，...，x_N以便最小化误差。

$$\min_{A,\{x_i\}} \|Ax_i - y_i\|_2^2 \tag{3.5}$$

一般来说，这个问题是极度过度确定的。 实际上，如果 x的维度大于 y的维度，那么可以轻松重构 A和 x_i，使得 $Ax_i = y_i$。
为了使问题有意义且可解，我们需要假设 x_i具有一些额外的结构。 在这里，我们假设样本 x是稀疏的。

恢复稀疏的 x_i有两个动机。 第一个动机是经验性的 - 生物神经元往往显示出稀疏的激活模式。 更广泛地说，稀疏性是捕捉高维数据中“潜在简单性”或“隐藏结构”的一种直观假设。 第二个动机是，在假设稀疏性的情况下，我们可以利用稀疏恢复和压缩感知的见解，对字典矩阵 A有一定的条件。 回想一下，具有低列内积的矩阵被称为不相干的：

定义 3.1.设 A为具有列 A_i的矩阵，使得 $\|A_i\| = 1$. 如果$|A_i^T A_j| \leq \frac{\mu}{\sqrt{n}}$ 不相干

现在，如果我们完全知道 A，并且 A足够不相干，那么我们有以下结果

定理 3.1(压缩感知，简要陈述).设 A为具有单位范数列的矩阵，使得 $|\langle A_i, A_j\rangle| \leq \frac{1}{2k}$. 假设给定 $y = Ax$其中 x是 k-稀疏的。 那么， x是唯一的 k-稀疏向量，使得 $y = Ax$. 因此， x可以在多项式时间内恢复。

指导性的洞察是，对于 $\mu/\sqrt{n}$不相干的字典， $A^T A \approx I$，因为对角线的值上界为 $\mu/\sqrt{n}$。 请注意，从谱意义上讲，这个近似不一定是很好的，因为 $A^T A - I$可能有大小为 $\Omega(\mu/\sqrt{n})$ 的$n^2 - n$个元素，因此 $\|A^T A - I\|$可能是 $\Omega(\mu\sqrt{n})$。

但仅仅看谱范数并不能充分利用稀疏性：实际上，$\|A^T A - I\| = \max_{\|z\|:1} z^T(A^T A - I)z$，如果 $A^T A - I$的元素都在 $\mu/\sqrt{n}$附近，那么这个最大值将在 $z^* \approx$处达到 $\frac{1}{\sqrt{n}}(1, \ldots, 1)$. 然而，如果我们强加条件 z^* 是 k-稀疏的，情况就有些不同。 定义半范数 $\|z\|_0 := \sum_i I(z_i = 0\}$, 并且让 $B_0(k) := \{z \in \mathbb{R}^n : \|z\| \leq 1, \|z\|_0 \leq k\}$。很容易证明

$$\sup_{z \in B_0(k)} z^T(A^T A - I)z \leq k\mu/\sqrt{n} \tag{3.6}$$

这种对 k=稀疏向量的限制引出了压缩感知文献中的“受限等距性质”。 实际上，如果 $k\mu/\sqrt{n} < 1/2$，那么对于所有$2k$稀疏向量 z，矩阵 A实际上是“可逆”的，如果 $k\mu/\sqrt{n} = o(1)$，那么$\langle Az, Az\rangle = \|z\|^2 + z^T(A^T A - I)z \approx \|z\|^2$对于所有 k-稀疏 z成立。 更准确地说，我们可以证明以下引理：

引理 3.2.设 z_1和 z_2是两个 k-稀疏向量，且 A的列具有单位范数，
则 $\langle Az_1, Az_2\rangle = \langle z_1, z_2\rangle \pm \frac{2k\mu}{\sqrt{n}}\|z_1\|\|z_2\|$.

证明。通过重新标记 A的列以及 z_1和 z_2的元素，我们可以假设
z_1和 z_2都支持在索引 $[2k] := \{1, \ldots, 2k\}$上。因此，

$$\begin{aligned} \langle Az_1, Az_2 \rangle &= \sum_{i\in[2k]} \|A_i\|^2 z_1(i)z_2(i) + \sum_{i\in[2k]} \sum_{j=1\in[2k]} z_1(i)z_2(j)\langle A_i, A_j\rangle & (3.7)\\ &= \langle z_1, z_2\rangle + E & (3.8)\end{aligned}$$

其中 $E := \sum_{i\in[2k]} \sum_{j\in[2k]} z_1(i)z_2(j)\langle A_i, A_j\rangle$.

$$\begin{aligned} |E| &\leq \sum_{i\in 2k} \sum_{j=i\in[2k]} |z_1(i)||z_2(j)| \cdot |\langle A_i, A_j\rangle| & (3.9)\\ &\leq \sum_{i\in 2k} \sum_{j\in[2k]} |z_1(i)||z_2(j)| \cdot |\langle A_i, A_j\rangle| & (3.10)\\ &\leq \frac{\mu}{\sqrt{n}} \sum_{i\in[2k]} \sum_{j\in} |z_1(i)||z_2(j)| \leq \frac{\mu}{\sqrt{n}} \|w_1 w_2^T\|_F & (3.11)\end{aligned}$$

其中 $w_1 \in \mathbb{R}^{2k}$具有 $w_1(i) = |z_1(i)|$对于所有 $i \in [2k]$，而 w_2类似地定义为 z_2，并且$\|\cdot\|_F$表示Frobenius范数。 因为 $w_1 w_2^T$是一个$2k \times 2k$矩阵，所以我们有 $\|w_1 w_2^T\|_F \leq 2k\|w_1 w_2^T\|$，其中$\|\cdot\|$表示谱范数。 但是$\|w_1 w_2^T\| = \|w_1\|\|w_2\| = \|z_1\|\|z_2\|$，因此

$$|E| \leq \frac{2k\mu}{\sqrt{n}} \|z_1\|\|z_2\| \quad (3.12)$$

□

3.1 字典学习的形式模型

为了鼓励稀疏性，Olshausen和Field [5]设计了一种交替梯度下降算法来最小化以下目标函数：

$$\min \sum_{i=1}^{N} |y_i - Ax_i|^2 + \sum_{i=1}^{N} \text{惩罚}_K(x) \quad (3.13)$$

正如我们所提到的，未经惩罚的字典学习是高度欠定的。
因此，Olshausen和Field引入了惩罚项 - 例如， l_1-正则化 - 来鼓励稀疏性并确保（或至少促进）模型的可识别性[5]。 在[3]中，Arora，Ge等人描述了一种基于Olshausen和Field的交替最小化算法，用于学习方程3.13中的目标函数。在这些笔记中，我们将限制我们的注意力在即将描述的“重叠社区方法”上。 在任何一种情况下，[3]中的交替最小化算法和[2]中的重叠社区检测方法将使用大致相同的假设，我们将其形式化如下：

1. 字典 $A \in \mathbb{R}^{n\times m}$具有单位范数列，并且具有$\frac{1}{\sqrt{n}}$不相干列，即： $|\langle A_i, A_j\rangle| \leq \frac{\mu}{\sqrt{n}}$.

2. 我们关注的是 $m \geq n$的情况，并且我们要求 $\|A\| = O(\sqrt{m}/\sqrt{n})$.

3. 每个 x具有恰好 k个非零坐标，均匀地从 $\{1, \ldots, m\}$中抽取（这可以稍微放松，如[3]中所述）

4. 对于每个坐标 x_i，其条件独立，并且 $x_i|x\ i=0$服从亚高斯分布，方差代理为 $O(1)$，并且存在一个常数 C- 对所有$_i \in [m]$都是通用的 - 使得 $|x_i||x_i=0 \geq C$几乎必然成立。. 例如，我们可以将 $x_i|x\mathrm{i}=0$视为从[1,10]或$[-10,-1] \cup [1,10]$均匀抽取。

5. 我们将从假设几乎必然成立的情况开始。 本文中的论证也适用于 $\mathbb{E}[\ x_i] = 0$的情况。 给定样本 $y_1 = Ax_1$和 $y_2 = Ax_2$，根据引理3.2可得

$$\langle y_1, y_2\rangle = \langle x_1, x_2\rangle \pm \|x_1\|\|x_2\|\frac{k\mu}{\sqrt{n}} \tag{3.14}$$

根据次高斯集中定理，高概率下成立 $\|x_1\|\|x_\| = \tilde{O}(k)$，因此只要 $k^2\mu/\sqrt{n}$大致为(1)，那么

$$\langle y_1, y_2\rangle = \langle x_1, x_2\rangle \pm(1) \tag{3.15}$$

如果我们假设 x_1和 x_2逐个非负，则

$$\begin{aligned}
\langle x_1, x_2\rangle &= \sum_{i\in \mathrm{Supp}(x_1)\cap \mathrm{Supp}(x_2)} x_1(i)x_2(i) && (3.16)\\
&\geq C|\mathrm{Supp}(x_1)\cap \mathrm{Supp}(x_2)| && (3.17)\\
&\geq CI(\mathrm{Supp}(x_1)\cap \mathrm{Supp}(x_2) = \emptyset) && (3.18)
\end{aligned}$$

因此，高概率下成立 $\langle y_1, y_2\rangle \geq C/2$ 如果且仅当 x_1和 x_2有一个非零的共同元素。 我们将在一个非正式引理中陈述这一点：

引理 **3.3.**如果 $k^2\mu/\sqrt{n}$大致等于 $o(\log n)$，那么非常高的概率下 $\langle x_1, x_2\rangle \geq C/2$ 如果且仅当 x_1 和 x_2 共享一个非零的元素。

这个观察使我们能够将问题从一个分析问题转化为一个组合问题。实际上，给定 N个观察值 $y_1, \ldots, y_N$, 让每个观察值 y_i对应于图 $G=(V,E)$ 中的一个顶点 i。 我们只有当 $\langle y\ i,\ y\ j\rangle \geq C/2$ 时，才在顶点$_i$和$_j$之间画一条边。 通过上面的讨论，高概率地，只有当 $x\ i$和 $x\ j$共享一个非零元素时，才会在$_i$和$_j$之间画边。 通过取并集的方式，以下命题成立：

引理**3.4**. 高概率下 $G \simeq \tilde{G}$，其中 $\tilde{G}$ 是由顶点$i \in [N]$连接所有索引i, j的边的图现在我们给出了一个更直观的方式来描述$\tilde{G}$：令 $C_1, \ldots, C_m$是定义的集合，使得 $C_j := \{_i \in [N] : xi($

$_j) = 0\}$. 我们将这些集合称为“社区”，因为所有$_i \in C_j$共享一个非零的共同元素。 根据每个样本 $x\ i$中恰好有 k个非零元素的假设，每个顶点$_i$被分配

到精确 k 个社区 $C_{j_1}, \ldots, C_{j_k}$。此外，根据集合 C_j的定义，如果x_i和x_j共享一个非零元素，则它们都属于同一个社区：$\tilde{G}$ 是通过在至少一个社区中的顶点之间绘制边来构建的图。因此，为了以很高的概率恢复x_i的稀疏模式，引理3.4告诉我们，图 G的边是由样本 y_i 和 y_j的内积构成的，其精确地由其顶点的社区分配生成。

3.1.1 均值为零的情况

如果x_i的元素均值为零，则论证略有不同：

$$\sum_{i \in \mathrm{Supp}(x_1) \cap \mathrm{Supp}(x_2)} x_1(i) x_2(i) \tag{3.19}$$

由于x_1和 x_2的元素具有相互抵消的符号，因此Cdue的绝对值可能远小于C。然而，以概率 $\Omega(k^2/m^2)$，$\mathrm{Supp}(x_1)$ 和 $\mathrm{Supp}(x_2)$最多只会有一个重叠的元素，因此如果我们愿意接受一个小的（但不像 $n^{-\omega(1)}$那样小）错过边的概率（这也与x_2x_1的公共支持集不独立），我们可以忽略这些相关性。另一方面，我们可以改进引理3.2中的界限，由于抵消的存在。实际上，我们有

$$\langle y_1, y_2 \rangle - \langle x_1, x_2 \rangle = \sum_{i=j} \langle A_i, A_j \rangle x_1(i) x_2(j) \tag{3.20}$$

使用界限 $|\langle A_i, A_j \rangle| \leq \mu/\sqrt{n}$，这个项的平均值为零，由于抵消，矩大致为 $O(k\mu/\sqrt{n})$，因此，$\langle y_1, y_2 \rangle = \langle x_1, x_2 \rangle + \tilde{O}(k\mu/\sqrt{n})$，因此我们的误差大致下降了一个因子 $\sqrt{k}$.

3.2 字典学习的归约到社区检测

鉴于我们的社区检测算法，我们已经给出了如何高效地恢复潜在样本 x_i的稀疏模式的概述。我们展示了如何使用这种技术来恢复字典 A，参考文献[2]。请注意，一旦 A被检索出来，我们可以使用更标准的稀疏恢复技术（近似地）恢复潜在信号向量 x。

基本思想是，矩阵 A的第 j列 A_j应该大致上等于所有样本$_i$的平均值，其中第 j个元素是活跃的，即$_i$的第j个元素大于等于0。因此，一个恢复 A的第一次尝试就是简单地计算以下平均值：

$$A_j := \frac{1}{|C_j|} \sum_{i : y_i \in C_j} y_i \tag{3.21}$$

不幸的是，在 x_i的均值为零的情况下，我们有 $\mathbb{E}[y_i] = \mathbb{E}[Ax_i] = A\mathbb{E}[x_i] = 0$。在 x_i的均值不为零的情况下，我们会从与 j不相等的索引处的样本的非零元素中得到许多虚假贡献，即$y_i = A_j x_i(j) + \sum_{j' \neq j} A_{j'} x_i(j')$.

一个更好的想法是，相反，我们可以看一下对于 $E[yy^T]$ 的最佳秩为1的近似值： $y \in C_j$。为了简单起见，我们首先处理均值为零的情况。首先注意，因为问题对 A的列的排列是不变的，所以只需证明一个能够恢复与社区 C_1对应的 A的列 A_1的算法。我们的策略是计算所有具有活跃第一列的样本y_i的经验协方差矩阵的最佳秩为1的近似值，即 $y_i \in C_1$。

$$M_1 := \frac{1}{\#y : y \in C} \sum_{y \in C} [yy^T] \tag{3.22}$$

也就是说，所有 $y \in C$的经验平均值 yy^T的近似值。首先，让我们证明 M_1是 $A_1 A_1^T$的一个很好的近似值，差距在一个常数因子内：

$$\begin{aligned} M_1 &= \mathbb{E}[x(1)^2 A_1 A_1^T] + \mathbb{E}[\sum_{i \geq 2} x(i)^2 A_i A_i^T] \\ &+ \mathbb{E}[\sum_{i \geq 2} x(i)x(j)(A_1 A_i + A_i A_1)] + E[\sum_{i,j \geq 2} x(i)x(j) A_i A_j] + \text{统计误差} \\ &\approx \Theta(A_1 A_1^T) + O(\frac{k}{m} \sum_{i \geq 1} A_i A_i^T) + \tilde{O}(k^2/\sqrt{N}) \end{aligned}$$

这里 N是使用的样本数量， $O(f)$ 表示一个其谱范数被 Cf限制的量，其中$C>0$，而 $O(M)$（或 $\Theta(M)$）表示一个根据半正定锥的规范顺序小于 CM（或小于 CM且大于 cM）的量。$\preceq$of the semidefinte cone . 第一项来自于 $\mathbb{E}[x(1)^2 | y \in C] = \Theta(1)$，第二项来自于 $\mathbb{E}[x(i)^2 | y \in C] = O(k/m)$。请注意，第二个误差项是系统性的 - 它不依赖于算法使用的样本数量。

剩余的误差项 $\tilde{O}(k^2/\sqrt{N})$ 是统计性质的，并且来自于所有项与它们的期望值的偏差。通过在所有项都很小的条件下进行条件概率计算，很容易建立起 $\tilde{O}(k^2/\sqrt{n})$ 的界限，然后使用Chernoff不等式来完成。这个界限可以改进为$\tilde{O}(k/\sqrt{N})$，但这个改进会影响算法的样本复杂度。另一方面， $O(\frac{}{mk} - \sum_i A_i A^T{}_i)$ 决定了在 k和 m满足最佳秩一逼近算法准确恢复底层字典的条件下。

首先，我们将在字典学习文献中使用一个标准假设，即$\|A\| = O(\sqrt{m}/\sqrt{n})$。在这个条件下，有$\|(\frac{}{mk} \sum_{i \geq 2} A_i A^T{}_i)\| = O(k/n)$。我们将假设样本数量足够大，以统计误差也受到 $O(k/n)$ 的支配。因此， $M_1 \propto A_1 A_1^T + E$，其中 E的范数为 $O(k/n)$。N现在让$\hat{A}_1$成为 M_1的顶部特征向量。我们可以通过引用Wedin定理来证明$\hat{A}_1$是 A_1的一个很好的估计值，Wedin定理是线性代数中的一个基本结果，它限制了PSD矩阵 A的顶部特征向量与 $A+E$的距离，其中 E是一个小的扰动。因为 A_1是 $A_1 A_1^T$的顶部特征向量，Wedin定理将帮助我们证明 M_1的顶部特征向量也应该接近 A_1：

定理3.5.设 v_1是*PSD*矩阵 A的顶部特征向量， v_2是 $A+E$的顶部特征向量。设 θ为 v_1和 v_2之间的角度。则 $\sin\theta \leq \frac{2\|E\|}{\sigma_1(A) - \sigma_2(A)}$。

作为推论，我们得到了 A和 $A+E$的（归一化的）顶部特征向量之间的欧几里德距离的有界性。

推论。设 A是一个范数为$_1$的秩一矩阵，其顶部特征向量为 v 1， v_2是 $A+E$的顶部特征向量。 只要 $\|E\| = o(1)$ $\|v_1 - v_2\| \leq \sqrt{1/2}$, $\|v_1 - v_2\| \leq 2\|E\|$

证明. 因为 $\sigma_1(A_1 A_1^T) = 1$，且 $\sigma_2(A_1 A_1^T) = 0$，所以 $\sin\theta(v_1, v_2) \leq 2\|E\|$.
因为 $v_1^T v_2 = 1 - \|v_1 - v_2\|^2$，所以我们有

$$\sin\theta(v_1, v_2) = \sin \arccos(v_1^T v_2) = \sqrt{1-(v_1^T v_2)^2} = \sqrt{2\|v_1 - v_2\|^2 - \|v_1 - v_2\|^4} \quad (3.23)$$

$$= \sqrt{2}\|v_1 - v_2\|\sqrt{1 - \|v_1 - v_2\|^2} \quad (3.24)$$

因为 $E = o(1)$，所以$\sin\theta(v_1, v_2)$，因此 $\|v_1 - v_2\|$也必须是(1)。
因此， $\|v_1 - v_2\| \leq \sqrt{2}\|E\| / \sqrt{1 - \|v_1 - v_2\|^2} \leq 2\|E\|$。 □

根据这个推论，可以立即得出 $\|\hat{A}_1 - A_1\| \leq O(k/n)$。 因此，给出足够多（但仍然是多项式数量的）样本，我们可以轻松地恢复 A的列，误差为 k/n。

4 深度神经网络的无监督学习

让我们从字典学习的受限设置返回到更一般的神经网络设置。 除了它们的成功之外，深度神经网络受欢迎的主要原因之一是最后一层似乎捕捉到了“有意义的特征”。 例如，在视觉问题中，神经网络学习到的对象的像素表示通常非常接近该对象的形状。 而且，在许多应用中，可以使用最后一层学习到的特征训练非常有效的分类器（例如使用SVM或逻辑回归）。 事实上，如果我们为一个分类任务（比如区分猫和狗）训练一个多层神经网络，然后重新训练最后一层来学习一个新任务（比如区分鸟和蜜蜂），而不重新训练大部分隐藏层的参数，重新训练后的网络在新的分类任务上仍然非常成功。 这表明神经网络的深层表示学习到了大部分相关信息，或者至少学习到了足够的信息来构建有效的分类器。

这表明深度神经网络在图像本身中捕捉到了一些固有结构，这让人们对隐藏层对应于可以从未标记数据中学习到的自然“特征”抱有希望。 （相比之下，最近的成功案例涉及利用大量标记图像。）深度神经网络的无监督训练是这个领域的一个圣杯，这个领域的主要研究人员一直试图定义与深度神经网络相对应的生成模型。 这个探索非常符合我们在之前的讲座中讨论的判别-生成对的精神（例如，朴素贝叶斯分类器是逻辑回归的生成模拟）。

如果我们从判别的角度转向生成的角度，我们可能会想知道神经网络可以从数据中提取出什么样的结构。 现在让我们考虑一个两层神经网络，其顶层编码为向量 x，底层编码为向量 y。

与其想象一个线性映射 A，它将稀疏输入 x映射到一个稠密输出 y，我们现在想象一个编码函数 $E(\cdot)$，它将稠密输出 y编码为一个稀疏输入 x。在线性情况下，我们有 $y = Ax$，所以 $x \approx A^T y$。 在一般情况下，我们建模$x = E(y) = h(A'y + b)$，其中 b是一个偏移函数， $h(\cdot)$是一个非线性映射，它对每个坐标都独立地起作用，例如， $h(\cdot)$可以是返回其参数的符号的函数。 同样，我们可以想象 x和 y的每个条目被视为二分图中的顶点， A是捕捉 x和 y之间边权重的邻接矩阵。

希望现在我们可以反转编码函数 $E(\cdot)$，并且实际上可以在存在噪声的情况下进行反转。 这激发了以下定义：

定义**4.1**(去噪自编码器)。给定一个邻接矩阵 A，自编码器由形式为 $E(y) = h(A'y + b)$的编码函数和形式为 $D(x) = h(Ax + b')$的解码函数组成。 如果自编码器对于噪声模型 $\xi \sim \mathcal{D}$具有解码鲁棒性，则称其为去噪自编码器，即：

$$E(D(h) + \xi) = h \text{ 的概率很高} \tag{4.25}$$

如果 $A' = A^T$，则称其为权重绑定。 这里 $D(h) + \xi$是 D受到噪声向量 ξ污染的简写。 这种污染不一定是加法的。

以下定理说明，如果逐元素非线性函数 $h(\cdot)$是符号函数，并且 A足够稀疏，那么[1]表明两层神经网络实际上是一个去噪自编码器:

定理**4.1**(松散陈述).考虑一个具有稀疏二分图 G和邻接矩阵 A的两层神经网络，边权重均匀分布在 $[-1, 1]$之间。 假设潜在样本 x是二进制的，支持集为 S。 最后，假设 $y = sign(Ax)$。
那么存在一个 b'，使得对于这对 $E(\cdot)$和 $D(\cdot)$来说，它们构成一个去噪自编码器，其中 $E(\cdot) = sign(A^T y + b')$。

事实上，我们可以以很高的概率学习编码/解码函数：

定理**4.2.**在一些正则性假设下，存在一个多项式时间算法来学习稀疏边权重的两层神经网络的编码和解码函数，这些边权重在 $[-1, 1]$中均匀抽取。

证明。为了保持直观，我们假设我们有一个无权二分图，该图从所有给定的顶点集合的x和 y的条目中均匀抽取。 我们假设 $E(\cdot)$ 和 $D(\cdot)$ 没有阈值函数，所以 $b = b' = 0$。 我们还假设 $\backslash x_i$是均匀抽取的、 k-稀疏的二进制向量，其中$\backslash x = \rho n$， ρ是一个小的值。

让我们从学习邻接矩阵 A，或者等价地，图 G开始。 什么是社区？ 它们是具有共同邻居的节点子集。 那么当两个节点有一个共同邻居时会发生什么。 如果u, v有一个共同邻居，那么$Pr[u, v$都是1$] \geq \rho$。 所以$Pr[$u, v都是1$] \leq (pd)^2$。 所以如果 $\rho \gg (\rho d)^2$，我们可以以很高的概率恢复社区。

现在让我们描述如何恢复样本的条目 x。 关键的直觉是，如果 x的一个条目，比如 x_1是活跃的，那么它的一些邻居 y_i的数量也会是活跃的。

好的。 因此，我们可以通过确定其邻居索引 y中是否有一定阈值以上的条目来恢复 x_1。 这些算法的保证来自以下观察：均匀抽取的稀疏二分图在很大概率上是扩展器。让我们更具体一些：

设 U表示与 x的条目对应的顶点集， V表示与 y的条目对应的顶点集。 给定 $u \in U$，设 $F(u)$表示它在 V中的所有邻居。最后，对于某个集合 $S \subset U$，设$U\,F(u, S)$为 u相对于 S的唯一邻居的集合，即

$$U\,F(u,\,S) := \{v \in V : v \in F(u),\, v \notin F(S - \{u\})\} \qquad (4.26)$$事实证明，对于一个随机生成的二分图和足够小的集合 S，对于每个 $u \in U$，在$U\,F(u, S)$中， u的邻居的总数至少是其总邻居数的$9d/10$。 因此，如果一个条目 x_i不活跃，我们预计在 V中，至多有$2d/10$的邻居是活跃的。 因此，我们可以通过设置

$$x_i = \text{阈值}_{2d/10}\ (\ x_i\text{的活跃邻居数}) \qquad (4.27)$$

□

也许更令人惊讶的是，[1]表明可以通过先学习最底层，然后向上移动来学习多层神经网络中的连接图 $G^{(l)}$：

定理4.3（推广到深度神经网络） 给定一个具有层x（^{l}）和加权连接图G（^{l}）的深度神经网络，其中期望度为d（^{l}），边权重均匀分布在[-1, 1]之间，并且顶层的样本是具有均匀稀疏支持大小为ρn的二进制向量。. 那么，如果ρ足够小，并且度数d（^{l}）增长不太快，则可以高概率地学习到真实的图G（^{l}）和相应的样本x。 事实上，可以通过从底层推断出次底层，然后逐层向上移动来学习它们。

参考文献

[1] Sanjeev Arora, Rong Ge, Aditya Bhaskara, Tengyu Ma. “Provable Bounds for Learning S ome Deep Representations.” 机器学习杂志, 卷32, 2014.

[2] Sanjeev Arora, Rong Ge, Ankur Moitra. “学习不连贯和过完备字典的新算法”学习理论会议, 2014年。

[3] Arora, Sanjeev等。 “稀疏编码的简单、高效和神经算法” *arXiv*预印本*arXiv:1503.00778*, 2015年。

[4] Candes, Emmanuel J. “受限等距性质及其对压缩感知的影响。”数学学报346.9 (2008): 5 89-592。

[5] Bruno A. Olshausen和David J. Field。 “使用过完备基函数的稀疏编码： $V1$所采用的策略。”视觉研究, 37:3311–3325, 1997a年。

[6]

张量分解+主题建模的另一种方法

讲师：Sanjeev Arora
记录员：Holden Lee

2015年4月13日

今天我们讨论张量分解，这是一种用于学习潜在变量模型的通用工具。 然后我们转换话题，讨论我们在之前的讲座中看到的主题建模算法的最新改进。

0.1 张量分解

张量分解是张量的谱分解的类比。

特征值/特征向量的好处是它们存在（好吧，在非对称矩阵的情况下是奇异值/向量），并且可以高效地计算它们。 对于一个对称的$n \times n$矩阵M，我们可以写成$M = \sum$

$$\lambda_i u_i u_i^T.$$

一个三维张量M是一个$n \times n \times n$数组。 将线性代数扩展到张量是非平凡的。许多关于张量的问题都是NP难的，比如秩（不容易定义）。

今天我们对张量感兴趣，我们保证有一个表示形式如$M = \sum \lambda_i u_i^{\otimes 3}$，其中 ui 是正交的。我们不知道 u_i，正在尝试恢复它们。 我们实际上可以类似于幂法恢复这些。 （回想一下，幂法重复设置 $x \leftarrow \frac{Mx}{\|Mx\|_2}$；如果前两个特征值之间有差距，则给出顶部特征向量。 运行时间与这个差距成反比。）

定义0.1：张量-向量乘积（也称为展平通过 x）定义如下：
Mx是矩阵，其中

$$(Mx)_{ij} = \sum_k M_{ijk} x_k.$$

现在

$$Mx = \sum \lambda_i (u_i \cdot x) u_i^{\otimes 2}.$$

这看起来像是一种谱分解：它将正交方向 u_i并通过 λ_i（$u_i \cdot x$）提升。 （在同构 $V \otimes V \sim = V \otimes V^*$, $u_i^{\otimes 2}$对应于 $u_i u^{T\,i}$。）

为什么这个有效？ 通过检查，$M\,x$的特征值是 $u_i \cdot x$，因为 u_i是正交的，谱分解是唯一的。 $M\,x$的分布近似为高斯分布，并且有很大的可能性$M\,x$有一个顶部特征值，与下一个特征值之间有显著的差距。

0.1.1 矩方法

在主题建模等领域，实际上我们使用的是矩方法。
一般的设置是我们进行采样

$$x \sim D := D(A)$$

其中 A是隐藏参数的矩阵；给定观测到的 X，我们尝试恢复 A。 我们可以考虑矩

$$\begin{aligned} \mathbb{E}X &= f_1(A) \\ \mathbb{E}(X^{\otimes 2}) &= f_2(A) \\ \mathbb{E}(X^{\otimes 3}) &= f_3(A) \\ &\vdots \end{aligned}$$

然后我们尝试解决这个非线性方程组。 很多机器学习可以这样思考。

数学家和统计学家研究过这样的问题：我们可以从第三阶矩或者第k阶矩中识别出什么分布？

回想一下，在主题建模中，在可分离性假设下，一个文档是从A中采样的，其中$w\in$Dir (α)。我们考虑了

$$XX^T = A\underbrace{\mathbb{E}[ww^T]}_{R}A^T$$

并使用了可分离矩阵分解。 我们确切地使用了二阶矩来恢复分布。

更多关于这个框架的内容，请参见[AGH+14]。
字典学习不是矩阵的方法；当$|\langle X, X'\rangle| \geq {}^{1}$时，我们在$X, X'$之间画边。
$\bar{2}$ 并对得到的图进行社区检测。

0.1.2 示例：相同球形高斯混合

考虑 k个高斯分布 $N(\mu_i, \sigma^2)$ 在 n维空间中 $(\mu_i \in \mathbb{R}^n)$，其中 σ^2已知。 令混合权重 w_i满足 $\sum^{k\ i}{}_{=1} w_i = 1$。. 为了选择一个样本，以概率 w_i选择 i，

并输出一个从 $N(\mu_i, \sigma^2)$ 中的样本。我们有

$$\mathbb{E}[X] = \sum_{i=1}^{k} w_i \mu_i$$
$$\mathbb{E}[X^{\otimes 2}] = \sum_{i=1}^{k} w_i \mu_i^{\otimes 2} + \sigma^2 I$$
$$\mathbb{E}[X^{\otimes 3}] = \sum_{i=1}^{k} w_i \mu_i^{\otimes 3}$$

假设我们将坐标移动，使得 $\mathbb{E}[X] = 0$，并且 μ_i是线性无关的。 如果我们可以对 $\mathbb{E}[X^{\otimes 3}]$进行张量分解，那么我们将得到 μ_i和权重 w_i。
然而，我们目前无法进行张量分解，因为 μ_i通常不正交。
我们必须先对向量进行白化处理。

0.1.3 白化处理

白化处理的思想是将形式为 $\sum w_i \mu_i^{\otimes 3}$的张量转换为 $\sum w_i \nu_i^{\otimes 3}$，其中 ν_i是正交的。令 $U = (\mu_1, \ldots, \mu_n)$，我们有

$$P = \sum_{i=1}^{k} w_i \mu_i^{\otimes 2} = U \operatorname{diag}(w_i) U^T.$$

（这不是谱分解，因为 U不是正交的。）谱分解是，假设

$$P = V D V^T$$

其中 V是正交的。假设 U, V是满秩的。 我们想要找到一个矩阵 A使得向量 $\nu_i := A\sqrt{w_i}\mu_i$是正交的，即 $AU \operatorname{diag}(\sqrt{w_i})$是正交的。这等价于

$$[AU \operatorname{diag}(\sqrt{w_i})][\operatorname{diag}(\sqrt{w_i}) U^T A^T] = 1 \iff APA^T = 1.$$

因此，取 $A = W^T$其中 $W = VD^{-\frac{1}{2}}$。然后

$$APA^T = D^{-\frac{1}{2}} V (VDV^T) V^T D^{-\frac{1}{2}} = I$$

需要的。
在高斯情况下，如果我们将 W应用于 $\sum_{i=1}^{k} w_i \mu_i^{\otimes 3}$，我们将得到

$$\sum w_i (W^T \mu_i)^{\otimes 3} = \sum \frac{1}{\sqrt{w_i}} \nu_i^{\otimes 3}.$$

（当然，我们实际上得到的是带有噪声的版本 $\sum_{i=1}^{k} w_i \mu_i^{\otimes 2}$， $\sum_{i=1}^{k} w_i \mu_i^{\otimes 3}$，所以如果我们想要进行适当的分析，我们必须考虑误差。）
（参见[BCMV13]，这是一个稍微不同的设置，超完备张量分解。）

0.2 基于SVD的主题模型方法（由安-德烈·里斯特斯基演示）

我们解释了Bansal、Bhattacharyya和Kannan [BBK]的一篇论文，该论文使用了SVD和其他一些技巧。 他们开发并证明了一种基于SVD的算法，该算法在某些假设下学习主题模型，并具有 L^1误差，包括捕捉词假设（锚词假设的一种弱化形式）。

我们建立符号表示。 设 k为主题数， n为单词数。 设 A为单词×主题的矩阵，给出每个主题的单词分布， W为主题×文档的矩阵。设 $M = AW$。 如果 $W_{\bullet i}$是 W的一列，则 M $\widetilde{\ }_{\bullet i}$根据 $M_{\bullet i}$给出的分布进行 m次抽样生成。 （m是每个文档中的单词数。）

目标是以 L^1误差恢复 A。 以前的研究，如Arora等人，以 L^2误差恢复。 请注意， L^2误差忽略了频率较小的单词，而且根据经验，很多单词的频率都很小。 此外，列是分布，因此自然范数是L^1。

0.2.1 假设

我们做出以下假设。 详细参数请参阅论文。

1. （主导主题）我们假设每个文档中都有一个主导主题：

   （a）对于每个文档 d，存在一个主题 t（d）使得 W_t（$_d$）$_{,d} > \alpha$。 对于所有其他主题 $t' = t$（d）， $W_{t',d} \leq \beta$，其中 $\beta - \alpha$足够大。

   (b) (每个主题出现为主导主题的次数足够多) 对于每个主题 t，存在 $\geq \varepsilon_0 w_0 s$文档 d，其中 $W_{t,d} \geq 1 - \delta$。

2. 定义 **0.2**： w是主题 t的一个关键词，如果对于所有的 $t' = t$， $A_{wt'} \leq \rho A_{wt}$，并且出现的概率不太小， $A_{wt} \geq \frac{8}{m\delta^2\alpha} \ln\left(\frac{20}{\varepsilon w_0}\right)$.

   主题 t的关键词占据了主题 t的词的很大比例

   $$\sum_{\text{关键词}w\text{是主题 }t\text{的一个关键词}} A_{wt} > \frac{1}{2}.$$

   (你可以替换 ^{1} $_{\overline{2}}$ 通过 p_0替换，后面的参数会依赖于 p_0。 为了简单起见，我们不这样做。 对于 p_0，存在一些绝对下界。

3. (几乎纯净的文档) 几乎纯净的文档只占很小一部分。 对于所有的 $i, \geq \varepsilon_0 w_0 D$，文档中 $Wt > 1 - \delta$的比例。

4. (无局部最小值假设) 设 $p_j(\zeta, t)$ 是文档中主题 t的概率，词 j出现 ζ次，即比例为 $\frac{\zeta}{m}$。那么

   $$p_j(\zeta, t) > \min(p_j(\zeta - 1, t), p_j(\zeta + 1, t)).$$

动机是有两种可能性：要么单词出现的概率随着 ζ的增大而衰减（例如，按幂律），要么它是一个流行词，直到某个频率后才衰减。

5. （主导混合）主题 i在文档中占主导地位的比例是 ^{D} $\overline{k}$，其中 k是文档的总数。

0.2.2 算法

直觉是主题模型就像是软聚类，软是因为每个文档不属于一个独占的聚类。

直观上，障碍是什么？ 假设某个词在簇1中的频率在$[0, \sigma]$之间，在簇2中的频率在[μ,1]之间，而在簇2中的波动要大得多。 然后聚类可能会将第二个簇分成两个部分。

这可以通过在聚类之前进行阈值处理来解决。 如果已知 μ，则通过 μ进行阈值处理：如果一个坐标$> \mu$，则将其设置为1，否则设置为0。 如果直接应用奇异值分解（SVD），则可以处理比先进行阈值处理时更少的噪声。

考虑以下问题。

问题**0.3**：给定一个随机的 $n \times n$矩阵 A，其中某个 $m \times m$子矩阵的概率 $\mathbb{P}(A_{ij} \geq \mu) \geq {}^{1}_{2}$，而其他元素为 $N(0, \sigma)$，找到该子矩阵（种植的团）。

解决方案。首先考虑朴素的奇异值分解解决方案。

思路是矩阵的谱范数明显大于其余部分的谱范数。

1. 设 C为子集（团），设 $\mathbf{1}_C$为特征向量。 然后（假设没有显著的负贡献）

$$\frac{\|A\mathbf{1}C\|}{\|\mathbf{1}_C\|} \sim \frac{\sqrt{K(K\frac{\mu}{2})^2}}{\sqrt{K}} = O(K\mu)$$

2. 随机部分的谱范数为 $\sqrt{n}\sigma$。 $^{-}$

只要 $K\mu \gg \sqrt{n}\sigma$，奇异值分解就能起作用。

$$\frac{\mu}{\sigma} \gg \frac{\sqrt{n}}{k}. \tag{1}$$

现在先考虑阈值处理：

1. 如果 $A_{ij} > \mu$，则将 $\widetilde{A}_{ij}$设为1；如果 $A_{ij} < \mu$，则将 $\widetilde{A}_{ij}$设为0。 在种植的团中，条目以概率 ^{1}为1。 $_{\overline{2}}$; 远离它的条目以概率1为$1 \sim e^{-}$ $^{\frac{\mu^2}{s\sigma^2}}$.

2. 现在我们将均值移回，使得非团体部分的均值为0。设 $\widetilde{A^{\sim}} = \widetilde{A} - e^{-}$ $^{\frac{\mu^2}{2\sigma^2}}J$, 其中 J为全1矩阵。

种植部分的谱范数为 $\left(\frac{1}{\sqrt{k}}\right)^2$ 随机部分的谱范数为
$\precsim \sqrt{n}e^{-\frac{\mu^2}{2\sigma^2}}$.

因此，在阈值处理后，我们可以在 $k \gg \sqrt{re^{-}}$的情况下解决问题 $^{-}$ $^{\frac{\mu^2}{2\sigma^2}}$, 即

$$e^{\frac{\mu^2}{\sigma^2}} \gg \frac{\sqrt{n}}{k}.$$

这比(1)中的范围更大。 □

算法如下（非正式）。

1. （选择阈值）对于所有的单词 j，按照以下方式选择阈值 ζ_j。 取 $\zeta_j \in \{0, 1, \ldots, m\}$,

$$\zeta_j = \text{argmax}_j \left\{ \left| \left\{ d : \widetilde{M_{wd}} > \frac{\zeta}{m} \right\} \right| \geq \frac{D}{k} \text{和} \quad \left| \left\{ d : \widetilde{f_{jd}} = \frac{\zeta}{m} \right\} \right| \leq \varepsilon \frac{D}{k} \right\}.$$

然后定义阈值矩阵

$$T_{wd} := \begin{cases} \sqrt{\zeta_w}, & \text{if } \widetilde{A_{wd}} > \frac{\zeta_w}{m} \text{ 并且 } \zeta_w \text{不太小} \\ 0, & \text{否则.} \end{cases}$$

2. 现在使用瑞士军刀 [KK10]。[1]

(a) 取 T, 进行排名 k-SVD, 并生成 $T^{(k)}$.

(b) 运行 k-均值的2-近似算法，得到临时聚类中心。

(c) 在列 S of B上运行Lloyd算法，起始点和中心如上所示。

3. 确定关键词。 (详见论文中的细节。)

4. 确定$(1-\delta)$-纯文档并获取主题-词混合。

分析中的一个关键点是要证明阈值化不会破坏聚类。 我们
需要使用非局部最小假设。

命题 0.4（引理 A1 在 [BBK] 中）：如果 $\sum_{\zeta \geq \zeta} {}_0 p_j(\zeta, i) \geq \nu$ 且 $\sum_{\zeta \leq \zeta_0} p_j(\zeta, i) \geq \nu$，那么 $p_j(\zeta_0, i) \succeq_{m\nu}$。

证明。令 $f(\zeta) := p_j(\zeta, i)$。 以下情况之一发生。

1。对于所有 $n \leq \zeta_i \leq \zeta_0$，有 $f(\zeta) \geq f(\zeta -$

2.$f(\zeta + 1) \leq f(\zeta)$ 对于所有 $m - 1 \geq \zeta \geq \zeta_0$.

[1]定理表明，当大于 $(1-\varepsilon)$ 的点满足接近条件时，算法有效。
M □ 在聚类 Tr 中满足接近条件，如果对于任意 $s = r$，A □ 在 μr 到 μs 线上的投影至少比 μs 更接近 μr。
这里 $\Delta rs = ck$ $\left(\frac{1}{\sqrt{n_r}+\sqrt{n_s}}\right)$ $\|M - C\|$其中 C由聚类中心组成
。

假设(1)。那么

$$\zeta_0 p_j(\zeta_0, i) \geq \sum_{\zeta \geq \zeta_0} p_j(\zeta, i) \geq \nu \implies p_j(\zeta_0, i) \geq \frac{\nu}{m}.$$

另一种情况类似。 □

引理0.5 (阈值化不能分离主导主题, 引理A3在[BBK]):
高概率下，对于固定的单词 w和主题 t，

$$\min(\mathbb{P}(\widetilde{A_{wd}} \leq \frac{\zeta_w}{m}; d \in T_t), \mathbb{P}(\widetilde{A_{wd}} > \frac{\zeta_w}{m}, d \in T_t) \leq O(m\varepsilon w_0).$$

其中 T_t由具有主导主题 t的文档组成。

参考文献

[AGH+14] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. 张量分解用于学习潜在变量模型。 ar*X*iv 预印本 *arXiv:1210.7559*, 15:1–55, 2014.

[BBK] Trapit Bansal，C Bhattacharyya和Ravindran Kannan。一种可证明的基于SVD的算法，用于学习主导混合语料库中的主题。第1-22页。

[BCMV13] Aditya Bhaskara，Moses Charikar，Ankur Moitra和Aravindan Vijayaraghavan。张量分解的平滑分析。 *arXiv:1311.3651 [cs，stat]*，2013年。

[KK10] Amit Kumar和Ravindran Kannan。具有谱范数和k-means算法的聚类。 计算机科学基础年度*IEEE*研讨会论文集，*FOCS*，第299-308页，2010年。