CWID: A20555359

CS422 Data Mining

Assignment – 1

1. Recitation Exercises

1.1 Chapter 1

(1) Show that the mean of the centered data matrix D in Eq. (1.9) is 0.

Reference:

Based on Zaki's "Data Mining and Machine Learning", 2nd Edition

Equation (1.9): **Centering the data matrix**

Given:

- $D \in R^{nxd}$ is the data matrix, where each row is a data point x_i
- $\mu \in R^d$ is the mean vector of the data points

$\mu = (1/n) \times (x_1 + x_2 + ... + x_n)$

- Centered data matrix:

$\tilde{D} = D - 1 \cdot \mu^T$

where:

- $1 \in R^n$ is a column vector with all entries equal to 1
- $\mu^T \in R^{1xd}$ is the transpose of the mean vector

To Prove:

The mean of the centered matrix $\tilde{D}$ is the zero vector:

$(1/n) \times$ (sum of all rows of $\tilde{D}$) $= 0 \in R^d$

Step 1: Expand the definition of $\tilde{D}$

Each row of $\tilde{D}$ is $x_i - \mu$

Mean of $\tilde{D}$:

$(1/n) \times [(x_1 - \mu) + (x_2 - \mu) + ... + (x_n - \mu)]$

Step 2: Distribute the summation

$= (1/n) \times [(x_1 + x_2 + ... + x_n) - n\mu]$

Step 3: Use the definition of μ

From $\mu = (1/n) \times (x_1 + x_2 + ... + x_n)$,

$\Rightarrow n\mu = (x_1 + x_2 + ... + x_n)$

So we substitute:

$= (1/n) \times [n\mu - n\mu]$

$= (1/n) \times 0$

$= 0$

Conclusion:

The mean of the centered data matrix $\tilde{D}$ is $0 \in R^d$.

This confirms that subtracting the mean vector from each data point centers the dataset at the origin.

1.2 Chapter 2

(1) True or False:

(a) Mean is robust against outliers.

Answer: False

Explanation: The mean is calculated by adding all values and dividing by the number of values. Because it takes into account every data point, a very large or very small value (an outlier) can significantly affect the mean, pulling it towards the outlier. So the mean is sensitive and not robust to outliers.

Example: Suppose the values are [2, 3, 4, 5, 100]. The mean is (2 + 3 + 4 + 5 + 100) / 5 = 114 / 5 = 22.8, which is much higher than most values because of the outlier 100. This shows the mean is affected by outliers.

(b) Median is robust against outliers.

Answer: True

Explanation: The median is the middle value when data is sorted. Since it depends only on the order and position of values, not their magnitude, extreme values (outliers) do not influence the median much. Therefore, the median is considered robust against outliers.

Example: For the same data [2, 3, 4, 5, 100], ordered it is the same. The median is the middle value, which is 4. This median stays the same even if the 100 becomes 1000 or 10000.

(c) Standard deviation is robust against outliers.

Answer: False

Explanation: Standard deviation measures the spread of data around the mean. Because the mean itself is sensitive to outliers, and because standard deviation squares the differences from the mean, outliers can greatly increase the standard deviation. Hence, standard deviation is not robust to outliers.

Example: For [2, 3, 4, 5, 100], the standard deviation is very large because 100 is far from the mean. If the outlier was smaller, say 10, the standard deviation would be much less.

(2) Given:

Random variables X (age) and Y (weight) with sample size

n = 20:

X = (69, 74, 68, 70, 72, 67, 66, 70, 76, 68, 72, 79, 74, 67, 66, 71, 74, 75, 75, 76)

Y = (153, 175, 155, 135, 172, 150, 115, 137, 200, 130, 140, 265, 185, 112, 140, 150, 165, 185, 210, 220)

(a) Find the mean, median, and mode for X.

Mean (μ) formula: $\mu = (1/n) \sum x_i$

Sum of X = 69 + 74 + 68 + ... + 75 + 76 = 1429

Mean μ = 1429 / 20 = 71.45

Median: Sort X and find middle value

Sorted X: [66, 66, 67, 67, 68, 68, 69, 70, 70, 71, 72, 72, 74, 74, 74, 75, 75, 76, 76, 79]

Middle two values (10th and 11th) are 71 and 72

Median = (71 + 72) / 2 = 71.5

Mode: Most frequent value in X

74 ppears 3 times, others less

Mode = 74

(b) What is the variance for Y?

Step 1: Calculate mean of Y (μ_Y)

Sum Y = 3284

Mean $\mu_Y = 3284 / 20 = 164.2$ (approx)

Step 2: Calculate squared differences $(y_i - \mu_Y)^2$

Examples:

$(153 - 164.2)^2 = 125.44$

$(175 - 164.2)^2 = 116.64$

Sum all squared differences = 27388.1 (approx)

Step 3: Variance formula:

$\sigma^2 = (1/(n - 1)) \sum (y_i - \mu_Y)^2$

Variance $\sigma^2 = 27388.1 / 19 \approx 1441.27$

(c) Plot the normal distribution for X.

Normal PDF formula:

$f(x) = (1/(\sigma \sqrt{2\pi})) \times \exp(-½ ((x - \mu) / \sigma)^2)$

Calculate standard deviation $\sigma_x = \sqrt{\text{variance}_x} = \sqrt{14.576} \approx 3.82$

(d) What is the probability of observing an age of 80 or higher?

Step 1: Calculate z-score:

$z = (80 - \mu) / \sigma$

$z = (80 - 71.45) / 3.82 \approx 2.23$

Step 2: Look up standard normal table or use function to find $P(Z \geq 2.23)$

$P \approx 0.0126$

(e) Find the 2-dimensional mean and the covariance matrix for these two variables.

Mean vector:

$\hat{\mu} = [\mu_X, \mu_Y] = [71.45, 164.7]$

Covariance matrix $\hat{\Sigma}$ calculation:

Covariance formula:

$\text{cov}(X,Y) = (1/(n - 1)) \sum (x_i - \mu_X)(y_i - \mu_Y)$

Calculate:

σ^2_X = variance of X = 14.57631579

σ^2_Y = variance of Y = 1441.27368421

cov(X,Y) = 128.87894737

Covariance matrix =

$$\begin{vmatrix} 14.5763 & 128.879 \\ 128.879 & 1441.274 \end{vmatrix}$$

(f) What is the correlation between age and weight?

Correlation coefficient ρ:

$\rho = \text{cov}(X,Y) / (\sigma_X \times \sigma_Y)$

$\sigma_X = \sqrt{14.5763} \approx 3.82$

$\sigma_Y = \sqrt{1441.274} \approx 37.95$

$\rho = 128.879 / (3.82 \times 37.95) \approx 0.889$

Strong positive correlation between age and weight

(3) Define a measure of deviation called mean absolute deviation for a random variable X as follows:
1/n Σ |x_i - μ| from i=1 to n
Is this measure robust? Why or why not?

The Mean Absolute Deviation (MAD), given by
MAD = $(1 / n)\ \Sigma\ |x_i - \mu|$ from i=1 to n,
measures the average distance between each data point and the mean μ. Compared to standard deviation, MAD is less sensitive to outliers because it uses absolute values instead of squaring the deviations. This makes it more resistant to the influence of extreme values. For example:

- Outliers affect MAD linearly, while in variance, the impact is quadratic.
- This linearity helps maintain stability when a dataset contains noisy or anomalous values.

However, MAD is not fully robust because it still relies on the mean, which can shift significantly due to outliers. When the mean moves, so does the MAD, reducing its reliability in heavily skewed or contaminated data. A more robust alternative is the Median Absolute Deviation, which replaces the mean with the median, offering:

- Better resistance to extreme values.
- Improved performance on non-normal or skewed distributions.

In summary, MAD provides a moderate level of robustness—better than variance, but not ideal for datasets with severe outliers. For high-resilience scenarios like anomaly detection, median-based MAD is preferred.

(4) Given the dataset in Table 2.2, compute the covariance matrix and the generalized variance.

Step 1: Compute mean vector μ

μ_j = (1/n) Σ x_ij   i=1 to n, n=3

μ = [
  (17 + 11 + 11)/3 = 13,
  (17 + 9 + 8)/3 = 11.3333,
  (12 + 13 + 19)/3 = 14.6667
]

Step 2: Center data Z = X - μ
Z =
[
  [4, 5.6667, -2.6667],
  [-2, -2.3333, -1.6667],
  [-2, -3.3333, 4.3333]
]

Step 3: Compute covariance matrix Σ
Σ = (1 / (n - 1)) * Z^T * Z

Elements:
Σ_11 = 12
Σ_22 = 24.3333
Σ_33 = 14.3333
Σ_12 = Σ_21 = 17
Σ_13 = Σ_31 = -8
Σ_23 = Σ_32 = -12.8333

Covariance matrix Σ =
```
[
  [12     17              -8     ]
  [17     24.33    -12.83 ]
  [-8     -12.83    14.33  ]
]
```

Step 4: Compute generalized variance = determinant of Σ
|Σ| ≈ 1.0110e-13 (near zero)

Interpretation: Low generalized variance indicates near collinearity or low effective dimensionality.

(5) Given the dataset below, assume X and Y are numeric and represent the entire population.
If the correlation between X and Y is zero, what can we infer about the values of Y?
Dataset:
X = [1, 0, 1, 0, 0]
Y = [a, b, c, a, c]

Step 1: Understand zero correlation
Correlation ρ between X and Y is zero means:
ρ = Cov(X, Y) / (σ_X * σ_Y) = 0
Since σ_X and σ_Y are positive (standard deviations), this implies:
Cov(X, Y) = 0
Covariance measures linear relationship; zero covariance => no linear dependence.

Step 2: Calculate means of X and Y
Mean of X (μ_X):
μ_X = (1 + 0 + 1 + 0 + 0) / 5 = 0.4

Mean of Y (μ_Y):
μ_Y = (a + b + c + a + c) / 5 = (2a + b + 2c) / 5

Step 3: Express covariance formula
Covariance formula for population:
Cov(X, Y) = (1/n) * Σ (X_i - μ_X) * (Y_i - μ_Y) for i=1 to n

Substituting values:
Cov(X, Y) =
(1/5) * [(1-0.4)(a - μ_Y) + (0-0.4)(b - μ_Y) + (1-0.4)(c - μ_Y)
+ (0-0.4)(a - μ_Y) + (0-0.4)(c - μ_Y)]

Step 4: Simplify covariance terms
Calculate each term:
(1-0.4) = 0.6
(0-0.4) = -0.4

Cov(X, Y) =
(1/5) * [0.6(a - μ_Y) - 0.4(b - μ_Y) + 0.6(c - μ_Y)
- 0.4(a - μ_Y) - 0.4(c - μ_Y)]

Group like terms:
Cov(X, Y) =
(1/5) * [(0.6 - 0.4)a + (-0.4)b + (0.6 - 0.4)c - 0.4μ_Y]

Cov(X, Y) =
(1/5) * [0.2a - 0.4b + 0.2c - 0.4μ_Y]

Step 5: Substitute μ_Y
Recall:
μ_Y = (2a + b + 2c) / 5

Therefore:
Cov(X, Y) =
(1/5) * [0.2a - 0.4b + 0.2c - 0.4 * ((2a + b + 2c) / 5)]

Multiply out:
Cov(X, Y) =
(1/5) * [0.2a - 0.4b + 0.2c - (0.08a + 0.08b + 0.16c)]

Simplify inside brackets:
Cov(X, Y) =
(1/5) * [(0.2 - 0.08)a + (-0.4 - 0.08)b + (0.2 - 0.16)c]

Cov(X, Y) =
(1/5) * [0.12a - 0.48b + 0.04c]

Step 6: Set covariance to zero for zero correlation condition

$0 = \mathrm{Cov}(X, Y) = (1/5) * (0.12a - 0.48b + 0.04c)$

Multiply both sides by 5:

$0 = 0.12a - 0.48b + 0.04c$

Multiply through by 25 to clear decimals:

$0 = 3a - 12b + c$

Conclusion:

For the correlation between X and Y to be zero, the values a, b, and c in Y must satisfy:
$3a - 12b + c = 0$

This means the values of Y are linearly related and constrained to satisfy this equation.
In other words, Y values cannot vary independently; their relationship with X balances out to produce zero linear correlation.

(6) Assume that we are given two univariate normal distributions, N_A and N_B, and let their mean and standard deviation be as follows: μ_A = 4, σ_A = 1 and μ_B = 8, σ_B = 2.

(a) For each of the following values x_i ∈ {5, 6, 7} find out which is the more likely normal distribution to have produced it.

Part (a): Likelihood comparison for x = 5, 6, 7 using Normal PDF formula

Given:

N_A ~ Normal(μ_A = 4, σ_A = 1)

N_B ~ Normal(μ_B = 8, σ_B = 2)

PDF formula:

$f(x|\mu, \sigma) = (1 / (sqrt(2 * \pi) * \sigma)) * exp(-(x - \mu)^2 / (2 * \sigma^2))$

Steps for each x:

(i) $x = 5$:

$P(N_A=5) = 1 / (sqrt(2 * \pi) * 1) * exp(- (5 - 4)^2 / (2 * 1^2))$

$= 1 / 2.5066 * exp(-0.5)$

≈ 0.24197

$P(N_B=5) = 1 / (sqrt(2 * \pi) * 2) * exp(- (5 - 8)^2 / (2 * 2^2))$

$= 1 / 5.0133 * exp(-1.125)$

≈ 0.06476

Conclusion: $P(N_A=5) > P(N_B=5) \Rightarrow x=5$ more likely from N_A

(ii) $x = 6$:

$P(N_A=6) = 1 / (sqrt(2 * \pi) * 1) * exp(- (6 - 4)^2 / (2 * 1^2))$

$= 1 / 2.5066 * exp(-2)$

≈ 0.05399

$P(N_B=6) = 1 / (sqrt(2 * \pi) * 2) * exp(- (6 - 8)^2 / (2 * 2^2))$

$= 1 / 5.0133 * exp(-0.5)$

≈ 0.12098

Conclusion: P(N_B=6) > P(N_A=6) ⇒ x=6 more likely from N_B

(iii) x = 7:

P(N_A=7) = 1 / (sqrt(2 * π) * 1) * exp(- (7 - 4)^2 / (2 * 1^2))

= 1 / 2.5066 * exp(-4.5)

≈ 0.00443

P(N_B=7) = 1 / (sqrt(2 * π) * 2) * exp(- (7 - 8)^2 / (2 * 2^2))

= 1 / 5.0133 * exp(-0.125)

≈ 0.17603

Conclusion: P(N_B=7) > P(N_A=7) ⇒ x=7 more likely from N_B

(b) Derive an expression for the point for which the probability of having been produced by both the normals is the same.

Part (b): Find x such that P_NA(x) = P_NB(x)

Given:

N_A ~ Normal(μ_A = 4, σ_A = 1)

N_B ~ Normal(μ_B = 8, σ_B = 2)

We want to find x where the PDFs are equal:

PDF formula for Normal distribution:

f(x|μ, σ) = (1 / (σ * sqrt(2π))) * exp(- (x - μ)^2 / (2σ^2))

Equate the two PDFs:

(1 / (1 * sqrt(2π))) * exp(- (x - 4)^2 / 2) =

(1 / (2 * sqrt(2π))) * exp(- (x - 8)^2 / 8)

Simplify by multiplying both sides by sqrt(2π):

(1 / 1) * exp(- (x - 4)^2 / 2) = (1 / 2) * exp(- (x - 8)^2 / 8)

Multiply both sides by 2:

2 * exp(- (x - 4)^2 / 2) = exp(- (x - 8)^2 / 8)

Take natural logarithm (ln) on both sides:

ln(2) + [- (x - 4)^2 / 2] = - (x - 8)^2 / 8

Rearrange:

ln(2) = (x - 4)^2 / 2 - (x - 8)^2 / 8

Multiply both sides by 8 to clear denominators:

8 * ln(2) = 4 * (x - 4)^2 - (x - 8)^2

Expand squares:

8 * ln(2) = 4 * (x^2 - 8x + 16) - (x^2 - 16x + 64)

= 4x^2 - 32x + 64 - x^2 + 16x - 64

= 3x^2 - 16x + 0

So,

3x^2 - 16x - 8 * ln(2) = 0

Substitute ln(2) ≈ 0.6931:

3x^2 - 16x - 5.544 = 0

Solve quadratic equation ax^2 + bx + c = 0:

a = 3

b = -16

c = -5.544

Discriminant Δ = b^2 - 4ac

= (-16)^2 - 4 * 3 * (-5.544)

= 256 + 66.528

= 322.528

Roots:

x = [16 ± sqrt(322.528)] / (2 * 3)

sqrt(322.528) ≈ 17.96

So,

$x_1 = (16 + 17.96) / 6 \approx 33.96 / 6 \approx 5.66$

$x_2 = (16 - 17.96) / 6 \approx -1.96 / 6 \approx -0.33$

Final result:

The PDFs of N_A and N_B are equal at approximately $x = -0.33$ and $x = 5.66$

(7) Under what conditions will the covariance matrix Σ be identical to the correlation matrix, whose (i, j) entry gives the correlation between attributes X_i and X_j? What can you conclude about the two variables?

The covariance matrix Σ consists of elements $\sigma_{ij} = \text{cov}(X_i, X_j)$, representing the covariance between attributes X_i and X_j. The correlation matrix R contains elements $\rho_{ij} = \text{cov}(X_i, X_j) / (\sigma_i \sigma_j)$, which normalize covariance by the product of the standard deviations of X_i and X_j, denoted σ_i and σ_j respectively.

For Σ to be identical to R, the following must hold for all i, j:

$$\sigma_{ij} = \rho_{ij}$$

Substituting the definition of correlation, this implies:

$$\text{cov}(X_i, X_j) = \text{cov}(X_i, X_j) / (\sigma_i \sigma_j)$$

Which is only true if:

$$\sigma_i \sigma_j = 1$$

Since σ_i and σ_j are positive standard deviations, this condition means:

$$\sigma_i = 1 \quad \text{for all } i$$

This indicates that each variable must have unit standard deviation.

Key points:

- Standardization (also called normalization) transforms each variable X_i to have mean zero and standard deviation one.

- After standardization, covariance matrix equals correlation matrix because the scale factors (standard deviations) are 1.

- This means the covariance matrix directly expresses the linear relationships without being affected by different scales or units.

- The equality Σ = R only holds for datasets where all variables are standardized.

- If variables are not standardized, the covariance matrix and correlation matrix generally differ because covariance depends on the scale of measurement.

Conclusion:

When the covariance matrix is identical to the correlation matrix, it implies all variables have been standardized to unit variance. This standardization is essential for meaningful comparison of variables measured in different units or scales, ensuring that the covariance matrix purely reflects correlation structure rather than scale differences.

1.3 Chapter 3

(1) Proof: Cosine Similarity Range for Categorical Vectors

Consider two vectors $x, y \in \{0,1\}^d$ representing categorical data encoded using one-hot encoding. Each vector has exactly one component equal to 1 per categorical attribute group.

Step 1: Dot Product

The dot product is defined as

$x^{\mathrm{T}}y \ = \sum_{i=1}^{d} xi\ yi = \ s_{\mathrm{i}}$

where s is the count of matching positions with 1s in both vectors. Since each component is binary, s ∈ {0, 1, ..., d}.

Step 2: Vector Norms

Because each vector contains exactly one 1 per group and zeros elsewhere,

$||x|| = \sqrt{(\sum_{i=1}^{d} x_i^2)} = \sqrt{d}$

$||y|| = \sqrt{d}$

Step 3: Cosine Similarity

Cosine similarity is given by

$\cos(\theta) = (x^{T}y) / (||x||\ ||y||) = s / (\sqrt{d} * \sqrt{d}) = s / d$

Step 4: Bounds on Cosine Similarity

Since s ranges between 0 and d,

$\cos(\theta) \in [0, 1]$

Step 5: Range of the Angle

Because cosine values are between 0 and 1, the corresponding angle θ lies within

$\theta \in [0°, 90°]$

Final conclusion:

$\cos(\theta) \in [0, 1]$ and $\theta \in [0°, 90°]$

This confirms that cosine similarity between one-hot encoded categorical vectors is always non-negative, ranging from 0 (orthogonal) to 1 (identical), and the angle between such vectors lies between 0° and 90°.

(2)

(a) What is the mean vector for this dataset?

Mean vector

Step 1: Compute the mean of X_1

Mean of X_1 = (0.30 – 0.30 + 0.44 – 0.60 + 0.40 + 1.20 – 0.12 – 1.60 + 1.60 – 1.32) / 10 = 0.00

Step 2: One-hot encode X_2

Let X_2a = 1 if X_2 = a, else 0

Let X_2b = 1 if X_2 = b, else 0

Count of a = 6 ⇒ mean of X_2a = 6 / 10 = 0.6

Count of b = 4 ⇒ mean of X_2b = 4 / 10 = 0.4

Mean vector μ = [0.00, 0.6, 0.4]

(b) What is the covariance matrix?

Covariance matrix

Use the formula

$S = (1 / (n - 1)) \times \Sigma (x_i - \mu)(x_i - \mu)^T$

Centered data rows =

(0.30, 0.4, –0.4)

(–0.30, –0.6, 0.6)

(0.44, 0.4, –0.4)

(–0.60, 0.4, –0.4)

(0.40, 0.4, –0.4)

(1.20, –0.6, 0.6)
(–0.12, 0.4, –0.4)
(–1.60, –0.6, 0.6)
(1.60, –0.6, 0.6)
(–1.32, 0.4, –0.4)

Compute entries:
S_{11} = variance of X_1
= $(1 / 9) \times \Sigma(x_{1i})^2 \approx 1.0234$

S_{22} = variance of X_2a
= $(1 / 9) \times (6\times0.4^2 + 4\times(-0.6)^2) = 0.2667$

S_{33} = variance of X_2b = 0.2667

S_{12} = covariance(X_1, X_2a) = $(1 / 9) \times \Sigma(x_{1i}\times x_2a_i) \approx -0.1000$

S_{13} = covariance(X_1, X_2b) = $(1 / 9) \times \Sigma(x_{1i}\times x_2b_i) \approx 0.1000$

S_{23} = covariance(X_2a, X_2b) = $(1 / 9) \times \Sigma(x_2a_i\times x_2b_i) = -0.2667$

Final covariance matrix:
[1.0234 –0.1000 0.1000]
[–0.1000 0.2667 –0.2667]
[0.1000 –0.2667 0.2667]

1.4 Chapter 6

(1) Given the gamma function in Eq. (6.12), show the following

(a) $\Gamma(1) = 1$

Given,

$\Gamma(\alpha) = \int x^{(\alpha-1)} e^{-x}\, dx$ from 0 to ∞, (a>0) Equation 6.12)

And the volume of a d-dimensional hypersphere of radius r is:

$\mathrm{vol}(S_n(r)) = K_n \cdot r^n = (\pi^{(n/2)} / \Gamma(n/2 + 1)) \cdot r^n$

Where $K_n = \pi^{(n / 2)} / \Gamma\,(n/2 + 1)$ and Γ is the gamma function.

Substitute α = 1 into the definition of the gamma function:

$\Gamma(1) = \int x^{(1-1)} e^{-x}\, dx$, from 0 to ∞

$= \int x^{0} e^{-x}\, dx$ from 0 to ∞

$= \int e^{-x}\, dx$ from 0 to ∞

The integral of e^{-x} is $-e^{-x}$, so:

$\Gamma(1) = [-e^{-x}]_0{}^{\wedge}\infty = [-e^{-\infty}] - [-e^{0}] = 0 - (-1) = 1$

Hence, Γ(1) = 1

(b) $\Gamma(½) = \sqrt{\pi}$

Substitute α = ½ into the eqn:

$\Gamma(½) = \int_0{}^{\wedge}\infty\, x^{(-½)} e^{-x}\, dx$

This is a standard result related to the Gaussian integral:

$\int_{-\infty}{}^{\wedge}\infty\, e^{-x^2}\, dx = \sqrt{\pi}$

Using substitution $x = \sqrt{t}$, $dx = dt/(2\sqrt{t})$, the integral transforms accordingly. Therefore:

$\Gamma(½) = \sqrt{\pi}$

(c) $\Gamma(\alpha) = (\alpha - 1) \cdot \Gamma(\alpha - 1)$

Assume α > 1. Use integration by parts:

Let $u = x^{(\alpha-1)}$, $dv = e^{-x}\, dx$

Then $du = (\alpha-1)\, x^{(\alpha-2)}\, dx$, $v = -e^{-x}$

Using $\int u\, dv = uv - \int v\, du$:

$\Gamma(\alpha) = [x^{(\alpha-1)} (-e^{-x})]_0{}^{\wedge}\infty - \int_0{}^{\wedge}\infty\, (-e^{-x})\, (\alpha-1)\, x^{(\alpha-2)}\, dx$

Boundary term:
- As x → ∞, $x^{(\alpha-1)} e^{-x}$ → 0 (since e^{-x} decays faster than $x^{(\alpha-1)}$ grows)
- As x → 0, $x^{(\alpha-1)}$ → 0 (since α−1 > −1)

Therefore, boundary term is zero.

Thus:

$\Gamma(\alpha) = (\alpha - 1) \int_0\text{^}\infty\, x^{(\alpha-2)} e^{-x}\, dx$
$\quad = (\alpha - 1)\, \Gamma(\alpha - 1)$

This holds for α > 1, and by induction for all α > 0.

(2) Show that the asymptotic volume of the hypersphere Sd (r) for any value of radius r eventually tends to zero as d increases.
The volume of a d-dimensional hypersphere is given by:
vol(S_d(r)) = (π^(d/2) / Γ(d/2 + 1)) * r^d

As d increases, we need to analyze the behavior of this expression. The gamma function Γ(d/2 + 1) grows rapidly with d because:
- Γ(n) ≈ (n-1)! for large integer n, and d/2 + 1 increases with d.
- For large d, Γ(d/2 + 1) can be approximated using Stirling's approximation:
Γ(n) ≈ √(2πn) (n/e)^n

Consider the ratio π^(d/2) / Γ(d/2 + 1):
- π^(d/2) grows exponentially with d/2.
- Γ(d/2 + 1) grows factorially with d/2 + 1, which outpaces exponential growth for large d.

Using Stirling's approximation for Γ(d/2 + 1) where n = d/2 + 1:
Γ(d/2 + 1) ≈ √(2π(d/2 + 1)) * ((d/2 + 1)/e)^(d/2 + 1)

The volume vol(S_d(r)) = (π^(d/2) / Γ(d/2 + 1)) * r^d
As d → ∞, the denominator Γ(d/2 + 1) grows much faster than π^(d/2), causing the fraction π^(d/2) / Γ(d/2 + 1) to approach 0.
Thus, vol(S_d(r)) → 0 as d increases, regardless of r, since r^d is dominated by the decaying coefficient.s

This confirms that the asymptotic volume tends to zero as d increases

1.5 Chapter 7

(1)

(a) Compute the mean μ and covariance matrix Σ for D.

Formula: $\mu = (1/n) \Sigma x_i$ where n = 5

$\mu_1 = (8 + 0 + 10 + 10 + 2) / 5 = 6$

$\mu_2 = (-20 + (-1) + (-19) + (-20) + 0) / 5 = -12$

$\mu = [6, -12]$

Formula: $\Sigma = (1/(n-1)) \Sigma (x_i - \mu)(x_i - \mu)^T$

Centered vectors:

$[2, -8], [-6, 11], [4, -7], [4, -8], [-4, 12]$

$\Sigma_{11} = (1/4)(4 + 36 + 16 + 16 + 16) = 22$

$\Sigma_{22} = (1/4)(64 + 121 + 49 + 64 + 144) = 110.5$

$\Sigma_{12} = \Sigma_{21} = (1/4)(-16 - 66 - 28 - 32 - 48) = -47.5$

Covariance matrix Σ:

| 22 –47.5 |

|–47.5 110.5|

(b) Compute the eigenvalues of Σ.

Formula: $\det(\Sigma - \lambda I) = 0$

Solve: $(22 - \lambda)(110.5 - \lambda) - (-47.5)^2 = 0$

$\rightarrow \lambda^2 - 132.5\lambda + (2431 - 2256.25) = 0$

$\rightarrow \lambda^2 - 132.5\lambda + 174.75 = 0$

Use quadratic formula:

$\lambda = [132.5 \pm \sqrt{}(132.5^2 - 4\times174.75)] / 2$

$\lambda \approx [132.5 \pm 129.81] / 2$

Eigenvalues: $\lambda_1 \approx 131.155$ $\lambda_2 \approx 1.345$

(c) What is the "intrinsic" dimensionality of this dataset (discounting some small amount of variance)?

The intrinsic dimensionality is the number of significant eigenvalues, discounting those contributing a small amount of variance (e.g., < 1% of total variance).

Total variance = λ1 + λ2 = 131.155 + 1.345 = 132.5

Proportion of λ2 = 1.345 / 132.5 ≈ 0.01015 (1.015%)

If we discount variance below 1%, only λ1 is significant.

Thus, the intrinsic dimensionality is 1.

(d) Compute the first principal component.

Solve: $(\Sigma - \lambda_1 I)v = 0$

Matrix:

$$\begin{vmatrix} -109.155 & -47.5 \\ -47.5 & -20.655 \end{vmatrix}$$

Equation 1: $-109.155v_1 - 47.5v_2 = 0$

$\rightarrow v_2 = -109.155/47.5 \times v_1 \approx -2.297v_1$

Normalize:

$v = [v_1, -2.297v_1]$

Length $= \sqrt{}(v_1^2 + (-2.297v_1)^2) = 1$

$\rightarrow v_1 \approx 0.399, \quad v_2 \approx -0.917$

First principal component: [0.399, –0.917]

(e) If the μ and Σ from above characterize the normal distribution from which the points were generated, sketch the orientation/extent of the 2-dimensional normal density function.

The normal density function is f(x) = (1 / (2π |Σ|^(1/2))) exp(-1/2 (x - μ)^T Σ^(-1) (x - μ)). The mean μ = [6, -12] centers the distribution. The covariance matrix Σ = 22 -47.5 -47.5 110.5 indicates a strong negative correlation (-47.5), suggesting the major axis (first principal component [0.399, -0.917]) is tilted.

The extent is elongated along this axis ($\lambda1 = 131.155$) and narrow along the minor axis ($\lambda2 = 1.345$).

Ellipse tilted, with the long axis sloping downward to the right.

Citations:

(1) Zaki, M. J., & Meira Jr, W. (2020). Data Mining and Machine Learning: Fundamental Concepts and Algorithms, 2nd Edition.

(2)https://en.wikipedia.org/wiki/Principal_component_analysis

(3)https://medium.com/@rgalvg/from-physics-to-data-science-the-beauty-and-power-of-cosine-similarity-f23e276afe29

(4) Assistance was taken from a large language model (ChatGPT by OpenAI) to understand concepts related to Principal Component Analysis (PCA) and for help in identifying relevant references.

(5) An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani