Mentor Help

Ask a mentor on our Q&A platform

Dimensionality Reduction and PCA - Lesson Topics

Phew! That was a ton of information - here is a quick recap!

1. Two Methods for Dimensionality Reduction

You learned that **Feature Selection** and **Feature Extraction** are two general approaches for reducing the number of features in your data. Feature Selection processes result in a subset of the most significant original features in the data, while Feature Extraction methods like PCA construct new latent features that well represent the original data.

2. Dimensionality Reduction and Principal Components

You learned that Principal Component Analysis (PCA) is a technique that is used to reduce the dimensionality of your dataset. The reduced features are called **principal components**, or **latent features**. These **principal components** are simply a linear combination of the original features in your dataset.

You learned that these components have two major properties:

1. They aim to capture the most amount of variability in the original dataset.
2. They are orthogonal to (independent of) one another.

3. Fitting PCA

Once you got the gist of what PCA was doing, we used it on handwritten digits within scikit-learn.

We did this all within a function called `do_pca`, which returned the PCA model, as well as the reduced feature matrix. You simply passed in the number of features you wanted back, as well as the original dataset.

4. Interpreting Results

You then saw there are two major parts to interpreting the PCA results:

1. The **variance explained** by each component. You were able to visualize this with scree plots to understand how many components you might keep based on how much information was being retained.
2. The **principal components** themselves, which gave us an idea of which original features were most related to why a component was able to explain certain aspects about the original datasets.

5. Mini-project

Finally, you applied PCA to a dataset on vehicle information. You gained valuable experience using scikit-learn, as well as interpreting the results of PCA.

With mastery of these skills, you are now ready to use PCA for any task in which you feel it may be useful. If you have a large amount of data, and are feeling afflicted by the curse of dimensionality, you want to reduce your data to a smaller number of latent features, and you know just the way to do it!

6. Do you think you understand PCA well enough yet to explain it in a way that would make sense to your grandmother?

Here is an interesting StackExchange post that does just that, and with animated graphics! https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues