Extra Credit

CMSC 691 — Introduction to Data Science

Faisal Rasheed Khan

VB02734

vb02734@umbc.edu

Principal Component Analysis is mostly used for reducing the dimensions and in the analysis of data in a lower dimension space.

Reasons to perform PCA:
Dimensionality Reduction:
With the help of PCA, we remove unnecessary dimensions of a dataset which are not useful, and transform the dataset into lower dimensional space, in order to overcome the problems which are faced by the dataset with large dimensions.

Feature Selection:
With the help of PCA, we can ignore unnecessary features and have only important features for the dataset which can explain the variance in the data.

Noise Reduction:
By focussing on the most significant features, PCA can handle the noise in the dataset.

Data Compression:
The unnecessary data is reduced and then the whole data set is compressed by maintaining necessary features.

Data Visualization:
PCA helps to understand the data in a lower dimensional space from a higher dimensional space.

Academic Papers:
Xueguang Ma, Minghan Li, Kai Sun, Ji Xin, and Jimmy Lin. 2021. Simple and Effective Unsupervised Redundancy Elimination to Compress Dense Vectors for Passage Retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2854–2859, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

The above paper explores approaches for compressing dense vectors for passage retrieval, which is relevant to PCA, a dimensionality reduction method. PCA is a known technique to minimize data dimensionality while maintaining variance. The paper's focus on eliminating duplication or compressing vectors aligns with PCA's dimensionality reduction objective.

Vilém Zouhar, Marius Mosbach, Miaoran Zhang, and Dietrich Klakow. 2022. Knowledge Base Index Compression via Dimensionality and Precision Reduction. In *Proceedings of the 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge*, pages 41–53, Dublin, Ireland and Online. Association for Computational Linguistics.

The study focuses on dimensionality and precision reduction in knowledge base index compression. PCA, which is as a way to reduce dimensionality method, could be useful for analyzing or evaluating techniques for reducing the dimensionality of knowledge sets. PCA is often utilized to reduce data dimensionality. In this context, the paper may explore approaches for compressing knowledge base indices, which are in line with PCA's goal of reducing dimensions while preserving important data.

The data set used to perform PCA is the iris dataset which is available in the sklearn library. I have performed the PCA on the dataset, by reducing its dimensions and performing data visualization (2D plot)

Inferences drawn after PCA Analysis: (PCA_IDS.ipynb)

After completing PCA Analysis on the iris dataset, I have plotted the 2D visualization of the 2 principal components. The plot shows the data of the principal components. The variance also is there to inference the information each principal component holds.
The variance ratio helps in understanding of how much variance in the original data is captured by the PCA components. So by maintaining the good features and excluding unnecessary features, we can still get a good model
The curse of dimensionality is reduced and the computations are speed up, reduced dimensions doesn’t sacrifice the relevant information. Reduced information helps in identifying good relations in the dataset for the variables

```
Explained variance ratio: [0.72962445 0.22850762]
```

1st principal component of iris dataset explains 0.72 of variance and the 2nd principal component of iris dataset explains 0.23 of variance, they together, the two principal components capture around 95% of entire variance for the iris dataset.
The accuracy of the PCA and without PCA model is almost nearer, therefore, it implies that PCA captures the important information by reducing the unnecessary features and reducing the dimensions.

References:

scikit-learn: machine learning in Python — scikit-learn 1.3.2 documentation

https://blackboard.umbc.edu/bbcswebdav/pid-6484632-dt-content-rid-72877506_1/xid-72877506_1

Xueguang Ma, Minghan Li, Kai Sun, Ji Xin, and Jimmy Lin. 2021. Simple and Effective Unsupervised Redundancy Elimination to Compress Dense Vectors for Passage Retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2854–2859, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vilém Zouhar, Marius Mosbach, Miaoran Zhang, and Dietrich Klakow. 2022. Knowledge Base Index Compression via Dimensionality and Precision Reduction. In *Proceedings of the 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge*, pages 41–53, Dublin, Ireland and Online. Association for Computational Linguistics.