Removing Redundancies From Data: Principle Component Analysis

COMS21202, Part III

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
Objectives

- Understand potential harm of high dimensionality of dataset
- Use Principle Component Analysis (PCA) to remove “redundant” dimensions from data.
High Dimensionality, Good? Bad?

- \( X = \{x_i\}_{i=1}^n, x \in R^d. \)

- Is a large \( d \) always a good thing?
  - ☺ We have more info as \( d \) grows!
  - ☹ LS does not work when \( d > n \)
  - ☹ Large \( d \) causes overfitting
  - More ☹ ?
Curse of Dimensionality (CoD)

- CoD is a generic term referring to the fact that many machine learning algorithms scale very poorly with $d$, in terms of performance.
  - Many geometry concepts work differently in higher dimensional space.
  - One of those concepts is “locality”.

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
The Vanishing Neighborhood

$V_0$: Neighborhood volume of $x_0$

$\frac{V_0}{V} = \frac{1}{k^2}$

$\lim_{d \to \infty} \frac{V_0}{V} = \frac{1}{k^d} = 0$

- Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
The Vanishing Neighborhood

- The neighborhood cube quickly vanishes as $d$ increases.
- As a result, your k-nearest neighbors are no longer in the neighborhood $V_0$.
- These neighbors are no longer good at predicting the label of $x_0$.
Reduce the Dimensionality using Feature Transform

- We want to find a feature transform $f(x) \in \mathbb{R}^m$, where $m \ll d$.
  - $f$ transforms original input $x$ to a subspace as $\mathbb{R}^m \subset \mathbb{R}^d$.
- We assume our dataset is centered:
  - $\frac{1}{n} \sum_{i=1}^{n} x_i = 0$
- If dataset $X'$ is not centered:
  - Centering: $\forall i \ x_i = x'_i - \frac{1}{n} \sum_{i=1}^{n} x'_i$

Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
Reduce the Dimensionality using Feature Transform

- What is the optimal strategy of selecting $f(x)$?
- Want to reduce dimension using $f$.
  - while preserving as much info as possible!
- Let’s look at this problem from data compression perspective!

- Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
Encoder and Decoder

\[ \begin{align*}
    &x \in \mathbb{R}^d \quad f(x) \quad \text{Encoding} \\
    &z \in \mathbb{R}^m \quad f'(z) \quad \text{Decoding} \\
    &x' \in \mathbb{R}^d
\end{align*} \]

Loss of Info: \( \| x_i - x'_i \|^2 \)

- Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
Codec

Original recording → Encoding → Spotify → Decoding → Music you hear

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
Linear Codec

- Suppose $f(x) = Bx^\top$, $B \in \mathbb{R}^{m \times d}$.
- Suppose $f'(z) = B'z^\top$, $B' \in \mathbb{R}^{d \times m}$.
- We can learn a codec by

$\min_{B,B'} \sum_{i=1}^{n} \left\| x_i^\top - B'Bx_i^\top \right\|^2$

- However, there are so many possible candidates $B$ and $B'$!
- Solving above problem is hard.

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
Linear Codec

- We need to put **constraints** on the $B$ and $B'$ to make our problem easier.
- One possible constraint is:
  - $B' = B^\top$
  - $BB' = BB^\top = I$
- Such a codec actually defines an orthogonal projection of $X$.
  - Show $B'B$ is an orth. projection matrix

- Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
Orthogonal Projection

\[ x_i' = B^\top B x_i^\top \]

\[ z_i = f(x_i) = B x_i^\top \]

is called an embedding of \( x_i \),

\( B \) is called embedding matrix.

- Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
A Pizza Topping Analogy of Embedding

- Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
Minimizing Projection Error

\[ \min_{B, BB^\top = I} \sum_{i=1}^{n} \left\| x_i^\top - B^\top B x_i^\top \right\|^2 \]

We minimize square error between original data points and its projection.

The above problem is equivalent to:

\[ \max_{B, BB^\top = I} \text{tr}(BX^\top XB^\top) \]

Live demonstration

---

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
Minimizing Projection Error

\[ \max_{B, BB^\top = I} \text{tr}(BX^\top XB^\top) \]

Remarkably, this seemingly complex optimization has an analytical solution:

Let \([(\lambda_1, v_1), \ldots, (\lambda_m, v_m)]\) be sorted eigenvalue and eigenvector of \(X^\top X\).

\(\lambda_1 \geq \lambda_2 \ldots \geq \lambda_m\)

\(\hat{B} = [v_1, v_2, \ldots, v_m]^\top\) is an optimal solution, suppose \(v_i\) is a column vector.

Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
Principle Component Analysis

- As $X$ is a centered dataset, $X^T X = n \cdot \text{cov}[x]$ (PC: show it!)

- Computing $\hat{B}$ via computing sorted eigenvectors of $\text{cov}[x]$ is called Principle Component Analysis (PCA).

- Finally, embedding $\hat{f}(x_i) = \hat{B} x_i^T \in R^m$ is called PCA embedding of $x_i$. $m$ dimensional “compression” we want!

- Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
Refresh: Eigenvectors and Eigenvalues

- Given a square $n \times n$ matrix $A$, if there exists non-zero vector $\mathbf{v}$ such that $A\mathbf{v} = \lambda \mathbf{v}$, $\mathbf{v} \in \mathbb{R}^n$

- Then $\lambda$ is an eigenvalue and $\mathbf{v}$ is an eigenvector of $A$. 

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
Example, One Cluster

$v_1$ always points at the direction where your dataset has the largest variance!

PC: Intuitively explain why.

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
Example, Embedding $z = \nu_1^T x^T$
Example, Two Clusters

However, PCA embedding does not necessarily preserve clustering information.

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.
Example, Embedding $z = \mathbf{v}_1^\top \mathbf{x}^\top$

Cluster information **lost** after embedding!
Will address this issue in the next lecture.

- Song Liu ([song.liu@bristol.ac.uk](mailto:song.liu@bristol.ac.uk)), Lecturer in Data Science and A.I.
Conclusion

- Curse of Dimensionality
  - $d$ increases, performance may decrease.

- Principle Component Analysis
  - Finding an embedding matrix $\hat{B}$ by computing sorted eigenvalue/vectors of $\text{cov}[x]$.
  - $f(x_i) = \hat{B}x^\top$.