DD2434 Machine Learning, Advanced Course

Exercise session on module 7

Aristides Gionis

argioni@kth.se

KTH Royal Institute of Technology

some useful mathematical background

vector norms

- for a vector $\mathbf{x} = \langle x_1, \ldots, x_d \rangle \in \mathbb{R}^d$ we define the Minkowski p-norm, for $p \geq 0$, by

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}}$$

  in particular, the Euclidean norm is the Minkowski 2-norm, i.e., $\|\mathbf{x}\|_2 = \left(\sum_{i=1}^{d} x_i^2 \right)^{\frac{1}{2}}$

- but what is the norm of a matrix ?

matrix norms

- there are many different definitions for matrix norms,

  but two particularly useful and popular definitions are the following :

- given an $m \times n$ matrix $\mathbf{A}$, we define

  – the spectral normal

$$\|\mathbf{A}\|_2 \;=\; \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \;=\; \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2$$

  – the Frobenius norm

$$\|\mathbf{A}\|_F \;=\; \left(\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2\right)^{\frac{1}{2}}$$

recall : the singular value decomposition (SVD)

- theorem : any $m \times n$ matrix $\mathbf{A}$, with $m \geq n$, can be factorized into

$$\mathbf{A} = \mathbf{U} \begin{pmatrix} \boldsymbol{\Sigma} \\ \mathbf{0} \end{pmatrix} \mathbf{V}^T$$

  where $\mathbf{U} \in \mathbb{R}^{m \times m}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$ are orthonormal (i.e., $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{m \times m}$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}_{n \times n}$) and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ is diagonal

$$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)\,, \quad \text{where} \quad \sigma_1 \geq \ldots \geq \sigma_n \geq 0$$

- let us write SVD as $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, where $\boldsymbol{\Sigma}$ is "appropriately" padded with 0s

SVD and matrix norms

- we can compute matrix norms using the SVD

  for any $m \times n$ matrix $\mathbf{A}$ (with $m \geq n$), with SVD $\mathbf{A} = \mathbf{U\Sigma V}^T$, we have

  - the spectral normal

$$\|\mathbf{A}\|_2 = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{Ax}\|_2}{\|\mathbf{x}\|_2} = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2 = \sigma_1$$

  - the Frobenius norm

$$\|\mathbf{A}\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2\right)^{\frac{1}{2}} = \left(\sum_{i=1}^{n} \sigma_i^2\right)^{\frac{1}{2}}$$

- we will prove these results as exercise.

low-rank matrix approximation

- theorem : let $\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T$ be the singular-value decomposition of $\mathbf{A}$, let $\mathbf{U}_k = (\mathbf{u}_1 \ldots \mathbf{u}_k)$, $\mathbf{V}_k = (\mathbf{v}_1 \ldots \mathbf{v}_k)$, $\mathbf{\Sigma}_k = \mathrm{diag}(\sigma_1 \ldots \sigma_k)$, and define

$$\mathbf{A}_k = \mathbf{U}_k\,\mathbf{\Sigma}_k\,\mathbf{V}_k^T$$

then,

$$\min_{\mathrm{rank}(\mathbf{B}) \leq k} \|\mathbf{A} - \mathbf{B}\|_2 = \|\mathbf{A} - \mathbf{A}_k\|_2 = \sigma_{k+1}$$

and

$$\min_{\mathrm{rank}(\mathbf{B}) \leq k} \|\mathbf{A} - \mathbf{B}\|_F = \|\mathbf{A} - \mathbf{A}_k\|_F = \left(\sum_{i=k+1}^{n} \sigma_i^2 \right)^{\frac{1}{2}}$$

in other words, $\mathbf{A}_k$ is the best rank-k approximation for the matrix $\mathbf{A}$ with respect to both the spectral norm and the Frobenius norm

a useful inequality

- we will use Von Neumann's trace inequality to prove low-rank matrix approximation
- first, define the matrix inner product

$$\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}\mathbf{B}^T), \quad \text{for } \mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$$

  and observe that

$$\langle \mathbf{A}, \mathbf{A} \rangle = \mathrm{tr}(\mathbf{A}\mathbf{A}^T) = \|\mathbf{A}\|_F^2$$

- Von Neumann's trace inequality (stated here without proof):

  consider the matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, with $m \geq n$, and with singular values
  $\sigma_1(\mathbf{A}) \geq \ldots \geq \sigma_n(\mathbf{A}) \geq 0$ and $\sigma_1(\mathbf{B}) \geq \ldots \geq \sigma_n(\mathbf{B}) \geq 0$
  then

$$\langle \mathbf{A}, \mathbf{B} \rangle \leq \sigma_1(\mathbf{A})\sigma_1(\mathbf{B}) + \ldots + \sigma_n(\mathbf{A})\sigma_n(\mathbf{B})$$

recall : eigen-decomposition

- let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a square matrix

  $\lambda \in \mathbb{C}$ is an eigenvalue of $\mathbf{A}$, and $\mathbf{v} \in \mathbb{C}^n$, $\mathbf{v} \neq \mathbf{0}$ is an eigenvector of $\mathbf{A}$, if

$$\mathbf{A}\,\mathbf{v} = \lambda\,\mathbf{v}$$

- if matrix $\mathbf{A}$ is symmetric, then its eigenvalues are real and its eigenvectors are orthogonal

- $\mathbf{A}$ is positive semi-definite if $\mathbf{x}^T \mathbf{A}\,\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$:

  a symmetric positive semi-definite real matrix has real and non negative eigenvalues

math questions

question 1. (spectral norm)

- let $\mathbf{A}$ be an $m \times n$ matrix with SVD $\mathbf{A} = \mathbf{U\Sigma V}^T$

  show that the spectral normal of $\mathbf{A}$ is equal to σ_1 :

$$\|\mathbf{A}\|_2 = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{Ax}\|_2}{\|\mathbf{x}\|_2} = \sigma_1$$

answer on question 1. (spectral norm)

$$
\begin{aligned}
\|\mathbf{A}\|_2 &= \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \\
&= \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \\
&= \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{\Sigma}\mathbf{V}^T\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \\
&= \max_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{\Sigma}\mathbf{y}\|_2}{\|\mathbf{V}\mathbf{y}\|_2} \\
&= \max_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{\Sigma}\mathbf{y}\|_2}{\|\mathbf{y}\|_2} \\
&= \max_{\mathbf{y}\neq\mathbf{0}} \frac{\left(\sum_i \sigma_i^2 y_i^2\right)^{\frac{1}{2}}}{\left(\sum_i y_i^2\right)^{\frac{1}{2}}} \;\leq\; \sigma_1
\end{aligned}
$$

and for $\mathbf{y} = \langle 1, 0, \ldots, 0\rangle$ the maximum is attained

question 2. (Frobenius norm)

- let $\mathbf{A}$ be an $m \times n$ matrix with SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$

  show that the Frobenius normal of $\mathbf{A}$ is equal to $\sqrt{\sigma_1^2 + \ldots + \sigma_n^2}$:

$$\|\mathbf{A}\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2 \right)^{\frac{1}{2}} = \left(\sum_{i=1}^{n} \sigma_i^2 \right)^{\frac{1}{2}}$$

answer on question 2. (Frobenius norm)

- we want to show that $\|\mathbf{A}\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2\right)^{\frac{1}{2}} = \left(\sum_{i=1}^n \sigma_i^2\right)^{\frac{1}{2}}$
- we start by showing that $\|\mathbf{A}\|_F = \sqrt{\mathrm{tr}(\mathbf{A}^T\mathbf{A})}$

  the diagonal elements of $\mathbf{A}^T\mathbf{A}$ are

$$(\mathbf{A}^T\mathbf{A})_{jj} = \sum_{i=1}^m A_{ji}^T A_{ij} = \sum_{i=1}^m A_{ij} A_{ij} = \sum_{i=1}^m A_{ij}^2$$

  thus,

$$\mathrm{tr}(\mathbf{A}^T\mathbf{A}) = \sum_{j=1}^n (\mathbf{A}^T\mathbf{A})_{jj} = \sum_{j=1}^n \sum_{i=1}^m A_{ij}^2 = \|\mathbf{A}\|_F^2$$

answer on question 2. (Frobenius norm) cont'd.

- next we show that multiplying with an orthonormal matrix $(\mathbf{U}^T\mathbf{U} = \mathbf{I})$ does not change the Frobenius norm

$$\|\mathbf{UA}\|_F^2 = \mathrm{tr}((\mathbf{UA})^T(\mathbf{UA})) = \mathrm{tr}(\mathbf{A}^T\mathbf{U}^T\mathbf{UA}) = \mathrm{tr}(\mathbf{A}^T\mathbf{A}) = \|\mathbf{A}\|_F^2$$

  and similarly for orthonormal matrix $\mathbf{V}$ with $\mathbf{V}^T\mathbf{V} = \mathbf{I}$

$$\|\mathbf{AV}^T\|_F^2 = \mathrm{tr}((\mathbf{AV}^T)^T(\mathbf{AV}^T)) = \mathrm{tr}(\mathbf{VA}^T\mathbf{AV}^T) = \mathrm{tr}(\mathbf{V}^T\mathbf{VA}^T\mathbf{A}) = \mathrm{tr}(\mathbf{A}^T\mathbf{A}) = \|\mathbf{A}\|_F^2$$

- now we can show $\|\mathbf{A}\|_F = \left(\sum_{i=1}^n \sigma_i^2\right)^{\frac{1}{2}}$

$$\|\mathbf{A}\|_F^2 = \|\mathbf{U\Sigma V}^T\|_F^2 = \|\mathbf{\Sigma V}^T\|_F^2 = \|\mathbf{\Sigma}\|_F^2 = \sum_{i=1}^n \sigma_i^2$$

question 3. (low-rank matrix approximation)

- let $\mathbf{A} = \mathbf{U}\,\boldsymbol{\Sigma}\,\mathbf{V}^T$ be the singular-value decomposition of $\mathbf{A}$,
  let $\mathbf{U}_k = (\mathbf{u}_1 \ldots \mathbf{u}_k)$, $\mathbf{V}_k = (\mathbf{v}_1 \ldots \mathbf{v}_k)$, $\boldsymbol{\Sigma}_k = \mathrm{diag}(\sigma_1 \ldots \sigma_k)$, and define

$$\mathbf{A}_k = \mathbf{U}_k\,\boldsymbol{\Sigma}_k\,\mathbf{V}_k^T$$

- show that

$$\min_{\mathrm{rank}(\mathbf{B}) \leq k} \|\mathbf{A} - \mathbf{B}\|_F = \|\mathbf{A} - \mathbf{A}_k\|_F = \left(\sum_{i=k+1}^{n} \sigma_i^2 \right)^{\frac{1}{2}}$$

  in other words, $\mathbf{A}_k$ is the best rank-k approximation for the matrix $\mathbf{A}$
  with respect to the Frobenius norm

question 4. (low-rank matrix approximation, auxiliary)

- consider the matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, with $m \geq n$, and with singular values $\sigma_1(\mathbf{A}) \geq \ldots \geq \sigma_n(\mathbf{A}) \geq 0$ and $\sigma_1(\mathbf{B}) \geq \ldots \geq \sigma_n(\mathbf{B}) \geq 0$
- show that

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \geq \sum_{i=1}^{n} |\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})|^2$$

- hint: use the Von Neumann's trace inequality

$$\langle \mathbf{A}, \mathbf{B} \rangle \leq \sum_{i=1}^{n} \sigma_i(\mathbf{A}) \sigma_i(\mathbf{B})$$

answer on question 4. (low-rank matrix approximation, auxiliary)

- we have

$$\begin{aligned}
\|\mathbf{A}-\mathbf{B}\|_F^2 &= \langle \mathbf{A}-\mathbf{B}, \mathbf{A}-\mathbf{B}\rangle \\
&= \|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 - 2\langle \mathbf{A}, \mathbf{B}\rangle \\
&\geq \sum_i \sigma_i^2(\mathbf{A}) + \sum_i \sigma_i^2(\mathbf{B}) - 2\sum_i \sigma_i(\mathbf{A})\sigma_i(\mathbf{B}) \\
&= \sum_i |\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})|^2
\end{aligned}$$

answer on question 3. (low-rank matrix approximation)

- recall again,

  let $\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T$ be the singular-value decomposition of $\mathbf{A}$,

  let $\mathbf{U}_k = (\mathbf{u}_1 \ldots \mathbf{u}_k)$, $\mathbf{V}_k = (\mathbf{v}_1 \ldots \mathbf{v}_k)$, $\mathbf{\Sigma}_k = \mathrm{diag}(\sigma_1 \ldots \sigma_k)$, and define

$$\mathbf{A}_k = \mathbf{U}_k\,\mathbf{\Sigma}_k\,\mathbf{V}_k^T$$

- we want to show that

$$\min_{\mathrm{rank}(\mathbf{B}) \leq k} \|\mathbf{A} - \mathbf{B}\|_F = \|\mathbf{A} - \mathbf{A}_k\|_F = \left(\sum_{i=k+1}^{n} \sigma_i^2 \right)^{\frac{1}{2}}$$

- hint : we will use the auxiliary inequality

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \geq \sum_{i=1}^{n} |\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})|^2$$

answer on question 3. (low-rank matrix approximation)

- for any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{m \times n}$ with $\operatorname{rank}(\mathbf{B}) \leq k$ we have

$$\begin{aligned} \|\mathbf{A} - \mathbf{B}\|_F^2 &\geq \sum_{i=1}^{n} |\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})|^2 \\ &= \sum_{i=1}^{k} |\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})|^2 + \sum_{i=k+1}^{n} |\sigma_i(\mathbf{A})|^2 \\ &\geq \sum_{i=k+1}^{n} |\sigma_i(\mathbf{A})|^2 \end{aligned}$$

and the minimum is achieved for $\mathbf{B} = \mathbf{A}_k$

question 5. (eigenvalues and eigenvectors of a symmetrix matrix)

- show that a real symmetric matrix has real eigenvalues and orthogonal eigenvectors

answer on question 5. (eigenvalues of a real symmetrix matrix)

- since $\mathbf{A}$ is real and symmetrix, then

$$\mathbf{A} = \mathbf{A}^T = \mathbf{A}^*$$

where, $\mathbf{A}^*$ is the conjugate transpose of $\mathbf{A}$

- let $\lambda \in \mathbb{C}$ be an eigenvalue of $\mathbf{A}$, then, there exists non-zero $\mathbf{x} \in \mathbb{C}^n$, such that

$$\mathbf{A}\,\mathbf{x} = \lambda\,\mathbf{x} \quad \Rightarrow \quad \mathbf{x}^T\mathbf{A}\,\bar{\mathbf{x}} = (\mathbf{A}\,\mathbf{x})^T\bar{\mathbf{x}} = \lambda\,\mathbf{x}^T\bar{\mathbf{x}}$$

by taking the complex conjugate of both sides of $\mathbf{A}\,\mathbf{x} = \lambda\,\mathbf{x}$, we have

$$\mathbf{A}\,\bar{\mathbf{x}} = \bar{\lambda}\,\bar{\mathbf{x}} \quad \Rightarrow \quad \mathbf{x}^T\mathbf{A}\,\bar{\mathbf{x}} = \mathbf{x}^T(\bar{\lambda}\,\bar{\mathbf{x}}) = \bar{\lambda}\,\mathbf{x}^T\bar{\mathbf{x}}$$

- therefore $\lambda\,\mathbf{x}^T\bar{\mathbf{x}} = \bar{\lambda}\,\mathbf{x}^T\bar{\mathbf{x}}$, and since $\mathbf{x} \neq \mathbf{0}$, then $\lambda = \bar{\lambda}$, which means $\lambda \in \mathbb{R}$

answer on question 5. (eigenvectors of a symmetrix matrix)

- let $(\lambda_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, be eigenvalue-eigenvector pairs of symmetric matrix $\mathbf{A}$

  consider two pairs $(\lambda_i, \mathbf{x}_i)$, $(\lambda_j, \mathbf{x}_j)$ that $\mathbf{x}_i$ and $\mathbf{x}_j$ are not colinear

- multiplying $\mathbf{A}\,\mathbf{x}_i = \lambda_i\,\mathbf{x}_i$ by $\mathbf{x}_j^T$ from the left gives

$$\mathbf{x}_j^T \mathbf{A}\,\mathbf{x}_i = \lambda_i\,\mathbf{x}_j^T \mathbf{x}_i \quad \text{and similarly} \quad \mathbf{x}_i^T \mathbf{A}\,\mathbf{x}_j = \lambda_j\,\mathbf{x}_i^T \mathbf{x}_j$$

- transposing the second equation, and using $\mathbf{A} = \mathbf{A}^T$, gives

$$\mathbf{x}_j^T \mathbf{A}\,\mathbf{x}_i = \lambda_j\,\mathbf{x}_j^T \mathbf{x}_i \quad \text{and therefore} \quad (\lambda_i - \lambda_j)\,\mathbf{x}_j^T \mathbf{x}_i = 0$$

- if $\lambda_i \neq \lambda_j$, then $\mathbf{x}_j^T \mathbf{x}_i = 0$, and thus, $\mathbf{x}_j \perp \mathbf{x}_i$
- if $\lambda_i = \lambda_j$, then any linear combination of $\mathbf{x}_j$ and $\mathbf{x}_i$ is an eigenvector, with the same eigevalue, so we can select two orthogonal ones from the linear subspace they span

question 6. (a simple form of positive semidefinite matrix)

- show that if an $n \times n$ matrix $\mathbf{A}$ can be written as $\mathbf{A} = \mathbf{B}^T\mathbf{B}$,
  for some matrix $\mathbf{B} \in \mathbb{R}^{m \times n}$,
  then $\mathbf{A}$ is positive semidefinite

answer on question 6. (a simple form of positive semidefinite matrix)

- recall, our definition : an $n \times n$ matrix $\mathbf{A}$ is called positive semidefinite if $\mathbf{x}^T \mathbf{A}\,\mathbf{x} \geq 0$, for all vectors $\mathbf{x} \in \mathbb{R}^n$
- since $\mathbf{A} = \mathbf{B}^T\mathbf{B}$, for any $\mathbf{x} \in \mathbb{R}^n$ it is

$$
\begin{aligned}
\mathbf{x}^T \mathbf{A}\,\mathbf{x} &= \mathbf{x}^T \mathbf{B}^T \mathbf{B}\,\mathbf{x} \\
&= (\mathbf{B}\,\mathbf{x})^T \mathbf{B}\,\mathbf{x} \\
&= \|\mathbf{B}\,\mathbf{x}\|^2 \\
&\geq 0
\end{aligned}
$$

questions with no formal math proof,

but having a short and concrete answer

recall : categorization of dimensionality-reduction methods

- linear vs. non-linear model
- continuous vs. discrete model
- integrated vs. external estimation of the dimensionality
- layered vs. standalone embedding
- batch vs. online algorithm
- exact vs. approximate optimization

question 7. (layered embedding)

- recall the definition of layered embedding
- argue that PCA is a layered dimensionality-reduction method

answer on question 7. (layered embedding)

- let $\mathcal{C}_k$ be the set of components computed for embedding into target dimension k
- a method is layered if $\mathcal{C}_k \subseteq \mathcal{C}_{k+1}$
- for PCA:
  - $\mathcal{C}_k$ is the set of principal k eigenvectors
  - we know that the principal eigenvectors are layered
  - $\mathcal{C}_{k+1}$ is obtained by simply adding $(k+1)$-th principal component into $\mathcal{C}_k$
  - therefore, PCA is layered

question 8. (PCA normalization)

- we mentioned (and you have to argue about this for Assignment 1) that in PCA we always work with "centered" data
- but, do we have to normalize each column (feature), by dividing with the standard deviation after centering?

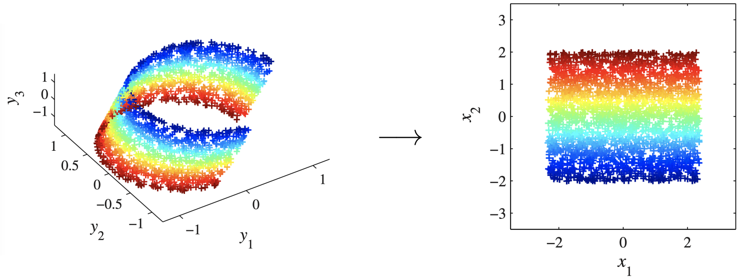
answer on question 8. (PCA normalization)

- it depends . . .
- normalization makes sense when attributes represent quantities in different units
  - temperature vs. distance vs. humidity
- normalization should not be done, when values between different attributes are comparable
  - e.g., in a documents $\times$ terms matrix, some terms may appear more frequently, leading to larger standard deviations, but this is important information to keep
- a trivial case when normalization should not be done
  - a column with standard deviation equal to 0

answer also discussed in [Lee, Verleysen] textbook, section 2.4.1

question 9. (PCA on “cardamon roll”)

- consider “cardamon roll” dataset, where data points lie on a 2-D manifold



- will PCA discover the hidden 2-D manifold?
  - if yes, why?
  - if no, how can we recover the 2-D maniford?

answer on question 9. (PCA on “cardamon roll”)

- no
  - PCA is a linear method
  - “cardamon roll” manifold is nonlinear

answer on question 9. (PCA on “cardamon roll”)

how to make it work?

1. isomap

2. kernel PCA

   - consider the kernel function $K(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2/2\sigma^2}$
   - $K(\mathbf{x}, \mathbf{y})$ can be seen as a similarity matrix between data points
     - identical points have value 1 and distant points have value 0
   - $K(\cdot, \cdot)$ goes to 0 expentially fast
     - set σ so that similarities of all non nearby points become 0
   - apply MDS with similarity matrix $\mathbf{K}$
   - selecting a different kernel function $K(\cdot, \cdot)$ gives different results

recall : example on Finnish dialects dataset

- data : 9000 dialect words, 500 counties

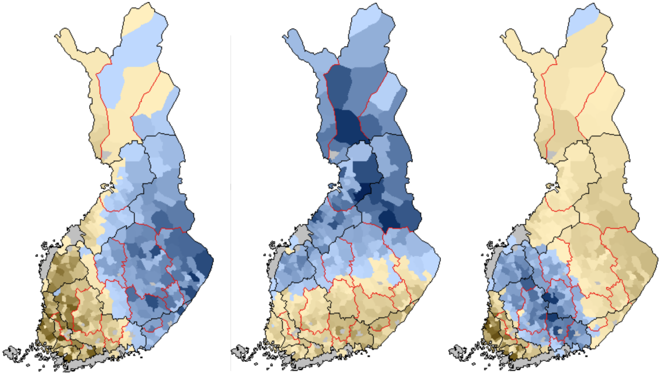
  points = words, dimensions = counties

  data matrix $\mathbf{Y}$, so that $y_{ij} = 1$ if word i appears in county j, and $y_{ij} = 0$ otherwise
- apply PCA to this data
- obtain principal component matrix $\mathbf{W} \in \mathbb{R}^{d \times k}$

example credited to Saara Hyvönen

question 10. (Finnish dialects dataset)

- we referred to the following figure as "visualization of first three components"



- but, what do the colors of the counties represent?
- why neighboring counties result in having similar colors?

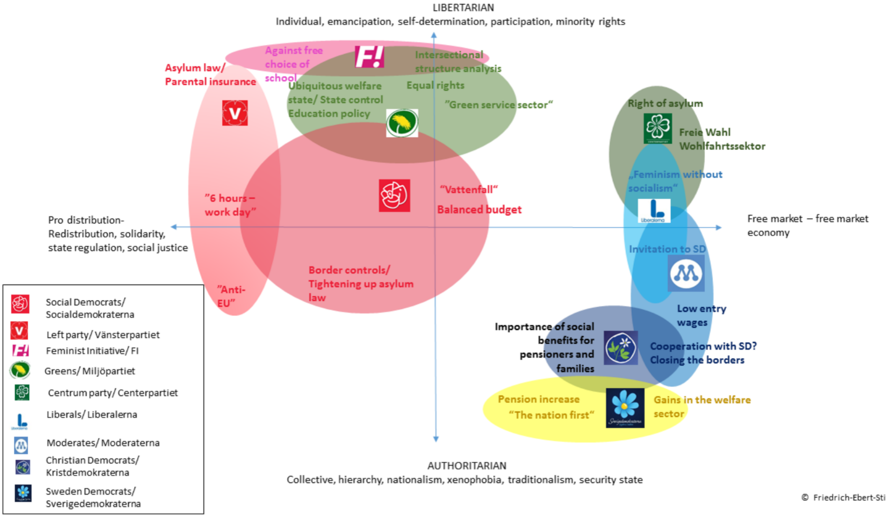
answer on question 10. (Finnish dialects dataset)

- data dimensionality (d) corresponds to counties
- a principal component is a d-dimensional vector
  - each county corresponds to a different coordinate of a component vector
- the color of a county in a component represents the value of the corresponding coordinate
- neighboring counties tend to use the same (or very similar) vocabulary
- coordinates corresponding to neighboring counties contribute in a similar manner to explaining words appearing in those counties

open-ended questions

question 11. (creating a political-compass visualization)

- how would you approach the problem of creating a "political compass" visualization of the political parties in Sweden, in a data-driven manner?
- what kind of help would you ask from a political scientist?

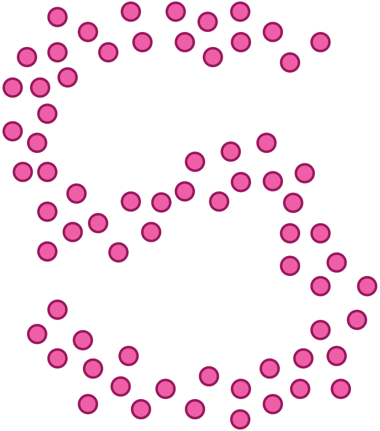


answer on question 11. (creating a political-compass visualization)

- identify a number (d) of key questions
  - e.g., economic freedom, personal freedom, foreign trade, ecology, immigration, etc.
  - $d \approx 20\text{-}30$ questions, which correspond to dimensions
  - help from political scientists here to identify the right questions
- conduct a survey over a few hundred people, obtaining answers to these questions as well as the political party they support
  - each person is a data point, in a d-dimensional space
- apply dimensionality projection to obtain a $k = 2$ dimensional embedding
  - use colors to represent political parties
  - apply statistical inference to represent political parties with probability distributions
  - transform the projected data to make the visualization more intuitive e.g., place left-leaning parties at the left, etc.

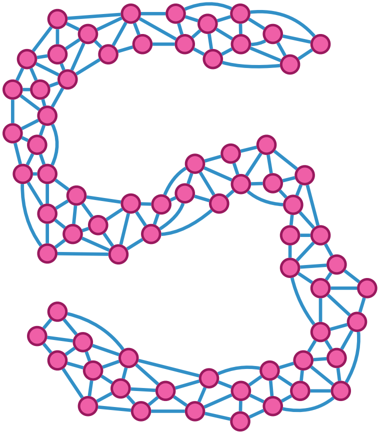
question 12. (robust Isomap)

- we want to embed the dataset below in 1-d using Isomap



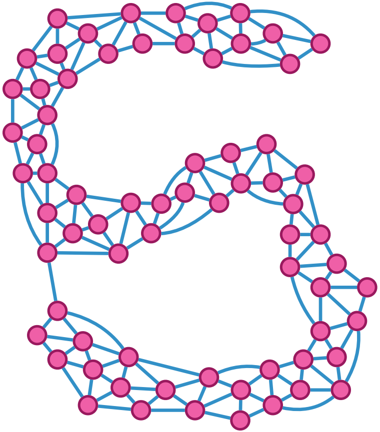
question 12. (robust Isomap)

- consider the k-nearest-neighbor graph constructed by Isomap



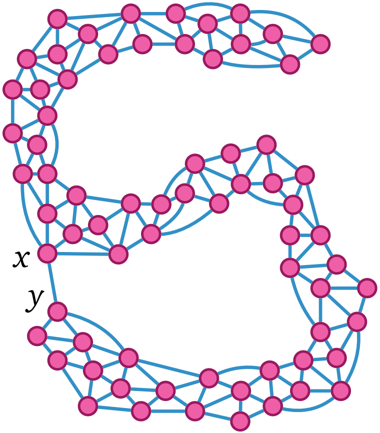
question 12. (robust Isomap)

- we may get some “undesirable” nearest neighbors



question 12. (robust Isomap)

- using shortest path distance may introduce erroneous distance evaluations



question 12. (robust Isomap)

- using shortest path distance on neighborhood graph is “brittle”
- how can we make it more robust?

answer on question 12. (robust Isomap)

- use other graph distances that are more robust
  - “commute time” distance
  - spectral graph embedding distance
- main idea for diffusion maps