

Data Science for Mathematicians

Exercises 1: The Data Science Landscape & The Geometry of Data

Instructions

Answer all exercises completely. Show all working, justify your answers, and state any assumptions you make. For computational exercises, carry out all intermediate steps explicitly. For proof exercises, clearly identify which definitions and theorems you are applying.

Exercises

Exercise 1. Consider the following two software systems used by a credit card company. Based on the distinctions provided in the lecture (Primary Goal, System Type, Mathematical Basis), determine which system constitutes “Data Science” and which does not. Justify your answer.

- (a) **System A:** A database trigger that automatically declines any transaction exceeding \$5,000 if the user resides in a specific list of countries.
- (b) **System B:** A fraud detection engine that analyzes transaction history, location, and time-of-day to generate a “suspicion score” for every swipe, flagging those in the top 1% of scores for review.

Exercise 2. You are analyzing a dataset of 3 used cars ($n = 3$). The features recorded are: Age (years), Mileage (thousands of miles), and Price (thousands of dollars).

- Car 1: 2 years old, 15k miles, \$20k.
- Car 2: 5 years old, 60k miles, \$12k.
- Car 3: 10 years old, 120k miles, \$5k.

- (a) Construct the Data Matrix $X \in \mathbb{R}^{3 \times 3}$.
- (b) Identify the vector $X_{:,2}$ and explain its semantic meaning in this context.
- (c) If we visualize this data as a point cloud, what is the dimensionality of the feature space?

Exercise 3. Two readers, User A and User B, rate two book genres: *Sci-Fi* and *Romance*. Their ratings are centered around their personal averages (positive means liked, negative means disliked).

$$u_A = \begin{bmatrix} 4 \\ -2 \end{bmatrix}, \quad v_B = \begin{bmatrix} -3 \\ 5 \end{bmatrix}$$

- (a) Calculate the dot product $u_A \cdot v_B$.
- (b) Based on the sign of the result, are the tastes of User A and User B aligned, opposing, or unrelated?
- (c) Explain the geometric relationship (angle θ) between these two vectors.

Exercise 4. Let x be a vector representing study hours and y be a vector representing exam scores for $n = 3$ students.

$$x = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}, \quad y = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$$

- (a) Compute the sample means \bar{x} and \bar{y} .
- (b) Construct the mean-centered vectors \tilde{x} and \tilde{y} .
- (c) Calculate the sample covariance $\text{Cov}(x, y)$ using the dot product of the centered vectors (as derived in the lecture).

Exercise 5. A delivery robot operates on a grid-based city map. Its current location is at coordinates $p = [2, 1]^T$ and the destination is $q = [5, 5]^T$.

- (a) Calculate the Euclidean distance (L_2) between p and q .
- (b) Calculate the Manhattan distance (L_1) between p and q .
- (c) Which metric accurately reflects the distance the robot must travel if it cannot move diagonally through city blocks?

Exercise 6. Consider the feature vector $w = [3, -4, 0]^T$.

- (a) Calculate $\|w\|_1$, $\|w\|_2$, and $\|w\|_\infty$.
- (b) Which norm gives the largest value and which gives the smallest? Is this ordering consistent with the general properties of L_p norms?

Exercise 7. Consider the point $v = [0.7, 0.6]^T$ in \mathbb{R}^2 .

- (a) Determine if v lies inside the L_2 unit ball ($B_2 = \{x : \|x\|_2 \leq 1\}$).
- (b) Determine if v lies inside the L_1 unit ball ($B_1 = \{x : \|x\|_1 \leq 1\}$).
- (c) Explain how the shape of the L_1 unit ball facilitates feature selection (forcing coefficients to zero) in optimization problems compared to the L_2 ball.

Exercise 8. In an image recognition task, you are working with 20×20 pixel grayscale images.

- (a) To apply linear algebra methods, these images are flattened. What is the dimension p of the resulting observation vector?
- (b) If you have a dataset of 500 such images, what are the dimensions of the Data Matrix X ?
- (c) If the dot product between two flattened image vectors is exactly 0, what does this imply about the pixels in the two images (assuming pixel values are non-negative)?

Exercise 9. Given two vectors $a = [1, 2, 3]^T$ and $b = [1, 0, -1]^T$:

- (a) Compute the Euclidean norms $\|a\|_2$ and $\|b\|_2$.
- (b) Compute the dot product $a \cdot b$.
- (c) Using the geometric interpretation of the dot product, calculate the cosine of the angle θ between a and b .
- (d) Is the angle acute, obtuse, or orthogonal?

Exercise 10. Suppose you have two centered feature vectors \tilde{u} and \tilde{v} such that $\tilde{u} \cdot \tilde{v} = 0$.

- (a) What is the sample covariance $\text{Cov}(u, v)$?
- (b) What statistical relationship corresponds to the geometric concept of orthogonality in the centered feature space?

Solutions

Solution 1. Classifying Data Science Systems

- (a) **System A is NOT Data Science.** It is a deterministic system executing a known rule (Boolean logic). There is no probabilistic modeling or extraction of unknown insights; the outcome is fixed and certain based on the input.
- (b) **System B IS Data Science.** It models a stochastic system (fraud is uncertain and complex). It uses statistical inference to assign a probability (suspicion score) and extracts insight from historical patterns rather than following a hard-coded rule.

Solution 2. The Data Matrix and Feature Space

- (a) The matrix is formed by stacking the transposes of the observations:

$$X = \begin{bmatrix} 2 & 15 & 20 \\ 5 & 60 & 12 \\ 10 & 120 & 5 \end{bmatrix}$$

- (b) The vector $X_{:,2} = [15, 60, 120]^T$. This column vector represents the **Mileage** feature for the entire collection of cars.
- (c) The feature space is \mathbb{R}^3 , corresponding to the 3 features (Age, Mileage, Price).

Solution 3. Dot Product as Similarity

- (a) $u_A \cdot v_B = (4)(-3) + (-2)(5) = -12 - 10 = -22$.
- (b) The tastes are **opposing**. The negative result indicates a negative correlation (User A likes what User B dislikes).
- (c) Since the dot product is negative, $\cos \theta < 0$, meaning the angle θ is **obtuse** (between 90° and 180°).

Solution 4. Covariance Calculation

(a) Means: $\bar{x} = \frac{1+3+5}{3} = 3$. $\bar{y} = \frac{2+4+6}{3} = 4$.

(b) Mean-centered vectors:

$$\tilde{x} = \begin{bmatrix} 1-3 \\ 3-3 \\ 5-3 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}, \quad \tilde{y} = \begin{bmatrix} 2-4 \\ 4-4 \\ 6-4 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}$$

(c) Using the formula $\text{Cov}(x, y) = \frac{1}{n-1}(\tilde{x} \cdot \tilde{y})$ with $n = 3$:

$$\tilde{x} \cdot \tilde{y} = (-2)(-2) + (0)(0) + (2)(2) = 4 + 0 + 4 = 8$$

$$\text{Cov}(x, y) = \frac{8}{3-1} = \frac{8}{2} = 4$$

Solution 5. Distance Metrics (L1 vs L2)

- (a) Euclidean (L_2): $\sqrt{(5-2)^2 + (5-1)^2} = \sqrt{3^2 + 4^2} = \sqrt{9+16} = \sqrt{25} = 5$.

- (b) Manhattan (L_1): $|5 - 2| + |5 - 1| = |3| + |4| = 3 + 4 = 7$.
- (c) The **Manhattan distance** (L_1) accurately reflects the travel distance for a grid-constrained robot.

Solution 6. Comparing Norms

- (a) $L_1 = |3| + |-4| + |0| = 7$.
 $L_2 = \sqrt{3^2 + (-4)^2 + 0^2} = \sqrt{25} = 5$.
 $L_\infty = \max(|3|, |-4|, |0|) = 4$.
- (b) The L_1 norm is the largest (7) and the L_∞ norm is the smallest (4). Yes, this is consistent with the general property $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$.

Solution 7. The Geometry of Sparsity

- (a) $\|v\|_2 = \sqrt{0.7^2 + 0.6^2} = \sqrt{0.49 + 0.36} = \sqrt{0.85} \approx 0.92$. Since $0.92 \leq 1$, v is **inside** the L_2 ball.
- (b) $\|v\|_1 = |0.7| + |0.6| = 1.3$. Since $1.3 > 1$, v is **outside** the L_1 ball.
- (c) The L_1 ball is a polytope (diamond in 2D) with “corners” on the axes. In optimization, the error contours are statistically more likely to touch these corners first, where one coordinate is exactly zero, thus inducing sparsity. The L_2 ball is smooth, so solutions rarely land on an axis.

Solution 8. High-Dimensional Operations

- (a) $p = 20 \times 20 = 400$ dimensions.
- (b) $X \in \mathbb{R}^{500 \times 400}$ (500 rows for images, 400 columns for pixels).
- (c) Since pixel values are non-negative, a dot product of 0 implies that for every position where one image has a non-zero pixel, the other implies must have a zero (black) pixel. They share **no overlapping active pixels**.

Solution 9. Calculating Angles in High Dimensions

- (a) $\|a\|_2 = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{1 + 4 + 9} = \sqrt{14} \approx 3.74$.
 $\|b\|_2 = \sqrt{1^2 + 0^2 + (-1)^2} = \sqrt{1 + 0 + 1} = \sqrt{2} \approx 1.41$.
- (b) $a \cdot b = (1)(1) + (2)(0) + (3)(-1) = 1 + 0 - 3 = -2$.
- (c) Using $\cos \theta = \frac{a \cdot b}{\|a\|_2 \|b\|_2}$:

$$\cos \theta = \frac{-2}{\sqrt{14}\sqrt{2}} = \frac{-2}{\sqrt{28}} = \frac{-2}{2\sqrt{7}} = \frac{-1}{\sqrt{7}} \approx -0.378$$
- (d) Since $\cos \theta < 0$ and $\cos \theta \neq -1$, the angle is **obtuse**.

Solution 10. Orthogonality and Independence

- (a) Since $\text{Cov}(u, v) \propto (\tilde{u} \cdot \tilde{v})$, if the dot product is 0, the covariance is **0**.
- (b) Orthogonality in the centered feature space corresponds to the features being **uncorrelated** (linearly independent).