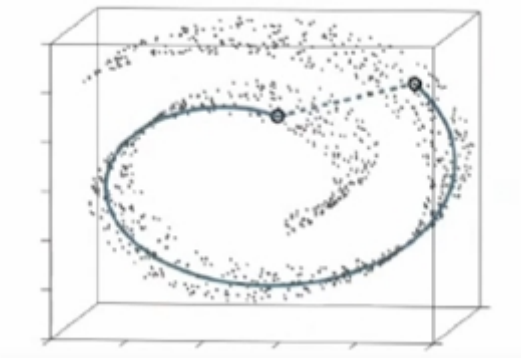
Что такое t-SNE?

Внедрение t-распределенных стохастических соседей (t-SNE) — это неконтролируемый нелинейный метод, который в основном используется для исследования данных и визуализации многомерных данных. Проще говоря, t-SNE дает вам ощущение или интуицию о том, как данные расположены в многомерном пространстве.

t-SNE vs PCA

Если вы знакомы с анализом основных компонентов (PCA), то, как и я, вам, вероятно, интересно, в чем разница между PCA и t-SNE. Прежде всего следует отметить, что PCA был разработан в 1933 году, а t-SNE — в 2008 году. С 1933 года в мире науки о данных многое изменилось, главным образом в области вычислений и размера данных. Во-вторых, PCA — это метод уменьшения линейной размерности, который стремится максимизировать дисперсию и сохраняет большие попарные расстояния. Другими словами, разные вещи оказываются далеко друг от друга. Это может привести к ухудшению визуализации, особенно при работе с нелинейными многообразными структурами. Думайте о многообразной структуре как о любой геометрической форме, такой как: цилиндр, шар, кривая и т. д.

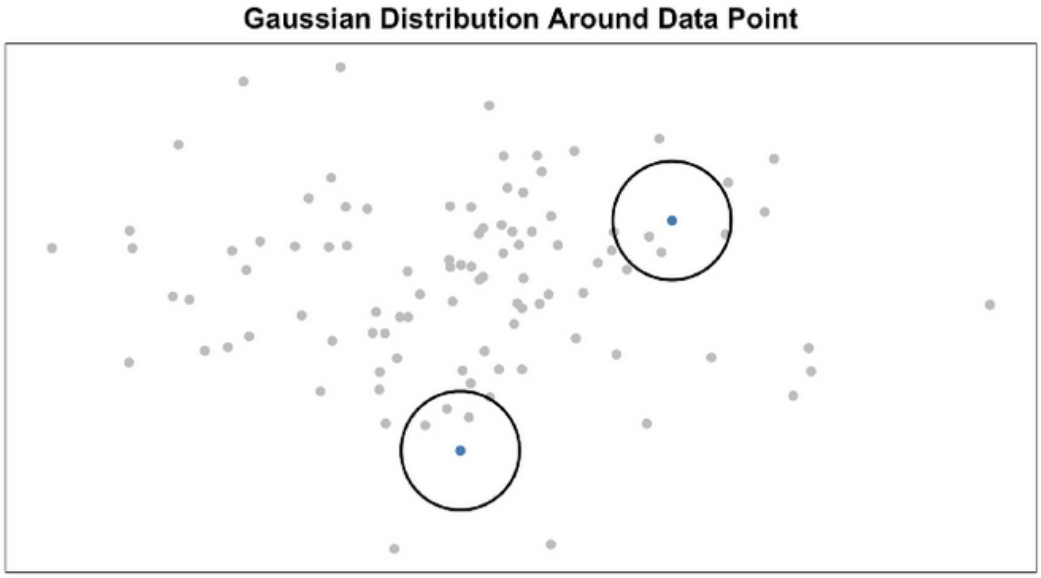
t-SNE отличается от PCA тем, что сохраняет только небольшие попарные расстояния или локальное сходство, тогда как PCA занимается сохранением больших попарных расстояний для максимизации дисперсии. Лоренс довольно хорошо иллюстрирует подход PCA и t-SNE, используя набор данных Swiss Roll на рисунке. Вы можете видеть, что из-за нелинейности этого игрушечного набора данных (многообразия) и сохранения больших расстояний PCA неправильно сохраняет структуру данных.



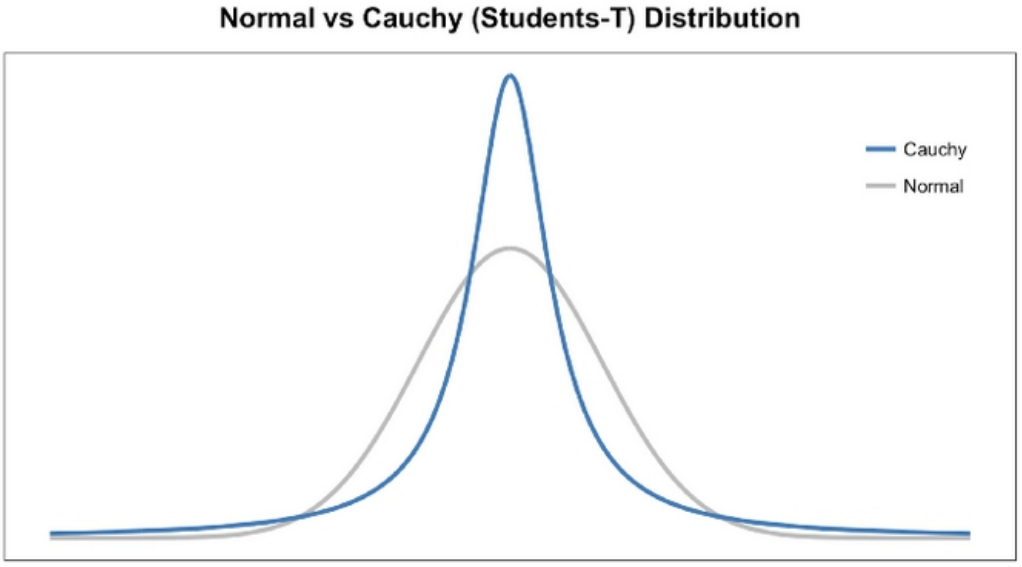
How t-SNE works

Теперь, когда мы знаем, почему мы можем использовать t-SNE вместо PCA, давайте обсудим, как работает t-SNE. Алгоритм t-SNE вычисляет меру подобия между парами экземпляров в пространстве высокой размерности и в пространстве низкой размерности. Затем он пытается оптимизировать эти две меры подобия, используя функцию стоимости. Давайте разобьем это на 3 основных шага.

Шаг 1. Измерьте сходство между точками в многомерном пространстве. Представьте себе набор точек данных, разбросанных по двумерному пространству (рис. 2). Для каждой точки данных (xi) мы центрируем гауссово распределение по этой точке. Затем мы измеряем плотность всех точек (xj) при этом распределении Гаусса. Затем перенормируйте для всех точек. Это дает нам набор вероятностей (Pij) для всех точек. Эти вероятности пропорциональны сходствам. Все это означает, что если точки данных x1 и x2 имеют равные значения под этим гауссовским кругом, то их пропорции и сходства равны, и, следовательно, у вас есть локальные сходства в структуре этого многомерного пространства. Гауссовым распределением или кругом можно управлять с помощью так называемого недоумения, которое влияет на дисперсию распределения (размер круга) и, по сути, на количество ближайших соседей. Нормальный диапазон недоумения составляет от 5 до 50 [2].



Шаг 2 аналогичен шагу 1, но вместо распределения Гаусса используется t-распределение Стьюдента с одной степенью свободы, также известное как распределение Коши (рис. 3). Это дает нам второй набор вероятностей (Qij) в маломерном пространстве. Как видите, t-распределение Стьюдента имеет более тяжелые хвосты, чем нормальное распределение. Тяжелые хвосты позволяют лучше моделировать большие расстояния.



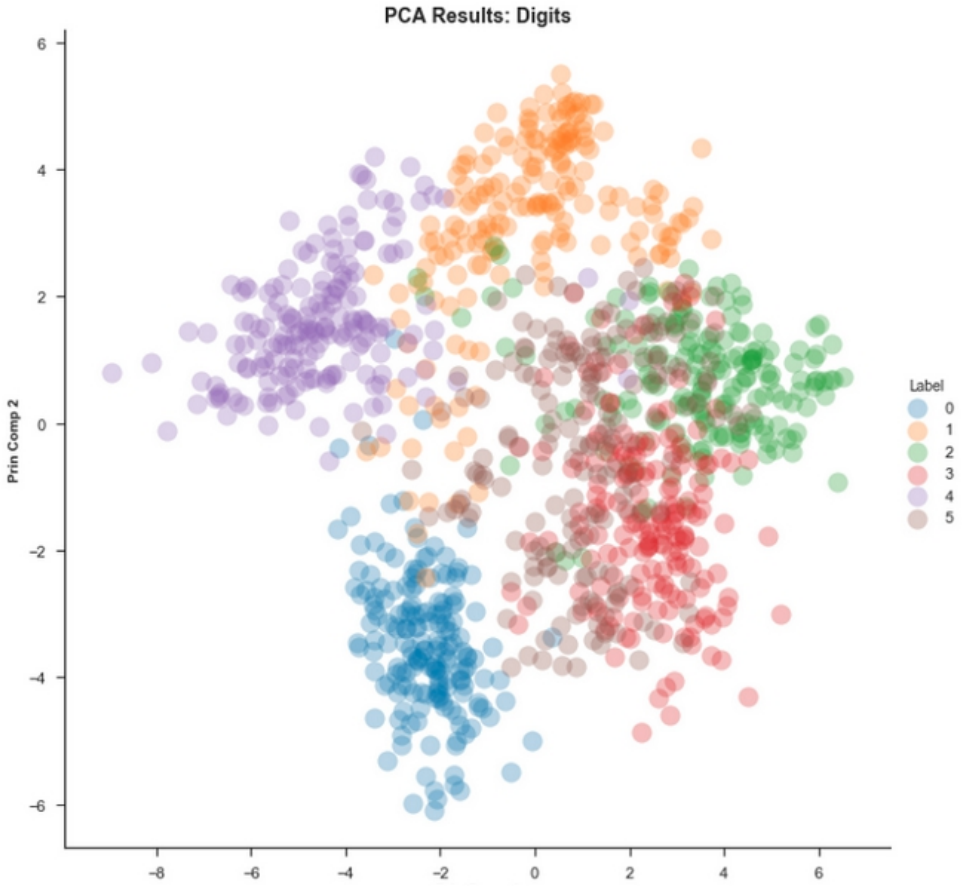
Последний шаг заключается в том, что мы хотим, чтобы этот набор вероятностей из низкоразмерного пространства (Qij) как можно лучше отражал вероятности из многомерного пространства (Pij). Мы хотим, чтобы две структуры карты были похожи. Мы измеряем разницу между вероятностными распределениями двумерных пространств, используя дивергенцию Кульбака-Либлера (KL). Я не буду слишком много вдаваться в KL, за исключением того, что это асимметричный подход, который эффективно сравнивает большие значения Pij и Qij. Наконец, мы используем градиентный спуск, чтобы минимизировать нашу функцию стоимости KL.

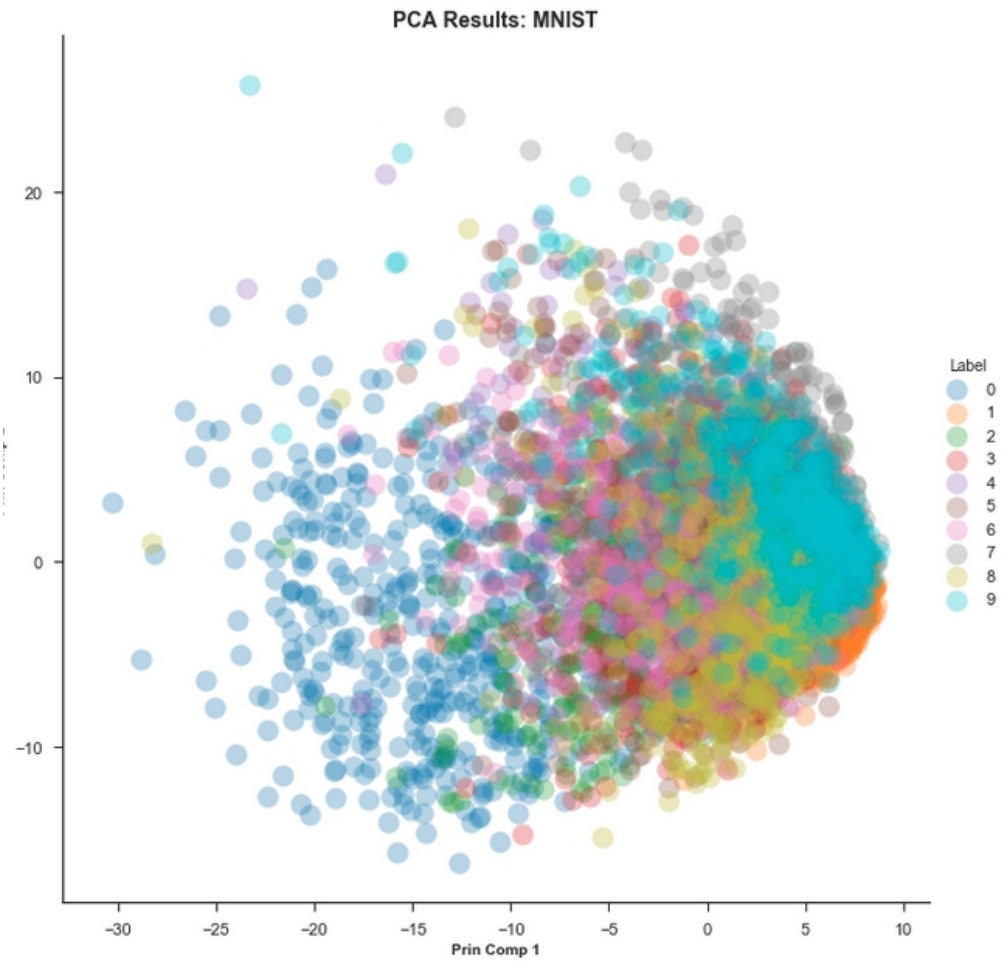
Пример использования t-SNE

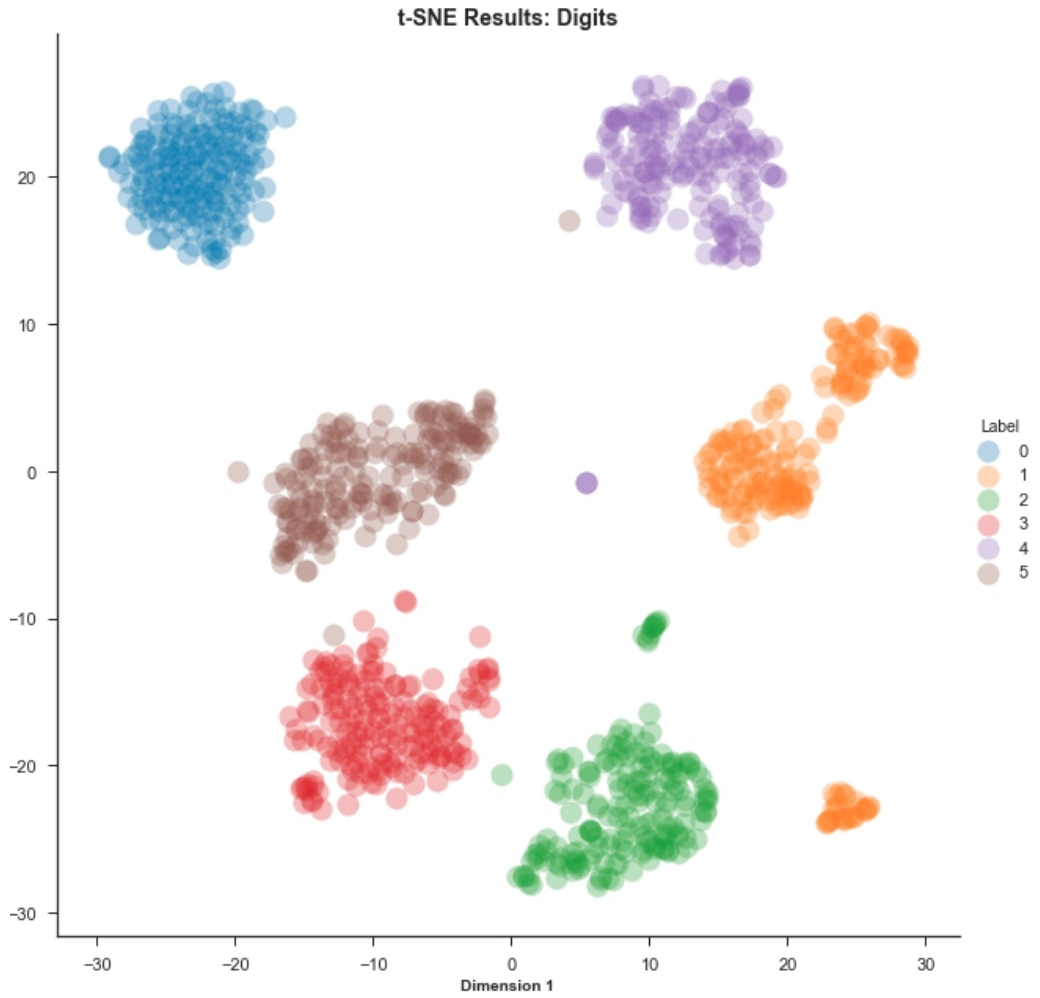
Теперь, когда вы знаете, как работает t-SNE, давайте быстро поговорим о том, где он используется. Мы можем использовать t-SNE в таких областях, как исследования климата, компьютерная безопасность, биоинформатика, исследования рака и т. д. t-SNE можно использовать для многомерных данных, а затем выходные данные этих измерений становятся входными данными для какой-либо другой модели классификации.

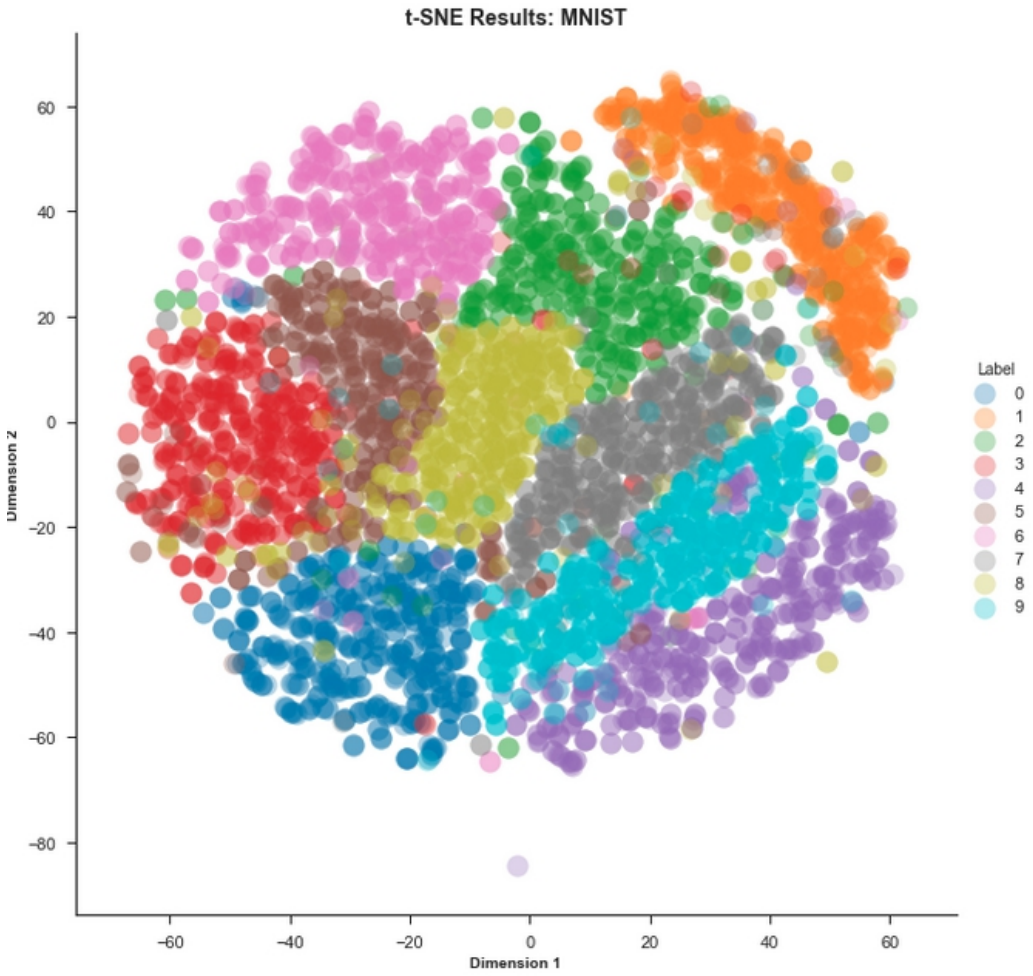
Кроме того, t-SNE можно использовать для исследования, изучения или оценки сегментации. Часто мы выбираем количество сегментов до моделирования или итерации после результатов. t-SNE часто может показывать четкое разделение в данных. Это можно использовать до использования вашей модели сегментации для выбора номера кластера или после, чтобы оценить, действительно ли ваши сегменты выдерживают. Однако t-SNE не является подходом к кластеризации, поскольку он не сохраняет входные данные, такие как PCA, и значения могут часто меняться между запусками, поэтому он предназначен исключительно для исследования.

Примеры




PCA Results: MNIST
20
10
0
−10
−30
−25
−20
−15
−10
−5
0
5
10
Prin Comp 1
Label
0
1
2
3
4
5
6
7
8
9


t-SNE Results: Digits
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30
Dimension 1
Label
0
1
2
3
4
5


t-SNE Results: MNIST
Dimension 2
Dimension 1
60
40
20
0
−20
−40
−60
−80
−60
−40
−20
0
20
40
60
Label
0
1
2
3
4
5
6
7
8
9