你有早上好现在你能听到我说话了谢谢麦克风与我合作。大家早上好我看到我们已经 我们已经失去了不少人。我认为大家做了正确的选择 因为我听到有人唠叨说有人生病了如果你们在我的讲座上生病了，我会很生气的。请呆在家里。请不要传染给你周围的人，也不要传染给我。我现在没时间生病。我很忙所以，如果你病了，请呆在家里。没人会生气。会有时间补习的。讲座将被录制下来，因此您可以在听完讲座后，在家里喝杯茶，舒适地观看讲座。好的。第八周，课程即将结束。大家都知道，这门课程只有11周的教材，所以我们还有8、9、10、11个正式讲座要上。现在我们正在讨论讨论不同的方法所以我们每堂课都会讲一种新的方法。今天的讲座会相对轻松一些，为下周的讲座做准备，下周的讲座会更紧张，时间也会更长一些。希望最后我们会有时间提问。因此，我们可能会通过讲座来复习迄今为止所涉及的一些材料，所以这次会比较轻松，不用担心。但首先，我想再次感谢大家参加上周的客座讲座。希望你们喜欢。事后，卡洛斯再次向我表示，他非常非常喜欢这次讲座。这对他来说是一次全新的讲座，所以他过去也做过客座讲座，但他重新修改了他所有的材料、幻灯片和讲座内容。所以，他让我为你们讲讲各种事情，感谢你们出席并提出了好问题，聆听了他的演讲。他还让我告诉你们，如果你们有问题或反馈意见，他非常希望听到。因此，我很乐意转发他的联系方式。我想你们可以在 LinkedIn 上找到他。他非常热衷于在 LinkedIn 上与大家联系，所以如果你们有兴趣，不要客气，可以联系他。我还想借此机会收集一下你们的意见或反馈。如果你们愿意匿名，或者愿意与全班分享。对嘉宾讲座有什么想法和反馈？什么？有什么不满意的？哪些地方不顺利，哪些地方你喜欢？总的来说好不好？有什么想法吗？是的，我很喜欢。我觉得他真的吗？我喜欢这种对建模的解释。很好如果你喜欢的话如果你喜欢的话我还喜欢这样的事实：当你不在现场时，不容易获得内部视角。所以很高兴知道计划是如何运作的，他们的思维过程是怎样的。是的，我觉得这真的很有趣，一方面，这个过程有多长，另一方面，建模过程又有多短。所以我想它提到了从开始谈论某事到真正得到结果大约需要六个月的时间，这一方面听起来相当长，另一方面又大致有多长。例如，本科生的毕业论文我认为是一年，但我们通常说是六个月左右。因此，我认为你将有一半的时间来完成毕业论文，这也可以告诉你毕业论文项目的大小。因此，我觉得能听到更多关于时间安排和他们必须经历的步骤的信息真的很有趣。显然，在公司里，你必须考虑到的是有更多的规定和步骤，以及说服人们等等。在产品小组工作中，你可能会有更多这样的经历，你必须说服小组成员，让他们相信某件事是正确的。我看到大家在争论时交换了眼神。还有其他反馈或意见吗？总体来说很满意。因此，我们应该推荐这样的课程，并将其保留在课程中，明年再做一次。非常好。我会让他知道的，我也会的。如果你们还有其他反馈意见，请告诉我，我会转达的。我相信，一点批评意见对他也没有坏处。我还想说的另一件事是，有人多次问我关于考试时间表的问题。在我的辩护中，我并没有制定考试时间表。我们学校有一个中央时间表编制单位。你可以想象，为我们这样一所规模庞大的大学安排考试日程是一项极其困难和复杂的任务。有时我也在想，他们是否会向我们寻求可以实施的优化方法。也许这是我们应该提供的，但没有。是的，第三件事就是时间表已经上线。这是链接。请自行查看时间表，以防有任何变动。如果考试改期了，请不要以我为借口。但我的讲座上说过，当时是这样的，请你自己去查，这是你的责任。但我最后一次查的时间是 12 月 18 日。所以你很幸运。我想我们的考试要到 23 日才结束。所以你可能在这里度过了半个圣诞节 下午两点到四点说实话，我也不太清楚那是什么地方。汤姆森的土地，我认为这是 某处波洛克持有。我知道有人查过是哪里吗在爱丁堡在市中心应该是应该可以步行。就像其他地方一样因此，这是每个人的主要考试地点。如果你有任何特殊安排，请查看你的个性化时间表。我想你可以通过相同的链接进入，如果你有任何住宿安排，那么你会被告知你自己的考试。例如，你可能需要比这更长的时间。在这种情况下，你可能会参加同样的主要考试，但你将有更多的时间在考试结束前完成考试。对考试有任何疑问？我已经上传了去年的考试范例，请大家看一看。我想这可以让你对考试有个大致的了解。有人问我是否会有复杂的计算。不，不会有计算，除非你想在解释概念时计算一个例子，我对此也非常满意。否则，它将与去年的考试类似。显然不一样，但也差不多。好的。让我们来回顾一下。我们的最后一堂课是在第五周。然后是阅读周和第六周 第七周是客座讲师课所以已经有一段时间了。但上次在第五周，我们围绕数据分割、重采样、类不平衡处理进行了一些回顾，这些内容我们在上周的教程中也有涉及。你应该还记得，数据拆分是将数据分为训练数据和测试数据，然后在其中一部分数据上训练模型，再进行测试和所有其他交叉验证，而我们的抽样是重复训练模型，在部分数据上重复运行模型。当我们在讨论类不平衡处理时，我们的想法是，如果你的类非常不平衡，尤其是如果这是你的结果变量。因此，如果这是你的预测变量之一，问题就会小得多。有时这确实是个问题，但通常情况下，如果这是你的预测变量之一，你就不会太在意类的不平衡。但如果你试图预测一个类别，而其中一个类别比另一个类别更频繁，那么模型自然会更适应更频繁的类别。所以在这种情况下，我们通常会使用不平衡法。是的，平衡法就是通过上采样、下采样、抽样等方法来均衡类别。天哪然后我们谈到了无监督机器学习。所以我们在讨论分区聚类算法。这就是你的K，也就是你的K流星体。我们还谈到了分层聚类算法，我们正在创建一个凹痕。所以，基本上是一个由多个聚类方案组成的结构，而不是一个单一的方案。本周，我们将从 k 和 n 开始，这是我们的近邻分类方法之一。因此，我们将告别无监督学习，进入分类领域。我们还将讨论特征选择、降维等问题。具体来说，我会简单谈谈因子分析。但大部分时间我们会花在主成分分析（PCA）和回归上。那么，让我们开始讨论卡农。我特意把它放在这个讲座中，是因为它是我们在第五周讲授的聚类和未来几周讲授的分类之间的一种自然联系，因为两者的想法极其相似。所以，你会觉得它有点像 K-均值法。我们仍在试图寻找模式，在这种情况下，我们试图通过寻找数据点之间的某种相似性或接近性来找到点的类别。我们仍然假设，处于同一组或上一次聚类中的点，在特定变量方面表现出某种相似性。这就是聚类的概念。这也是近邻分类的理念。这两者之间的主要区别在于，在这种情况下，我们仍然有新的未分类数据点，我们试图将它们归入某个组、聚类或类别，而我们是通过分析相邻数据点来做到这一点的。然而，最大的区别在于，在这种情况下，这是一种分类方法。因此，这是一种有监督的方法，与聚类不同，在聚类中，我们只是将所有数据扔给算法，希望算法能产生一些结果。因此，我们在讨论聚类有一个很大的缺点，那就是你无法或很难控制从聚类中得到什么样的结果。它非常适合无监督的情况。它只是在做自己的事情，你希望结果对你有用。有一些方法，你可以选择簇的数量，也可以根据特定的标准或联系或所有这些东西进行优化，你可以引导算法找到一个好的解决方案。但在聚类问题中，如何定义一个好的解决方案是非常困难的。所以有时有人问我，聚类的好结果是什么？这几乎是不可能的。你可以使用一些评估标准，例如，你可以看看各组的区别有多大？它们的大小是否相等？所有这些事情。但这并不能告诉你一个具体的数字，比如有多少元素在正确的群组中，因为并不存在正确的群组。你并不知道每个点应该属于哪一组。现在，在监督学习中，我们知道正确的分类应该是什么，至少对于我们的训练数据来说是这样。因此，我们有某种预先标记好的训练数据，在这些数据中，我们选择了一些数据点，并清楚地知道它们属于哪一组。这就是我们对每个数据的答案。现在，我们的目标是对新的组别、新的成员进行分类，将它们归入我们已知的已有类别中。因此，在这种情况下，我们实际上可以看到我们的算法在分类结果方面表现如何。这与逻辑回归非常相似，我们可以计算出被正确分类的元素。好的。这可能是最流行的邻域分类方法了。我会努力告诉你另一种方法。我知道有很多不同的邻域分类方法，但卡根在这一领域占据主导地位，以至于我们很难记住其中任何一种，因为总是有一些听起来像 "N "的东西，应该叫 "N"，但为了听起来更花哨或更新颖而换了个新名字，诸如此类。因此，"K "和 "N "的工作分为四个相对简单的步骤。首先，我们要有训练数据，也就是一些特征组合。即一些变量和一个类标签。因此，你有某种 x 和某种 y，你试图预测它们，就像逻辑回归一样。然后，对于数据中任何新的、未标记的数据点，你要查看该数据点周围的 k 个近邻点，以及它们属于哪个类别。现在，在二维空间中，这一点相对容易想象，因为基本上就是如果你有一堆点，你看一个点，你不知道它属于哪一类，它周围的点又属于哪一类。这是非常简单的逻辑。显然，你必须再次定义 "关闭 "的含义。所以，我们又回到了 "距离测量 "这个概念。这也是我们与聚类的另一种联系。这其实是同一个概念。根据你的应用，你可以使用欧几里得距离，定义未标记点与所有相邻点的距离，然后查看其中最近的 k 个点。这样，你就可以按照距离的递减顺序排列这些点。然后，你基本上会找出与之最接近的 k 个点。然后进行多数表决。所以，如果有三个点，三个最近的邻居，其中两个说是一类，一个说是二类，那么你就把它归为一类。这就是一个例子。在这里，你还可以看到它的影响取决于你为数据点选择了多少个近邻。因此，你可以看到这里绿色的数据点我们还没有标记。我们不知道这个点应该属于哪个类别。我们有两个类别，蓝色正方形和红色三角形。问题是，我们的绿色点应该归入哪一类呢？现在，这里的内圈向你展示了如果你决定你的三个近邻应该属于哪一类会发生什么。我们看一下这三个最近的点，可以看到有两个红色三角形和一个蓝色正方形。很明显，如果这三个点投票，我们就会给绿色的点贴上红色三角形的标签。现在，如果我们决定我们的情况应该是五个，会发生什么呢？换句话说，我们会以更大的距离、更大的半径来观察这个点。突然间，我们的投票发生了变化，因为突然间我们有三票支持 "不"，应该是蓝色正方形，而只有两票支持 "应该是红色三角形"。因此，在这种情况下，这将变成一个蓝色正方形。这已经告诉你，k-NN 最棘手的地方就是我们已经在努力解决的问题。我上次说过，K 会成为你最不喜欢的参数名称，原因就在于此，因为你在选择它时总是很纠结，而且对结果影响很大。因此，选择 "蛋糕 "可能极具挑战性，其根本原因在于，如果你选择一个大到 2 的值，基本上就会抹平任何可能存在的微小动态变化。因此，在这个例子中，如果你选择一个非常非常大的蛋糕，涵盖所有这些数据点，那么中间红色三角形的这种小动态就会被完全抹平。你根本不会注意到这一点，但如果你选择的 k 非常小，就像之前的 k 3，那么你可能会把这个点分配给红色三角形，尽管如果你从整体上看，也许蓝色正方形与这个点的契合度并不差。这也有点棘手，因为我之前说过，K 和很简单，因为你可以计算出一个点是否被正确分类。但当我们说分类正确与否时，这完全取决于你对相邻点的选择。那么这个点的正确分类到底是什么意思呢？它可能是，嘿，我们想要捕捉这种小动态，因为我们对这些较小的群体行为感兴趣。我们对子群体感兴趣，比如消费者或诸如此类的群体。实际上，也可能不是这样。我们更喜欢更广阔的视野。我们对其中的异常值或个体并不感兴趣。所以，在你选择 K 之前，或者至少在你决定你想要的商业结果是什么，或者你想如何从商业角度来解决这个问题之前，这并不是一个真正正确的分类。所以，是的。太大。过于平滑。太小。你对任何异常值都很敏感。因此，如果你在任何地方发现了一个奇异的红色三角形，那么你可能会将一个点分配到该组中，即使它只是一个位于更大的蓝色方形组边缘的异常值。因此，当 K 影响你的建模结果时，你实际上是可以调整的，因此我们可以将整个问题调整到非常不同的应用环境中。因此，我们可以称其为 "调整参数"，因为我们可以根据自己的应用偏好来调整参数。现在，在实践中。明天，我们将在计算机实验室中使用交叉验证，测试 k 的多个值。因此，如果你知道在你的训练数据中，你已经对所有数据进行了预先标记。因此，你可以准确地知道每个点的标签。因此，你可以在训练数据上使用多个 k 值进行交叉验证测试，这样你就能知道自己的准确率，然后在测试数据上使用该 k 值。因此，这是一种与 k 表示法非常相似的方法，你还记得在 k 表示法中，我们讨论了优化和寻找最佳 k 值的不同方法，我们还讨论了肘部标准。我们还讨论了剪影评分和所有这些不同的衡量标准。因此，K 作为一个有监督的问题，会让问题变得简单一些，因为它更容易量化算法的性能，因为我们可以真正计算出训练数据的准确性。另一个我认为对肯来说非常重要的考虑因素是，距离度量的选择非常关键。因此，我认为很多人都犯了一个错误，他们只是把欧氏距离丢给他们的数据，并希望它能起作用，而在某些情况下，它能起很好的作用。如果你的数据都是数字，而且非常漂亮。那么无论如何，欧氏距离都是一种方法。我想这也是我们在计算机实验室的例子中会用到的距离，因为我们只需查看两个维度的数值数据，明确欧氏距离，然后就大功告成了。不要想太多。但如果你有混合数据和社会科学，我们有很多这样的数据，突然间就会变得困难得多。因此，如果你回想一下我们关于聚类算法的讲座，我们曾讨论过这些不同的距离度量，我们可以根据具体情况使用这些度量。因此，我们有一些方法更适合混合数据类型或序数数据，或者只有一种数据类型，即数值型，等等。因此，所有这些预处理、变量缩放对 k 和 n 非常非常重要，然后再选择距离度量。在开始考虑最佳邻域数之前，所有这些都是非常重要的。是的，不同的数据需要不同的距离测量方法。关于时间序列数据和空间数据，我想我已经说得差不多了。距离到底意味着什么？这就成了一个哲学问题。所以，如果你想多思考这个问题，请继续，我有时觉得我是世界上唯一一个会思考 "在不同情况下，距离到底意味着什么 "的人。大多数人都是抛出一些方法，然后就能解决了。所以我想这很好。我想这就是你能做的。是的，这也很有趣。在某些情况下，我们会给邻居分配一个与距离成正比的权重。因此，如果离群值真的很远，这可以减少离群值的影响。所以，基本上我们不会只看邻居的数量。例如，在我们之前的例子中，这个红色三角形的邻居离我们的未知点非常近，我们可以说这是一个非常重要的点，它决定了我们这个未标记点的类别。这个点离哪里很远呢？我们可能会说，它离我们太远了，以至于对它的类别没有太大影响。因此，有一种对距离进行加权的方法，可以降低对异常值的敏感度。例如，如果点的密度分布很不均匀，它也会有所帮助。因此，我的研究领域之一就是思考在我的案例中，点的密度如何决定聚类的分类。因此，如果你对聚类感兴趣，如果你想在特定的国际区域内找到数据点群，那么你可以从地理空间的角度来考虑。不一定非要看某物是否是一个聚类，也不一定非要看某组或某点是否属于一起，k 和 n 不仅取决于某物的位置，还取决于其周围有哪些点，以及这些点与周围的点相比，聚类的密度有多大。例如，在密度较低的区域，聚类并不一定是点的超级密集区，因为自然而然地，该区域的点就不那么密集了，这是一个非常有趣的概念。因此，这是一个相关的概念，我们要研究的是距离如何与相邻点的数量相互影响，而相邻点的数量是我们为新点分配成员资格时需要考虑的。没错。因此，有时我们会根据距离的远近来分配相邻点的数量。最后但并非最不重要的一点是，不幸的是，在高维空间中，"不幸 "的情况总是发生。如果只有几个变量、几个维度，k 和 n 的效果会很好，但在高维情况下就很难使用了。因此，在有大量特征、大量变量的情况下。其中部分原因是，如果你回想一下计算与邻居距离的想法，如果你有很多数据点，这已经变得很困难了，因为对于每一个新数据点，你都必须计算与整个数据集中其他数据点的距离，以确定谁实际上离它们最近，因为你还不知道这一点。实际上，你必须计算每个数据点的距离。现在，维度越多，变量越多，距离计算就越复杂，计算成本也就越高。因此，Kanan 对于中小型数据集来说非常出色，但一旦你扩展到大型数据、大量维度时，它就会非常吃力。这主要是因为它需要计算与每个邻居的距离。因此，有一些方法可以解决这个问题。从理论上讲，例如，你可以将它与不同的方法结合起来，这样就可以在一定程度上划分空间。例如，你可以使用聚类，将空间划分为子空间。然后，在这些子空间内，你可以使用 k 和 n，因为这将限制你实际寻找点的邻居的距离。因此，有一些方法可以克服这个问题。但通常情况下，我们不会使用 k n，只有在低维情况下，通常是在中小型数据集中。我对这个双关语感到非常自豪。所以，回想一下聚类以及 k 和 N，你可以和你的近邻讨论一下，K 是你的选择，以及聚类的一些优缺点，因为我们在选举之后并没有讨论过这个问题，我认为这真的很重要。我们刚刚讨论了 K 和 N 的一些优缺点。如果你愿意，可以将其与聚类进行对比。然后，我认为这是一个有趣的问题。你认为什么时候你会选择其中一个？你认为什么时候聚类方法更适合你的数据？或者K和n更适合你的数据。好吧，那就花几分钟，也许十分钟，和你的邻居谈谈，然后再收集。我们要一起收集几个点吗？好的，我们开始吧？也许吧我们该怎么说？聚类的优缺点然后是k和n的优缺点。然后你什么时候会选择其中一个？所以，让我们做几个优缺点吧。我真的应该这样做我不应该这样做。我的字写得很糟糕我的字在最好的时候也写得不好这个这个好多了我相信你知道12月的考试是手写的 就像，对，就像一张纸和一支铅笔你知道铅笔是什么吗？知道所以我建议你们开始练习用手写笔记 因为我知道如果你们没有练习过 突然要写两个小时的话 会把手弄脏的所以现在就开始练习吧。我记得去年有几个学生抱怨手写很疼，不习惯手写。所以别说我没提醒你们。好吗？话虽如此，但我不会因为你的字迹而打倒你。是的实际上，我看到作业的问题有不同的标记。所以，有多少页。我们到底应该怎么写我不知道这次是否会是同样的分数分配。所以这是第一件事。一般来说，我的建议是总分100分。你知道，总共分配了两个小时。这样，你就可以计算出多少分钟、多少时间应该花在多少分上。如果 100 分需要两个小时，那么每 100 分大概需要一个小时。这样，你就可以计算出每道题应该花费多少空间或时间。我对一个好的答案应该有多长没有任何想法。这完全取决于你自己。我认为一个经过深思熟虑、简明扼要的答案和长篇大论的答案同样好，只要它们能涵盖相同的要点。是啊。35 分问题的达标标准是什么？我是按照评分标准来做的。你会觉得有吗？有啊因此，我构建考试的方式是，我有一个想法，我想读什么。所以，基本上就是你应该提到什么样的分数才能得到什么样的分数。是啊。拿满分。很显然，批判性思维、讨论技巧和细节等也是得分点。因此，这更像是一种全面的整体观。但我认为，它与课业评分的主要区别在于。很明显，我不指望你们做研究。因此，记住书名或类似的东西是不会得分的。但要有批判性思维和讨论的想法 然后我有一份要点清单，我想读一读。好的让我们看看这个镜头。足够大我们要做什么？一些额外的阅读这是个很难回答的问题，因为你会注意到，这些讲座的结构是强调从整体角度理解一个概念或技巧，然后再应用。因此，你如何理解技巧的作用对我来说并不重要。如果你觉得我的解释很直观，而且你也明白了这种方法是如何实现的，那就没问题。如果你更喜欢教科书，你也可以使用它，因为它比我解释得更好。这样也行。如果你喜欢不同的教科书，也没关系。我只是希望你们能养成理解的习惯。不同的方法是如何工作的，因为这可以让你选择正确的方法来解决正确的问题。所以，如果你回到我们的学习目标，我认为这基本上是你能做的最好的事情。如果你想知道评分标准，这也是你能做的最好的事情，因为学习目标是我构建课程、构建论文或考试的指导。好了，最后我们会有更多的时间来回答更多的考试问题，所以不用担心。不过，现在让我们把注意力集中在我们的朋友聚类上。我的最爱你们讨论过的聚类有哪些优缺点？优点发现数据中的模式或关系没错基本上，聚类的重点在于发现。所以，模式、关系都是如此。都是关于数据的探索性视角。是的，非常好。非常好另一个优点或缺点时间太久远了是的这就像就像使用聚类。很有意思所以基本上是用聚类来检测异常值 So basically use clustering to detect outliers.就像你可以聚类发现每组的平均值如何使用你的外部数据？啊那么啊所以，基本上是因为你不是在看单个数据点和异常值，而是在看整体结构或群体，而不是个人，这基本上会。说得好。我认为有些方法对异常值非常敏感。我们刚才说到了分层聚类。分层聚类在这方面是出了名的。我每次用它都很费劲，因为它往往会把这些异常值抓得很紧。因此，分层聚类对异常值相当敏感，尤其是在使用单一关联作为关联标准的情况下。如果使用完全关联，情况会好一些，但仍然相当敏感。但从分析的角度来看，你说得没错。所以你基本上是把数据点组合在一起。这就是为什么聚类被用于市场调研中的市场细分的原因之一，在市场细分中，你可以通过 "角色 "来描述客户群体。因此，如果你们当中有人有市场营销方面的背景，你们就会知道 "角色 "这个概念。"角色 "基本上是一个人格化的消费者，是你利用聚类等方法构建的人工消费者。因此，你要研究一个群组中的普通人，并从中构建一个角色，这有助于你创建营销市场沟通工具。如何才能更好地利用这一工具呢？这也许就是个人数据点与群体、点与群体之间的作用。然后，我把离群值加上引号，你们都会记住我说的话。有了我的话，你们就能比我的书面笔记更好地记住我的话。希望如此希望如此还有别的吗？有什么缺点吗？有我认为聚类与黑盒模型的区别在于聚类。你知道为什么要把点分组，所以你会明白点是根据彼此的距离分组的。举例来说，这样更直观。而在黑盒模型中，你真的不知道为什么某个点会被分配到某个特定的组中。所以这两者是有区别的。但我觉得你说得很好，因为它确实是。它的一个缺点是它本质上是不受监督的。因此，你无法真正地将它引向一个方向，你无法真正地影响或引导它。这可能是一个优势，因为有时你根本不知道数据中发生了什么。所以，你只需要能用的东西。但有时，你对数据中应该有什么有具体的想法，而你在尝试聚类时却得到了其他结果。这可能会让人非常沮丧。还有其他优点和缺点吗？就像我说的，我有偏见。这是不是意味着引入的数据越多？我的意思是我是说有了所以你指的是这个偏差方差权衡的问题，基本上就是你是否过度拟合的问题。我的意思是，聚类的好处在于你已经给了它所有的数据。因此，与分类法相比，并没有额外的数据让你过度拟合。有时，你会过度拟合数据，导致新数据无法真正分类。尽管如此，对于聚类，我想我们已经讨论过，如果你有非常小的聚类、小群体，你就会过分强调小模式。在这种情况下，是的，基本上就是过度拟合，因为如果你确实创建了新数据，比如新客户，那么这种方法就很难运行。基本上，你必须重新启动。我们启动该方法，而不是将点重新分配到聚类中。我的意思是，你可以采用混合方法。你可以先进行聚类，然后使用 k 和 n 将新的点分配到已有的聚类中。但这样可能会导致过度拟合。是这样吗？是的。那就把它放在这里吧。所以可能会过度强调模式，这意味着无法推广到新的数据中。强调。哦，好吧不，它只是不喜欢我的英式英语，我想。好吧 好吧 Okay.好吧 然后呢 Okay.And then.好的你不需要一个训练阶段的男人。我的意思是，是的，你不能。你不必。是的所以，基本上我们通常会做的是，我们会调整它的K。 所以，集群的数量的数量K。 是的。但这有点像A11的方法。你有了 K，基本上就可以把它扔进去了。所以是的。所以，我觉得对训练的依赖更少了。还有所以，我们将在计算机实验室看到这一点。我们说的 "训练 "指的是什么？"训练 "与神经网络相比，神经网络需要经过很多很多轮的调整，不仅是超参数的调整，还有模型本身的每个参数的调整，而 "k "和 "n "的调整则更快一些。是的。还有关于 KNM 的其他信息吗？还有选蛋糕还有我的意思是，我们之前在选择K的时候也遇到过这个问题，我们在聚类的时候也遇到过这个问题，我们可以把它看作聚类的另一个缺点，就是同样的问题。只要你从研究人员那里得到任何类型的输入，你就会遇到一些麻烦。例如，使用 k 表示你必须选择 k 和距离度量。对于分层聚类，你必须选择一个距离度量和联系度量。因此，在所有这些情况下，选择最适合数据的方法可能会很棘手。k 和 N 也是同样的道理。我们之前在讨论这个正方形和三角形的例子时也遇到过同样的问题。当我们在讨论 k 的作用以及什么是最佳 k 时，这取决于你的应用。所以，是的，选择k很棘手。是的在大数据点上是的，我们正在讨论这个问题。又大又高的维度现在稍后再说所以，是的，它深受其害。当然有些方法在大型数据集上效果更好。有一些方法，我的意思是，它也可以成为一种优势。它在小数据上效果非常好，这也是我们正在讨论的问题。逻辑回归。逻辑回归在小数据和 k 数据上的效果非常好。但也有一些方法在大数据上效果更好。因此，神经网络只适用于大型数据集。支持向量机对大型数据集非常有效。随机森林非常适合大型数据集？是的。我不知道我不知道如果这是个缺点，但使用 GNN 时，我们不需要对数据进行标准化吗？是的。因此，K 和 N 对尺度很敏感，这是因为它使用了距离度量。因此，为了定义两点之间的距离，你必须将数据标准化，否则其中一个尺度就会使整个数据出现偏差。没有距离度量，就没有距离。如果你有纯粹的数字数据，漂亮的基本数字数据，用欧几里得距离来标准化就可以了。对，就是这样。但如果有混合数据你可以在我们学过的这些美丽的名字上 使用像沙卡系数这样的东西我想问一个问题我们什么时候可以考虑数据集取决于你的电脑有多好不，真的取决于过去我们认为是大数据集的数据，现在已经不算大了。所以现在我认为的小数据集是一两百个或者两三千个数据点。对我来说，中型数据集是数千到一万个数据点，大型数据集是十万到数百万行。但这只是经验之谈。这只是基于我的笔记本电脑。如果你有一台超级计算机，一切都会变得非常小。我想我们还有一个问题，对吧？是的你可以使用更大的数据集。这取决于方法取决于方法比如K法，计算效率很高。所以在大型数据集上效果很好。而分层聚类的成本就比较高了。所以这取决于算法。但一般来说，聚类适用于中等数据集。至少有些方法也适用于大型数据集。可以。你说在较小的数据集上效果更好。但如果数据集太小。这就是问题所在。大数据集的计算成本高？对如果有必要，可以更精确没错所以大型数据集的问题纯粹是计算上的 所以大型数据集 因为我们要计算每个点之间的距离 巴黎太贵了你说得太对了。如果一个数据集太小，那么很明显，它实际上能告诉我们多少数据。因此，我们总是要在两者之间取得平衡。如果你只有 20 个数据点，你真的能通过 20 个数据点发现分类规则吗？另一方面，你还记得在校长的讲座中，我们谈到什么是小数据，什么是大数据，这取决于你的人群有多大。因此，样本可以很大。就其人口而言，即使样本很小。如果你只想研究一个非常非常小的群体，那么你的样本可能很小，但它仍然显示了整个人口的一个重要比例。在这种情况下，一个小的数据集就足够了。因此，我们要考虑的是，你实际上要测量的是什么人群，以及你能从样本中概括出多少人群。还有卡登的优势。你显然不喜欢是啊，我觉得它的计算时间比凯登要短一些。是的，它相对简单，也就是说，对于中小型数据集来说，它的计算效率相对较高。所以不需要通过所有这些训练大小。所以计算效率很高计算效率高而且而且易于理解和解释 我认为这与此有关所以如果我们在讨论交流 解释你的模型是做什么的 我们之前讨论的是黑箱模型所有这些都没有太大问题。我能在 15 分钟内向你解释清楚吗？我想你已经明白了。也许你不能理解每一个细节，但你能理解其中的想法。这真的很直观。你可以很好地向经理解释。你基本上可以解释说，好吧，这个点与其他点类似，因此它属于同一组。就是这样。这就是 k 和 n 的真正含义。因此，与神经网络相比，它听起来非常花哨和漂亮，但实际上如何解释它的作用呢？这非常棘手。我们会花一整堂课的时间来讲解，而不是只讲一半。好吧，我觉得可以因为它可以购买中心点为了塞恩还是为了选择中心点当你计算距离的时候不，我们没有我们并不真正计算中心点。所以我们要做的就是计算这个绿点到这个绿点、这个绿点和这个绿点的距离。所有这些点。然后从中选出三个最小的距离。因此，我们并不像 K-means 中的 k 和 n 那样计算中心点。在使用 K-means 时，我们有一个聚类中心点，从这里我们可以计算出所有的距离。然后我们反复移动这个中心点。而使用 k 和 n 时，我们只需计算所有点对的距离。嗯？所以中心点只是被收集起来。随机选择。随着 k 的数量增加，我们就没有了 K 的中心点，这就是一个未分类的新数据点。它看起来像一个中心点，但其实不是。它只是一个新的数据点，我们把它插入其中，想知道它是红色的还是蓝色的？它看起来像一个中心。点，对吗？我想之所以看起来像一个中心点，是因为它周围的圆圈，但周围的圆圈只是为了向你展示我们是如何寻找相邻点的。小圆圈是 k3，大圆圈是 k5。这只是我们在新数据点周围寻找点的搜索半径。好的另一个问题嗯观察结果，用户必须贴标签吗？可以贴标签。虽然标签是一个类，所以这不是一个回归问题，但它的标签必须是一个类标签。因此，它可以是一个字符串或数字，但必须是一个标签。例如，好客户或坏客户。这就不是回归问题了。在回归中，我们试图预测一个连续的数字，而在分类中，我们预测的是一个标签。是啊，是啊。是的 - 是的是的，是的我们可以休息一下。休息之后，我们将简要讨论一下关于比较的持续讨论。然后我们再做一下特征选择好的 五到十分钟好的来吧可以欢迎回来 欢迎回来听我说 我在说话To you.好吗？希望你休息得愉快我有几个我有一个关于 K 的直观例子以及如何使用 K 的问题，尤其是与聚类的比较，我想这可能会很有趣或对你们很有帮助。很抱歉，我总是把这些市场营销的例子拿出来。这绝对是我的背景。这也是我感觉最舒服的原因。所以你必须面对它。但我相信，你一定能根据自己的兴趣和背景加以调整。所以，如果你把自己的公司想象成一家银行，就像卡洛斯说的那样，我上次投过票，他说的是如何针对不同的客户群体进行不同类型的沟通。例如，在不同的时间点通过不同的渠道提供抵押贷款服务或类似服务。因此，为了实现这一目标，你必须了解数据库中的不同客户群体。有不同的方法可以找到并实现这一点。如果你完全不知道你的数据库中有哪些类型的不同客户，那么聚类绝对是个好办法。举例来说，你可以使用客户行为的一些维度，比如他们的收入、性别、年龄、所在地等。使用这些维度进行聚类，然后根据这些维度创建不同的客户群体。然后，你就可以用平均值来描述这些组别了。但归根结底，这些分组仍然是由算法创建的。因此，它们可能对你有帮助，也可能没有帮助。这有点取决于你的运气，但基本上取决于你输入的数据质量、你的目标以及所有这些不同的维度。现在，K 和 N 也有类似的想法。你仍然对不同的客户群体感兴趣，但这次你已经对其中一些群体进行了预先标记。例如，你有一组客户，他们是收入丰厚的年轻人，你想为他们提供抵押贷款。还有一类客户年龄稍长。他们收入很高，你想跟他们谈谈退休计划。因此，你有这些不同的群体，你知道他们的存在。你也许有针对他们的策略。你已经给他们起了名字。基本上，你的数据库中已经有了这些细分客户群，然后你又有了新客户，但你还不一定知道要如何对待他们。因此，您可以使用 K 和 N 来检查数据库中的其他客户与哪些客户最接近？他们是最接近想要首次抵押贷款的年轻人，还是更接近年纪稍长、收入不错的退休人士？诸如此类。所以，你基本上是在寻找每一个新的数据点，即进入你银行的新客户，你会把他们分配到哪个细分市场？你这样做是有监督的，因为你知道你想从中获得哪些细分市场。因此，我希望这两种方法的比较能让人更好地理解，更好地了解谁。好的，我想再花五分钟来讨论这个问题，因为我认为思考最后一个问题真的很重要。你认为你会在什么时候选择其中一个？刚才我已经提到了一点，那就是我们是否知道我们想从中得到什么？或者我们不知道我们想从中得到什么？所以，如果你把 k 的含义与 k 和 n 相比较，就会发现这一点尤其正确。不过，我们也在讨论聚类。你认为分层聚类在什么时候会比Canyon更好？例如，分层聚类和 K-means 的主要区别是什么？我的意思是，聚类算法的结果是什么？如果你想在你的数据上得到这个结果的话。我听到了一些杂音。你不能这么做你很勇敢说出来是个词吗少了一个字凹痕或克你还记得树状结构吧所以你得到的不是一个聚类结果 而是由多个结果组成的嵌套系统现在，我们正在讨论如何决定在哪里砍掉这棵树。但你不必这样做。你不必砍掉这棵树。你可以原封不动地使用整个结构，用树枝图、树本身来解释数据中的结构。这就是分层聚类、分区聚类和聚类之间的主要区别之一，也是分层聚类与 k 和 n 之间的区别之一。它是切割。就是把人分成几组。这就是你的解决方案。对于分层聚类，你不一定只对一种解决方案感兴趣。你感兴趣的是数据多个阶段的解决方案。因此，如果你想在数据中传达一些关于结构的信息，这可能会非常有用。好了，以上就是我们对聚类的讨论。我们的讨论到此结束。现在，本讲座的第二部分将讨论通过降维来选择特征。这很费口舌。它与我们之前谈论回归时的内容有关。你应该还记得，当时我们在讨论正则化，以及我们如何对回归模型进行正则化？当时的想法是，我们只想为模型选择最重要的特征。因此，我们不想使用所有的变量，而只想使用其中的几个，特别是那些最能解释模型中关系的变量。显然，这样做有很多好处。提高可解释性，因为你不必解释 200 个变量中的每一个。你可以只解释其中的十个变量。这样也能提高准确性，因为你不会把水搅浑。所以这有点违反直觉，因为你可能会认为，参数越多，模型就越准确，但实际上并非如此。这样做的好处是，参数在减少。因此，即使你增加了参数的数量，在某些时候，你也无法获得更高的精确度。在某些情况下，即使是线性回归中相互影响的参数也会影响模型的准确性。因此，我曾有过这样的模型，去掉参数后，模型的准确性大大提高，因为突然间，一个与其他参数相互影响的参数被从等式中去掉了。因此，有时去掉回归变量是有意义的，而且由于显而易见的原因，优化这些变量也更容易。参数越少，计算压力越小。因此，我们在讨论套索回归和弹性网回归时，它们都非常相似。它们的理念都是对回归模型中的回归因子数量进行惩罚，从而试图减少回归因子的数量，或者通过套索回归完全去除回归因子，或者通过回归和弹性网络提升减少回归因子的数量并使其变小。所以，我们现在遵循的是非常相似的逻辑。是的。使用性别或国家等变量实际上会在我们选择这些变量时产生偏差。或者说，在数据分析中，我们是否应该将其包括在内，比如说，如果是我们说的给某人提供孩子的情况，包括他们的性别，这是否会对我们提供抵押贷款的对象产生偏差，或者说，这是否有助于我们消除偏差。这样更好我认为这是违法的我很确定你不能在模型中加入性别因素所以，比方说，我们不讨论抵押贷款。比方说，我们正在谈论 向他们发送电子邮件 别的东西。我没有权利，也不知道立法。所以，要回答你的问题，其实是个很有趣的问题，包括性别这个变量。例如，监管机构不允许进行信用评分。因此，你不能使用性别作为变量来决定是否给某个人，比如说，是否给他信用卡，或者是否给他贷款。你可能会认为这样做有好处，但据我所知，这其实是几年前出台的。据我所知，我的一位同事对此要了解得更多。如果你对这个话题感兴趣，可以去看看加琳娜-安德烈娃教授的出版物。她在商学院工作。她从事信用评分研究，对信用评分中的性别变量很感兴趣。如果我没记错的话，她在几个月前或几年前的一次演讲中说过，将性别作为信用评分模型中的一个变量，可以提高模型的准确性，而且它实际上特别改善了女性的状况，你可能会认为这是为女性提供的更好的支持。因此，我们的想法是将性别作为一个变量，因为在过去，女性在获得贷款时可能会处于不利地位。例如，我们现在谈论的是 50 或 60 年代。实际上，如果把性别作为一个变量，就会增加女性获得贷款的机会，因为她们往往更善于偿还贷款。因此，从理论上讲，将性别作为变量会更好，但出于监管原因，你不能这样做。是的。因此，当我们谈论偏见时，问题的关键在于我们的模式是否偏向谁，偏向什么方向，有哪些优缺点，以及我们所处的监管框架是什么。希望能对你有所帮助。回答你的问题。我之前举的例子可能考虑不周。所以我们还是把 h 去掉吧。另一方面，将年龄作为变量也是一个有趣的问题。因此，H 实际上也会使模型产生偏差。举例来说，如果你考虑到劳动力市场中年龄稍大的人往往很难找到新工作，因为他们年龄较大，人们认为他们可能很快就会退休。因此，有些变量是很难包括在内的。是的，这是底线。好的。那我们来谈谈降维。这可以看做是通过正则化进行特征选择的另一种方法，我们刚刚也谈到过，但它的主要设置非常不同。我们现在不再应用正则化来选择预测因子，而是尝试将预测因子进行组合。因此，我们通过组合现有的预测器，特别是通过创建线性组合，来创建新的预测器。我们试图通过这种方式来解释数据中最大的变异性。现在，这听起来有点奇怪，但我稍后详细解释时会更清楚。我们通常将因子分析和主成分分析区分为两类。从基本的数学概念来看，它们的工作原理非常相似。因此，如果你理解了其中一个，你就理解了这两个。但它们的使用环境不同。具体来说，因子分析通常用于市场营销文献和许多社会科学文献，而 PCA 主成分分析通常更多用于工程、Stem、科学等领域，但两者的基本逻辑是相同的。因此，让我们简单谈谈因子分析。这里的思路是，我们试图减少预测因子的数量或数据的维度，因为我们认为有某种潜在的因素在驱动着这些变量中的多个变量。因此，我们认为存在一个潜在的、未观察到的或潜在的因素，通过它可以解释所有这些变量。我们在调查中经常会遇到这个问题。因此，如果你曾经填写过一份调查问卷，你会发现很多问题听起来都很奇怪地相似。因此，他们可能会多次用不同的词来询问你认为是同一个概念或同一件事，有时甚至会调换量表。我们在市场营销中就是故意这样做的。其中一些原因是我们认为它们是同一个概念。这是因素分析，但我们并不完全确定。或者，我们认为它们是同一个概念，但我们不知道你是否也有同样的想法。在整个调查过程中，我们也会多次测量同一件事，因为你可能会觉得无聊，然后就开始测量中间的方框。因此，我们希望检查你的注意力。这也是为什么我们会把量表换个方向，看看你是否还在关注，或者你是否开始在你的量表上选择所有的四格或三格或其他。所以，我们的想法是，我们认为有一个概念。我们试图通过多个不同的变量来衡量这个概念，例如，调查中的多个问题。这个信念系统是非常重要的因素分析。这也是 PCA 和 PCA 因子分析的主要区别。我们不在乎。我们不关心是否存在一个通过相同变量描述的潜在因子。我们只想知道能否将它们合二为一，因为它们测量的是数据中的同一种东西，也就是数据中的同一种变异性。这就是我们所关心的。我们不关心信仰体系。我们不关心潜在的因素。我们只是试图组合并减少维度。所以，这就是为什么基本上他们在做同样的事情。事情。但原因不同。这就是为什么我们有时会将两者区分开来。是的，它们的工作原理是一样的。我们在这里只谈 PCA，因为因子分析的原理是一样的。但我们首先要做一点线性代数，才能理解我们在说什么。因此，我们今天将讨论两个奇妙的概念--特征值和特征向量。我终于可以自信地说出这两个词了。因为 "特征 "这个词是德语。所以，eigen 在德语中是 "大约自己的 "或 "自己的 "的意思。所以我们会说，这基本上就是特征值，它自己的值，或者特征向量，它自己的向量。所以，这终于是一个概念了，实际上我可以用类似高斯Gaussian的发音来发音。我说 "高斯 "是因为 "高斯 "也是德语。就是这样就是这样那么什么是特征向量的特征值呢？如果我们有一个正方形矩阵 A 乘以某个向量 x 等于某个标量。也就是说，一个数字乘以同一个向量。那么我们就可以称 lambda 为 A 的特征值，x 为特征向量。现在这就做到了。这看起来有点平淡无奇，但实际上非常有趣。它基本上意味着，如果你将这个矩阵与某个向量相乘，就等于将同一个向量与一个奇异数相乘，而不是与整个矩阵相乘。所以这其实是个非常有趣的概念。是的，我们可以通过寻找同一矩阵、乘以同一矩阵、求解同一矩阵并不是真正的解等等等等。真的很有趣。如果你想实际计算特征值和特征向量。举个例子，我们这里有一个数字示例，让计算变得更简单。如果我们这里有一个正方形矩阵，我们想计算它的特征值，我们可以按如下方法求解。我们将它与它的特征矩阵相乘，不是与它的特征矩阵相乘，而是与它的特征矩阵相乘。就是这样。概念不同我们可以求出 lambda。我就不一一列举了，因为我相信你可以自己求解。你基本上是在乘对角线。你在求和往上这是我的这是我的不，我的。是啊，是啊，是啊。她会没事的好吧 是的如果我们求解它 就能得到两个特征值 本例中是3和2如果我们把这些特征值 插到之前的等式中 就能计算出特征向量了这对你来说非常有趣我相信我相信你们都已经看过了。我把它放在这里是为了完整起见，因为它的计算。并不重要，但我认为更重要的是理解它。下一个好的我们先来了解一下什么是协方差矩阵，我相信你们都知道。你们都看到了，它基本上是一个相关性和协方差矩阵。是的，线性组合。我们也见过。那么我们来谈谈 PCA。我们已经讨论过 PCA 了。基本上，我们想要寻求简化。我们希望降低数据的维度，用更少的维度来涵盖数据中发生的事情。因此，这背后的想法是，它实际上可以让我们重新定位数据，从而减少解释数据所需的维度。在本专题中，我们还将讨论投影的概念。我认为这基本上就是 PCA 的核心。举个例子PCA 允许我们聚合在你的变量中成分。罗盘或彗星对不起，让我重新开始这句话，因为我写得很别扭。PCA 可以将原始变量中的共同信息聚合到新变量中。基本上，我们正在创建新的变量，我们称之为我们的成分，我们的主成分。但我们仍然希望数据中的信息量与之前的相同。因此，我们将原始变量组合在一起，然后根据它们实际传达的信息进行排序。在这种情况下，我们可以看到这里有三个原始变量。而不同的颜色基本上就是隐藏在它们之间的信息。因此，如果我们考虑因子分析，这就是这些不同变量所包含的不同概念。举例来说，这三个不同的问题和一份问卷，它们测量的是什么样的概念。因此，这三个问题所测量的概念是相同的，但数量不同。如果我们还想展示信息中的大部分变化和发生的大部分情况，我们首先要把这些数据分成三个不同的部分。我们可以看到，通过这些变量捕捉到的信息中，有 60% 是绿色的，30% 是橙色的，10% 是蓝色的。因此，这是相同的信息。我们只是把这些信息整合起来，把绿色部分、橙色部分和蓝色部分整合起来。现在，PCA 背后的理念是，与其把所有这些成分都提取出来，为什么不把最后一个成分去掉呢？因为只有 10%。如果只去掉最后一个部件，我们的损失并不会太大。我们基本上只损失了原始信息的 10%。但如果将数据的维度降低一维，突然间就只有两维数据、两个分量了。这就是 PCA 的基本思想。我们首先要看看有哪些可用信息。我们如何才能将这些信息组合成线性组合，从而只测量数据的一个方向，即数据的一个因子。然后，我们可以思考一下，我们的数据中到底需要多少这样的信息？有没有办法减少这些因素，从而减少数据的维度？那么，为什么我们之前要介绍这背后的数学逻辑呢？原因在于，如果我们有一个数据空间。我们有 n 个观测值，有一些相关向量。这些基本上就是我们的变量。因此，我们就有了某种点基。然后我们可以计算协方差矩阵。这样我们就可以看看我们是如何布局的。这些不同的数据点是否存在差异。我们可以将其计算为协方差矩阵。现在我们通过主成分法寻找线性组合。我们把这些向量中的 Z 称为不相关向量。这就是 "不相关 "的概念。因此，这个绿色空间与这个橙色空间没有任何关系。之前它们都是混在一起的。我们可以做到这一点。找到我们协方差矩阵的特征向量。注意，这就是我们的协方差矩阵。然后你。创建我们的主成分。这听起来很奇怪，我也不喜欢用这个概念来解释，因为在教我这个概念的时候，我从来没有理解过 PCA，直到有人给我看了一张图片，才让我明白了这个概念，其实很简单。想象一下，我们有一个二维空间。我们有一个变量 XX1 和两个变量 X2。这里有一个点模式。这些就是我们从数据中收集到的点。如果我问你，这些数据中是否存在线性趋势，你其实已经能看出来了。你会说是的，显然有。就像这个方向没错。是这样的模式。那如果我告诉你，我们要找的主成分就是这条线呢？就是这条线，我们要穿过它。因为通过这条线我们能看到什么？数据方差的最大值。所以我们感兴趣的是数据的分布，数据的方差。怎样才能捕捉到数据的最大方差呢？就是这个方向。因此，我们基本上是在绘制一个新的系统，一个新的坐标系，通过这样的方式来捕捉数据。它的差异。然后我们在绘制，因为我们知道我们的组成部分必须是相互独立的。你的在线对话就是通过这一点进行的。然后我们就有了一个新的坐标系，通过数据。现在你会问我，这和降维有什么关系？这仍然是二维的。即使你把这个空间稍微转一下，它还是二维的。它仍然是二维的。但如果你再想，好吧，这将是我们捕捉到的最大方差，而这是我们捕捉到的剩余方差。如果我们说，好吧，让我们忘掉这个。让我们忘掉上下的差值。我们感兴趣的是从左到右的方差。因为这才是数据中真正有趣的部分。我们并不关心它周围的这种微小的差值。我们关心的是这个大的分布。这样一来，你就可以通过查看沿其中一个轴的分布来降低数据的维度。这就是 PCA 的真正理念。因此，你要寻找的是数据中最大的分布、最大的方差。你可以通过数据的特征向量找到它，因为它是协方差矩阵的特征向量。如果你还记得，协方差矩阵只不过是对数据分布的描述。所以我认为，我认为这背后的想法要简单得多，你可以想想数据是如何扩散的？我们怎样才能通过一个新的坐标系最好地描述数据分布？没错。所以，如果你考虑一下特征向量间距，描述一下扩散的方向。这就是这条线。其特征值就是扩散的强度。数据沿着这条线扩散的范围有多大。因此，如果我们想知道我们到底需要多少个 PCA 或 PC 主成分，我们就要看这些特征向量的扩散程度或特征值有多大。我们会选择那些最大的特征向量。因此，最大的扩散、最大的特征值。我们基本上对特征值进行排序。我们寻找最大的特征值，通常是 70% 或 80%。我们说，好吧，我们不在乎 10 个或 20 个。我们不关心这里的小差值。我们关心的是大的。因此，我们将其缩小维度，只捕捉数据中 7% 到 80% 的分布。在二维空间中，这看起来更容易一些，但在三维空间中也是一样的。所以，如果你仍然认为还是要把数据投射到更低的维度。所以当我说我们忘掉这个轴的时候，我们只关心这个轴。这些点都可以投影到这个轴上。所以我们不在乎这个。我们说只有一个维度。我们把它全部缩小。所以我们要经常思考 "挤压效应"。我们基本上是把数据压缩成一个部分。如果你有两个数据点，你也可以把三个数据点分成两个部分。我们甚至可以将其可视化。这里有两个数据点。你可以将它们缩小到平面上的两个维度。所以，我希望大家记住的关键是，我们感兴趣的是，用比原始变量更少的分量，能解释多少方差？我之前提到过，一个经验法则是我们希望能解释 80% 的数据。可以是 70%，也可以是 90%。这取决于你的应用。但是，你在寻找我需要解释多少个成分、多少个维度、多少个特征向量？数据总方差的 80%。你可以通过特征值看到这一点。所以，这就是这些 "树枝图 "中的一个。这个名字很奇怪吧？X轴上是成分的数量，Y轴上是解释方差，你一定能认出这个图，因为它和我们的K-means的肘图非常相似。因此，我们又在寻找某种参数的某种特征。在这种情况下，它不是 K。所以耶！我们欢呼雀跃，但我们要找的仍然是一些成分，类似于 K-means，类似于 K。明天你们就能在计算机实验室里看到它的实际效果了。我们将在计算机实验室里进行演示。你们会看到如何绘制这些图。然后查找成分的数量。但只要你理解了这个逻辑，这个寻找如何解释大部分方差的想法，我们就能确定坐标系的方向。坐标系。就这样只要你能理解 我就很高兴了我很高兴。保险成分是的所以，让我想想。嗯。它应该在变体中丢失吗？让我检查一下这是非常奇怪的。让我考虑五分钟 Hmm.让我考虑五分钟然后再回来找你好的 Yes.好了，让我把话说完 因为我觉得我们快说完了关于 PCA，有一点很重要。它不是规模不变的。这是有道理的。我们刚刚看到了我们的坐标系。我们对传播差异很感兴趣。所以它显然不是比例不变的。所以你必须重新调整变量的比例，才能使用它。此外，如果尺度方差之间存在较大差异，那么这将决定你使用哪个分量。如果你对差异和很多显示出很大方差的变量感兴趣，这也是有意义的。这些变量将成为早期成分的驱动力。因此，一定要将变量标准化。因此，我们在很早很早的时候就提到过，我想是在第一讲中，当我们讨论有监督学习和无监督学习时，PCA 在技术上是无监督的，因为它不考虑任何因变量。因此，如果你做 PCA，你只需要考虑你的 X，基本上就是你的预测变量，然后你试图减少这些变量的数量或矩阵的维数，但你并不一定能创建出非常适合预测的成分，因为模型并不考虑实际的预测值或类别。还有一个最大的问题是，这些成分可能不太好解释。所以，我们之前在看这张图的时候就讨论过这个问题。这个成分一。这真的很不错。它解释了 60% 的变异性。但它到底是什么呢？我猜是这个的一部分和这个的一部分。但如果这是问卷中的两个问题，比如说，你并不真正知道那是什么。你可以试着解释它，你可以试着解释它。但归根结底，成分、主成分是无法准确解释的。你可以谈论载荷。这就是变量对每个成分的载荷。每个变量对每个成分的影响有多大。并试着这样解释。但最终还是很难解释。我还认为，这也是为什么因子分析没有脱离 PCA，而是作为一个讨论点发展起来的原因之一，因为它试图通过思考潜在的概念和驱动因素，来解释这些因子的想法，在这种情况下，这些因子就是我们的成分。但最终，你还是不确定。你无法像以前那样解释它们。战争变量你有战争变量解释 PCA 非常容易。维度越少，可解释性越低。让我们简单谈谈主成分回归。这是一个非常简单的概念。我们先简单介绍一下，因为它的原理是，我们先使用 PCA 来减少可能的预测因子的数量，然后使用主成分而不是变量来建立回归模型，而不是使用所有的预测因子。因此，举例来说，你不用再使用 100 个变量，而是使用 PCA。你会发现 20 个成分完全没问题。因此，这些只是我们之前拥有的所有变量的组合。然后你就有了一个只有 20 个因子的回归模型，这更容易训练，也更容易建立。因为我们说过，这少量的因子可以解释 80% 的数据，所以你的模型仍然具有很高的准确性。显然，我们说过线性回归的最大优势之一就是可解释性。因此，在这一点上，你会损失不少。但在很多情况下，如果你认为这些主成分的含义或解释对你来说并不重要，那么线性回归就真的很有优势。你更关心的是准确性，尤其是计算效率。所以是的，如果我们的假设成立，那么最小平方模型拟合实际上会带来更好的结果。如果我们将其拟合到成分中，那将是我们的 Z，而不是我们的 X，因为数据中的大部分或全部信息都在我们的成分中。这就是我们创建 PCA 的方法。方差的 80% 左右应该在成分中。因此，这可以很好地减少过度拟合。