# ASSIGNMENT - 1

## Applied Data Science

NAME - SOURAV BISWAS

ROLL NO - 224CS1016

SPECIALIZATION - COMPUTER SCIENCE

---

1. Given a binary classification data set with three input features. If the 2 classes are separated by a sphere with centre at (0,0,0) and radius of 5, then devise a mathematical model to classify the dataset.

Soln:

Let the 3 input features be $x_1, x_2, x_3$.

Centre at (0,0,0)

Radius = 5 units

The mathematical model to classify the dataset will be distance calculation from the origin, represented as

$$d = \sqrt{x_1^2 + x_2^2 + x_3^2}$$

For binary classification we need 2 classes 0 and 1. Let us consider the point lying inside the sphere that is the distance of the point from the origin / centre of the sphere is less than equal to 5 as Class 1.

And the point outside the sphere that is distance of the point from origin is greater than 5 as Class 0.

The equation of the decision bouridary is given as

$$\sqrt{x_1^2 + x_2^2 + x_3^2} = 5$$

$$\Rightarrow x_1^2 + x_2^2 + x_3^2 = 25 \quad [\text{Squaring both sides}]$$

Final Mathematical model to classify is

$$y(x_1, x_2, x_3) = \begin{cases} 1, & \text{if } x_1^2 + x_2^2 + x_3^2 \leq 25 \\ 0, & \text{if } x_1^2 + x_2^2 + x_3^2 > 25 \end{cases}$$

where $y = 1$ represents class 1
$y = 0$ represent class 0

2. How hyper-planes be interpreted in a binary classification problem?

Soln. In a binary classification problem, hyper-planes are commonly used to separate the two classes in a feature space.

A hyperplane is a flat subspace 1 dimension less than the feature space that is if input exist in an $n$-dimensional space, then a hyperplane is an $(n-1)$-dimensional geometric object. It generalizes the concept of a line in 2-D feature space and a plane in 3-D feature space.
It is mathematically represented as:

$$w_1 x_1 + w_2 x_2 \ldots w_n x_n + b = 0$$

- $w_1, w_2, w_3$ are ~~input features~~ weights corresponding to input feature.
- $x_1, x_2, x_3$ are input features
- $b$ is the bias value.

In a 2D plane, if $w_1 x_1 + w_2 x_2 + b = 0$

then $w_1 x_1 + w_2 x_2 + b > 0 \longrightarrow$ Class A

$w_1 x_1 + w_2 x_2 + b < 0 \longrightarrow$ Class B

3. Mention four plots used in univariate analysis.

Soln. Four plots used in univariate analysis are:-

- Histogram
- Probability Density Function
- Box Plot
- Violin Plot

4. Given a pair plot with 4 features. How many distinct scatter plots will this pair plot have?

Soln. The no. of distinct scatter plot if we take 2 features into consideration will be in the form of ${}^4C_2$ where 4 is the no. of features.

$\therefore$ Total no. of distinct plots $= {}^4C_2 = \frac{4 \times 3}{2 \times 1}$

$= 6$ distinct scatter plots.

5. What cumulative distribution function signifies?

Soln. Cumulative Distribution Function signifies the probability that a random variable is less than or equal to a given value. It provides the cumulative probabilities ~~and~~ instead probability densities.

6. Given below the CDF of petal length of Setosa flowers? How many percentage of Setosa flowers petal length is less than 1.5?

Soln. From the graph the probabilities (cumulative) of Setosa flower's petal length less than 1.5 is 0.6.

$\therefore$ The percentage = 60%.

7. Given a classification problem dataset with a single feature which may have outliers. How can you develop a simple model to predict the class labels quantitatively?

Soln: In order to develop a simple model for a classification problem with a single feature which may have outliers, we need to visualize the data & figure out the outliers.

Using 1) Box plot or
2) Scatter plot outliers can easily be detected.

Outlier can be handled by correcting them manually or can be removed from the dataset. Various methods like clipping or winsorizing can be used.

To develop a simple model we can choose either of the one for predicting the class labels -

① Logistic Regression - Predict the class labels using a sigmoid function

② Decision Tree - Splits the feature space into intervals and more robust to outliers.

③ K-Nearest Neighbour - Classifies based on the nearest neighbours, but can be sensitive to outliers. Using robust distance metrics like Manhattan distance can be optimal.

8. What is a random variable? Write a short note on types of random variable based on number of outcomes.

Soln. A random variable is a variable that takes the outcomes of the random experiment as its value.

In probability, a real-valued, defined over the sample space of a random experiment, is called a random experiment.

There are 2 types of random variable -

1) Discrete random variable
   - It can take only a finite number of discrete values (Sample space is finite)
   - Probability mass function (PMF) is used to describe the probabilities
   - Eg:- Tossing a coin, Rolling a die

2) Continuous random variable
   - It can take infinite number of values. (Sample space has uncountable values)
   - Probability Density Function (PDF) is used to describe the probabilities of continuous random variables.
   - Eg:- Height and weight of a person.

9. Suppose the random variable height of a person follows normal distribution with mean ($\mu$) = 150 cm and variance $\sigma^2 = 25$. What is the range within which 68% person's height lie.

Soln.

Mean ($\mu$) = 150 cm

Variance ($\sigma^2$) = 25

Standard deviation ($\sigma$) = $\sqrt{25}$

= $\pm 5$

In a normal distribution, 68% of the data lies within 1 standard deviation of the mean, i.e

$\mu - \sigma \leq h \leq \mu + \sigma$ $\quad$ $h$ = height

$\Rightarrow 150 - 5 \leq h \leq 150 + 5$

$\Rightarrow 145 \leq h \leq 155$

$\therefore$ The ranges starts from 145 cm to 155 cm.

10. Explain the Central Limits Theorem.

Soln. Let $X$ be a random variable with finite population mean $\bar{X}$ and variance $\sigma^2$ which follows a distribution.

Let $S_1, S_2, S_3 \ldots S_m$ be $m$ sample spaces with each containing $n$ observations and $\bar{S_1}, \bar{S_2} \ldots \bar{S_m}$ be their sample means resp.

Then the distribution formed by $\bar{S_1}, \bar{S_2}, \bar{S_3} \ldots \bar{S_m}$ is called sampling distribution of sample means which follows a normal distribution with mean $\bar{X}$ and variance $\sigma^2/n$ as long as the sample size is large enough.

11. Why other distributions are tried to be transformed to normal distribution? Justify your answer.

12. What covariance between 2 variables signifies? How to compute covariance between 2 variables?

Soln. Covariance between 2 variables signifies the joint variability of the 2 variables. That is with the change in 1 variable, how the other variable changes.

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x) * (y_i - \mu_y)$$

where $n$ = no. of samples.
$\mu_x$ = mean of $x$-sample
$\mu_y$ = mean of $y$-sample
$x_i$ = $i^{th}$ sample in $x$
$y_i$ = $i^{th}$ sample in $y$

$cov(X,Y) > 0 \longrightarrow$ Both variables move in same direction
$cov(X,Y) < 0 \longrightarrow$ Both variables move in opposite direction
$cov(X,Y) = 0 \longrightarrow$ No relation between the 2 variables.

13. Compute the pearson's correlation coefficient and spearman's rank correlation coefficient between the following 2 variables: $X = \{10, 15, 13, 6\}$ and $Y = \{3, 9, 5, 1\}$.

Soln: $X = \{10, 15, 13, 6\}$ $Y = \{3, 9, 5, 1\}$

$\mu_x = 11$ $\mu_y = 4.5$

$$cov(X,Y) = \frac{1}{4}[(10-11)(3-4.5) + (15-11)(9-4.5) + (13-11)(5-4.5) + (6-11)(1-4.5)]$$

$$cov(X,Y) = \frac{38}{4} = 9.5$$

$$\sigma_x = \sqrt{\frac{1}{n}\Sigma(x_i - \mu_x)^2} = \sqrt{\frac{1}{4}[(10-11)^2 + (15-11)^2 + (13-11)^2 + (6-11)^2]}$$

$$= \sqrt{\frac{1}{4}[(-1)^2 + 4^2 + 2^2 + (-5)^2]} = \sqrt{\frac{46}{4}} = 3.39$$

$$\sigma_y = \sqrt{\frac{1}{n}[(3-4.5)^2 + (9-4.5)^2 + (5-4.5)^2 + (1-4.5)^2}$$

$$= \sqrt{\frac{1}{4}[(-0.5)^2 + (4.5)^2 + (0.5)^2 + (-3.5)^2}$$

$$= \sqrt{\frac{33}{4}}$$

$$= 2.87$$

$\therefore$ Pearson's Correlation coeff $\rho(X,Y) = \dfrac{cov(X,Y)}{\sigma_x \sigma_y}$

$$= \frac{9.5}{3.39 \times 2.87}$$

$$= \frac{9.5}{9.7293}$$

$$= 0.9764 \text{ (Ans)}$$

(Positive correlation)

| X | Y | $R_x$ | $R_y$ | $d = R_x - R_y$ | $d^2$ |
|---|---|---|---|---|---|
| 10 | 3 | 2 | 2 | 0 | 0 |
| 15 | 9 | 4 | 4 | 0 | 0 |
| 13 | 5 | 3 | 3 | 0 | 0 |
| 6 | 1 | 1 | 1 | 0 | 0 |
| | | | | | $\sum d^2 = 0$ |

$n = 4$

Spearank's Correlation Coefficient $(\rho) = 1 - \dfrac{6 \sum d_i^2}{n(n^2-1)}$

$$= 1 - \frac{6 \times 0}{4(16-1)}$$

$$= 1 - 0$$

$$= 1 \quad \text{(Ans)}$$

14. What is an attribute? How many types of attributes are there? Differentiate between them.

Soln. Attribute is a property or characteristic of an object that may vary, either from one object to another or from one time to another.

Types of attributes in total are 4 types.
They are:

1) Nominal - Values of a nominal attribute are just different names. (=, ≠) Eg:- Zip codes, gender

2) Ordinal - Provide enough information to ~~object~~ order objects (<, >). Eg:- Grades, {Good, better, best}

3) Interval - The difference between values are meaningful i.e unit of measurement existence. (+, -)
Eg:- calender dates, temperature.

4) Ratio - Both differences and ratios are meaningful (*, /)
Eg:- Temperature in Kelvin, counts, age.

15. What do you mean by Curse of Dimensionality Problem? How can it be overcomed?

Soln. Curse of Dimensionality Problem states that when dimensionality increases, data becomes increasingly sparse. Density and distance between points, which is critical to clustering, outlier analysis, become less meaningful. The possible combination of supspaces will grow exponentially.

Curse of dimensionality can be overcomed by:
1. Avoid including irrelevant feature and eliminate if any and reduce noise
2. Reduce time and space required.
3. Allow easier visualization.

16. What is the difference between feature selection and dimensionality reduction techniques?

Soln.

| Feature Selection | Dimensionality Reduction |
|---|---|
| 1. Involves selecting a subset of the most relevant features from the original dataset | 1. Transforms the original features into a lower-dimensional space. |
| 2. It evaluates the importance of each feature based on criteria such as statistical tests, correlation with the target variable, or model-based importand scores. | 2. It combines the features into new ones, often by capturing the underlying structure of the data. |
| 3. The output retains the original features but reduces their number. | 3. The output consists of new features that are combinations of the original features. |
| 4. Eg:- Chi-squared test, Recursive feature elimination, Lasso regression. | 4. Eg:- Principal Component Analysis (PCA), t-Distribution Schochastic Neighbour embedding, Wavelength Transform |

17. Explain the steps of Principal Component Analysis used for dimensionality Reduction.

Soln. The steps of Principal Component Analysis used for dimensionality reduction are:-

1. Standardize the range of continuous initial variables.
   - so that each one of them contributes equally to the analysis
   - If there are larger differences between the ranges of initial variables, those variables will dominate over those with small ranges.

- Mathematically, this can be done by subtracting the mean and dividing by the standard deviation of each variable.

2) Compute the covariance matrix to identify correlations.
   - understand how the variables of the input data set are varying from the mean with respect to each other, i.e
   - we compute the covariance matrix.

3) Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components
   - Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine principal components of the data.
   - Principal components are the new variables that are constructed as linear combinations or mixture of the initial variables.

4) Create a feature vector to decide which pricipal components to keep.
   - Feature vector is simply a matrix that has as columns the eigenvectors of the components that we decide to keep.
   - This makes it the first step towards dimentionality reduction, because if we choose to keep only p eigenvectors out of n, the final data set will have only p dimensions.

5) Recast the data along the principal components axes.
   - Aim to use the feature vector formed using the eigenvectors of the covariance matrix, to reorient the data from the original axes to the ones represented by the principal components.
   - Can be done by multiplying the transpose of the original data set by the transpose of the feature vector.

18. What is an outlier? Give 2 examples where outliers are the goal of analysis.

Soln. Outliers are data objects with characteristics that are considerably different than most of the other data objects in the dataset.

2 examples where outliers are goal of analysis -

- Credit card fraud
- Intrusion detection.

19. Given two objects having binary attributes only:

x = 1 0 0 0 0 0 0 0 0 0

y = 0 0 0 0 0 0 1 0 0 1

What is the Simple Matching Coefficient of the 2 objects.

Soln. x = 1000000000

y = 0000001001

$f_{01}$ → Value of x = 0 and y = 1 → 2

$f_{10}$ → Value of x = 1 and y = 0 → 1

$f_{00}$ → Value of x = 0 and y = 0 → 7

$f_{11}$ → Value of x = 1 and y = 1 → 0

Simple Matching Coefficient (SMC)

$$= \frac{f_{00} + f_{11}}{f_{00} + f_{11} + f_{01} + f_{10}}$$

$$= \frac{0+7}{7+0+1+2}$$

$$= \frac{7}{10}$$

$$= 0.7 \text{ (Ans)}$$

20. Given two documents d1 = 4205003204 and d2 = 1000000102. Compute the cosine similarity.

Soln: d1 = 4205003204
d2 = 1000000102

Inner product $\langle d1, d2 \rangle$ =

$4 \times 1 + 2 \times 0 + 0 \times 0 + 5 \times 0 + 0 \times 0 + 0 \times 0 + 3 \times 0 + 2 \times 0 + 0 \times 0 + 4 \times 2$

$= 4 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 8$

$= 4 + 2 + 8$

$= 14$

$\|d_1\| = (4^2 + 2^2 + 5^2 + 3^2 + 2^2 + 4^2)^{1/2}$

$= \sqrt{74}$

$= 8.602$

$\|d_2\| = (1^2 + 1^2 + 2^2)^{1/2}$

$= \sqrt{6}$

$= 2.449$

$\therefore \cos(d_1, d_2) = \dfrac{\langle d_1, d_2 \rangle}{\|d_1\| \|d_2\|}$

$= \dfrac{14}{8.602 \times 2.449}$

$= 0.6645$ (Ans)

21. What do you mean by scaling and translation? Fill the following table.

Soln:

| Property | Invariant to scaling | Invariant to Translation |
|---|---|---|
| Cosine | Yes | No |
| Correlation | Yes | Yes |

Scaling - Multiplication by value

Translation - Adding a constant.

Eg:-

Scaling → $y_s = y * 2$

Translation → $y_t = y + 5$

22. What is the entropy of a fair 4 sided die?

Soln: 4 sided die.

Let the event be X.

For a 4 sided die, there are 4 possible outcomes = 1, 2, 3, 4.

Each outcome has a probability of = $\frac{1}{4}$ $P_1 = P_2 = P_3 = P_4 = \frac{1}{4}$

$$H(x) = -\sum_{i=1}^{4} P_i \log_2 P_i$$

$$= - 4 \times \frac{1}{4} \log_2 \frac{1}{4}$$

$$= - 4 \times 2^{-2} \log_2 2^{-2}$$

$$= (-4) \times (-2) \times 2^{-2} \times 1$$

$$= 2 \quad \underline{(Ans)}$$

23. Given the following data (in non-decreasing order) for the attribute age : 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

a) Use smoothing by bin means to smooth these data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

b) How might you determine outliers in the data?

c) What ~~are~~ ~~the~~ other methods are there for data smoothing?.

Soln : a) Smoothing by Bin Means

Step 1 - Organize data with 3 bin size

Bin 1 - (13, 15, 16)
Bin 2 - (18, 19, 20)
Bin 3 - (20, 21, 22)
Bin 4 - (22, 25, 25)
Bin 5 - (25, 25, 30)
Bin 6 - (33, 33, 35)
Bin 7 - (35, 35, 35)
Bin 8 - (36, 40, 45)
Bin 9 - (46, 52, 70)

Step 2 - Calculate the bin mean

Bin 1 = 14.67
Bin 2 = 18.33
Bin 3 = 21
Bin 4 = 24
Bin 5 = 26.67
Bin 6 = 33.67
Bin 7 = 35
Bin 8 = 40.33
Bin 9 = 56

Step 3 - Replace each value with bin mean

Bin 1 = (14.67, 14.67, 14.67)
Bin 2 = (18.33, 18.33, 18.33)
Bin 3 = (21, 21, 21)
Bin 4 = (24, 24, 24)
Bin 5 = (26.67, 26.67, 26.67)
Bin 6 = (33.67, 33.67, 33.67)
Bin 7 = (35, 35, 35)
Bin 8 = (40.33, 40.33, 40.33)
Bin 9 = (56, 56, 56)

Smoothed Data

14.67, 14.67, 14.67, 18.33, 18.33, 18.33, 21, 21, 21, 24, 24, 24, 26.67, 26.67, 26.67, 33.67, 33.67, 33.67, 35, 35, 35, 40.33, 40.33, 40.33, 56, 56, 56

Effect of Smoothing by Bin Means -

- Good scaling of data.
- Managing categorical attributes can tricky.

b) To Determine Outliers in the Data -

1. Z-Score - Data points with Z-score greater than 3 are considered outliers

2. Interquartile Range (IQR) Method -
   Data points below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ are considered outliers.

3. Boxplot Visualization.

c) Other methods of Data Smoothing -

1. Histogram analysis
2. Clustering analysis
3. Decision tree analysis
4. Correlation

24. Using the data for age given in Question 23, answer the following:

a) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].

b) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.

c) Use normalization by decimal scaling to transform the value 35 for age.

d) comment on which method you would prefer to use for the given data, giving reasons as to why.

Soln.

a) $new\_min_A = 0.0$

$new\_max_A = 1.0$

$$v' = \frac{35 - 13}{70 - 13}(1.0 - 0.0) + 0$$

$$= 0.4583$$

b) $$v' = \frac{v - \mu}{\sigma} \qquad \mu = \frac{809}{27}$$

$$= 29.96$$

$$= \frac{35 - 29.96}{12.94}$$

$$= 5.02$$

c) $v' = \frac{v}{10^j}$

$= \frac{35}{100}$

$= \frac{35}{10^2}$

$= 0.35 < 1$

$\therefore j = 2$

d) In order to visualize data within bounded range, min-max normalization is used.

Otherwise, z-score normalization is preferred as it follows a Gaussian distribution.

25. Suppose a group of 12 sales price records has been sorted as follows: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215. Partition them into 3 bins by each of the following methods:

a) equal-frequency (equal depth) partitioning.

b) equal-width partitioning

Soln. a) Equal Depth

12 records / 3 = 4 record per bin

Bin 1 – (5, 10, 11, 13)

Bin 2 – (15, 35, 50 55)

Bin 3 – (72, 92, 204, 215)

b) Equal Width Partitioning

Range = 215 - 5 = 210

For 3 bins

$$\text{Width} = \frac{\text{Range}}{\text{Bin}} = 70$$

Bin boundaries are -

Bin 1 - [5, 75)

Bin 2 - [75, 145)

Bin 3 - [145 - 215]

∴ Bin 1 - (5, 10, 11, 13, 15, 35, 50, 55, 72)

Bin 2 - (92)

Bin 3 - (204, 215)

26. Explain the Huber loss. When will you prefer to use huber loss instead of squared loss? Justify your answer.

Soln. Huber loss is a loss function used in robust regression that combines the properties of squared loss and absolute loss. It is defined as:

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{if } |y - f(x)| \leq \delta \\ \delta \cdot (|y - f(x)| - \frac{1}{2}\delta) & \text{otherwise} \end{cases}$$

where $y$ is true value,

$f(x)$ is the predicted value,

$\delta$ is the threshold parameter

- Quadratic for small errors (with a threshold $\delta$)
- Linear for Large Errors

We prefer Huber Loss instead of squared loss because -

- It is more robust against outliers.
- Effectively minimizes small errors.
- Differentiable everywhere, making it stable for optimization

Justification

1. Robustness - Reduces the impact of outliers in real-world datasets.
2. Flexibility - The parameter $\delta$ allows tuning for sensitivity based on data characteristics.
3. Empirical Performance - Often provides better predictive performance in datasets with both typical observations and outliers.

27. What is overfitting? How can you control overfitting in Linear Regression? Explain two different ways.

Soln. Overfitting is the condition when the training error is considerably less but the testing error may be high. This is because every input sample including the outliers can be mapped to the model for the training data.

Overfitting can be controlled in Linear Regression in various ways like -

- Clean dataset training
- Performing cross validation
- Regularization.
- Ensembling.

Two ways -

1. Regularization - It is a collection of training/optimization techniques that seek to reduce overfitting. These methods try to eliminate those factors that do not impact the prediction outcomes by grading features based on importance.

2) Ensembling - It combines predictions from several separate models. Some models are weak learners because their results are often inaccurate. Ensemble methods combine all the weak learners to get more accurate result.

28. Explain Data Science Process with a neat Diagram.

Soln.