Machine Learning and Data Mining

11. May 2021.

Principal Components Analysis (PCA)

Principal Components Analysis (PCA)

- (algorithm) that transforms a set of correlated variables (p) into smaller $k (k << p)$ number of uncorrelated variables called principal components
- The subset you select constitutes a new space that is small in size compared to the original space.
- Uncorrelated random variables have a Pearson correlation coefficient of zero
- It maintains as much of the complexity of data as possible.

Uncorrelated random variables

- If two variables are uncorrelated, there is no linear relationship between them.
- In probability theory and statistics, two real-valued random variables X, Y, are said to be uncorrelated if their covariance,$\operatorname{cov}[X, Y] = \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y]$, is zero.
- Uncorrelated random variables have a Pearson correlation coefficient of zero
- If X and Y are independent, with finite second moments, then they are uncorrelated. However, not all uncorrelated variables are independent.
- orthogonality means "uncorrelated." An orthogonal model means that all independent variables in that model are uncorrelated. If one or more independent variables are correlated, then that model is non-orthogonal.

Principal Components Analysis (PCA)

Principal Components Analysis (PCA)

- a linear dimensionality reduction technique
- an Unsupervised dimensionality reduction technique,
- you can cluster the similar data points based on the feature correlation between them without any supervision (or labels)

Principal Components Analysis (PCA) Application

- Data Visualization
- Speeding Machine Learning (ML) Algorithm

Principal Components Analysis (PCA) Application

- Principal Components captures (or holds) most of the variance (information) of your data.
- Principal components have both direction and magnitude

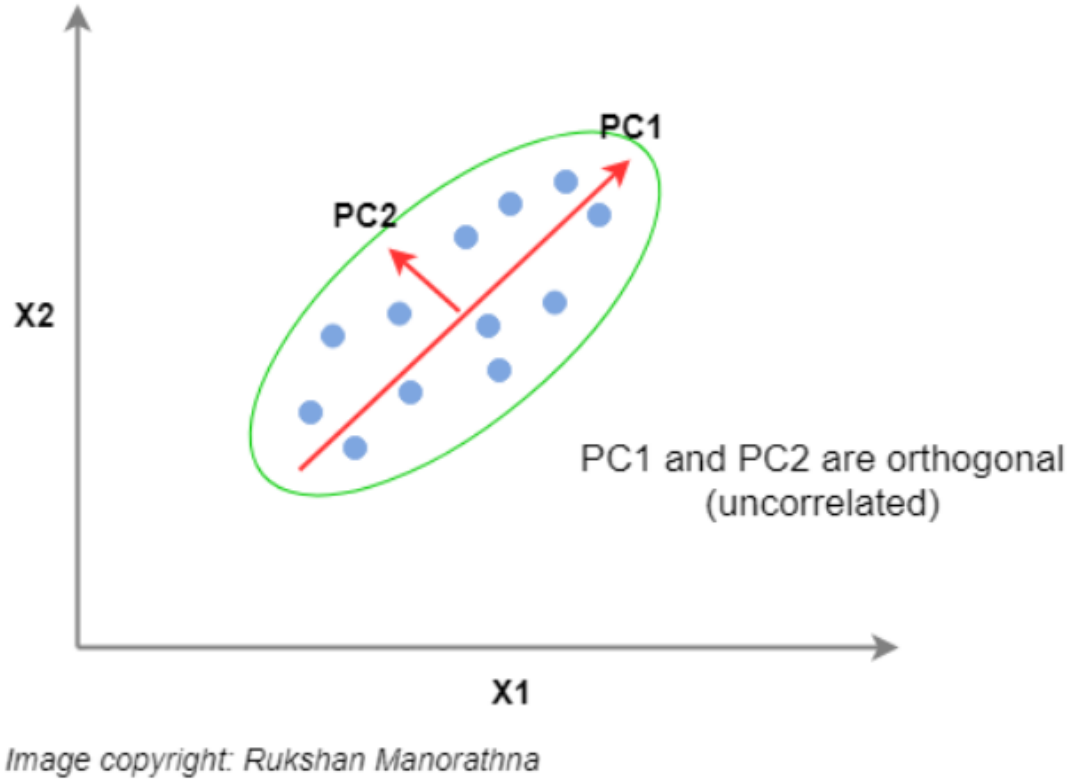

Figure 1: The PCA considers the correlation among variables.

Understanding the Data

- The Breast Cancer data set is a real-valued multivariate data that consists of two classes
- The malignant class has 212 samples, whereas the benign class has 357 samples.
- $https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)$
- an easy way is by loading it with the help of the sklearn library.
- This dataset contains breast cancer data of 569 females (observations). The dimensionality of the dataset is 30. It means that there are 30 attributes (characteristics) for each female (observation) in the dataset.

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst radius	worst texture	worst perimeter	worst area	worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension
0	17.990	10.38	122.80	1001.0	0.11840	0.27760	0.300100	0.147100	0.2419	0.07871	...	25.380	17.33	184.60	2019.0	0.16220	0.66560	0.71190	0.26540	0.4601	0.11890
1	20.570	17.77	132.90	1326.0	0.08474	0.07864	0.086900	0.070170	0.1812	0.05667	...	24.990	23.41	158.80	1956.0	0.12380	0.18660	0.24160	0.18600	0.2750	0.08902
2	19.690	21.25	130.00	1203.0	0.10960	0.15990	0.197400	0.127900	0.2069	0.05999	...	23.570	25.53	152.50	1709.0	0.14440	0.42450	0.45040	0.24300	0.3613	0.08758
3	11.420	20.38	77.58	386.1	0.14250	0.28390	0.241400	0.105200	0.2597	0.09744	...	14.910	26.50	98.87	567.7	0.20980	0.86630	0.68690	0.25750	0.6638	0.17300
4	20.290	14.34	135.10	1297.0	0.10030	0.13280	0.198000	0.104300	0.1809	0.05883	...	22.540	16.67	152.20	1575.0	0.13740	0.20500	0.40000	0.16250	0.2364	0.07678
5	12.450	15.70	82.57	477.1	0.12780	0.17000	0.157800	0.080890	0.2087	0.07613	...	15.470	23.75	103.40	741.6	0.17910	0.52490	0.53550	0.17410	0.3985	0.12440
6	18.250	19.98	119.60	1040.0	0.09463	0.10900	0.112700	0.074000	0.1794	0.05742	...	22.880	27.66	153.20	1606.0	0.14420	0.25760	0.37840	0.19320	0.3063	0.08368
7	13.710	20.83	90.20	577.9	0.11890	0.16450	0.093660	0.059850	0.2196	0.07451	...	17.060	28.14	110.60	897.0	0.16540	0.36820	0.26780	0.15560	0.3196	0.11510
8	13.000	21.82	87.50	519.8	0.12730	0.19320	0.185900	0.093530	0.2350	0.07389	...	15.490	30.73	106.20	739.3	0.17030	0.54010	0.53900	0.20600	0.4378	0.10720
9	12.460	24.04	83.97	475.9	0.11860	0.23960	0.227300	0.085430	0.2030	0.08243	...	15.090	40.68	97.65	711.4	0.18530	1.05800	1.10500	0.22100	0.4366	0.20750
10	16.020	23.24	102.70	797.8	0.08206	0.06669	0.032990	0.033230	0.1528	0.05697	...	19.190	33.88	123.80	1150.0	0.11810	0.15510	0.14590	0.09975	0.2948	0.08452
11	15.780	17.89	103.60	781.0	0.09710	0.12920	0.099540	0.066060	0.1842	0.06082	...	20.420	27.28	136.50	1299.0	0.13960	0.56090	0.39650	0.18100	0.3792	0.10480
12	19.170	24.80	132.40	1123.0	0.09740	0.24580	0.206500	0.111800	0.2397	0.07800	...	20.960	29.94	151.70	1332.0	0.10370	0.39030	0.36390	0.17670	0.3176	0.10230

Figure 2: The sample from the breast cancer dataset

Correlation matrix and variance-covariance matrix

Correlation matrix and variance-covariance matrix

- PCA can be performed using either correlation or variance-covariance matrix
- A correlation matrix is a table showing correlation coefficients between variables
- A variance-covariance matrix is a matrix that contains the variances and covariances associated with several variables.

Eigenvalues and eigenvectors

- Let A be an $n \times n$ matrix.
- A scalar λ is called an eigenvalue of A if there is a non-zero vector x satisfying the equation $Ax = \lambda x$.
- The eigenvectors of the correlation matrix or variance-covariance matrix represent the principal components (the directions of maximum variance).