Introduction to Machine Learning

Lecture 3
Data Reduction

A. S. M. Sanwar Hosen

Email: sanwar@wsu.ac.kr

Date: 12 Oct., 2023

Data Reduction

- **Data Reduction:** In ML model, there are often too many variables (features) to work with. The higher the number of features, the more difficult it is to model them. The data reduction refers to techniques that reduce the number of input variables in a dataset that makes the model more efficient in prediction. The common ways of data reduction techniques are:

- ✓ Feature Selection and Elimination
- ✓ Data Decomposition or Feature Extraction

Data Reduction

- **Feature Selection and Elimination:** It is used to find the best set of features that allows to build a useful ML model. The feature selection and elimination methods are based on different techniques that are as follows:

- ✓ **Unsupervised Technique:** Do not use the target variable and remove redundant variables.

  1) Correlation: Finds the association/relationship between two variables in a dataset.

- ✓ **Supervised Technique:** It takes advantage of labeled target variables or class labels to make decisions about which features to include or exclude in a dataset.

  1) Wrapper: Search for well-performing subsets of features.

  2) Filter: Select subsets of features based on their relationship with the target variable.

  3) Intrinsic: It performs automatic feature selection during training the ML model.

Data Reduction

- **Correlation:** A statistical technique determines how one variable changes in relation with the other variables. It gives us the direction of the relationship of the two variables. It defines the following relationship:

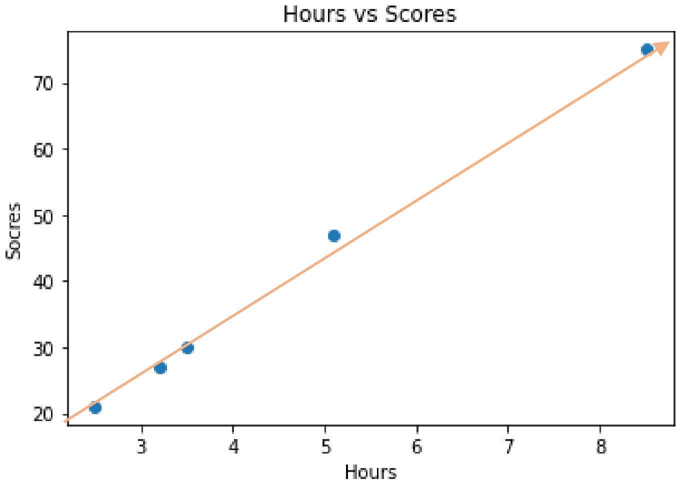
- ✓ **Positive Correlation:** Two variables (features) can be positively correlated. It means that when the value of one variable increases then the value of the other variable(s) also increases.
- ✓ **Negative Correlation:** Two variables (features) can be negatively correlated. It means that when the value of one variable increases then the value of the other variable(s) decreases.
- ✓ **No Correlation:** Two variables (features) are not correlated. It means that when the value of one variable increases or decreases then the value of the other variable(s) doesn’t increase or decrease.

Data Reduction

❑ Correlation (cont..)

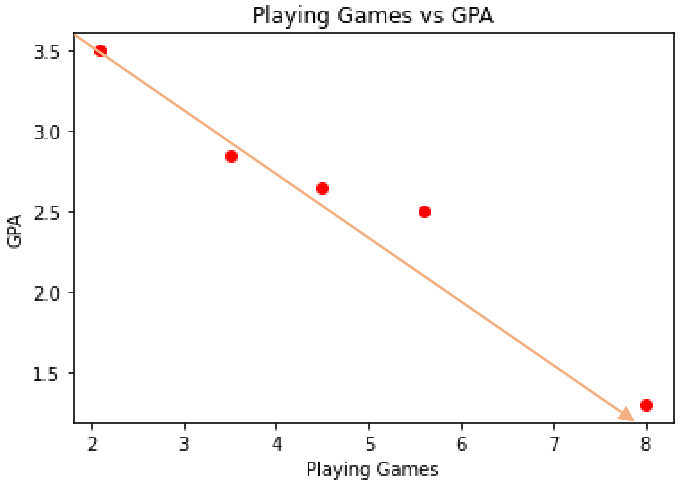
Examples of Correlations:

	Hours	Scores
0	2.5	21
1	5.1	47
2	3.2	27
3	8.5	75
4	3.5	30



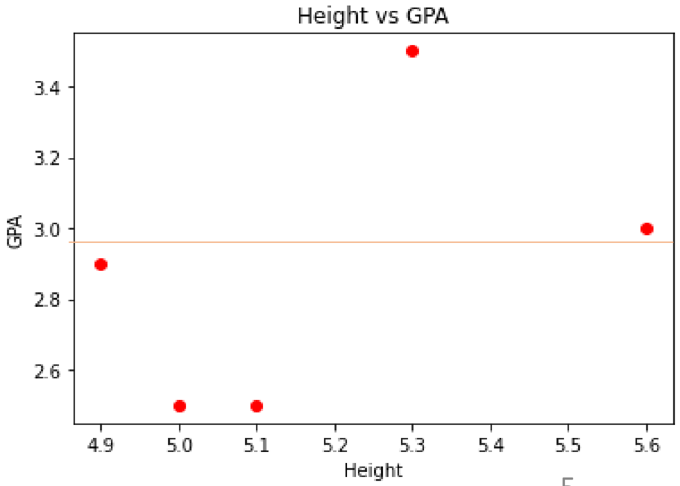
Positive Correlation: $x(hours)$ increases, $y(scores)$ also increa

	Playing Games	GPA
0	5.6	2.50
1	3.5	2.85
2	4.5	2.65
3	8.0	1.30
4	2.1	3.50



Negative Correlation: $x(playing\ games)$ increases, $y(GPA)$ decreases

	Height	GPA
0	5.3	3.50
1	5.1	2.50
2	5.6	3.80
3	4.9	2.90
4	5.0	3.25



No Correlation: $x(height)$ increases or decreases, $y(GPA)$ increases or decreases

Data Reduction

- **Correlation (cont..)**

✓ **Correlation Coefficient:** It finds the strength of the relationship between two variables. The correlation coefficient techniques commonly used in ML are:

1) Pearson Correlation Coefficient (PCC)

2) Spearman's Correlation Coefficient (SCC)

Data Reduction

- **Pearson Correlation Coefficient (PCC):** It finds the direction and strength of a relationship (r) between two variables in a dataset. The range of the possible results of this PCC is -1 to 1, where:

  1) 0 indicates no correlation.

  2) 1 indicates a strong positive correlation.

  3) -1 indicates a strong negative correlation.

  The value of r is calculated using the following equation:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

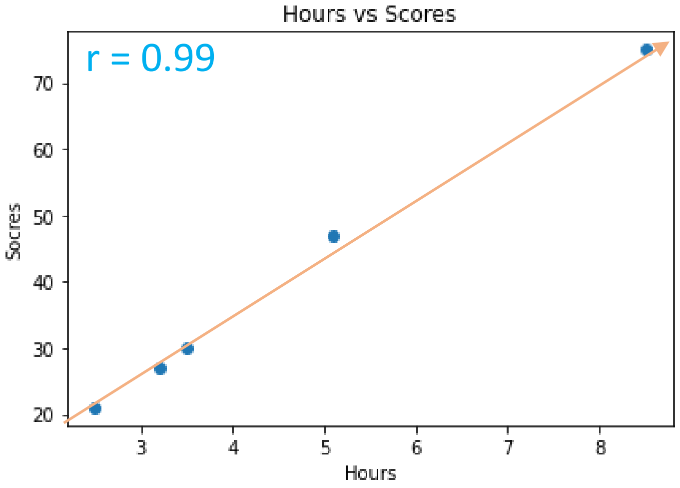
Where, r is the relationship degree of PCC, and n is the number of values in x and y.

Data Reduction

❑ Pearson Correlation Coefficient (PCC) (cont..)

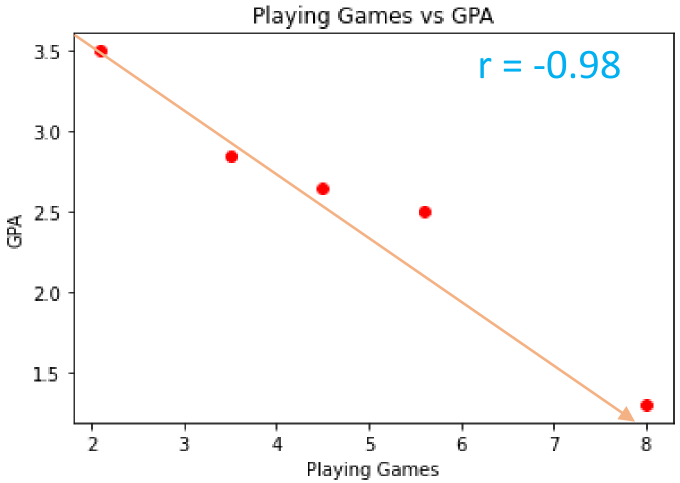
Examples of PCC:

	Hours	Scores
0	2.5	21
1	5.1	47
2	3.2	27
3	8.5	75
4	3.5	30



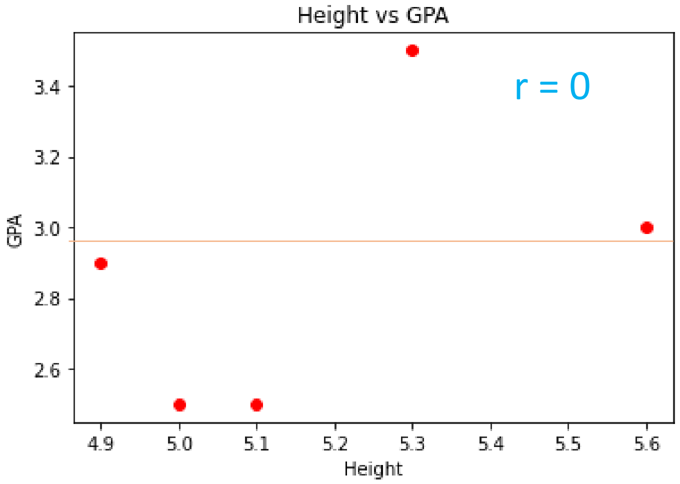
Positive Correlation: $x(hours)$ increases, $y(scores)$ also increas

	Playing Games	GPA
0	5.6	2.50
1	3.5	2.85
2	4.5	2.65
3	8.0	1.30
4	2.1	3.50



Negative Correlation: $x(playing\ games)$ increases, $y(GPA)$ decreases

	Height	GPA
0	5.3	3.50
1	5.1	2.50
2	5.6	3.80
3	4.9	2.90
4	5.0	3.25



No Correlation: $x(height)$ increases or decreases, $y(GPA)$ increases or decreases

1096323.1999999997

Data Reduction

Pearson Correlation Coefficient (PCC) (cont..)

Example: Find the correlation value r of the variables in the given dataset using PCC. Given $x(Hours) = [2.5, 5.1, 3.2, 8.5, 3.5]$ and $y(Scores) = [21, 47, 27, 75, 30]$.

	Hours	Scores
0	2.5	21
1	5.1	47
2	3.2	27
3	8.5	75
4	3.5	30

Solution: We know the formula of PCC is:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

x	y	xy	x^2	y^2
2.5	21	52.50	6.25	441
5.1	47	239.70	26.01	2209
3.2	27	86.40	10.24	729
8.5	75	637.50	72.25	5625
3.5	30	105	12.25	900

Here,
$\sum x$ = 22.80; $\sum y$ = 200; $\sum xy$ = 1121.10; $\sum x^2 = 127$; $\sum y^2 = 9904$ and $n = 5$ (no. of samples)

$$r = \frac{5\times1121.10-(22.80\times200)}{\sqrt{[(5\times127)-(22.80)^2]\times[(5\times9904)-(200)^2]}} = \frac{1045.5}{\sqrt{1096323.199999}} = \frac{1045.5}{1047.05453} = 0.9985$$

Data Reduction

- **Data Decomposition and Feature Extraction:** Decomposing of data is a statistical technique that reforms the data dimensionality. For example, it transposes higher dimension data into a lower dimension data. Principal Component Analysis (PCA) is one of the most widely used decomposition technique in ML.

- ✓ **Principal Component Analysis (PCA):** It is a technique for feature extraction that combines the input variables in a specific way and drops the least important variables from a dataset.

Advantages of PCA:

1) Reduces the number of variables.

2) Extracts the variables are most independent of one another.

3) Makes independent variables less interpretable.

Principal Component Analysis (PCA) (cont..)

✓ **Step 1 (Data Standardization):** The range of variables is calculated and standardized in this process to analyze the contribution of each variable equally by using the following equation:

$$Z = \frac{X - \mu}{\sigma}$$

Where, X is the values in dataset, μ is the mean of the variables, and σ is the standard deviation.

Dataset: An Example (Input Features)

	x_1 (F1)	x_2 (F2)	x_3 (F3)	x_4 (F4)
	1	5	3	1
	4	2	6	3
	1	4	3	2
	4	4	1	1
	5	5	2	3
μ	3	4	3	2
σ	1.87	1.22	1.87	1

Dataset: Normalized

x_1 (F1)	x_2 (F2)	x_3 (F3)	x_4 (F4)
−1.0695	0.8196	0	−1
0.5347	−1.6393	1.6042	1
−1.0695	0	0	0
0.5347	0	−1.0695	−1
1.0695	0.8196	0.5347	1

Data Reduction

Principal Component Analysis (PCA) (cont..)

✓ **Step 2 (Covariance Matrix Computation):** A covariance matrix is a $N \times N$ symmetrical matrix that contains the covariances of all possible datasets. The covariance matrix of two-dimensional data is calculated as:

$$Cov(x, y) = \frac{\sum_{i=1}^{n}(x-\mu_x)(y-\mu_y)}{n}$$

Where, n is the number of data points/samples.

Dataset: Normalized

x_1 (F1)	x_2 (F2)	x_3 (F3)	x_4 (F4)
-1.0695	0.8196	0	-1
0.5347	-1.6393	1.6042	1
-1.0695	0	0	0
0.5347	0	-1.0695	-1
1.0695	0.8196	0.5347	1

Covariance Matrix

	F1	F2	F3	F4
F1	$Var(F1)$	$Cov(F1, F2)$	$Cov(F1, F3)$	$Cov(F1, F4)$
F2	$Cov(F2, F1)$	$Var(F2)$	$Cov(F2, F3)$	$Cov(F2, F4)$
F3	$Cov(F3, F1)$	$Cov(F3, F2)$	$Var(F3)$	$Cov(F3, F4)$
F4	$Cov(F4, F1)$	$Cov(F4, F2)$	$Cov(F4, F3)$	$Var(F4)$

Covariance Matrix

	F1	F2	F3	F4
F1	0.78	-0.8586	-0.055	0.424
F2	-0.8586	0.78	-0.607	-0.326
F3	-0.055	-0.607	0.78	0.426
F4	0.424	-0.326	0.426	0.78

Consider, $\mu = 0$, and $\sigma = 1$

$$Var(F1) = \frac{(-1.0695-0)^2 + (0.5347-0)^2 + (-1.0695-0)^2 + (0.5347-0)^2 + (1.069-0)^2}{5} = 0.78$$

$$Cov(F1, F2) = \frac{((-1.0695-0)(0.8196-0) + (0.5347-0)(-1.6393-0) + (-1.0695-0)(0.0000-0) + (0.5347-0)(0.0000-0) + (1.0695-0)(0.8196-0))}{5} = -0.8586$$

Data Reduction

❑ Principal Component Analysis (PCA) (cont..)

✓ **Step 3 Feature Vector (Eigenvalues and Eigenvectors Computation):**

1) Eigenvalues: Eigenvalues are scalar values that represent the scaling factor by which an eigenvector is stretched or shrunk when a linear transformation is applied to it. Given a square matrix A, an eigenvalue (λ) and its corresponding eigenvector (v) satisfy the equation: $Av = \lambda v$.
2) Eigenvectors: Eigenvectors are non-zero vectors that remain in the same direction but may be scaled when a linear transformation (represented by the matrix A) is applied to them.

According to the definition:
$Av = \lambda v$, then λ is called eigenvalue associated with eigenvector v of A. Upon substituting the values in $det(A - \lambda I) = 0$, we get the following matrix. Then substitute each eigen value in $(A - \lambda I)V = 0$ and apply Cramer's rule.

	F1	F2	F3	F4
F1	$0.78 - \lambda$	-0.8586	-0.055	0.424
F2	-0.8586	$0.78 - \lambda$	-0.607	-0.326
F3	-0.055	-0.607	$0.78 - \lambda$	0.426
F4	0.424	-0.326	0.426	$0.78 - \lambda$

⇨

E1	E2	E3	E4
0.515514	-0.623012	0.0349815	-0.58726
-0.616625	0.113105	0.452326	-0.634336
0.399314	0.744256	-0.280906	-0.455767
0.441098	0.212477	0.845736	0.212173

$\lambda = 2.11691\,, 0.855413\,, 0.481689\,, 0.334007$

For $\lambda = 2.11691$, solving the above equation using Cramer's rule, the values for the v vector are $v1 = 0.515514, v2 = -0.616625, v3 = 0.399314, v4 = 0.441098$

Principal Component Analysis (PCA) (cont..)

Example of Cramer’s Rule: Solve the following system of 3 equations in 3 variables using Cramer's rule:

Solution: The given system can be written in the matrix form $AV = B$

Where, $A = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix}$, $V = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$, and $B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$

Step 1: Compute the determinant of A

$$D = det(A)$$

Step 2: Compute the determinant of v_1, v_2, v_3, and v_4

$$D_{v_1} = det(v_1)$$

$$D_{v_2} = det(v_2)$$

$$D_{v_3} = det(v_3)$$

$$D_{v_4} = det(v_4)$$

Step 3: Apply the following formulas

$$v_1 = \frac{D_{v1}}{D},\ v_2 = \frac{D_{v2}}{D},\ v_3 = \frac{D_{v3}}{D},\ \text{and } v_4 = \frac{D_{v4}}{D}$$

Principal Component Analysis (PCA) (cont..)

✓ **Step 3 Feature Vector (Eigenvalues and Eigenvectors Computation) (cont..):** Determine the principal components by first calculating the sum of the eigenvalues in each column, arranging them in descending order, and selecting the highest eigenvalues from the list.

E1	E2	E3	E4
0.515514	-0.623012	0.0349815	-0.58726
-0.616625	0.113105	0.452326	-0.634336
0.399314	0.744256	-0.280906	-0.455767
0.441098	0.212477	0.845736	0.212173

E1	E2	E3	E4
0.51068	0.44682	0.4185745	-0.553655

Where, $E1 > E2 > E3 > E4$

E1	E2
0.515514	-0.623012
-0.616625	0.113105
0.399314	0.744256
0.441098	0.212477

Selected Top Most Eigen Values

Data Reduction

Principal Component Analysis (PCA) (cont..)

✓ **Step 4 Final Dataset (Data Decomposition):** Compress the dataset into a small dataset without any loss of data. It is calculated as follows:

$$Final\ Dataset = Standardized\ Orinigal\ Dataset\ * Feature\ Vector$$

Dataset: Standardized Original Dataset

x_1 (F1)	x_2 (F2)	x_3 (F3)	x_4 (F4)
-1.0695	0.8196	0	-1
0.5347	-1.6393	1.6042	1
-1.0695	0	0	0
0.5347	0	-1.0695	-1
1.0695	0.8196	0.5347	1

×

Vector: Feature Vector

E1	E2
0.515514	-0.623012
-0.616625	0.113105
0.399314	0.744256
0.441098	0.212477

Dataset: Transformed Data

x_1 (F1)	x_2 (F2)
0.4268066978	0.0011611400000
-0.4234920968	-0.39182856
-0.551342223	0.79559219
0.4652040128	0.89996329
0.08272794800	0.28653037

Data Reduction

- **Lecture Overview**
- ✓ **Data Reduction**
- ✓ **Feature Selection and Elimination**
  - Correlation
  - Correlation Coefficient (Pearson Correlation Coefficient)
- ✓ **Data Decomposition and Feature Extraction**
  - Principal Component Analysis (PCA)