M3M4 - PCA and Weather Analysis

1 Lectures: Covariance and PCA

2 Lectures: Visualizing PCA Coefficients I

3 Lectures: Visualizing PCA Residuals

4 Lectures: Visualizing PCA Residuals II

总结:

1 Lectures: Covariance and PCA

1.1 High Dimensional Vectors

详见 jupyter notebook: 1.FunctionsAsVectors.ipynb

POLL
How can we visualize a 5-dimensional vector?

RESULTS

f(1, 2, 3, 4, 5)	74%
[1 0]	1%
As a d-1 dimensional array	18%
Other	7%

FEEDBACK
We can visualize any d-dimensional vectors as a function from 1, 2, ..., d to the reals. In this case, we can visualize a 5-dimensional vector as a function from 1, 2, 3, 4, 5 to the reals.

POLL
What is a Fourier basis?

RESULTS

An arbitrary set of functions	2%
A set of constant functions	1%
A set of vectors in a vector space	3%
a set of orthonormal functions made or sines and cosines	92%
any set of orthogonal functions	1%

总结：

1.2 Computing PCA Using RDD

详见 jupyter notebook: 2.PCA_computation per state.ipynb

POLL

Why do we use an RDD instead of DataFrames

RESULTS

RDD's treat nan values correctly	69%
DataFrames have little memory capacity	4%
DataFrames will be too slow on big data operations	18%
None of the above - we want to avoid using RDD's	9%

POLL

Why would we perform Principal Component Analysis?

RESULTS

It gives us a way to expand our dataset based on the patterns of the original set	10%
It gives us a way to analyze covariance correlations between elements of the vector	82%
It allows us to add dimensions to our set to make calculations easier	5%
We do not want to perform PCA because we do not want to reduce dimensions	3%

总结:

1.3 Loading the Data

详见 jupyter notebook: 2.PCA_computation per state.ipynb

POLL

Why is it unwise to perform early-stage measurements on sets for which there is sparse and little data?

RESULTS

It makes dimensionality reduction techniques such as PCA more difficult	16%
In the early stages of analysis, we are just trying to get a global picture of the distribution	47%
Any analysis we would perform would yield NaN values	26%
Even if there is little data for some measurements, we want to analyze them to get the fullest picture of the data we can get	11%

总结:

2 Lectures: Visualizing PCA Coefficients I

2.1 Weather Data Visualization

详见 jupyter notebook: 3. Weather Analysis - Initial Visualisation.ipynb

POLL
How can we register a DataFrame as a table in Spark?

RESULTS

sqlContext.registerDataFrameAsTable(...)	90%
sqlContext.parquetFile(...).dtypes	5%
sqlContext.DataFrameReader(...)	4%
DataFrameReader.load(...)	1%

2.2 Variance Explained by Eigenvectors

详见 jupyter notebook: 3. Weather Analysis - Initial Visualisation.ipynb

POLL
Which of the following gives the best representation of a set of data?

RESULTS

An eigenvector explaining a large variance	81%
An eigenvector explaining a small variance	13%
An eigenvector with explaining a zero variance	6%

总结:

Quiz

Properties of orthogonal and orthonormal sets I

0.0/10.0 points (graded)

Mark all true statements

You may want to refer this great tutorial before you attempt this question.

- [x] Any orthogonal set of non-zero vectors is linearly independent ✔
- [x] Any orthonormal set of non-zero vectors is linearly independent ✔
- [ ] Any linearly independent set of vectors is orthogonal
- [ ] Any linearly independent set of vectors is orthonormal
- [x] The set of vectors {e1 = (1, 0, 0), e2 = (0, 1, 0), e3 = (0, 0, 1)} forms an orthonormal basis of R^3 ✔

Fourier Series

0.0/15.0 points (graded)

Mark all true statements

- [x] A Fourier series is a way to represent a periodic function as the sum of simple sine and cosine waves ✔
- [ ] To get a better approximation of the periodic function, we should decrease the number of terms in its Fourier series representation
- [x] The first term in a Fourier series is the average value of the function being approximated ✔
- [x] Fourier series is one of the many applications of orthonormal bases ✔
- [ ] Coefficients of all terms in the Fourier Series should be the same
- [x] The magnitude of the coefficient of a particular term in the Fourier Series describes its influence in approximating the periodic function under consideration ✔

总结：

Dealing with NaN values

0.0/15.0 points (graded)

Mark all true statements

- [ ] `numpy.mean(x)` only considers non-nan values of x while calculating the mean.
- [x] `numpy.nanmean(x)` only considers non-nan values of x while calculating the mean. ✔
- [ ] It is recommended to use `numpy.nanmean(x)` everytime we wish to calculate mean of x, irrespective of the distribution of the missing values of x (which have been replaced by `nan`)
- [x] When dealing with big data, it is often the case that our dataset has a lot of missing values which can adversely affect the operation of machine learning algorithms on that dataset ✔
- [ ] `numpy.mean([1,2,np.nan]) = 1.5`
- [x] `numpy.nanmean([1,2,np.nan]) = 1.5` ✔
- [ ] `numpy.nanmean([1,2,np.nan]) = 1`

computeStatistics

0.0/10.0 points (graded)

Read the all the python files in the Section2-PCA/PCA/lib/ directory from the git repository. Which of the following statistics are calculated by the computeStatistics function on the given data?

- [ ] Mean of the given data
- [x] Mean of a sample of the given data ✔
- [ ] Median of a sample of the given data
- [x] Standard Deviation ✔
- [ ] Eigen Values and Eigen Vectors using the Co-variance computed from the given data ✔
- [x] The set of values that fall between the 99th and 100th percentile of a sample of the given data ✔

总结:

3 Lectures: Visualizing PCA Residuals

3.1 Spectral Analysis of Snow Depth in NY Part 1

详见 jupyter notebook: 4. Weather Analysis - reconstruction SNWD.ipynb

POLL
What is the residual norm?

RESULTS

the crossproduct of the eigenvectors	4%
A 0 value for the residual	4%
the squared sum of the square norm of each residual	44%
the square norm of the residual vector	49%

FEEDBACK
After we've found the residuals, we compute their square norms. Smaller values for the residual norm mean a better approximation.

POLL
As we add more coefficients, what happens to the residual norm?

RESULTS

It gets smaller	79%
It gets larger	14%
The value of the residual norm stays the same	7%

FEEDBACK
As we add coefficients, the value for your residual norm gets better (smaller).

总结:

3.2 Spectral Analysis of Snow Depth in NY Part 2

详见 jupyter notebook: 4. Weather Analysis - reconstruction SNWD.ipynb

POLL

What would you expect to see as you move from one to two coefficients?

RESULTS

We see a better approximation of the data	95%
We see a worse approximation of the data	4%
We see a faster runtime	0%
No changes	0%

Submit

Results gathered from 212 respondents.

FEEDBACK

The more coefficients you add, the better fit your model has on the data. You can only get *better* approximations as you add more components! Also notice how the more components you have, the smaller the change in approximation.

3.3 Spectral Analysis of Snow Depth in NY Part 3

详见 jupyter notebook: 4. Weather Analysis - reconstruction SNWD.ipynb

POLL

How can we get a better value of the residual?

RESULTS

Change the dataset	3%
Run your model again to see if anything changes	5%
Take away coefficients	2%
Add more coefficients	89%

Submit

Results gathered from 203 respondents.

FEEDBACK

Remember that the value of the residual decreases as coefficients are added. This means that as we add coefficients, our model *follows the data closer.*

总结:

4 Lectures: Visualizing PCA Residuals II

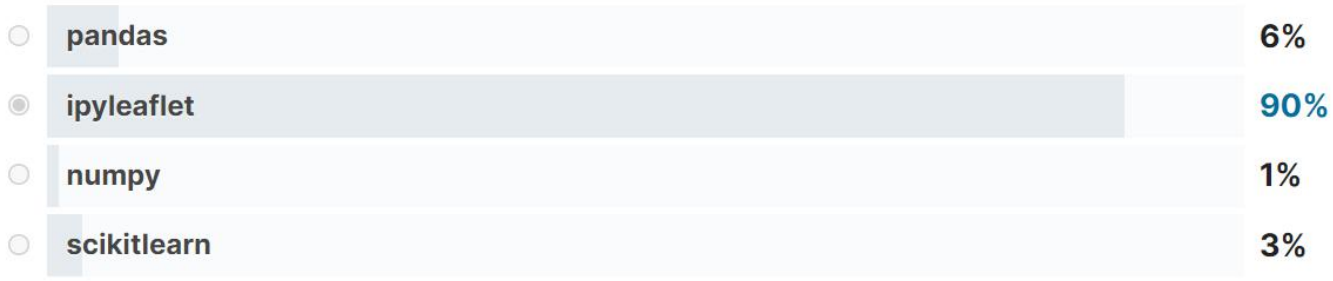
4.1 Spectral Analysis of Snow Depth in NY Part 4

详见 jupyter notebook: 5. maps using iPyLeaflet.ipynb

POLL

Which of the following can we use to visualize data on top of maps in our Jupyter notebook?

RESULTS



Submit

Results gathered from 200 respondents.

FEEDBACK

Leaflet is a library for interactive maps. ipyleaflet is a Jupyter notebook extension for Leaflet maps.

4.2 Snow Changes from Yr to Yr

详见 jupyter notebook: 5. maps using iPyLeaflet.ipynb

总结:

Quiz 4

Box Plot

0.0/20.0 points (graded)

Which of these statements are true about Box Plots?

- [x] It is also known as the box and whiskers diagram. ✔
- [ ] Box plots are a powerful means of visualizing bimodal data.
- [ ] A Box Plot does not reveal any information about any skew in the data.
- [ ] The top of the box represents the 25th percentile in a Box Plot.
- [ ] The line in the middle represents the mean of the data in a Box Plot. ✗ 是：中位数
- [x] The Inter-Quartile range is depicted in a box plot. ✔

PCA

0.0/20.0 points (graded)

Which of the following statements are true? Check all that apply.

- [x] Given an input $x \in \mathbb{R}^n$, PCA compresses it to a lower-dimensional vector $z \in \mathbb{R}^k$. ✔
- [ ] PCA is susceptible to local optima; trying multiple random initializations may help.
- [x] Even if all the input features are on very similar scales, we should still perform normalization(so that each feature has zero mean) before running PCA. ✔
- [ ] PCA can be used only to reduce the dimensionality of the data by 1 (such as 3D to 2D, 2D to 1D).

总结：