Mid 2 Practice Questions

By Mannan Ul Haq

Data Normalization:

Normalization is a data transformation that scales numeric variables to a common range, often [0, 1] or [-1, 1]. It is useful for ensuring that variables are on a consistent scale, which is important for machine learning algorithms that are sensitive to feature magnitudes.

Example:

1. Min-Max Normalization:

Min-Max scaling is a common normalization technique that normally re-scales features to a distribution value between 0 and 1. The minimum value transforms into 0, and the maximum value transforms into 1. This implies the data is more concentrated around the mean with Max-Min Normalization. If there are outliers in your features, normalizing data scales most of it to a small interval, providing uniform scales but not handling outliers well. Standardization is more robust to outliers, making it preferable in many cases.

$$x' = \left(\frac{x - \min}{\max - \min}\right) \times (\text{new max} - \text{new min}) + \text{new min}$$

Question#2: Apply **min-max normalization (0, 5)** on attribute "**Income**" of a dataset. Show working. **[5 marks]**

Income	45000	60000	75000	50000	80000	55000	70000	90000	65000	75000
Normalized Income	0	1.67	3.33	0.56	3.89	1.11	2.78	5	2.22	3.33

$$x' = \left(\frac{x - \min}{\max - \min}\right) \times (\text{new max} - \text{new min}) + \text{new min}$$

From given data,

new max = 5
new min = 0
min = 45000
max = 90000

For 45000,

$$x' = \left(\frac{45000 - 45000}{90000 - 45000}\right) \times (5 - 0) + 0$$

$$= 0$$

2. Standardization (Z-Score Normalization):

Rescales features to ensure mean and standard deviation are 0 and 1, respectively. The range is from -3 to +3.

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

3. Decimal Scaling:

Decimal Scaling is a normalization technique used to scale data by moving the decimal point of values. This method ensures that all values fall within a specific range, typically between -1 and 1. The scaling factor is determined by the maximum absolute value of the attribute, and the decimal point is shifted accordingly.

Steps for Decimal Scaling:

1. Identify the maximum absolute value in the data.
2. Determine the number of decimal places required to scale the maximum value between **-1** and **1** (let this number be **j**).
3. Divide all values by **10^j**, where **j** is the smallest integer such that the maximum absolute value becomes less than 1.

Example:

Suppose the values of attribute A range from -986 to 917.

1. **Maximum absolute value**: (max(|-986|, |917|) = **986**)
2. To normalize by decimal scaling, we divide all values by **10^3** (since, j = **3**).

- -986 becomes **-0.986**
- 917 becomes **0.917**

Student_ID	Math	Science	English
1	80	90	85
2	75	88	92
3	89	94	76
4	90	85	88
5	82	78	80

Formula: $x' = \frac{x}{10^j}$

For each subject, the highest value determines j

Math: Highest absolute value = 90 → $j = 2$

Science: Highest absolute value = 94 → $j = 2$

English: Highest absolute value = 92 → $j = 2$

Apply Decimal scaling normalization:

Student ID	Math (normalized)	Science (normalized)	English (normalized)
1	0.80	0.90	0.85
2	0.75	0.88	0.92
3	0.89	0.94	0.76
4	0.90	0.84	0.88
5	0.82	0.78	0.80

Binning Data:

It is a technique used to smooth data and handle outliers by grouping data into bins or intervals. It helps in identifying and mitigating the impact of outliers by transforming the data into a form that reduces the effect of extreme values.

Equal Frequency Binning:

In Equal Frequency Binning, the dataset is divided into bins where each bin contains approximately the same number of data points.

Equal Width Binning:

In Equal Width Binning, the dataset is divided into bins where each bin has the same width, but the number of data points in each bin can vary.

no. of bins = sqrt(no. of rows)

$$\frac{\text{max - min}}{\text{no. of bins}} = \text{width}$$

intervals = width + min , 2*width + min , ...

Data = 4, 8, 15, 21, 21, 24, 25, 28, 34

width = 10

intervals = 4 - 14, 15 - 24, 25 - 34

Question#3: Discretize the **"Income"** attribute using **equi-width** binning. Create three bins with labels **"Low"**,**"Medium"**, and **"High"**. Show working. **[9 Marks]**

Income	45000	60000	75000	50000	80000	55000	70000	90000	65000	75000

Number of bins : 3

Bin 1 = low

Bin 2 = Medium

Bin 3 = High

Bin Width = $\frac{\text{max} - \text{min}}{\text{no. of bins}} = \frac{75000 - 45000}{3} = 15000$

Bins Range :

Bin 1 : (min , width + min) = (45000 , 60000)

Bin 2 : (width + min + 1, width*2 + min) = (60001, 75000)

Bin 3 : (width*2 + min + 1, width*3 + min) = (75001, 90000)

Bin 1 (Low) : 45000, 50000, 55000, 60000

Bin 2 (Medium) : 65000, 70000, 75000, 75000

Bin 3 (High) : 80000 , 90000

Other Formula:

bin#1 range={min , min+w-1}
bin#2 range={min+w , min+2w-1}
bin#3 range ={min+2w , max}

- After creating the bins, the data smoothened using
- Bin means
- Bin medians
- Bin boundaries
- Example:
  - Bin1 = [2,10,18], Bin2 = [18,19,20], Bin3 = [22,25,28]
  - Smooth by mean:
    - Bin1=[10,10,10]
    - Bin2=[19,19,19]
    - Bin3=[25,25,25]

Smooth by median:
Bin1 = [10,10,10]
Bin2 = [19,19,19]
Bin3 = [25,25,25]

Smooth by boundaries:
Bin1 = [2,2,18]
Bin2 = [18,18,20]
Bin3 = [22,22,28]

In this example, used minimum boundary in case of tie

Feature Selection

Dimensionality Reduction

Dimensionality reduction is a process used to reduce the number of features or attributes in a dataset while retaining as much information as possible.

Attribute Subset Selection

Reduces the data set size by removing irrelevant or redundant attributes (or dimensions).

Common Methods:

- **Filter Methods**: Evaluate features based on statistical measures like correlation, mutual information, or variance.
- **Wrapper Methods**: Use a predictive model to evaluate subsets of features and select the best performing subset.
- **Embedded Methods**: Incorporate feature selection as part of the model training process, such as regularization techniques in linear models and decision tree induction.

Filter Methods:

Definition: The Filter Methods involve selecting features based on their various statistical scores with the output column.

- The filter method ranks each feature based on some uni-variate metric and then selects the highest-ranking features. We identify and retain the most relevant features based on statistical measures or predefined criteria.
- The selection of features is independent of any Machine Learning algorithm.

Rules:

- The more the features are correlated with the output column or the column to be predicted, the better the performance of the model.
- Features should be least correlated with each other. If some of the input features are correlated with some additional input features, this situation is known as **Multicollinearity**. It is recommended to get rid of such a situation for better performance of the model.

Filter Selection Select independent features with:

- No constant variables

- No duplicate rows
- High correlation with the target variable
- Low correlation with another independent variable
- Higher information gain or mutual information of the independent variable

Common Filter Methods:

1. Removing features with low variance (Low variance filter):

Definition: The Low Variance Filter is a feature selection method that focuses on identifying and eliminating features with minimal variance across the dataset. Features with low variance often carry limited information and may not contribute significantly to the modeling process.

Key Components:

Variance:

- Measures the spread or variability of values within a feature.
- High variance indicates a wide range of values, while low variance suggests a more constant or uniform feature.

Threshold:

- A predetermined threshold is set to determine the acceptable level of variance.
- Features with variance below the threshold are considered low-variance features.

Steps:

To remove features with low variance from our dataset, follow these steps:

1. **Remove features with zero variance (constants).**
2. **Normalize the data** and compute the variance for each feature.
3. **Choose a threshold value** that defines the minimum acceptable variance.
4. **Identify and remove features** with variance below the selected threshold.

ID	season	holiday	workingday	weather	f5	temp	atemp	humidity	windspeed	count
AB101	1	0	0	1	7	9.84	14.395	81	0.0000	16
AB102	1	0	0	1	7	9.02	13.635	80	0.0000	40
AB103	1	0	0	1	7	9.02	13.635	80	0.0000	32
AB104	1	0	0	1	7	9.84	14.395	75	0.0000	13
AB105	1	0	0	1	7	9.84	14.395	75	0.0000	1
AB106	1	0	0	2	7	9.84	12.880	75	6.0032	1
AB107	1	0	0	1	7	9.02	13.635	80	0.0000	2
AB108	1	0	0	2	7	9.02	12.880	86	0.0000	3
AB109	1	0	0	1	7	9.84	14.395	75	0.0000	8
AB110	1	0	0	1	7	13.12	17.425	76	0.0000	14

2. Pearson's Correlation Coefficient:

Definition: Pearson's Correlation Coefficient, often denoted as r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables.

Objective: To determine the degree of linear association between two variables, with values ranging from **-1** (perfect negative correlation) to **1** (perfect positive correlation), and **0** indicating no linear correlation.

Formula:

The Pearson correlation coefficient is calculated using the following formula:

$$r_{A,B} = \frac{\Sigma(A-\bar{A})(B-\bar{B})}{(n-1)\sigma_A\sigma_B}$$

OR

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

Example: A researcher in a scientific foundation wished to evaluate the relation between annuals salaries of mathematicians (Y, in thousand dollars) and an index of work quality (X1), number of years of experience (X2), and an index of publication success (X3).

Index of Work Quality (X1)	Number of Years of Experience (X2)	Index of Publication Success (X3)	Annual Salaries (in thousand dollars) Y
3.5	9	6.1	33.2
5.1	18	7.4	38.7
6	13	5.9	37.5
3.1	5	5.8	30.1
4.5	25	5	38.2

Heat Map:

A Heat Map is a graphical representation that visualizes the strength and direction of relationships between multiple variables, often using color gradients.

- Calculate Pearson's correlation coefficient for all pairs of variables in the dataset.
- Represent the correlation matrix as a heat map, where each cell color corresponds to the correlation value.

- **Handling Multicollinearity**:
  - Address multicollinearity issues, especially when features are highly correlated.
  - Choose features with the least redundancy.

3. Mutual Information (Information Gain):

Definition: Mutual Information measures the amount of information obtained about one variable through the observation of another variable. In the context of feature selection, it quantifies the degree of dependence between a feature and the target variable.

The feature having the most information is considered important by the algorithm and is used for training the model.

The effort is to reduce the entropy and maximize the information gain.

Entropy:

- A measure of uncertainty or randomness in a variable.
- High entropy indicates greater uncertainty.
- Information gain uses entropy to make decisions. If the entropy is less, information will be more.

Formula:

The Mutual Information between two variables *X* and *Y* is calculated using the following formula:

IG (X|Y) = E(S) - E (X|Y)

Example:

Example1 : Consider a dataset about whether a student will play a sport based on weather conditions:

Day	Weather	Temperature	Humidity	Wind	PlaySport
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes

14	Rain	Mild	High	Strong	No

Step 1: Calculate Entropy for the Target Variable

- Target variable: Play Tennis (Yes/No)
- Total instances: 14

p(Yes) = 9/14

p(No) = 5/14

Entropy(Play Tennis) = - [p(Yes) * log2(p(Yes)) + p(No) * log2(p(No))]

= - [(9/14) * log2(9/14) + (5/14) * log2(5/14)]

≈ 0.940

Step 2: Calculate Information Gain for Each Attribute

We'll calculate the information gain for each attribute (Outlook, Temp, Humidity, Wind) to determine the best attribute to split the dataset.

Information Gain (IG) = Entropy(parent) - [Weighted Average * Entropy(children)]

Information Gain for Weather

Sunny:

- p(Yes) = 2/5
- p(No) = 3/5
- Entropy(Sunny) = -[(2/5) * log2(2/5) + (3/5) * log2(3/5)] ≈ 0.971

Overcast:

- p(Yes) = 4/4 = 1
- p(No) = 0/4 = 0
- Entropy(Overcast) = 0

Rain:

- p(Yes) = 3/5
- p(No) = 2/5
- Entropy(Rain) = -[(3/5) * log2(3/5) + (2/5) * log2(2/5)] ≈ 0.971

Weighted Average Entropy:

- (5/14) * 0.971 + (4/14) * 0 + (5/14) * 0.971 ≈ 0.693

Information Gain:

- 0.940 - 0.693 ≈ 0.247

Hence,

Information Gain for Temp: 0.0289

Information Gain for Humidity: 0.1516

Information Gain for Wind: 0.0478

In this case, **Weather** has the highest information gain.

Advantages of Filter methods:

- Filter methods are model agnostic(compatible)
- Rely entirely on features in the data set
- Computationally very fast
- Based on different statistical methods

Disadvantages of Filter methods:

- The filter method looks at individual features for identifying it's relative importance. A feature may not be useful on its own but may be an important influencer when combined with other features. Filter methods may miss such features.
- One thing that should be kept in mind is that the filter method does not correctly remove multicollinearity. So, you must deal with the multicollinearity of features as well before training models for your data.

Wrapper Methods:

Definition: In Wrapper Methods the problem of Feature Selection is reduced to a search problem.

A model is built using a set of features and its accuracy is recorded.

Based on the accuracy, more features are added or removed, and the process is repeated.

Common Wrapper Methods:

1. Forward Selection:

- Forward Selection is an iterative method.
- In this method, we start with one feature and we keep on adding features until no improvement in the model is observed.
- The search is stopped after a pre-set criteria is met.
- This is a greedy approach because it always targets the features in a forward fashion, which gives a boost to the performance.

- If the number of features are large, it can be computationally expensive.

2. Backward Elimination:

- This process is the opposite of the Forward Selection Method.
- It starts initially with all the features and keeps on removing features until no improvement is observed.

3. Exhaustive Search:

- This Feature Selection Method tries all the possible combinations of features to select the best model.
- This method is quite computationally expensive.
- For example, if we have five features, we will be evaluating 2^5 = 32 models before finalizing a model with good accuracy.

4. Recursive Feature Elimination (RFE):

- Recursively Feature Elimination (RFE) involves the following steps:
  - We train the model on the initial set of features, and the importance of each feature is calculated.
  - In the second iteration, a model is built again using the most important features and excluding the least important features.
  - These steps are repeated recursively until we are left with the most important features for the problem under consideration.
- Scikit-learn library provides functions for Recursive Feature Elimination.

Forward Feature Selection:

Forward Feature Selection is a stepwise process for selecting important features for a predictive model. It begins with an empty model and iteratively adds features that improve the model's performance the most. Here's how you can perform forward feature selection:

- **Select the Best Single Feature**: Add each feature to the model individually and select the one that improves the model's performance the most.
- **Add the Next Best Feature**: Add each of the remaining features to the current model (one at a time) and select the one that provides the best performance improvement.
- **Repeat the Process**: Continue adding features one by one, selecting the one that improves the model the most each time, until adding more features does not significantly improve the model's performance or a stopping criterion is met.
- **Final Model**: The process stops when no additional features improve the model significantly or when a predefined number of features is reached.

Example:

Let's use the given dataset to perform forward feature selection to predict the **Fitness Level**.

Initial Dataset:

ID	Calories Burnt	Gender	Plays Sport?	Fitness Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Step-by-Step Forward Feature Selection:

Step 1: Add the Best Single Feature:

Evaluate each feature's performance:

- Fitness Level ~ Calories Burnt
- Fitness Level ~ Gender
- Fitness Level ~ Plays Sport?

Let's assume adding **Calories Burnt** results in the best performance improvement and we got an accuracy of **87%**.

Step 2: Add the Next Best Feature:

Evaluate adding each of the remaining features to the model with **Calories Burnt**:

- Fitness Level ~ Calories Burnt + Gender
- Fitness Level ~ Calories Burnt + Plays Sport?

Suppose adding **Plays Sport?** improves the model performance the most and we get an accuracy of **91%**.

Step 3: Repeat the Process:

Evaluate adding the remaining feature to the model:

- Fitness Level ~ Calories Burnt + Plays Sport? + Gender

Assume adding **Gender** does not significantly improve the model further.

Final Model:

The final model after forward feature selection: **Fitness Level ~ Calories Burnt + Plays Sport?**

Backward Elimination:

Backward Elimination is a stepwise feature selection process where you start with all features and iteratively remove the least significant ones. This method aims to identify the most important features by eliminating the least contributing ones.

- **Start with All Features**: Begin with a model that includes all available features.
- **Fit the Model**: Fit the model to the data and evaluate the accuracy.
- **Identify the Least Significant Feature**: Identify the feature with the low accuracy.
- **Remove the Least Significant Feature**: Remove the feature identified in the previous step.
- **Refit the Model**: Fit the model again with the remaining features.
- **Repeat the Process**: Continue removing the least significant feature, refitting the model each time, until all remaining features are significant, or a stopping criterion is met.
- **Final Model**: The process stops when no further features can be removed without a significant loss in model performance.

Example:

Using the same dataset to perform backward elimination to predict the **Fitness Level**.

Initial Dataset:

ID	Calories Burnt	Gender	Plays Sport?	Fitness Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Step-by-Step Backward Elimination:

Step 1: Start with All Features:

- Initial model: Fitness Level ~ Calories Burnt + Gender + Plays Sport?
- Suppose we get an accuracy of **92%** using all three independent variables.

Step 2: Identify the Least Significant Feature:

Model without 'Gender':

- **Features:** Calories Burnt, Plays Sport?
- **Accuracy:** 91.60%

Model without 'Plays Sport?':

- **Features:** Calories Burnt, Gender
- **Accuracy:** 88%

Model without 'Calories Burnt':

- **Features:** Gender, Plays Sport?
- **Accuracy**: 90%

Step 3: Remove the Least Significant Feature:

If you see gender has produced the smallest change in the performance in the model first, it was **92%** when we took all the variables and when we dropped gender, it was **91.6%**. So, we can infer that gender does not have a high impact on the Fitness_Level variable. And hence it can be dropped.

Step 4: Repeat the Process:

We will repeat all these steps until no more variables can be dropped.

Multicollinearity

Multicollinearity is a statistical phenomenon that occurs when two or more independent variables in a regression model are highly correlated with each other.

Variance Inflation Factor (VIF)

Variance Inflation Factor (VIF) is a metric used to detect the presence and severity of multicollinearity in regression models. It determines the strength of the correlation between independent variables.

Key Points:

- VIF measures how much the variance of a regression coefficient is inflated due to multicollinearity with other independent variables.
- A high VIF indicates that the variable has a strong linear relationship with other predictors, which may distort the model.

VIF Formula:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where:

- VIF_i is the Variance Inflation Factor for the ith predictor.
- R_i^2 is the coefficient of determination obtained by regressing the ith predictor on all the other predictors.

VIF Interpretation:

VIF Value	Interpretation
1	No correlation between the variable and others (no multicollinearity).
1 to 5	Moderate correlation, acceptable level of multicollinearity.
> 5	High correlation, potential multicollinearity problem.
> 10	Severe multicollinearity that requires correction.

Addressing Multicollinearity:

- Removing highly correlated variables.
- Combining correlated variables into a single feature.
- Applying dimensionality reduction techniques like Principal Component Analysis (PCA).

Example: Calculating VIF

Given the dataset with the following features:

	Gender	Age	Years of service	Education level	Salary
0	0.0	27.0	1.7	0.0	39343.0
1	1.0	26.0	1.1	1.0	43205.0
2	1.0	26.0	1.2	0.0	47731.0
3	0.0	27.0	1.6	1.0	46525.0
4	0.0	26.0	1.5	1.0	40891.0

Steps to Calculate VIF:

1. For each independent variable, fit a regression model with the variable as the dependent variable and all other independent variables as predictors.
2. Calculate the value for each regression.
3. Apply the VIF formula.

	variables	VIF
0	Gender	2.207155
1	Age	13.706320
2	Years of service	10.299486
3	Education level	2.409263

Feature Extraction using Principle Component Analysis:

Given the data in Table, reduce the dimesion from 2 to 1 using the Principal Component Analysis (PCA) algorithm.

Feature	Example 1	Example 2	Example 3	Example 4
X_1	4	8	13	7
X_2	11	4	5	14

Step 1: Calculate Mean:

$$\bar{X}_1 = \frac{1}{4}(4+8+13+7) = 8$$

$$\bar{X}_2 = \frac{1}{4}(11+4+5+14) = 8.5$$

Step 2: Calculate of the covariance matrix:

$$S = \begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) \\ Cov(X_2, X_1) & Cov(X_2, X_2) \end{bmatrix}$$

$$Cov(X_1, X_1) = \frac{1}{n-1} \sum_{k=1}^{n} (X_{1k} - \bar{X}_1)(X_{1k} - \bar{X})$$

$$= \frac{1}{3}\left((4-8)^2 + (8-8)^2 + (13-8)^2 + (7-8)^2\right)$$

$$= 14$$

$$Cov(X_1, X_2) = \frac{1}{3}\left((4-8)(11-8.5) + (8-8)(4-8.5) + (13-8)(5-8.5) + (7-8)(14-8.5)\right)$$

$$= -11$$

$$Cov(X_2, X_1) = -11$$

$$Cov(X_2, X_2) = 23$$

$$S = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

Step 3: Eigenvalues of the covariance matrix.

$$0 = \det(S - \lambda I)$$

$$= \begin{vmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{vmatrix}$$

$$= (14-\lambda)(23-\lambda) - (-11)(-11)$$

$$= \lambda^2 - 37\lambda + 201$$

By using calculator,

$$\lambda_1 = 30.3849$$

$$\lambda_2 = 6.6151$$

Step 4: Computation of the eigenvectors:

$$U = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \qquad \begin{bmatrix} 0 \\ 0 \end{bmatrix} = (S - \lambda I)U$$

$$= \begin{bmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$= \begin{bmatrix} (14-\lambda)u_1 - 11u_2 \\ -11u_1 + (23-\lambda)u_2 \end{bmatrix}$$

$$\Rightarrow (14-\lambda)u_1 - 11u_2 = 0$$

$$(14-\lambda)u_1 = 11u_2$$

$$\frac{u_1}{11} = \frac{u_2}{(14-\lambda)} = t$$

$u_1 = 11t \quad , \quad u_2 = (14-\lambda)t$

Hence,

$$u = \begin{bmatrix} 11 \\ 14-\lambda \end{bmatrix}$$

To find the first Principal Components, we need only compute the eigenvector corresponding to the largest eigenvalue.

First,

$$\|u_1\| = \sqrt{11^2 + (14-\lambda_1)^2}$$
$$= \sqrt{11^2 + (14-30.3849)^2}$$
$$= 19.7348$$

Now,

$$e_1 = \begin{bmatrix} 11/19.7348 \\ (14-30.3849)/19.7348 \end{bmatrix}$$
$$= \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}$$

Step 5: Computation of first Principal components:

$$e_1^T \begin{bmatrix} X_{1k} - \bar{X}_1 \\ X_{2k} - \bar{X}_2 \end{bmatrix}$$
$$= \begin{bmatrix} 0.5574 & -0.8303 \end{bmatrix} \begin{bmatrix} X_{11} - \bar{X}_1 \\ X_{21} - \bar{X}_2 \end{bmatrix}$$
$$= 0.5574(4-8) - 0.8303(11-8.5)$$
$$= -4.30535$$

X_1	4	8	13	7
X_2	11	4	5	14
First Principal Components	-4.3052	3.7361	5.6928	-5.1238

Correlation Analysis:

Correlation analysis helps identify and measure how strongly one attribute implies or predicts another attribute based on the available data.

- For **nominal data** (categorical data), the **χ2 (chi-square) test** is often used.
- For **numeric attributes**, two common measures are:
  - The **correlation coefficient**
  - **Covariance**

χ2 Correlation Test:

The chi-square test of independence is a statistical method used to determine if there's a relationship between two categorical variables.

Example: Suppose that a group of 1500 people was surveyed. The gender of each person was noted. Each person was polled as to whether his or her preferred type of reading material was fiction or nonfiction.

Step 1: Create a contingency table for any two attributes

	Male	Female	Total
fiction	250	200	450
non_fiction	50	1000	1050
Total	300	1200	1500

Step 2: State Hypothesis

Null Hypothesis: H_0 : Independent

Alternative Hypothesis: H_1 : Dependent

Here, **v** = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1

Step 3: State Alpha

alpha = 0.001

Step 4: Calculate Expected Values

$$e_{11} = \frac{count(male) \times count(fiction)}{n} = \frac{300 \times 450}{1500} = 90,$$

Step 5: Calculate χ2 formula

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.$$

Step 6: If calculated value chi-square is less than or equal to the critical/tabular value, then the attributes are not correlated

$$critical\ value\ of\ \chi^2\ >=\ calculated\ value\ of\ \chi^2$$

10.828 < 507.93

So we can infer that correlation exist.

Correlation Coefficient:

To quantify the strength and direction of the correlation between two numerical variables, we use a number called the **correlation coefficient**. This number ranges between -1 and 1.

- A correlation coefficient of **1** indicates a perfect positive correlation.
- A correlation coefficient of **-1** indicates a perfect negative correlation.
- A correlation coefficient close to **0** suggests little to no correlation.

$$r_{A,B} = \frac{\Sigma(A-\bar{A})(B-\bar{B})}{(n-1)\sigma_A\sigma_B}$$

Covariance:

Covariance is a measure used to determine the direction of the linear relationship between two variables. Unlike correlation, which standardizes the measure, covariance provides the actual measure of the joint variability of two random variables.

- **Positive Covariance**: Indicates that the variables tend to move in the same direction. When one variable increases, the other also tends to increase.
- **Negative Covariance**: Indicates that the variables tend to move in opposite directions. When one variable increases, the other tends to decrease.
- **Covariance close to zero**: Suggests that there is little to no linear relationship between the variables.

$$E(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n}$$

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}.$$

Stock Prices for *AllElectronics* and *HighTech*

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

$$E(AllElectronics) = \frac{6+5+4+3+2}{5} = \frac{20}{5} = \$4$$

$$E(HighTech) = \frac{20+10+14+5+5}{5} = \frac{54}{5} = \$10.80.$$

$$Cov(AllElectroncis, HighTech) = \frac{6\times 20 + 5\times 10 + 4\times 14 + 3\times 5 + 2\times 5}{5} - 4\times 10.80$$
$$= 50.2 - 43.2 = 7.$$

Student_ID	Study_Hours	Test_Score
1	2	50
2	3	60
3	5	75
4	7	85
5	8	90

Covariance Calculation

Formula:

$$Cov(A,B) = E(A.B) - \bar{A}.\bar{B}$$

Where,

A = Study_Hours
B = Test_Score

Now,

$$\bar{A} = \frac{2+3+5+7+8}{5} = 5$$

$$\bar{B} = \frac{50+60+75+85+90}{5} = 72$$

A	B	A.B
2	50	100
3	60	180
5	75	375
7	85	595
8	90	720

$$E(A.B) = \frac{100+180+375+595+720}{5} = 394$$

$$Cov(A,B) = (394) - (5)(72)$$
$$= 34$$

Correlation Coefficient:

Formula: $r = \frac{Cov(A,B)}{\sigma_A . \sigma_B}$

Here,

$$\sigma_A = \sqrt{\frac{\Sigma(A-\bar{A})^2}{n-1}} = \sqrt{\frac{1}{4}((2-5)^2+\ldots(8-5)^2)} = 2.5495$$

$$\sigma_B = \sqrt{\frac{\Sigma(B-\bar{B})^2}{n-1}} = \sqrt{\frac{1}{4}((50-72)^2+\ldots(90-72)^2)}$$
$$= 16.8077$$

Hence,

$$r = \frac{34}{2.5495 \times 16.8077} = 0.9918$$