DSC 257R: Unsupervised learning

Homework 2

Distances and similarities

Note : Recall that for vectors in $\mathbb{R}^d$, the ℓ_1, ℓ_2, and ℓ_∞ *norms* are defined as follows.

- The ℓ_1 norm: $\|x\|_1 = \sum_{i=1}^d |x_i|$.
- The ℓ_2 (Euclidean) norm: $\|x\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$.
- The ℓ_∞ norm: $\|x\|_\infty = \max_i |x_i|$.

The ℓ_p *distance* between two points $x, x' \in \mathbb{R}^d$ is then the norm of $x - x'$, that is, $\|x - x'\|_p$.

1. For the point $x = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}$ in $\mathbb{R}^3$, compute the following.
   (a) $\|x\|_1$
   (b) $\|x\|_2$
   (c) $\|x\|_\infty$
2. Consider the following two points in $\mathbb{R}^4$:
$$x = \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}, \quad x' = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$
   (a) What is the ℓ_2 distance between them?
   (b) What is the ℓ_1 distance between them?
   (c) What is the ℓ_∞ distance between them?
3. *Comparing the ℓ_1, ℓ_2, and ℓ_∞ norms.*
   (a) Of all points $x \in \mathbb{R}^d$ with $\|x\|_\infty = 1$, which has the largest ℓ_1 norm? The largest ℓ_2 norm?
   (b) Of all points $x \in \mathbb{R}^d$ with $\|x\|_2 = 1$, which has the largest ℓ_1 norm? The largest ℓ_∞ norm?

   Here are some useful relationships between these three norms: for any $x \in \mathbb{R}^d$,
$$\|x\|_1 \;\geq\; \|x\|_2 \;\geq\; \|x\|_\infty$$
$$\|x\|_1 \;\leq\; \|x\|_2 \cdot \sqrt{d} \;\leq\; \|x\|_\infty \cdot d$$
   Something to think about if you have time (not for turning in): why do these inequalities hold? It should be possible to derive the first using algebra alone. For the second, one useful fact is the Cauchy-Schwarz inequality: that is, $|a \cdot b| \leq \|a\|_2 \|b\|_2$ for any vectors a, b.

4. *Weighted ℓ_2 norm.* Let $w_1, \ldots, w_d \geq 0$ be any non-negative numbers. Let $w = (w_1, \ldots, w_d)$ and define

$$\|x\|_w = \sqrt{\sum_{i=1}^{d} w_i x_i^2},$$

a weighted version of the ℓ_2 norm on $\mathbb{R}^d$. Sketch the region $\|x\|_w \leq 1$ (we would describe this as the *unit ball of the* $\|\cdot\|_w$ *norm*) for $d = 2$ and $w = (1, 4)$.

5. The following table specifies a distance function on the space $\mathcal{X} = \{A, B, C, D\}$. Is this a metric? Justify your answer.

	A	B	C	D
A	0	2	1	5
B	2	0	4	3
C	1	4	0	2
D	5	3	2	0

6. *KL divergence properties.* The KL divergence between two distributions p and q over a discrete (countable) set of outcomes $\mathcal{X}$ is given by

$$K(p, q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

(a) What is the largest this distance could be in the case $|\mathcal{X}| = 2$?

(b) Show by means of a small example, with $|\mathcal{X}| = 2$, that this distance function is *not* symmetric.

7. *Jaccard example.* What is the Jaccard similarity between the following two sets?

- $A = \{1, 3, 5, 7, 9\}$
- $B = \{2, 3, 5, 7\}$

8. *Jaccard similarity for text data.* When using the Jaccard similarity on text data, it is common to map a piece of text to the *set* of *bigrams* or *trigrams* in it. A *bigram* is a pair of words that appear consecutively in the text; a *trigram* is a triple of words that appear consecutively.

Consider, for example, the sentence $x =$ "a rose is a rose is a rose". It has

- bigrams $B(x) = \{$(a, rose), (rose, is), (is, a)$\}$ and
- trigrams $T(x) = \{$(a, rose, is), (rose, is, a), (is, a, rose)$\}$.

To compute the similarity between two sentences x and x', we could use the Jaccard similarity between $B(x)$ and $B(x')$, or between $T(x)$ and $T(x')$.

Compute the bigram-based Jaccard similarity between the following two sentences: "Napoleon was born in 1769" and "Napoleon was born when?". (You may assume the question mark is discarded when processing the second sentence.)

9. *Cosine similarity.*

(a) Compute the cosine similarity between $x = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ and $x' = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$.

(b) When is the cosine similarity between two vectors equal to zero? Give a precise characterization in terms of the angle between the vectors.

(c) Suppose that data lie in $\mathbb{R}^2$. Sketch the set of points whose cosine similarity to $x = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ is at least 0.9. In your picture, mark x; the rest of the sketch can be very rough, as long as it gives approximately the correct *shape* of the region.

Simple summary statistics

10. *Loaded dice.* A six-sided dice is loaded so that

$$\Pr(1) = \Pr(2) = \frac{1}{3}, \ \Pr(3) = \Pr(4) = \Pr(5) = \Pr(6) = \frac{1}{12}.$$

(a) Let the random variable X denote the outcome of rolling the dice. Compute the mean and median of X.

(b) Compute the variance and standard deviation of X.

(c) The dice is rolled 10 times, with outcomes

$$2, 5, 1, 4, 2, 2, 5, 6, 1, 2.$$

What is the *empirical distribution* corresponding to these observations?

(d) Continuing from (c), what are the mean, median, variance, and standard deviation of the empirical distribution?

11. For which of the following random variables do you think the mean might be significantly different from the median? Give a brief justification in each case; there is quite a bit of subjectivity in this problem, so what matters is your reasoning.

(a) Pick a person at random from a big city (e.g., New York), and let H denote their height.

(b) Let C denote the cost of their house.

(c) Let G denote their high school GPA.

(d) Let S denote their salary.

12. A random variable Z has mean -1 and standard deviation 2. What is $\mathbb{E}[Z^2]$?

13. Two random variables X, Y take values in $\{1, 2, 3\}$ and have a joint distribution given by the following table.

		Y 1	2	3
	1	0.1	0.2	0.05
X	2	0.1	0.1	0.1
	3	0.1	0.2	0.05

(a) Determine whether X and Y are independent or not. Justify your answer.

(b) Compute the covariance and correlation between X and Y.

14. *Correlation between linearly related variables.* Let X be any random variable that takes values in $\mathbb{R}$, and let $Y = aX + b$ for some constants a, b.
    (a) Give a formula for the covariance between X and Y, in terms of the variance of X.
    (b) What is the correlation between X and Y?
15. *Do deterministic relationships imply correlation?* Suppose $X \in \{-1, 0, 1\}$ takes each value with probability exactly 1/3. Specify a function f on $\{-1, 0, 1\}$ such that $Y = f(X)$ is uncorrelated with X.