维数灾难

维基百科，自由的百科全书

维数灾难（英语：curse of dimensionality，又名**维度的詛咒**）是一个最早由理查德·贝尔曼（Richard E. Bellman）在考虑优化问题时首次提出来的术语[1][2]，用来描述当（数学）空间维度增加时，分析和组织高维空间（通常有成百上千维），因体积指数增加而遇到各种问题场景。这样的难题在低维空间中不会遇到，如物理空间通常只用三维来建模。

举例来说，100个平均分布的点能把一个单位区间以每个点距离不超过0.01采样；而当维度增加到10后，如果以相邻点距离不超过0.01小方格采样一单位超正方体，则需要10^{20} 个采样点:所以，这个10维的超正方体也可以说是比单位区间大10^{18}倍。（这个是理查德·贝尔曼所举的例子）

在很多领域中，如采样、组合数学、机器学习和数据挖掘都有提及到这个名字的现象。这些问题的共同特色是当维数提高时，空间的体积提高太快，因而可用数据变得很稀疏。稀疏性对于任何要求有统计学意义的方法而言都是一个问题，为了获得在统计学上正确并且有可靠的结果，用来支撑这一结果所需要的数据量通常随着维数的提高而呈指数级增长。而且，在组织和搜索数据时也有赖于检测对象区域，这些区域中的对象通过相似度属性而形成分组。然而在高维空间中，所有的数据都很稀疏，从很多角度看都不相似，因而平常使用的数据组织策略变得极其低效。

“维数灾难”通常是用来作为不要处理高维数据的无力借口。然而，学术界一直都对其有兴趣，而且在继续研究。另一方面，也由于本征维度的存在，其概念是指任意低维数据空间可简单地通过增加空余（如复制）或随机维将其转换至更高维空间中，相反地，许多高维空间中的数据集也可削减至低维空间数据，而不必丢失重要信息。这一点也通过众多降维方法的有效性反映出来，如应用广泛的主成分分析方法。针对距离函数和最近邻搜索，当前的研究也表明除非其中存在太多不相关的维度，带有维数灾难特色的数据集依然可以处理，因为相关维度实际上可使得许多问题（如聚类分析）变得更加容易。另外，一些如马尔科夫蒙特卡洛或共享最近邻搜索方法[3]经常在其他方法因为维数过高而处理棘手的数据集上表现得很好。

目录

组合学

在一些问题中，每个变量都可取一系列离散值中的一个，或者可能值的范围被划分为有限个可能性。把这些变量放在一起，则必须考虑很多种值的组合方式，这后果就是常说的组合爆炸。即使在最简单的二元变量例子中，可能产生的组合总数就已经是在维数上呈现指数级的$O(2^d)$。一般而言，每个额外的维度都需要成倍地增加尝试所有组合方式的影响。

采样

当在数学空间上额外增加一个维度时，其体积会呈指数级的增长。如，点间距离不超过$10^{-2}=0.01$，$10^2=100$个均匀间距的样本点足够采样到一个单位区间（“一个维度的立方体”）；一个10维单元超立方体的等价采样，其相邻两点间的距离为$10^{-2}=0.01$则需要10^{20}个样本点。一般而言，点距为10^{-n}的10维超立方体所需要的样本点数量，是1维超立方体这样的单元区间的$10^{n(10-1)}$倍。在上面的n=2的例子中：当样本距离为0.01时，10维超立方体所需要的样本点数量会比单元区间多10^{18}倍。这一影响就是上面所述组合学问题中的组合结果，距离函数问题将在下面介绍。

优化

当用数值逆向归纳法解决动态优化问题时，目标函数针对每个可能的组合都必须计算一遍，当状态变量的维度很大时，这是极其困难的。

机器学习

在机器学习问题中，需要在高维特征空间（每个特征都能够取一系列可能值）的有限数据样本中学习一种“自然状态”（可能是无穷分布），要求有相当数量的训练数据含有一些样本组合。给定固定数量的训练样本，其预测能力随着维度的增加而减小，这就是所谓的*Hughes影响*[4]或*Hughes现象*（以Gordon F. Hughes命名）。[5][6]

贝叶斯统计

在贝叶斯统计中维数灾难通常是一个难点，因为其后验分布通常都包含着许多参数。

然而，这一问题在基于模拟的贝叶斯推理（尤其是适应于很多实践问题的马尔科夫蒙特卡洛方法）出现后得到极大地克服，当然，基于模拟的方法收敛很慢，因此这也并不是解决高维问题的灵丹妙药。

距离函数

当一个度量，如欧几里德距离使用很多坐标来定义时，不同的样本对之间的距离已经基本上没有差别。

一种用来描述高维欧几里德空间的巨型性的方法是将超球体中半径r和维数d的比例，和超立方体中边长$2r$和等值维数的比例相比较。这样一个球体的体积计算如下：$\frac{2r^d\pi^{d/2}}{d\Gamma(d/2)}$

立方体的体积计算如下：$(2r)^d$

随着空间维度d的增加，相对于超立方体的体积来说，超球体的体积就变得微不足道了。这一点可以从当d趋于无穷时比较前面的比例清楚地看出：$\frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \to 0$

当$d \to \infty$。 因此，在某种意义上，几乎所有的高维空间都远离其中心，或者从另一个角度来看，高维单元空间可以说是几乎完全由超立方体的“边角”所组成的，没有“中部”，这对于理解卡方分布是很重要的直觉理解。 给定一个单一分布，由于其最小值和最大值与最小值相比收敛于0，因此，其最小值和最大值的距离变得不可辨别。$\lim_{d\to\infty} \frac{\mathrm{dist}_{\max} - \mathrm{dist}_{\min}}{\mathrm{dist}_{\min}} \to 0$.

这通常被引证为距离函数在高维环境下失去其意义的例子。

最近邻搜索

最近邻搜索在高维空间中影响很大，因为其不可能使用其中一个坐标上的距离下界来快速地去掉一个候选项，因为该距离计算需要基于所有维度。[7][8]

然而，最近的研究表明仅仅一些数量的维度不一定会必然导致该问题，[3]因为相关的附加维度也能增加其相反项。另外，结果排序的方法仍然有助于辨别近处和远处的邻居。然而，不相关（“噪声”）维度也如期望一样会减少相反项，在时间序列分析中，数据一般都是高维的，只要信噪比足够高的话，其距离函数也同样能够可靠地工作。[9]

*k*近邻分类

高维度在距离函数的另一个影响例子就是*k*近邻（*k*-NN）图，该图使用一些距离函数从数据集构造。当维度增加时，*k*-NN有向图的入度分页将会向右倾斜，从而导致中心的出现，很多的数据实例出现在其他许多实例（比预期多得多）的*k*-NN列表中。这一现象对很多技术，如分类（包括最近鄰居法、半监督学习，和聚类分析都有很大的影响。[10]，同时它也对信息检索问题有影响。[11]

延伸閱讀

- 组合爆炸
- 相似度集中
- 降维
- 傅立葉變換列表
- 高维数据聚类
- 动态规划
- 贝尔曼方程
- 逆向归纳法
- 主成分分析
- 最小二乘法
- 准随机
- 聚类分析
- 小波分析
- 时间序列
- 奇异值分解

参考資料

1. Richard Ernest Bellman; Rand Corporation. Dynamic programming Princeton University Press. 1957. ISBN 978-0-691-07951-6.
   Republished: Richard Ernest Bellman. Dynamic Programming Courier Dover Publications. 2003. ISBN 978-0-486-42809-3.
2. Richard Ernest Bellman. Adaptive control processes: a guided tour Princeton University Press. 1961.
3. doi: 10.1007/978-3-642-13818-8_34(https://dx.doi.org/10.1007%2F978-3-642-13818-8_34)
   本引用來源將會在數十分鐘後自動完成。您可以检查英文对应模板或手動擴充
4. doi:10.1007/s11004-008-9156-6(https://dx.doi.org/10.1007%2Fs11004-008-9156-6)
   本引用來源將會在數十分鐘後自動完成。您可以检查英文对应模板或手動擴充
5. Hughes, G.F., 1968. "On the mean accuracy of statistical pattern recognizers", IEEE Transactions on Information Theory, IT-14:55-63.
6. Not to be confused with the unrelated, but similarly named *Hughes effect in electromagnetism* (named after Declan C. Hughes (http://spiedl.aip.org/vsearch/servlet/VerityServlet?KEY=SPIEDL&possible1=Hughes%2C+Declan+C.&possible1zone=author&maxdisp=25&smode=strresults&pjournals=OPEGAR%2CJBOPFO%2CPSISDG%2CJEIME5%2CJMMMGF%2CJARSC4%2CJNOACQ&deliveryType=spiedl&aqs=true)) which refers to an asymmetry in the hysteresis curves of laminated cores made of certain magnetic materials, such as permalloy or mu-metal, in alternating magnetic fields.
7. R. B. Marimont and M. B. Shapiro, "Nearest Neighbour Searches and the Curse of Dimensionality", *Journal of the Institute of Mathematics and its Applications*, 24, 1979, 59-70.
8. E. Chavez et al., "Searching in Metric Spaces", *ACM Computing Surveys*, 33, 2001, 273-321.
9. doi: 10.1007/978-3-642-22922-0_25(https://dx.doi.org/10.1007%2F978-3-642-22922-0_25)
   本引用來源將會在數十分鐘後自動完成。您可以检查英文对应模板或手動擴充
10. Radovanovi?, Milo?; Nanopoulos, Alexandros; Ivanovi?, Mirjana. Hubs in space: Popular nearest neighbors in high-dimensional data (PDF). Journal of Machine Learning Research. 2010. **11**: 2487–2531
11. doi: 10.1145/1835449.1835482(https://dx.doi.org/10.1145%2F1835449.1835482)
    本引用來源將會在數十分鐘後自動完成。您可以检查英文对应模板或手動擴充

- Bellman, R.E. 1957. *Dynamic Programming* Princeton University Press, Princeton, NJ.
  - Republished 2003: Dover, ISBN 0486428095
- Bellman, R.E. 1961. *Adaptive Control Processes* Princeton University Press, Princeton, NJ.
- Powell, Warren B. 2007. *Approximate Dynamic Programming: Solving the Curses of Dimensionality* Wiley, ISBN 0470171553.

取自“https://zh.wikipedia.org/w/index.php?title维数灾难&oldid=46355783”