Non-linear Dimensionality Reduction

t-Distributed Stochastic Neighbor Embedding

Xiaoyu Xue

Microsoft BingAds UCM

Table of contents

Dimensionality Reduction

Dimensionality Reduction

Definition of dimensionality reduction:

Given a set of data $X = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$, $\mathbf{x}_i \in \mathbb{R}^n$, find a map $f : \mathbb{R}^n \to \mathbb{R}^d$, make $\mathbf{y}_i = f(\mathbf{x}_i)$ and $d \ll n$. Where $f = (f_1, \ldots, f_n)$, $f_i : \mathbb{R}^n \to \mathbb{R}$

Linear Dimensionality Reduction

If f_i is a linear map, $f = P = [\mathbf{p}_1, \ldots, \mathbf{p}_n]$, $\mathbf{y}_i = P^T \mathbf{x}_i$

Dimensionality reduction: Some Assumptions

1. High-dimensional data often lies on or near a much lower dimensional, curved manifold

2. A good way to represent data points is by their low-dimensional coordinates.

3. The low-dimensional representation of the data should capture information about highdimensional pairwise distances.

Dimensionality Reduction

- **Linear Dimensionality Reduction**: PCA(Principal Components Analysis), LDA(Linear Discriminant Analysis), MDS(Classical Multidimensional Scaling)

- **None-Linear Dimensionality Reduction**: Isomap(Isometric Mapping), LLE(Locally Linear Embedding), LE(Laplacian Eigenmaps), **tSNE**(t-Distributed Stochastic Neighbor Embedding)

Stochastic Neighbor Embedding

Define the similarity of data point $\mathbf{x}_i$ in original space as conditional probability $p_{j|i}$. It is the probability that $\mathbf{x}_i$ would pick *bolx$_j$* as its neighbor under a Gaussian centered at $\mathbf{x}_i$

$$p_{j|i} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2\right)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)}$$

In low-dimensional space, define the similarity $q_{j|i}$

$$q_{j|i} = \frac{\exp\left(-\|\mathbf{y}_i - \mathbf{y}_j\|^2\right)}{\sum_{k \neq i} \exp(-\|\mathbf{y}_i - \mathbf{y}_k\|^2)}$$

Cost function of SNE

If the map points $\mathbf{y}_i$ and $\mathbf{y}_j$ correctly model the similarity between the high-diminsional datapoints $\mathbf{x}_i$ and $\mathbf{x}_j$, the conditional probability $p_{j|i}$ and $q_{j|i}$ will be equal. Use the Kullback-Leibler divergences to minimize the mismatch:

$$Cost = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

If the map points $\mathbf{y}_i$ and $\mathbf{y}_j$ correctly model the similarity between the high-diminsional datapoints $\mathbf{x}_i$ and $\mathbf{x}_j$, the conditional probability $p_{j|i}$ and $q_{j|i}$ will be equal. Use the Kullback-Leibler divergences to minimize the mismatch:

$$Cost = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

To minimize the cost function

$$\frac{\partial Cost}{\partial \mathbf{y}_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(\mathbf{y}_i - \mathbf{y}_j)$$

Metropolis

The **metropolis** theme is a Beamer theme with minimal visual noise inspired by the HSRM Beamer Theme by Benjamin Weiss.

Enable the theme by loading

```
\documentclass{beamer}
\usetheme{metropolis}
```

Note, that you have to have Mozilla's *Fira Sans* font and XeTeX installed to enjoy this wonderful typography.

Sections

Sections group slides of the same topic

```
\section{Elements}
```

for which **metropolis** provides a nice progress indicator . . .

Titleformats

metropolis supports 4 different titleformats:

- Regular
- SMALLCAPS
- ALLSMALLCAPS
- ALLCAPS

They can either be set at once for every title type or individually.

Small caps

This frame uses the `smallcaps` titleformat.

Potential Problems

Be aware, that not every font supports small caps. If for example you typeset your presentation with pdfTeX and the Computer Modern Sans Serif font, every text in smallcaps will be typeset with the Computer Modern Serif font instead.

This frame uses the `allsmallcaps` titleformat.

Potential problems

As this titleformat also uses smallcaps you face the same problems as with the `smallcaps` titleformat. Additionally this format can cause some other problems. Please refer to the documentation if you consider using it.

As a rule of thumb: Just use it for plaintext-only titles.

This frame uses the `allcaps` titleformat.

Potential Problems

This titleformat is not as problematic as the `allsmallcaps` format, but basically suffers from the same deficiencies. So please have a look at the documentation if you want to use it.

Elements

Typography

```
The theme provides sensible defaults to
\emph{emphasize} text, \alert{accent} parts
or show \textbf{bold} results.
```

becomes

The theme provides sensible defaults to *emphasize* text, accent parts or show **bold** results.

Font feature test

- Regular
- *Italic*
- SMALLCAPS
- **Bold**
- **Bold Italic**
- **Bold SmallCaps**
- `Monospace`
- *`Monospace Italic`*
- `Monospace Bold`
- *`Monospace Bold Italic`*

Lists

Items	Enumerations	Descriptions
• Milk	1. First,	**PowerPoint** Meeh.
• Eggs	2. Second and	**Beamer** Yeeeha.
• Potatos	3. Last.	

- This is important

- This is important
- Now this

- This is important
- Now this
- And now this

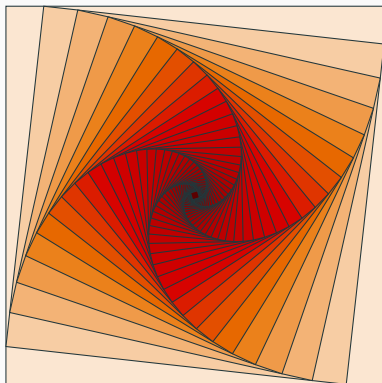- This is really important
- Now this
- And now this

Figure 1: Rotated square from texample.net.

Table 1: Largest cities in the world (source: Wikipedia)

City	Population
Mexico City	20,116,842
Shanghai	19,210,000
Peking	15,796,450
Istanbul	14,160,467

Blocks

Three different block environments are pre-defined and may be styled
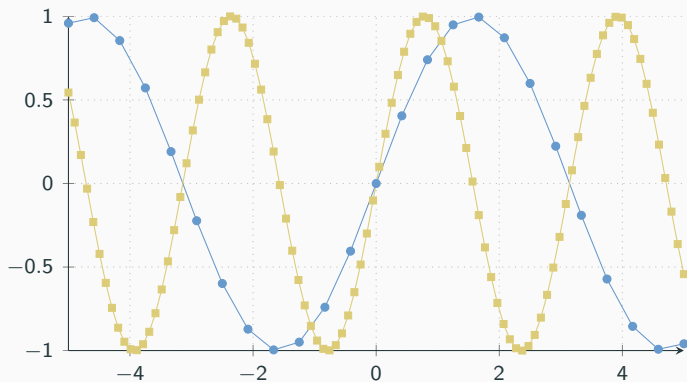with an optional background color.

Default	Default
Block content.	Block content.

Alert	Alert
Block content.	Block content.

Example	Example
Block content.	Block content.

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n$$
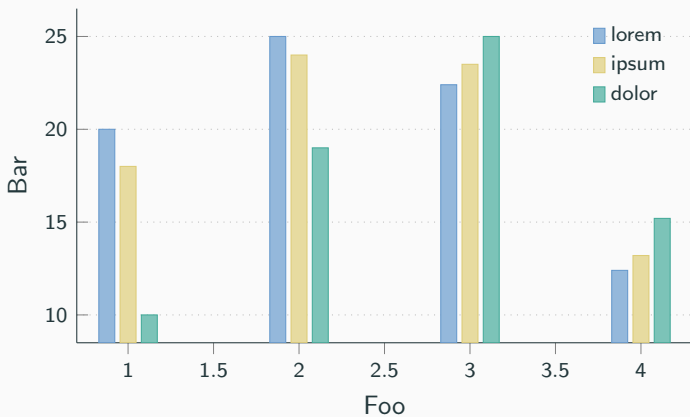
Veni, Vidi, Vici

Frame footer

metropolis defines a custom beamer template to add a text to the footer. It can be set via

`\setbeamertemplate{frame footer}{My custom footer}`

References

Some references to showcase [allowframebreaks] [4, 2, 5, 1, 3]

Conclusion

Summary

Get the source of this theme and the demo presentation from

github.com/matze/mtheme

The theme *itself* is licensed under a Creative Commons
Attribution-ShareAlike 4.0 International License.

Questions?

Backup slides

Sometimes, it is useful to add slides at the end of your presentation to refer to during audience questions.

The best way to do this is to include the appendixnumberbeamer package in your preamble and call \appendix before your backup slides.

metropolis will automatically turn off slide numbering and progress bars for slides in the appendix.

P. Erdős.
A selection of problems and results in combinatorics.
In *Recent trends in combinatorics (Matrahaza, 1995)*, pages 1–6.
Cambridge Univ. Press, Cambridge, 1995.

R. Graham, D. Knuth, and O. Patashnik.
Concrete mathematics.
Addison-Wesley, Reading, MA, 1989.

G. D. Greenwade.
The Comprehensive Tex Archive Network (CTAN).
TUGBoat, 14(3):342–351, 1993.

D. Knuth.
Two notes on notation.
Amer. Math. Monthly, 99:403–422, 1992.

H. Simpson.
Proof of the Riemann Hypothesis.
preprint (2003), available at
http://www.math.drofnats.edu/riemann.ps, 2003.