Dimensionality reduction

Examples of data that often require dimensionality reduction either for visualization or for modeling purposes include images, texts, signal processing data, astronomical data, and health data. The sklearn.decomposition module includes a number of matrix decomposition algorithms including PCA, NMF and ICA. Matrix decomposition has been used for a long time to enable dimensionality reduction. One major drawback to using PCA is that non-linear or curved surfaces tend to not be well-explained by the approach. Manifold learning for dimensionality reduction has gained a lot of traction recently. In particular the t-distributed stochastic neighbor embedding (tSNE) family of approaches have become a viable alternative to PCA. It is also worth noting that feature selection techniques like using an ANOVA to select K features (see SelectKBest) is also a valid form of dimensionality reduction.

Topic models

Latent Dirichlet Allocation (LDA) and non-negative matrix factorization (NMF) are both commonly used in the context of topic modeling. Generally, these approaches use a bag-of-words representation. These models are in practice a form of dimensionality reduction. The embedding approach tSNE is often used to visualize the results of topic model representations in lower dimensional space to both tune the model as well as gather insights from the data. The package pyLDAvis is specifically purposed with visualizing the results of these models.

Topic modeling has a number of use cases apart from feature engineering for supervised learning. There is utility in being able to organize a large corpus of data. Take for example a law firm that has used the same types of forms for decades. Before the form was electronic it was simply put in a folder. Now that the forms are electronic they are organized into categories. LDA can be used to model the new corpus before the trained model is fed the historical documents. The trained model would then make probability estimates for membership in the categories.

Additional resources

- Review paper about topic model applications
- LDA was used to extract information from clinical notes in this example about sleep medication prescriptions and clinical decision-making