Solution 1

Solution 1 (a)

Define ℓ_1

The ℓ_1 or $||x||_1$ is defined as:

$$\ell_1 = ||x||_1 = \sum_{i=1}^{d} |x_i|$$

Compute ℓ_1

Let $x = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}$

$$\begin{aligned} ||x||_1 &= \sum_{i=1}^{3} |x_i| \\ &= |x_1| + |x_2| + |x_3| \\ &= |1| + |-2| + |3| \\ &= 1 + 2 + 3 \\ &= 6 \end{aligned}$$

$\therefore ||x||_1 = 6$

Solution 1

Solution 1 (b)

Define ℓ_2

The ℓ_2 or $||x||_2$ is defined as:

$$\ell_2 = ||x||_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$$

Compute ℓ_2

Let $x = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}$

$$\begin{aligned} ||x||_2 &= \sqrt{\sum_{i=1}^{3} x_i^2} \\ &= \sqrt{x_1^2 + x_2^2 + x_3^2} \\ &= \sqrt{1^2 + (-2)^2 + 3^2} \\ &= \sqrt{1 + 4 + 9} \\ &= \sqrt{14} \end{aligned}$$

$\therefore ||x||_2 = \sqrt{14}$

Solution 1

Solution 1 (c)

Define ℓ_∞

The ℓ_∞ or $||x||_\infty$ is defined as:

$$\ell_\infty = ||x||_\infty = \max_i |x_i|$$

Compute ℓ_∞

Let $x = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}$

$$\begin{aligned} ||x||_\infty &= \max(\{|x_1|, |x_2|, |x_3|\}) \\ &= \max(\{|1|, |-2|, |3|\}) \\ &= \max(\{1, 2, 3\}) \\ &= 3 \end{aligned}$$

$\therefore ||x||_\infty = 3$

Solution 2

Solution 2 (a)

Define ℓ_2 distance

The ℓ_2 distance is defined as:

$$d(x, x')_{\ell_2} = \sqrt{\sum_{i=1}^{n} (x_i - x'_i)^2}$$

Compute ℓ_2 distance

Let $x = \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}$ and $x' = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

$$\begin{aligned} d(x, x')_{\ell_2} &= \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2 + (x_3 - x'_3)^2 + (x_4 - x'_4)^2} \\ &= \sqrt{((-1) - 1)^2 + (1 - 1)^2 + ((-1) - 1)^2 + (1 - 1)^2} \\ &= \sqrt{(2)^2 + (0)^2 + (2)^2 + (0)^2} \\ &= \sqrt{4 + 0 + 4 + 0} \\ &= \sqrt{8} \end{aligned}$$

$\therefore d(x, x')_{\ell_2} = \sqrt{8}$

Solution 2

Solution 2 (b)

Define ℓ_1 distance

The ℓ_1 distance is defined as:

$$d(x, x')_{\ell_1} = \sum_{i=1}^{n} |x_i - x'_i|$$

Compute ℓ_1 distance

Let $x = \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}$ and $x' = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

$$\begin{aligned} d(x, x')_{\ell_1} &= |x_1 - x'_1| + |x_2 - x'_2| + |x_3 - x'_3| + |x_4 - x'_4| \\ &= |-1 - 1| + |1 - 1| + |-1 - 1| + |1 - 1| \\ &= |-2| + |0| + |-2| + |0| \\ &= 2 + 0 + 2 + 0 \\ &= 4 \end{aligned}$$

$\therefore d(x, x')_{\ell_1} = 4$

Solution 2

Solution 2 (c)

Define ℓ_∞ distance

The ℓ_∞ distance is defined as:

$$d(x, x')_{\ell_\infty} = \max_i |x_i - x'_i|$$

Compute ℓ_∞ distance

Let $x = \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}$ and $x' = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

$$\begin{aligned} d(x, x')_{\ell_\infty} &= \max\{|x_1 - x'_1|, |x_2 - x'_2|, |x_3 - x'_3|, |x_4 - x'_4|\} \\ &= \max\{|-1-1|, |1-1|, |-1-1|, |1-1|\} \\ &= \max\{|-2|, |0|, |-2|, |0|\} \\ &= \max\{2, 0, 2, 0\} \\ &= 2 \end{aligned}$$

$\therefore d(x, x')_{\ell_\infty} = 2$

Solution 3

Solution 3 (a)

(i): Maximize $\|x\|_1$ given $\|x\|_\infty = 1$

We are given the constraint $\|x\|_\infty = \max_i |x_i| = 1$. This implies that $|x_i| \leq 1$ for all components $i = 1, \ldots, d$. The objective is to maximize the ℓ_1-norm, which is the sum of the absolute values of the components, $\|x\|_1 = \sum_{i=1}^{d} |x_i|$. To make this sum as large as possible, each term $|x_i|$ in the sum must be maximized. The constraint allows each $|x_i|$ to be at most 1. Therefore, the maximum is achieved when $|x_i| = 1$ for all i.

$$\begin{aligned} \|x\|_1 &= \sum_{i=1}^{d} |x_i| \\ &\leq \sum_{i=1}^{d} 1 \quad (\text{since } |x_i| \leq \|x\|_\infty = 1) \\ &\leq d \end{aligned}$$

This maximum value is achieved by any vector x where all components are either $+1$ or -1.

$\therefore$ The vector $x = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$ maximizes the norm, with a value of $\|x\|_1 = d$.

(ii): Maximize $\|x\|_2$ given $\|x\|_\infty = 1$

Given the same constraint $\|x\|_\infty = 1$, which implies $x_i^2 \leq 1$ for all i, we seek to maximize the ℓ_2-norm, $\|x\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$. This is equivalent to maximizing the squared norm, $\|x\|_2^2 = \sum_{i=1}^{d} x_i^2$. To maximize this sum of squares, each term x_i^2 must be maximized. Given the constraint, the maximum value for any x_i^2 is $1^2 = 1$. This occurs when $|x_i| = 1$ for all i.

$$\begin{aligned} \|x\|_2^2 &= \sum_{i=1}^{d} x_i^2 \\ &\leq \sum_{i=1}^{d} 1^2 \quad (\text{since } x_i^2 \leq \|x\|_\infty^2 = 1) \\ &\leq d \\ \implies \|x\|_2 &\leq \sqrt{d} \end{aligned}$$

This maximum is achieved by the same set of vectors as in the previous case, where all components are ± 1.

$\therefore$ The vector $x = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$ maximizes the norm, with a value of $\|x\|_2 = \sqrt{d}$.

Solution 3

Solution 3 (b)

(i): Maximize $\|x\|_1$ given $\|x\|_2 = 1$

We are given the constraint $\|x\|_2 = 1$, or $\sum_{i=1}^{d} x_i^2 = 1$. We seek to maximize $\|x\|_1 = \sum_{i=1}^{d} |x_i|$. By the Cauchy-Schwarz inequality, for vectors $u = (|x_1|, \ldots, |x_d|)$ and $v = (1, \ldots, 1)$, we have $(\sum |x_i|)^2 \leq (\sum |x_i|^2)(\sum 1^2)$. This is precisely $(\|x\|_1)^2 \leq (\|x\|_2^2)(d)$.

$$
\begin{aligned}
(\|x\|_1)^2 &= \left(\sum_{i=1}^{d} |x_i|\right)^2 \\
&\leq \left(\sum_{i=1}^{d} x_i^2\right)\left(\sum_{i=1}^{d} 1^2\right) \quad \text{(Cauchy-Schwarz Inequality)} \\
&\leq (1)(d) \\
\implies \|x\|_1 &\leq \sqrt{d}
\end{aligned}
$$

Equality holds when the vectors are linearly dependent, meaning $|x_1| = |x_2| = \cdots = |x_d| = c$. The constraint $\sum x_i^2 = 1$ implies $d \cdot c^2 = 1$, so $c = 1/\sqrt{d}$.

$\therefore$ The vector $x = \begin{bmatrix} \pm 1/\sqrt{d} \\ \vdots \\ \pm 1/\sqrt{d} \end{bmatrix}$ maximizes the norm, with a value of $\|x\|_1 = \sqrt{d}$.

(ii): Maximize $\|x\|_\infty$ given $\|x\|_2 = 1$

Given the constraint $\sum_{i=1}^{d} x_i^2 = 1$, we seek to maximize $\|x\|_\infty = \max_i |x_i|$. Let $|x_k|$ be the component with the maximum absolute value. From the constraint, we can write $x_k^2 + \sum_{i \neq k} x_i^2 = 1$. Since the sum of squares is non-negative, it must be that $x_k^2 \leq 1$, which implies $|x_k| = \|x\|_\infty \leq 1$.

$$
\begin{aligned}
\|x\|_\infty^2 &= (\max_i |x_i|)^2 \\
&= \max_i (x_i^2) \\
&\leq \sum_{i=1}^{d} x_i^2 \quad \text{(as all terms are non-negative)} \\
&\leq 1 \\
\implies \|x\|_\infty &\leq 1
\end{aligned}
$$

This maximum value of 1 is achieved when one component has a magnitude of 1, which forces all other components to be zero to satisfy the constraint.

$\therefore$ Any standard basis vector e_k (e.g., $x = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$) maximizes the norm, giving $\|x\|_\infty = 1$.

Graduate Level Explanation This exercise demonstrates the geometric relationship between the unit balls of different ℓ_p norms. The task of maximizing one norm subject to a constraint on another is equivalent to finding the point on the surface of one unit ball that is "farthest" from the origin as measured by the other norm's metric. The results confirm the well-known norm inequalities for $x \in \mathbb{R}^d$:

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq \sqrt{d}\|x\|_2 \leq d\|x\|_\infty$$

Our findings show that these bounds are tight:

- When constrained to the ℓ_∞ unit ball (a hypercube), the points that maximize the ℓ_1 and ℓ_2 norms are the corners of the hypercube, such as the vector of all ones. At these points, the inequalities $\|x\|_1 \leq d\|x\|_\infty$ and $\|x\|_2 \leq \sqrt{d}\|x\|_\infty$ become equalities.
- When constrained to the ℓ_2 unit ball (a hypersphere), the point that maximizes the ℓ_1 norm is where the energy is distributed equally among all components (e.g., $(1/\sqrt{d}, \ldots, 1/\sqrt{d})$). This point makes the inequality $\|x\|_1 \leq \sqrt{d}\|x\|_2$ an equality. Conversely, the ℓ_∞ norm is maximized when all energy is concentrated in a single component (e.g., a basis vector like $(1, 0, \ldots, 0)$), which corresponds to the points where the hypersphere intersects the coordinate axes and makes the inequality $\|x\|_\infty \leq \|x\|_2$ an equality.

Explanation for 5 year old Imagine a big, flat playground. We're going to draw two different "play zones" you have to stay inside.

1. A perfect **square** play zone.
2. A perfect **circle** play zone.

Now, we want to find the spot inside your zone that is the "farthest" from the very center of the playground. But we have different ways to measure "farthest"!

- **Walking Distance:** How many steps you take along the grid lines (like city blocks) to get back to the center.
- **Flying Distance:** The straight line distance if you could fly like a bird.

Here is what we find:

- If you are in the **SQUARE** zone: The farthest place you can be, for *both* walking and flying, is always at one of the four **corners** of the square!
- If you are in the **CIRCLE** zone: It gets tricky!
  - To get the biggest **Walking Distance**, you should stand exactly in the middle of the curvy edge, halfway between North and East. You have to walk a medium amount in two directions.
  - To get the biggest "single-step" distance (the longest part of your walk), you should stand right at the top of the circle (the "North Pole"). Here, you put all your effort into one big step North and took zero steps East or West.

So, the "best" spot depends on both the shape of your play zone and how you decide to measure the distance!

Solution 4

Solution: Derivation of the Unit Ball Equation

We are asked to find and sketch the unit ball for the weighted norm $\|x\|_w = \sqrt{\sum_{i=1}^{d} w_i x_i^2}$ in $d = 2$ dimensions with the weight vector $w = (w_1, w_2) = (1, 4)$. The unit ball is the set of all points $x = (x_1, x_2)$ such that $\|x\|_w \leq 1$. The boundary of this set is defined by the equation $\|x\|_w = 1$. We derive the explicit equation for this boundary below.

$$\begin{aligned} \|x\|_w &= 1 \\ \sqrt{w_1 x_1^2 + w_2 x_2^2} &= 1 \\ \sqrt{1 \cdot x_1^2 + 4 \cdot x_2^2} &= 1 \\ x_1^2 + 4x_2^2 &= 1 \quad \text{(Squaring both sides)} \end{aligned}$$

This equation can be rewritten in the standard form for an ellipse, $\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} = 1$, as:

$$\frac{x_1^2}{1^2} + \frac{x_2^2}{(1/2)^2} = 1$$

This shows the unit ball is an ellipse centered at the origin.

Solution: Sketch of the Unit Ball

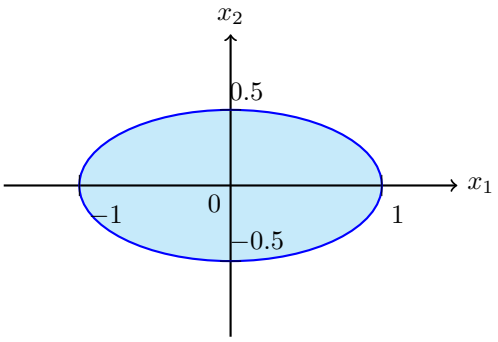

Figure 1: The unit ball $\|x\|_w \leq 1$ for the weight vector $w = (1, 4)$ is an ellipse centered at the origin with semi-major axis of length 1 along the x_1-axis and semi-minor axis of length 1/2 along the x_2-axis.

$\therefore$ The unit ball is an ellipse defined by $x_1^2 + 4x_2^2 \leq 1$, with a semi-major axis of length 1 and a semi-minor axis of length 1/2.

Graduate Level Explanation

The weighted ℓ_2 norm, $\|x\|_w$, is a specific instance of a norm induced by a quadratic form. The squared norm, $\|x\|_w^2 = x^T W x$, where W is a diagonal matrix of the weights ($W = \text{diag}(w_1, \ldots, w_d)$), defines the geometry of the vector space. The unit ball, defined by $x^T W x \leq 1$, is an ellipsoid whose principal axes are aligned with the coordinate axes.

This structure is directly related to the **Mahalanobis distance**, where the distance from a point x to the origin is given by $\sqrt{x^T \Sigma^{-1} x}$, with Σ being the covariance matrix. Our weighted norm corresponds to a Mahalanobis distance where the features are uncorrelated, and the covariance matrix Σ is diagonal with entries $\Sigma_{ii} = 1/w_i$.

A large weight w_i corresponds to a small variance $\sigma_i^2 = 1/w_i$ in that direction. Geometrically, this means the space is "tighter" or less variant along that axis. Consequently, the unit ball is contracted along any axis with a weight $w_i > 1$ and expanded along any axis with $w_i < 1$. In our case, $w_2 = 4$ implies a variance of $1/4$, leading to an axis length of $\sqrt{1/4} = 1/2$. The weights effectively warp the standard Euclidean space, transforming the spherical ℓ_2 unit ball into an ellipsoid.

Explanation for 5 year old

Imagine you have a perfectly round balloon. That's like the normal way we measure distance (the regular unit circle).

Now, let's think about the "weights". A weight is like a special rule for stretching the balloon's rubber.

- For the side-to-side direction (x_1), the weight is 1. That's a normal rule, so we don't stretch the balloon at all in that direction. It stays as wide as it was.
- For the up-and-down direction (x_2), the weight is 4. That's a super strong rule! It's like the rubber in that direction is four times tighter and harder to stretch.

Because the up-and-down rubber is so tight, the balloon can't puff out very far in that direction. It gets **squished vertically**!

So, our perfectly round balloon gets squished into a short, wide oval shape (an ellipse). It's still just as wide as it was before, but it's only half as tall because the "heavy" weight of 4 squashed it from the top and bottom.

$$\ell_2 = \|p - q\|_2 = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Solution 5

Problem Statement

We are given a set of points $\mathcal{X} = \{A, B, C, D\}$ and a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by the following distance table. We will determine if d is a metric by checking the required axioms.

$d(x, y)$	A	B	C	D
A	0	2	1	5
B	2	0	3	4
C	1	3	0	2
D	5	4	2	0

Axiom 1: Non-negativity and Identity of Indiscernibles

This axiom requires that $d(x, y) \geq 0$ for all $x, y \in \mathcal{X}$, and that $d(x, y) = 0$ if and only if $x = y$.

- **Non-negativity:** All entries in the table are non-negative, so this condition holds.
- **Identity:** The diagonal entries are all zero, so $d(x, x) = 0$. All off-diagonal entries are strictly positive, so $d(x, y) > 0$ when $x \neq y$.

Conclusion: The first axiom is satisfied.

Axiom 2: Symmetry

This axiom requires that $d(x, y) = d(y, x)$ for all $x, y \in \mathcal{X}$. This corresponds to the distance matrix being symmetric across its main diagonal. By inspection:

- $d(A, B) = 2 = d(B, A)$
- $d(A, C) = 1 = d(C, A)$
- $d(A, D) = 5 = d(D, A)$
- $d(B, C) = 3 = d(C, B)$
- $d(B, D) = 4 = d(D, B)$
- $d(C, D) = 2 = d(D, C)$

Conclusion: The symmetry axiom is satisfied.

Axiom 3: The Triangle Inequality

This axiom requires that for any three points $x, y, z \in \mathcal{X}$, the inequality $d(x, z) \leq d(x, y) + d(y, z)$ must hold. We search for a counterexample. Let's test the path from point A to point D by way of point C. We set $x = A$, $y = C$, and $z = D$.

$$\begin{aligned} d(A, D) &\leq d(A, C) + d(C, D) \\ 5 &\leq 1 + 2 \\ 5 &\leq 3 \quad \text{(This is FALSE)} \end{aligned}$$

Conclusion: Since we have found a set of points $\{A, C, D\}$ for which the triangle inequality does not hold, the function d fails this axiom.

$\therefore$ The function d is not a metric because it fails the triangle inequality.

Graduate Level Explanation The triangle inequality is the most crucial axiom for establishing a coherent geometric structure on a set. It formalizes the intuitive concept that the length of a direct path between two points, $d(x, z)$, cannot be greater than the length of an indirect path that goes through an intermediate point y. When this axiom fails, as it does here with $d(A, D) > d(A, C) + d(C, D)$, the function does not induce a proper metric space. The set $\{A, B, C, D\}$ equipped with this function d is not a metric space.

Such a function, which satisfies non-negativity, identity, and symmetry but not the triangle inequality, is sometimes called a **semimetric**. The failure implies a non-intuitive geometry where "shortcuts" exist that are longer than the direct route, making standard topological concepts like open balls and convergence behave unpredictably. In applied contexts, this could represent a flawed cost matrix where, for instance, a direct flight from A to D is inexplicably more expensive than two separate flights (A to C, then C to D).

Explanation for 5 year old Imagine you have a treasure map that shows four cities: A, B, C, and D. The numbers on the map tell you how many minutes it takes to walk between the cities.

The map must follow one very important rule: a shortcut can't be longer than going straight! This is called the "Triangle Rule."

Let's look at the path from city **A** to city **D**.

- The map says the direct road from **A** to **D** takes **5 minutes**.

But what if we take a shortcut through city **C**?

- The map says going from **A** to **C** takes **1 minute**.
- Then, going from **C** to **D** takes **2 minutes**.

If you add those up, the shortcut through C only takes $1 + 2 =$ **3 minutes** in total!

Wait a minute... that can't be right! How can the direct path take 5 minutes when a shortcut only takes 3 minutes? The direct path is supposed to be the fastest. Since this map has a silly rule-breaking path, it's a broken map! That's why we say it's **not a metric**.

Solution 6

Solution 6 (a)

The Largest Possible Value of $K(p, q)$

The KL divergence becomes unbounded and approaches infinity if there is an outcome $x \in \mathcal{X}$ that is possible under the true distribution p (i.e., $p(x) > 0$) but is considered impossible by the approximating distribution q (i.e., $q(x) = 0$). This situation violates the condition of absolute continuity. The term $p(x) \log \frac{p(x)}{q(x)}$ becomes infinite, driving the entire sum to infinity. We demonstrate this with two deterministic but opposing distributions.
Let $p = (1, 0)$ and $q = (0, 1)$. This means $p(x_1) = 1, p(x_2) = 0$ and $q(x_1) = 0, q(x_2) = 1$.

$$\begin{aligned} K(p, q) &= p(x_1) \log \frac{p(x_1)}{q(x_1)} + p(x_2) \log \frac{p(x_2)}{q(x_2)} \\ &= 1 \cdot \log \frac{1}{0} + 0 \cdot \log \frac{0}{1} \\ &= \infty + 0 \\ &= \infty \end{aligned}$$

$\therefore$ The largest possible value of the KL divergence is ∞.

Solution 6

Solution 6 (b)

(b): Proof of Non-Symmetry

To prove that the KL divergence is not a symmetric function, we must provide a concrete example where $K(p,q) \neq K(q,p)$. We choose one distribution to be uniform and the other to be deterministic. This highlights the asymmetry in how KL divergence penalizes impossible events.
Let $p = (1/2, 1/2)$ and $q = (1, 0)$. First, we compute $K(p,q)$.

$$\begin{aligned} K(p,q) &= p(x_1)\log\frac{p(x_1)}{q(x_1)} + p(x_2)\log\frac{p(x_2)}{q(x_2)} \\ &= \frac{1}{2}\log\frac{1/2}{1} + \frac{1}{2}\log\frac{1/2}{0} \\ &= \frac{1}{2}\log\frac{1}{2} + \infty \\ &= \infty \end{aligned}$$

Next, we compute the reverse divergence, $K(q,p)$.

$$\begin{aligned} K(q,p) &= q(x_1)\log\frac{q(x_1)}{p(x_1)} + q(x_2)\log\frac{q(x_2)}{p(x_2)} \\ &= 1 \cdot \log\frac{1}{1/2} + 0 \cdot \log\frac{0}{1/2} \\ &= \log(2) + 0 \\ &= \log(2) \end{aligned}$$

$\therefore$ Since $K(p,q) = \infty$ and $K(q,p) = \log(2)$, we have shown that $K(p,q) \neq K(q,p)$, and thus KL divergence is not symmetric.

Graduate Level Explanation The Kullback-Leibler divergence is a measure of **relative entropy**, not a distance metric. Its failure to be a metric is fundamental to its interpretation. The non-symmetry shown above is a key property. $K(p, q)$ measures the information lost when q is used to approximate p; this is an inherently directional concept.

The unboundedness of $K(p, q)$ is a direct consequence of the violation of **absolute continuity**. The divergence is finite if and only if the support of p is a subset of the support of q (denoted $\text{supp}(p) \subseteq \text{supp}(q)$ or $p \ll q$). If this condition fails, there exists an event x for which $p(x) > 0$ but $q(x) = 0$. The model q assigns zero probability to an event that can actually occur, leading to an infinite "surprise" or error. This property is crucial in variational inference and Bayesian model comparison, where an infinite divergence signals a catastrophic failure of the approximating distribution to cover the posterior's support. KL divergence is a member of the broader classes of *f-divergences* and *Bregman divergences*, none of which are required to be symmetric.

Explanation for 5 year old Imagine a guessing game. The "Truth" knows the right answer (p), and you are making a guess (q). The KL divergence is a number that tells us how surprised the Truth is by your guess. A bigger number means more surprise!

- **Why it can be infinite:** The Truth is "The secret animal is a Dog" (p is 100% Dog). You guess, "I am 100% certain it's a Cat!" (q is 100% Cat). You said a Dog was impossible! When the Truth reveals it's a Dog, your surprise is infinite because you were completely wrong about something that was certain.
- **Why it's not the same backwards:** Let's look at two cases.
  1. The Truth is "The secret animal could be a Dog or a Cat, 50/50 chance" (p is 50/50). You guess, "It's definitely a Dog!" (q is 100% Dog). The Truth is infinitely surprised because you said a Cat was impossible, but it was possible.
  2. Now let's switch! The Truth is "It's definitely a Dog" (q is 100% Dog). You guess, "It could be a Dog or a Cat, 50/50 chance" (p is 50/50). Is the Truth surprised? A little bit! The Truth is surprised you weren't more confident, but not infinitely surprised, because you at least said a Dog was possible.

Since the surprise is **infinite** in the first game but just a **little bit** in the second game, the "surprise-o-meter" doesn't work the same forwards and backwards!

Solution 7

Solution: Jaccard Similarity Calculation

We are asked to compute the Jaccard similarity, J(A, B), for the following two sets:
A = {1, 3, 5, 7, 9}
B = {2, 3, 5, 7}
The Jaccard similarity is defined as the size of the intersection of the sets divided by the size of their union.

Step 1: Compute the Intersection

The intersection of A and B, denoted A intersect B, is the set of elements common to both sets.
A intersect B = {3, 5, 7}
The size of the intersection is |A intersect B| = 3.

Step 2: Compute the Union

The union of A and B, denoted A union B, is the set of all unique elements present in either set.
A union B = {1, 2, 3, 5, 7, 9}
The size of the union is |A union B| = 6.

Step 3: Compute the Jaccard Similarity

We now apply the formula J(A, B) = |A intersect B| / |A union B|.

$$\begin{aligned} J(A,B) &= |\{3,5,7\}|/|\{1,2,3,5,7,9\}| \\ &= 3/6 \\ &= 0.5 \end{aligned}$$

$\therefore$ The Jaccard similarity J(A, B) is 0.5 or 50%.

Graduate Level Explanation

The Jaccard index is a statistic used for gauging the similarity and diversity of sample sets. It is particularly effective for sparse binary data, where the number of shared attributes (the intersection) is compared against the total number of unique attributes observed across both sets (the union). For binary vectors representing sets, this is also known as the Tanimoto coefficient. It is widely used in applications like document similarity analysis (where sets are bags of words), image segmentation, and bioinformatics. The corresponding Jaccard distance, defined as d_J(A, B) = 1 - J(A, B), is a metric and measures dissimilarity. In this case, the Jaccard distance would be 1 - 0.5 = 0.5.

Explanation for 5 year old

Imagine you have a box of toys (Box A) and your friend has a box of toys (Box B). We want to see how similar your toy boxes are!

1. First, we look at all the toys you BOTH have. This is the "intersection". Let's say you both have a car, a block, and a teddy bear. That's 3 shared toys.
2. Next, we pour all the toys from both boxes into one big pile and count every DIFFERENT toy just once. This is the "union". Maybe in the big pile, there are 6 unique toys in total.
3. To get the similarity score, we just divide the number of shared toys by the total number of unique toys. So, 3 divided by 6 is 1/2. Your toy boxes are 50% similar!

Solution 8

Solution: Bigram Jaccard Similarity

We are asked to compute the Jaccard similarity based on word bigrams for the following two sentences, where x' is pre-processed to remove punctuation:
x: "Napoleon was born in 1769"
x': "Napoleon was born when"
The Jaccard similarity for the bigram sets $B(x)$ and $B(x')$ is defined as $J(B(x), B(x')) = \frac{|B(x) \cap B(x')|}{|B(x) \cup B(x')|}$.

Step 1: Derive Bigram Sets

First, we generate the sets of consecutive word pairs (bigrams) for each sentence.

- $B(x) = \{(\text{Napoleon, was}), (\text{was, born}), (\text{born, in}), (\text{in, 1769})\}$
- $B(x') = \{(\text{Napoleon, was}), (\text{was, born}), (\text{born, when})\}$

The sizes of the sets are $|B(x)| = 4$ and $|B(x')| = 3$.

Step 2: Compute Intersection and Union of Sets

Next, we find the intersection (common bigrams) and the union (all unique bigrams) of the two sets.

- Intersection: $B(x) \cap B(x') = \{(\text{Napoleon, was}), (\text{was, born})\}$
- Union: $B(x) \cup B(x') = \{(\text{Napoleon, was}), (\text{was, born}), (\text{born, in}), (\text{in, 1769}), (\text{born, when})\}$

The sizes are $|B(x) \cap B(x')| = 2$ and $|B(x) \cup B(x')| = 5$.

Step 3: Compute the Jaccard Similarity

Finally, we apply the Jaccard similarity formula using the sizes of the intersection and union.

$$\begin{aligned} J(B(x), B(x')) &= \frac{|B(x) \cap B(x')|}{|B(x) \cup B(x')|} \\ &= \frac{2}{5} \\ &= 0.4 \end{aligned}$$

$\therefore$ The bigram-based Jaccard similarity between the two sentences is 0.4

Graduate Level Explanation

Using Jaccard similarity over n-grams (in this case, bigrams) allows for a measure of text similarity that captures local syntactic and semantic structure, which is a significant improvement over simple bag-of-words models. Unlike methods that treat words as independent, n-grams preserve word order within a small window. This makes the metric sensitive to phrasal correspondence but robust to larger-scale sentence reordering. It is particularly effective for tasks like plagiarism detection, identifying near-duplicate documents, and record linkage, where detecting overlapping text chunks is more important than understanding deep semantic meaning. While it is computationally simpler than embedding-based methods like cosine similarity on BERT vectors, it effectively balances structural awareness with efficiency.

Explanation for 5 year old

Imagine you and your friend are building sentences with Lego bricks, where each brick is a word.
Instead of just counting how many of the same color bricks you both used, we're going to look at how you connected them. We'll look at every pair of bricks stuck together. A pair of word-bricks is a "bigram".

1. First, we find all the two-brick connections that are exactly the same in both of your sentences. Let's say you find **2** matching connections.

2. Next, we count all the unique connections you both made. Maybe in total, there are **5** different connections.

To see how similar your sentences are, we just divide the number of matching connections by the total number of connections: 2 divided by 5. That gives a similarity score! It tells us how much of the sentences were built in the same way.

Solution 9

Solution 9 (a)

(a): Calculating Cosine Similarity

We compute the cosine similarity for the vectors $x = [1, 2, 3]^T$ and $x' = [3, 2, 1]^T$ using the formula $\cos(\theta) = \frac{x \cdot x'}{\|x\|_2 \|x'\|_2}$. First, we compute the dot product and the ℓ_2-norms.
Dot Product: $x \cdot x' = (1)(3) + (2)(2) + (3)(1)$
Norms: $\|x\|_2 = \sqrt{1^2 + 2^2 + 3^2}$ and $\|x'\|_2 = \sqrt{3^2 + 2^2 + 1^2}$

$$\begin{aligned}
\cos(\theta) &= \frac{(1)(3) + (2)(2) + (3)(1)}{\sqrt{1^2 + 2^2 + 3^2}\sqrt{3^2 + 2^2 + 1^2}} \\
&= \frac{3 + 4 + 3}{\sqrt{1 + 4 + 9}\sqrt{9 + 4 + 1}} \\
&= \frac{10}{\sqrt{14}\sqrt{14}} \\
&= \frac{10}{14} \\
&= \frac{5}{7}
\end{aligned}$$

$\therefore$ The cosine similarity between x and x' is $5/7$.

Solution 9

Solution 9 (b)

(b): Characterization of Zero Similarity

The cosine similarity between two non-zero vectors x and x' is zero if and only if the numerator of the formula is zero.

$$\cos(\theta) = 0 \iff x \cdot x' = 0$$

In a Euclidean vector space, a dot product of zero signifies that the two vectors are **orthogonal** (perpendicular) to each other. The angle θ between them is 90° or $\pi/2$ radians. Geometrically, they form a right angle.

$\therefore$ A cosine similarity of zero means the vectors are orthogonal.

Solution 9

Solution 9 (c)

(c): Sketch of a High-Similarity Region

We want to sketch the region in $\mathbb{R}^2$ where any vector x' has a cosine similarity of at least 0.9 with the vector $x = [1, 2]^T$. The condition is $\cos(\theta) \geq 0.9$, which implies that the angle θ between x and x' must satisfy $|\theta| \leq \arccos(0.9) \approx 25.84°$.

Description of the Sketch:

- Draw the standard Cartesian axes, x_1 and x_2.
- Draw the reference vector x as an arrow from the origin $(0, 0)$ to the point $(1, 2)$.
- The region of high similarity is a double-sided cone (a pair of sectors in 2D) centered on the line passing through the origin and the point $(1, 2)$.
- The cone opens with an angle of $2 \times 25.84°$. Any vector x' that originates at $(0, 0)$ and whose tip lies within this cone satisfies the condition.

$\therefore$ The region is a cone centered on the vector x with an angular radius of $\arccos(0.9)$.

Graduate Level Explanation

Cosine similarity is a measure of orientation, not magnitude. By normalizing the vectors to unit length via the ℓ_2-norm, it isolates the angle between them, making it an effective measure of similarity in applications where vector length is a confounding variable. For instance, in Natural Language Processing, documents are often represented as high-dimensional TF-IDF vectors. A longer document might have a larger Euclidean distance from a shorter one, even if they discuss the same topic. Cosine similarity correctly identifies them as similar by disregarding their "length" (i.e., word count) and focusing solely on their "direction" (i.e., topic, as defined by the distribution of word frequencies). This property is also critical in recommender systems, where it can measure the similarity of user preferences irrespective of the total number of items they have rated.

Explanation for 5 year old

Imagine you and your friend are each pointing with an arrow. Cosine similarity doesn't care how long your arrows are, only which way they're pointing!

- If you both point in the **exact same direction**, the similarity is **1**. Perfect match!
- If your arrows make a perfect "L" shape (a right angle), the similarity is **0**. You're not pointing in similar directions at all.
- If you point in the **exact opposite direction** of your friend, the similarity is **-1**. You completely disagree!

It's just a score from -1 to 1 that tells you how much your arrows line up.

Solution 10

Solution 10 (a)

1 (a): Mean and Median (Theoretical)

The mean (Expected Value) $E[X]$ is calculated as the sum of each outcome weighted by its probability: $E[X] = \sum_i x_i P(X = x_i)$. The median is the value x_m for which the cumulative probability $P(X \leq x_m)$ first equals or exceeds 0.5.

$$\begin{aligned} E[X] &= (1)\left(\frac{1}{3}\right) + (2)\left(\frac{1}{3}\right) + (3)\left(\frac{1}{12}\right) + (4)\left(\frac{1}{12}\right) + (5)\left(\frac{1}{12}\right) + (6)\left(\frac{1}{12}\right) \\ &= \frac{4}{12} + \frac{8}{12} + \frac{3}{12} + \frac{4}{12} + \frac{5}{12} + \frac{6}{12} \\ &= \frac{30}{12} = \frac{5}{2} = 2.5 \end{aligned}$$

For the median, we find the cumulative probabilities:
$P(X \leq 1) = 1/3 \approx 0.333$
$P(X \leq 2) = 1/3 + 1/3 = 2/3 \approx 0.667$
Since $P(X \leq 2)$ is the first cumulative probability to exceed 0.5, the median is 2.

$\therefore$ The theoretical mean is $E[X] = 2.5$ and the median is 2.

Solution 10

Solution 10 (b)

Variance is calculated as $\text{Var}(X) = E[X^2] - (E[X])^2$, where $E[X^2] = \sum_i x_i^2 P(X = x_i)$. The standard deviation is the square root of the variance, $\text{SD}(X) = \sqrt{\text{Var}(X)}$.

$$\begin{aligned} E[X^2] &= (1^2)\left(\frac{1}{3}\right) + (2^2)\left(\frac{1}{3}\right) + (3^2)\left(\frac{1}{12}\right) + (4^2)\left(\frac{1}{12}\right) + (5^2)\left(\frac{1}{12}\right) + (6^2)\left(\frac{1}{12}\right) \\ &= \frac{4}{12} + \frac{16}{12} + \frac{9}{12} + \frac{16}{12} + \frac{25}{12} + \frac{36}{12} \\ &= \frac{106}{12} = \frac{53}{6} \approx 8.833 \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= E[X^2] - (E[X])^2 = \frac{53}{6} - \left(\frac{5}{2}\right)^2 \\ &= \frac{53}{6} - \frac{25}{4} = \frac{106}{12} - \frac{75}{12} = \frac{31}{12} \approx 2.5833 \end{aligned}$$

$$\text{SD}(X) = \sqrt{\frac{31}{12}} \approx 1.607$$

$\therefore$ The theoretical variance is $\text{Var}(X) = 31/12 \approx 2.5833$ and the standard deviation is $\text{SD}(X) \approx 1.607$.

Solution 10

Solution 10 (c)

1 (c): Empirical Probability Distribution

Given the set of $N = 10$ observations $D = \{2, 5, 1, 4, 2, 2, 5, 6, 1, 2\}$, the empirical probability $\hat{P}(X = x)$ for each outcome is its observed frequency divided by the total number of observations, N.
The counts for each outcome are:
Count(1) = 2
Count(2) = 4
Count(3) = 0
Count(4) = 1
Count(5) = 2
Count(6) = 1

$\therefore$ The empirical probabilities are: $\hat{P}(1) = 0.2$, $\hat{P}(2) = 0.4$, $\hat{P}(3) = 0.0$, $\hat{P}(4) = 0.1$, $\hat{P}(5) = 0.2$, $\hat{P}(6) = 0.1$.

Solution 10

Solution 10 (d)

1 (d): Mean, Median, Variance, and SD (Empirical)

The empirical mean is $\bar{x} = \frac{1}{N}\sum x_i$. The median is the middle value of the sorted data. The empirical variance is $s^2 = \frac{1}{N}\sum(x_i - \bar{x})^2$, and the standard deviation is $s = \sqrt{s^2}$.

$$\bar{x} = \frac{1}{10}(2+5+1+4+2+2+5+6+1+2) = \frac{30}{10} = 3.0$$

Sorted data: $\{1, 1, 2, 2, \mathbf{2,\ 2}, 4, 5, 5, 6\}$. The median is the average of the 5th and 6th values: $\frac{2+2}{2} = 2$.

$$\begin{aligned}
s^2 &= \frac{1}{10}\sum_{i=1}^{10}(x_i - 3.0)^2 \\
&= \frac{1}{10}\left[(1-3)^2 \times 2 + (2-3)^2 \times 4 + (4-3)^2 \times 1 + (5-3)^2 \times 2 + (6-3)^2 \times 1\right] \\
&= \frac{1}{10}\left[(-2)^2 \times 2 + (-1)^2 \times 4 + (1)^2 \times 1 + (2)^2 \times 2 + (3)^2 \times 1\right] \\
&= \frac{1}{10}\left[4 \times 2 + 1 \times 4 + 1 \times 1 + 4 \times 2 + 9 \times 1\right] \\
&= \frac{1}{10}\left[8 + 4 + 1 + 8 + 9\right] = \frac{30}{10} = 3.0
\end{aligned}$$

$$s = \sqrt{3.0} \approx 1.732$$

$\therefore$ The empirical mean is $\bar{x} = 3.0$, median is 2, variance is $s^2 = 3.0$, and standard deviation is $s \approx 1.732$.

Graduate Level Explanation: Theoretical vs. Empirical

The theoretical distribution describes the true, underlying probability model of the random variable X. Its moments, such as the mean $\mu = E[X]$ and variance $\sigma^2 = \text{Var}(X)$, are fixed population parameters derived from this model. The empirical distribution, conversely, is constructed from a finite sample of observations. Its statistics, like the sample mean $\bar{x}$ and sample variance s^2, are estimates of the true parameters. The **Law of Large Numbers (LLN)** states that as the sample size N approaches infinity, the sample mean $\bar{x}$ converges in probability to the theoretical mean μ. Similarly, other empirical moments converge to their theoretical counterparts. The discrepancy between our calculated theoretical values ($\mu = 2.5, \sigma^2 \approx 2.58$) and empirical values ($\bar{x} = 3.0, s^2 = 3.0$) is expected due to random sampling variation in our small sample ($N = 10$).

Explanation for a 5-Year-Old

Imagine you have a magic cookie jar. The **plan** (the theoretical part) says that for every 12 cookies you pull out, you *should* get 4 chocolate chip, 4 oatmeal, 1 sugar, 1 peanut butter, 1 ginger, and 1 snickerdoodle. The plan's "average cookie" is a mix between a chocolate chip and an oatmeal cookie.

But then, you actually pull out just 10 cookies. This is your **handful** (the empirical part). In your handful, you got a lot of oatmeal cookies and no sugar cookies at all! What happened in your one small handful is a little different from the big plan for the whole jar. If you kept pulling out cookies all day (thousands of them!), your handful would start to look a lot more like the original plan.

Solution 11

Solution 11 (a)

Justification: The distribution of human height for a given population (e.g., adult males) is famously well-approximated by a symmetric, bell-shaped curve (a Normal or Gaussian distribution). In a perfectly symmetric distribution, the mean, median, and mode coincide. Therefore, the center of mass (mean) and the geometric center (median) are expected to be nearly identical.

$\therefore$ We expect $\text{Mean}(H) \approx \text{Median}(H)$. No significant difference.

Solution 11

Solution 11 (b)

Justification: The distribution of housing costs is almost always characterized by a strong **right-skew (positive skew)**. While most houses fall within a certain price range, there is a long tail of extremely expensive properties (mansions, luxury estates). These high-value outliers pull the arithmetic average significantly upward, while the median remains a more robust measure of the "typical" house price, unaffected by these extreme values.

$\therefore$ We expect $\text{Mean}(C) > \text{Median}(C)$. A significant difference.

Solution 11

Solution 11 (c)

Justification: The distribution of GPAs is often **left-skewed (negative skew)**. This is due to a "ceiling effect," where a large number of students achieve high grades clustered near the maximum possible GPA (e.g., 4.0), while fewer students have very low GPAs. This clustering at the high end pulls the median towards the right, while the lower-end scores pull the mean to the left. However, this skew is often less pronounced than in financial data.

$\therefore$ We expect $\text{Mean}(G) < \text{Median}(G)$, but the difference may not be as significant as for cost or salary.

Solution 11

Solution 11 (d)

Justification: Similar to housing costs, salary data exhibits a pronounced **right-skew**. The vast majority of people earn salaries within a relatively modest range. However, a small number of individuals (CEOs, top athletes, etc.) have extraordinarily high incomes. These outliers exert a strong influence on the mean, pulling it far to the right of the median. The median salary is therefore a much more accurate representation of a typical worker's earnings.

$\therefore$ We expect $\text{Mean}(S) > \text{Median}(S)$. A significant difference.

Graduate Level Explanation

The **mean** (μ) is the first moment of a distribution, representing its center of mass. It is defined as the expected value of the random variable X, so that $\mu = E[X]$. The mean minimizes the expected squared error, i.e., $\mu = \arg\min_c E[(X-c)^2]$. Its sensitivity to the magnitude of every data point makes it susceptible to being heavily influenced by outliers.

The **median** (m) is the value that separates the higher half from the lower half of a data set. More formally, it is a value that minimizes the L1-norm or the expected absolute error, $m = \arg\min_c E[|X-c|]$. This property makes the median a **robust** estimator of central tendency, as it is insensitive to the magnitude of extreme outliers.

In a **symmetric distribution**, the center of mass and the 50th percentile coincide, thus $\mu = m$. In a **skewed distribution**, the mass in the tail pulls the mean in its direction. For a right-skewed distribution, the long tail of high-value outliers pulls $\mu > m$. Conversely, for a left-skewed distribution, the tail of low-value outliers pulls $\mu < m$. The median remains a better indicator of the central location of the bulk of the probability mass.

Explanation for a 5-Year-Old

Imagine five kids are sharing candy. Four kids get 2 pieces of candy each. But one super-lucky kid gets 50 pieces!

If we use the **mean** (the "average"), we add all the candy up ($2+2+2+2+50=58$) and divide by the number of kids (5). The mean is almost 12 pieces per kid! That doesn't sound right, because most kids only got 2. The one kid with 50 pieces pulled the average way up.

If we use the **median** (the "middle"), we line the kids up by how much candy they have: 2, 2, **2**, 2, 50. The kid in the middle has 2 pieces. The median is 2. This is a much better description of what a typical kid got.

So, the **mean is sensitive** to that one kid with a huge amount of candy (an outlier), while the **median just looks at the middle** and doesn't care how much the richest kid has.

Solution 12

---

nesto ovdje

Question 1: Derivation of the Second Raw Moment

The second raw moment, denoted as $E[Z^2]$, can be derived from the first raw moment (the mean, $\mu = E[Z]$) and the second central moment (the variance, $Var[Z] = \sigma^2$) using the fundamental relationship:

$$Var[Z] = E[Z^2] - (E[Z])^2$$

Rearranging this formula allows us to solve for $E[Z^2]$:

$$E[Z^2] = Var[Z] + (E[Z])^2 = \sigma^2 + \mu^2$$

Given the values $E[Z] = -1$ and the standard deviation $SD[Z] = \sigma = 2$, we can calculate the variance as $\sigma^2 = 2^2 = 4$.

$$\begin{aligned} E[Z^2] &= \sigma^2 + \mu^2 \\ &= (SD[Z])^2 + (E[Z])^2 \\ &= (2)^2 + (-1)^2 \\ &= 4 + 1 \\ &= 5 \end{aligned}$$

$\therefore$ The second raw moment $E[Z^2]$ is 5.

Graduate Level Explanation The relationship $Var(Z) = E[Z^2] - (E[Z])^2$ provides a direct conversion between the second central moment ($Var(Z)$) and the first two raw moments ($E[Z]$ and $E[Z^2]$). In general, central moments, $E[(Z - \mu)^k]$, describe the shape of a distribution relative to its mean, while raw moments, $E[Z^k]$, describe the distribution relative to the origin. This specific formula is a fundamental identity derived from the linearity of expectation and is crucial when working with moment-generating functions (MGFs) or characteristic functions, as raw moments are easily derived from them via differentiation. Physically, in signal processing, if Z represents a random signal, $E[Z]$ is its DC component, and $E[Z^2]$ is related to its total average power. The variance, representing the second central moment, is then interpreted as the AC power of the signal.

Explanation for a 5-Year-Old Imagine you are playing a game. Your **average score** is your mean ($E[Z]$). Let's say your average is -1. How much your scores jump around from that average is the **wildness** (standard deviation, σ). Let's say the wildness is 2. We want to find the "average of the scores squared" ($E[Z^2]$). To get this, you just follow a simple rule: combine the **wildness squared** with the **average score squared**.

$$\text{Average Score Squared} = (\text{Wildness})^2 + (\text{Average Score})^2$$

So, we calculate $(2 \times 2) + (-1 \times -1) = 4 + 1 = 5$. It's a fundamental statistical rule that connects how spread out your scores are with what your average score is.

Solution 13

Solution 13 (a)

Part (a): Are X and Y Independent?

Two discrete random variables X and Y are independent if and only if $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all possible pairs (x, y). A single counterexample is sufficient to prove dependence. First, we compute the marginal probability mass functions, $P(X = x)$ and $P(Y = y)$, by summing the rows and columns of the joint PMF table, respectively.

$$P(X = 1) = 0.10 + 0.20 + 0.05 = 0.35$$
$$P(X = 2) = 0.10 + 0.15 + 0.05 = 0.30$$
$$P(X = 3) = 0.10 + 0.15 + 0.10 = 0.35$$

$$P(Y = 1) = 0.10 + 0.10 + 0.10 = 0.30$$
$$P(Y = 2) = 0.20 + 0.15 + 0.15 = 0.50$$
$$P(Y = 3) = 0.05 + 0.05 + 0.10 = 0.20$$

We test the independence condition for the pair $(X = 1, Y = 1)$:

$$P(X = 1, Y = 1) \stackrel{?}{=} P(X = 1)P(Y = 1)$$
$$0.10 \stackrel{?}{=} (0.35)(0.30)$$
$$0.10 \neq 0.105$$

$\therefore$ Since the joint probability $P(X = 1, Y = 1)$ does not equal the product of the marginal probabilities $P(X = 1)P(Y = 1)$, the random variables X and Y are NOT independent.

Solution 13

Solution 13 (b)

Part (b): What are the Covariance and Correlation of X and Y?

To find the covariance and correlation, we must first compute the expectations $E[X]$, $E[Y]$, and $E[XY]$. The covariance is then given by $\text{Cov}[X,Y] = E[XY] - E[X]E[Y]$, and the correlation is $\rho_{X,Y} = \frac{\text{Cov}[X,Y]}{\sigma_X \sigma_Y}$.

1. **Calculate Expectations:**

$$\begin{aligned} E[X] &= \sum_x xP(X=x) \\ &= 1(0.35) + 2(0.30) + 3(0.35) \\ &= 0.35 + 0.60 + 1.05 = \mathbf{2.0} \end{aligned}$$

$$\begin{aligned} E[Y] &= \sum_y yP(Y=y) \\ &= 1(0.30) + 2(0.50) + 3(0.20) \\ &= 0.30 + 1.00 + 0.60 = \mathbf{1.9} \end{aligned}$$

$$\begin{aligned} E[XY] &= \sum_{x,y} xyP(X=x, Y=y) \\ &= (1)(1)(0.10) + (1)(2)(0.20) + (1)(3)(0.05) \\ &\quad + (2)(1)(0.10) + (2)(2)(0.15) + (2)(3)(0.05) \\ &\quad + (3)(1)(0.10) + (3)(2)(0.15) + (3)(3)(0.10) \\ &= 0.10 + 0.40 + 0.15 + 0.20 + 0.60 + 0.30 + 0.30 + 0.90 + 0.90 \\ &= \mathbf{3.85} \end{aligned}$$

2. **Calculate Covariance:**

$$\begin{aligned} \text{Cov}[X,Y] &= E[XY] - E[X]E[Y] \\ &= 3.85 - (2.0)(1.9) \\ &= 3.85 - 3.80 = \mathbf{0.05} \end{aligned}$$

3. **Calculate Variances and Standard Deviations:**

$$\begin{aligned} E[X^2] &= 1^2(0.35) + 2^2(0.30) + 3^2(0.35) = 0.35 + 1.20 + 3.15 = 4.7 \\ \text{Var}[X] &= E[X^2] - (E[X])^2 = 4.7 - 2.0^2 = 0.7 \\ \sigma_X &= \sqrt{0.7} \approx 0.8367 \end{aligned}$$

$$\begin{aligned} E[Y^2] &= 1^2(0.30) + 2^2(0.50) + 3^2(0.20) = 0.30 + 2.00 + 1.80 = 4.1 \\ \text{Var}[Y] &= E[Y^2] - (E[Y])^2 = 4.1 - 1.9^2 = 4.1 - 3.61 = 0.49 \\ \sigma_Y &= \sqrt{0.49} = 0.7 \end{aligned}$$

4. Calculate Correlation Coefficient:

$$\begin{aligned}\rho_{X,Y} &= \frac{\text{Cov}[X,Y]}{\sigma_X \sigma_Y} \\ &= \frac{0.05}{\sqrt{0.7} \cdot \sqrt{0.49}} \\ &= \frac{0.05}{(0.8367)(0.7)} \\ &\approx \frac{0.05}{0.5857} \approx \mathbf{0.0854}\end{aligned}$$

$\therefore$ The covariance is $\text{Cov}[X,Y] = 0.05$, and the correlation coefficient is $\rho_{X,Y} \approx 0.0854$. This indicates a very weak positive linear relationship between X and Y.

Graduate Level Explanation

Two random variables X and Y are independent if and only if the σ-algebras they generate, $\sigma(X)$ and $\sigma(Y)$, are independent. For discrete variables, this is equivalent to the factorization of the joint PMF, $p_{X,Y}(x, y) = p_X(x)p_Y(y)$, for all (x, y) in the support.

Covariance and correlation are measures of the **linear dependence** between two random variables. A non-zero correlation implies dependence, but the converse is not true. Two variables can be functionally dependent yet have zero correlation if the relationship is non-linear. For example, if $X \sim \text{Uniform}\{-1, 0, 1\}$ and $Y = X^2$, then X and Y are clearly dependent. However, $E[X] = 0$ and $E[XY] = E[X^3] = 0$, so $\text{Cov}[X, Y] = 0$. In this problem, because $\text{Cov}[X, Y] \neq 0$, we have a second confirmation that X and Y are dependent. The joint PMF is the most fundamental description, as it fully defines the probability measure for the random vector (X, Y) and allows for the computation of any property, including marginals, conditionals, and moments.

Explanation for a 5-Year-Old

Imagine two friends, Alex (who is variable X) and Ben (who is variable Y). They each have a bag with balls numbered 1, 2, and 3.

Independence: If they are "independent," it means that when Alex picks a number, it tells you *nothing* about what number Ben is likely to pick. Knowing Alex picked a '3' doesn't make it any more or less likely that Ben will pick a '1', '2', or '3'. In our problem, we found out they are **not** independent. This means that knowing what number Alex picks gives you a small hint about what Ben might pick.

Correlation: This is a number that tells us *how much* Alex's choice is connected to Ben's choice.

- If the number is close to **+1**, it means they are best buddies. When Alex picks a high number (like 3), Ben is very likely to pick a high number too.
- If the number is close to **-1**, they are opposites. When Alex picks a high number, Ben is very likely to pick a low number (like 1).
- If the number is close to **0**, it means there isn't a clear "high-high" or "high-low" pattern.

We found a number that is very, very close to 0 (it was 0.0854). This means that while Alex and Ben are not totally independent, there is only a tiny, tiny hint of a pattern where if one picks a higher number, the other one might also pick a slightly higher number. But the connection is so weak, it's almost not there.

Solution 14

Solution 14 (a)

Part (a): Deriving the Covariance Cov$[X, Y]$

We begin by finding the expectation of Y, $E[Y]$, using the linearity of expectation:

$$E[Y] = E[aX + b] = aE[X] + b$$

The covariance between X and Y is defined as $\text{Cov}[X, Y] = E[(X - E[X])(Y - E[Y])]$. Substituting the expression for $E[Y]$, we derive the covariance as follows:

$$\begin{aligned}
\text{Cov}[X, Y] &= E\left[(X - E[X])(Y - E[Y])\right] \\
&= E\left[(X - E[X])((aX + b) - (aE[X] + b))\right] \\
&= E\left[(X - E[X])(aX + b - aE[X] - b)\right] \\
&= E\left[(X - E[X])(aX - aE[X])\right] \\
&= E\left[(X - E[X])a(X - E[X])\right] \\
&= a \cdot E\left[(X - E[X])^2\right] \\
&= a \cdot \text{Var}[X]
\end{aligned}$$

$\therefore$ The covariance between X and Y is $\text{Cov}[X, Y] = a \cdot \text{Var}[X]$.

Solution 14

Solution 14 (b)

Part (b): Deriving the Correlation $\rho_{X,Y}$

The correlation coefficient $\rho_{X,Y}$ is defined as $\rho_{X,Y} = \frac{\text{Cov}[X,Y]}{\text{SD}[X]\text{SD}[Y]}$. We first need to find the standard deviation of Y, $\text{SD}[Y]$.

$$\text{Var}[Y] = \text{Var}[aX + b] = a^2\text{Var}[X]$$

$$\text{SD}[Y] = \sqrt{\text{Var}[Y]} = \sqrt{a^2\text{Var}[X]} = |a|\sqrt{\text{Var}[X]} = |a| \cdot \text{SD}[X]$$

Now we substitute the known quantities into the correlation formula:

$$\begin{aligned}
\rho_{X,Y} &= \frac{\text{Cov}[X,Y]}{\text{SD}[X]\text{SD}[Y]} \\
&= \frac{a \cdot \text{Var}[X]}{\text{SD}[X] \cdot (|a| \cdot \text{SD}[X])} \\
&= \frac{a \cdot (\text{SD}[X])^2}{|a| \cdot (\text{SD}[X])^2} \\
&= \frac{a}{|a|} \\
&= \text{sgn}(a)
\end{aligned}$$

$\therefore$ The correlation between X and Y is $\rho_{X,Y} = \text{sgn}(a)$, which is 1 if $a > 0$, -1 if $a < 0$, and undefined if $a = 0$.

Graduate Level Explanation The result $\rho_{X,Y} = \pm 1$ provides a foundational insight into the nature of the Pearson correlation coefficient. Correlation measures the strength and direction of a **linear relationship** between two variables. Since $Y = aX + b$ is, by definition, a perfect linear function of X, the magnitude of the correlation must be maximal, which is 1. The sign of the correlation is determined entirely by the slope of the line, a. A positive slope ($a > 0$) means X and Y move in the same direction, yielding $\rho_{X,Y} = 1$. A negative slope ($a < 0$) means they move in opposite directions, yielding $\rho_{X,Y} = -1$.

It is critical to note that the additive constant, b, has no impact on the final correlation. This is because correlation, like covariance and variance, is a measure of spread and co-movement, not of location. The term b simply shifts the mean of Y ($E[Y] = aE[X] + b$) but does not alter its variance ($\text{Var}[Y] = a^2\text{Var}[X]$). Since the correlation coefficient is standardized by the variables' standard deviations, it is invariant to shifts in location.

Explanation for a 5-Year-Old Imagine you have a magic growing toy (X). Every year it gets older, it also gets taller in a very specific way. Let's say for every 1 year it gets older (X), it grows exactly 3 inches taller (Y). So, $Y = 3X$.

If you know the toy's age, you know its height perfectly! They are perfectly linked together. When its age goes up, its height **must** go up. This perfect link is what we call a correlation of **1**. It's the highest score you can get!

Now, what if we started measuring the toy from a small stool that is 10 inches tall? The rule becomes $Y = 3X + 10$. The toy is always 10 inches taller than before, but the way it grows each year doesn't change. It still grows exactly 3 inches for every 1 year. The link between its age and growth is still perfect. The +10 is like the stool (b)—it changes the starting point but not the relationship.

If the magic toy got shorter as it got older (a weird toy!), the link would still be perfect, but it would be going in the opposite direction. That would be a perfect negative link, which is a correlation of **-1**.

Solution 15

---

nesto ovdje

Let X be a discrete random variable with the sample space $\Omega_X = \{-1, 0, 1\}$ and a uniform probability mass function $P(X = x) = \frac{1}{3}$ for all $x \in \Omega_X$. Find a non-trivial function $Y = f(X)$ such that X and Y are deterministically dependent but have zero covariance, i.e., $\text{Cov}[X, Y] = 0$.

Part 1: Determine E[X] and Simplify the Covariance Condition

Two random variables X and Y are uncorrelated if their covariance is zero. The covariance is defined as:

$$\text{Cov}[X, Y] = E[XY] - E[X]E[Y]$$

We first calculate the expected value of X, $E[X]$.

$$\begin{aligned} E[X] &= \sum_{x \in \Omega_X} x \cdot P(X = x) \\ &= (-1) \cdot P(X = -1) + (0) \cdot P(X = 0) + (1) \cdot P(X = 1) \\ &= (-1) \cdot \frac{1}{3} + (0) \cdot \frac{1}{3} + (1) \cdot \frac{1}{3} \\ &= -\frac{1}{3} + 0 + \frac{1}{3} \\ &= 0 \end{aligned}$$

Since $E[X] = 0$, the covariance formula simplifies significantly:

$$\text{Cov}[X, Y] = E[XY] - (0) \cdot E[Y] = E[XY]$$

Therefore, for X and Y to be uncorrelated, we only need to find a function f such that $\mathbf{E}[\mathbf{X}\mathbf{f}(\mathbf{X})] = \mathbf{0}$.

Part 2: Derive and Test the Function $f(X)$

The condition $E[Xf(X)] = 0$ expands to:

$$E[Xf(X)] = \sum_{x \in \Omega_X} xf(x)P(X = x) = \frac{1}{3}\left((-1)f(-1) + (0)f(0) + (1)f(1)\right) = 0$$

This implies that we need $-f(-1)+f(1) = 0$, or $\mathbf{f}(-\mathbf{1}) = \mathbf{f}(\mathbf{1})$. A simple, non-linear function that satisfies this condition is the quadratic function $f(x) = x^2$. Let's define $Y = X^2$ and verify that $E[XY] = 0$.

$$\begin{aligned} E[XY] &= E[X \cdot X^2] = E[X^3] \\ &= \sum_{x \in \Omega_X} x^3 \cdot P(X = x) \\ &= (-1)^3 \cdot P(X = -1) + (0)^3 \cdot P(X = 0) + (1)^3 \cdot P(X = 1) \\ &= (-1) \cdot \frac{1}{3} + (0) \cdot \frac{1}{3} + (1) \cdot \frac{1}{3} \\ &= -\frac{1}{3} + 0 + \frac{1}{3} \\ &= 0 \end{aligned}$$

Since $E[XY] = 0$ and $E[X] = 0$, we have $\text{Cov}[X, Y] = 0$. The variable $Y = X^2$ is perfectly determined by X, yet they are linearly uncorrelated.

$\therefore$ Final Answer

A suitable function is $\mathbf{Y} = \mathbf{f}(\mathbf{X}) = \mathbf{X^2}$.

Graduate Level Explanation

Covariance and its normalized counterpart, the Pearson correlation coefficient ($\rho_{X,Y}$), are measures of **linear dependence** only. They quantify the strength and direction of a linear relationship between two variables. In this case, the relationship $Y = X^2$ is perfectly deterministic but also perfectly non-linear (it is parabolic). The distribution of X is symmetric about the origin, which means its first moment, $E[X]$, is zero. The covariance calculation hinges on the third moment, $E[X^3]$, because $E[XY] = E[X \cdot X^2] = E[X^3]$. Due to the symmetry of the underlying distribution of X, the negative contribution from $X = -1$ (i.e., $(-1)^3 = -1$) exactly cancels the positive contribution from $X = 1$ (i.e., $1^3 = 1$). This results in $E[X^3] = 0$ and thus $\text{Cov}[X, Y] = 0$. This is a canonical example of **uncorrelated dependence**, illustrating that a lack of linear correlation does not imply statistical independence. Independence implies zero covariance, but the converse is not true.

Explanation for a 5-Year-Old

Imagine two friends, Xavier (X) and Yasmin (Y). They have a secret rule: Yasmin's number is always Xavier's number multiplied by itself ($Y = X \times X$). So if Xavier picks -1, Yasmin's number must be 1. If Xavier picks 1, Yasmin's number must also be 1. Yasmin's number is perfectly predictable from Xavier's.

Now, "correlation" is a way to ask: when Xavier's number goes up, does Yasmin's number also tend to go up? Or does it go down? We are only checking if they follow a straight line together.

But look at their game. When Xavier goes from -1 to 0, his number goes up, but Yasmin's number goes down (from 1 to 0). But when Xavier goes from 0 to 1, his number goes up, and Yasmin's number also goes up (from 0 to 1). They don't stick to a simple "go up together" or "one goes up, one goes down" rule. Because their relationship isn't a straight line, the correlation calculator gets confused and says "zero correlation!" even though they are perfectly linked by their secret rule.