Lab1 Report

Mohamed Elmaghraby Mohamed 55

Prepare data:

- Import segmentation data and test then merged them.
- Skip the first two rows as they contain info about data.
- Merge both data sets.
- Add index column and class column.
- Explore data:
 - Number of instances = 2310
 - o number of attributes = 19
 - o class names : ['BRICKFACE' 'SKY' 'FOLIAGE' 'CEMENT' 'WINDOW' 'PATH' 'GRASS']
 - o number of classes: 7

Data Exploration:

- Plot histogram with bins [1, 5, 10, 15]
 - o Each figure contain 7 images, one for each class
 - Each image show all attributes with different colors
- Plot Box plot for the data
- Use .corr(method = 'Pearson') to get the Pearson's correlation between each 2 attributes.

o Draw the correlation using heatmap.

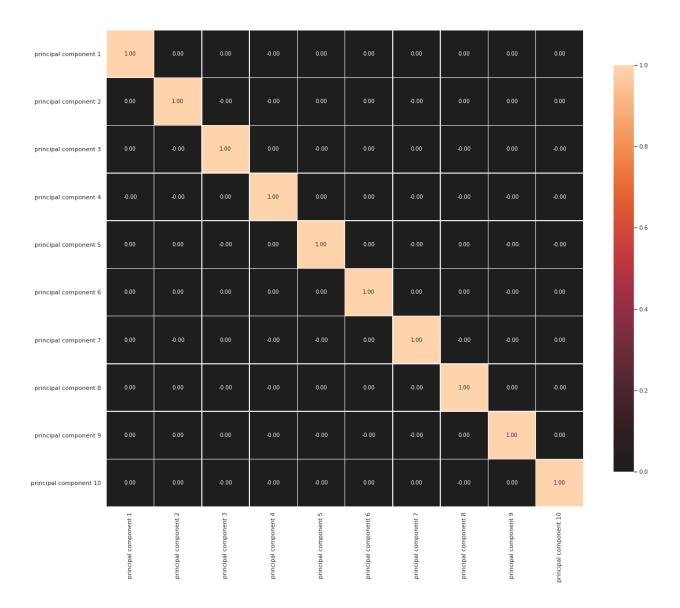
REGION-CENTROID-COL	1.00	0.03		-0.05	-0.02	-0.01	0.02	-0.02	-0.00	0.06	0.05	0.06	0.06	-0.09	0.04	0.01	0.06	-0.11	0.04
	0.03	1.00		0.06	0.04	0.03	-0.05	0.11	-0.02	-0.47	-0.47	-0.48	-0.44	0.35	-0.49	0.48	-0.46	0.08	0.59
REGION-CENTROID-ROW	0.03	1.00		0.06	0.04	0.03	-0.05	0.11	-0.02	-0.47	-0.47	-0.48	-0.44	0.35	-0.49	0.48	-0.46	0.08	0.59
REGION-PIXEL-COUNT																			
SHORT-LINE-DENSITY-5	-0.05	0.06		1.00	-0.01	-0.02	-0.03	-0.02	-0.04	-0.02	-0.02	-0.02	-0.02	0.03	-0.04	0.03	-0.02	-0.04	0.11
SHORT-LINE-DENSITY-2	-0.02	0.04		-0.01	1.00	0.26	0.19	0.30	0.24	-0.01	-0.01	0.00	-0.01	-0.04	0.06	-0.06	-0.00	0.02	-0.08
VEDGE-MEAN	-0.01	0.03		-0.02	0.26	1.00	0.64	0.56	0.49	0.01	-0.01	0.02	-0.00	-0.10	0.11	-0.08	0.02	-0.06	-0.10
VEDGE-SD	0.02	-0.05		-0.03	0.19	0.64	1.00	0.47	0.70	0.00	-0.00	0.01	0.00	-0.05	0.03	0.00	0.00	0.00	-0.06
HEDGE-MEAN	-0.02	0.11		-0.02	0.30	0.56	0.47	1.00	0.67	0.03	0.03	0.04	0.03	-0.10	0.09	-0.06	0.04	-0.13	-0.09
HEDGE-SD	-0.00	-0.02		-0.04	0.24	0.49	0.70	0.67	1.00	0.01	0.01	0.02	0.01	-0.06	0.03	-0.00	0.01	-0.02	-0.07
INTENSITY-MEAN	0.06	-0.47		-0.02	-0.01	0.01	0.00	0.03	0.01	1.00	1.00	1.00	1.00	-0.83		-0.51	1.00	-0.61	-0.33
RAWRED-MEAN	0.05	-0.47		-0.02	-0.01	-0.01	-0.00	0.03	0.01	1.00	1.00	0.99	0.99	-0.79		-0.51	0.99	-0.62	-0.33
RAWBLUE-MEAN	0.06	-0.48		-0.02	0.00	0.02	0.01	0.04	0.02	1.00	0.99	1.00	0.98	-0.86	0.84	-0.57	1.00	-0.60	-0.38
RAWGREEN-MEAN	0.06	-0.44		-0.02	-0.01	-0.00	0.00	0.03	0.01	1.00	0.99	0.98	1.00	-0.83	0.74	-0.43	0.99	-0.61	-0.26
EXRED-MEAN	-0.09	0.35		0.03	-0.04	-0.10	-0.05	-0.10	-0.06	-0.83	-0.79	-0.86	-0.83	1.00	-0.85	0.43	-0.86	0.42	0.28
EXBLUE-MEAN	0.04	-0.49		-0.04	0.06	0.11	0.03	0.09	0.03	0.79	0.77	0.84	0.74	-0.85	1.00	-0.85	0.83	-0.41	-0.64
EXGREEN-MEAN	0.01	0.48		0.03	-0.06	-0.08	0.00	-0.06	-0.00	-0.51	-0.51	-0.57	-0.43	0.43	-0.85	1.00	-0.54	0.28	0.80
VALUE-MEAN	0.06	-0.46		-0.02	-0.00	0.02	0.00	0.04	0.01	1.00	0.99	1.00	0.99	-0.86	0.83	-0.54	1.00	-0.60	-0.34
SATURATION-MEAN	-0.11	0.08		-0.04	0.02	-0.06	0.00	-0.13	-0.02	-0.61	-0.62	-0.60	-0.61	0.42	-0.41	0.28	-0.60	1.00	-0.06
HUE-MEAN	0.04	0.59		0.11	-0.08	-0.10	-0.06	-0.09	-0.07	-0.33	-0.33	-0.38	-0.26	0.28		0.80	-0.34	-0.06	1.00
	REGION-CENTROID-COL	EGION-CENTROID-ROW	REGION-PIXEL-COUNT	SHORT-LINE-DENSITY-5	SHORT-LINE-DENSITY-2	VEDGE-MEAN	VEDGE-SD	HEDGE-MEAN	HEDGE-SD	INTENSITY-MEAN	RAWRED-MEAN	RAWBLUE-MEAN	RAWGREEN-MEAN	EXRED-MEAN	EXBLUE-MEAN	EXGREEN-MEAN	VALUE-MEAN	SATURATION-MEAN	HUE-MEAN

Data Preprocessing:

- Normalization:
 - Min-max scalar:
 - Calculate min-max scalar using preprocessing.MinMaxScaler() from the scikit-learn library
 - Then plot box plot for resulting data
 - **■** Conclusion:
 - The result data spread on a very small scale, so values are more close to each other.
 - With min-max normalization, we were guaranteed to reshape both of our features to be between 0 and 1.
 - doesn't handle outliers.
 - Z-score:
 - Calculate z-score using zscore(x) from scipy library
 - Then plot box plot for resulting data
 - Conclusion:
 - The result data spread on less scale, so values are more close to each other.
 - A value is exactly equal to the mean of all the values of the feature, it
 will be normalized to 0. If it is below the mean, it will be a negative
 number, and if it is above the mean it will be a positive number.

• Dimensionality reduction

- o Feature Projection:
 - Use PCA from the scikit-learn library seeking to sum of variance ratio we need of 0.95.
 - Number of PCA components = 10
 - Variance ratio = [0.42341135, 0.16203649, 0.09959451, 0.05857283, 0.05197997, 0.05050372, 0.04041415, 0.03120143, 0.02999802, 0.02195028]
 - Plot correlation matrix for result data.



■ Conclusion

- Principal component analysis convert a set of observations of correlated variables into a set of values of linearly uncorrelated variables called principal components
- less than half of the attributes can describe the data with 0.95 percent.

- o Feature selection
 - Use SelectKBest from the scikit-learn library with k = 10.

■ Plot correlation matrix for result data.

Feture 1	100	-0.47	-0.47	-0.48	-0.44	0.35	-0.49	0.48	-0.46	0.59
Feture 2	-0.47	1.00	1.00	1.00	1.00	-0.83		-0.51	1.00	-0.33
Feture 3	-0.47	1.00	1.00	0.99	0.99	-0.79	0.77	-0.51	0.99	-0.33
Feture 4	-0.48	1.00	0.99	1.00	0.98	-0.86	0.84	-0.57	1.00	-0.38
Feture 5	-0.44	1.00	0.99	0.98	1.00	-0.83		-0.43	0.99	-0.26
Feture 6	0.35	-0.83	-0.79	-0.86	-0.83	1.00	-0.85	0.43	-0.86	0.28
Feture 7	-0.49	0.79	0.77	0.84	0.74	-0.85	1.00	-0.85	0.83	-0.64
Feture 8	0.48	-0.51	-0.51	-0.57	-0.43	0.43	-0.85	1.00	-0.54	0.80
Feture 9	-0.46	1.00	0.99	1.00	0.99	-0.86	0.83	-0.54	1.00	-0.34
Feture 10 F	0.59	-0.33	-0.33	-0.38	-0.26	0.28	-0.64	0.80	-0.34	100
Fet	Feture 1	Feture 2	Feture 3	Feture 4	Feture 5	Feture 6	Feture 7	Feture 8	Feture 9	Feture 10

Conclusion

- Feature Selection selects those features which contribute most to your prediction variable or output in which you are interested in.
- These features are highly correlated.