CSE 151: Homework 2 Solutions

1. *Norms.*

   (a) $||x||_1 = \sum_{i=1}^{d} |x_i| = 1 + 2 + ... + d = \frac{d(d+1)}{2}$

   (b) $||x||_2 = \sqrt{\sum_{i=1}^{d} x_i^2} = \sqrt{1^2 + 2^2 + ... + d^2} = \sqrt{\frac{d(d+1)(2d+1)}{6}}$

   (c) $||x||_\infty = \max_i |x_i| = d$

2. The points in $\mathbb{R}^d$ whose ℓ_1 and ℓ_2 norm are both 1 are $\pm e_1, \pm e_2, \ldots, \pm e_d$, where e_i is the ith coordinate vector (i.e. all coordinates zero, except for a 1 in the ith position): a total of $2d$ points.

3. *Properties of metrics.* Recall that d is a distance metric if and only if it satisfies the following properties:

   (P1) $d(x, y) \geq 0$

   (P2) $d(x, y) = 0 \iff x = y$

   (P3) $d(x, y) = d(y, x)$ (symmetry)

   (P4) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

   (a) Hamming distance is a metric.

   (P1) $d(x, y) \geq 0$ because number of positions at which two strings differ can't be negative.

   (P2) $d(x, x) = 0$ because a string differs from itself at no positions. Also, if $x \neq y$, there will be at least one position where x and y differ and hence $d(x, y) > 0$.

   (P3) $d(x, y) = d(y, x)$ because x differs from y at exactly the same positions where y differs from x.

   (P4) Pick any $x, y, z \in \Sigma^m$. Let A denote the positions at which x, y differ: $A = \{i : x_i \neq y_i\}$, so that $d(x, y) = |A|$. Likewise, let B be the positions at which y, z differ and let C be the positions at which x, z differ.
   Now, if $x_i = y_i$ and $y_i = z_i$, then $x_i = z_i$. Thus $C \subseteq A \cup B$, whereupon
   $d(x, z) = |C| \leq |A| + |B| = d(x, y) + d(y, z)$.

   (b) Squared Euclidean distance is not a metric as it does not satisfy the triangle inequality. Consider the following three points in $\mathbb{R}$: $x = 1, y = 4, z = 5$.

   $$\begin{aligned} d(x, z) &= (1 - 5)^2 = 16 \\ d(x, y) &= (1 - 4)^2 = 9 \\ d(y, z) &= (4 - 5)^2 = 1 \end{aligned}$$

   Here $d(x, z) > d(x, y) + d(y, z)$.

4. If d_1 and d_2 are metrics, then so is $g(x, y) = d_1(x, y) + d_2(x, y)$. All four properties can be verified directly.

   (P1) $g(x, y) \geq 0$ because it is the sum of two nonnegative values.

   (P2) Pick any x, y.

   $$\begin{aligned} g(x, y) = 0 &\iff d_1(x, y) + d_2(x, y) = 0 \\ &\iff d_1(x, y) = 0 \text{ and } d_2(x, y) = 0 \text{ (since both nonnegative)} \\ &\iff x = y \end{aligned}$$

   (P3) $g(x, y) = d_1(x, y) + d_2(x, y) = d_1(y, x) + d_2(y, x) = g(y, x)$.

(P4) For any x, y, z,

$$\begin{aligned} g(x,z) &= d_1(x,z) + d_2(x,z) \\ &\leq (d_1(x,y) + d_1(y,z)) + (d_2(x,y) + d_2(y,z)) \\ &= (d_1(x,y) + d_2(x,y)) + (d_1(y,z) + d_2(y,z)) \\ &= g(x,y) + g(y,z) \end{aligned}$$

5. *Classification or Regression Problem*

   (a) Classification problem since dependent variable 'state of person' takes categorical values.

   (b) Regression problem since dependent variable 'speed of car' takes continuous values.

   (c) Regression problem since dependent variable 'GPA' takes continuous values.

   (d) Classification problem since dependent variable takes binary values.

6. *Probability space*

   (a) $\Omega = \{H, T\}$, $\Pr(H) = \Pr(T) = \frac{1}{2}$

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

   (c) $\Omega = \{H, T\}^{10}$, each outcome in Ω has probability $1/2^{10}$

7. *Probabilities*

   (a) $\Pr(d) = 1 - (\Pr(a) + \Pr(b) + \Pr(c)) = \frac{1}{8}$

   (b) $\Pr(A) = \Pr(a) + \Pr(b) + \Pr(c) = \frac{7}{8}$

   (c) $A \cap B = \{a, c\}$, so $\Pr(A \cap B) = \Pr(a) + \Pr(c) = \frac{3}{4}$

8. *Two fair dice*

   (a) For sum to be 10 given that first roll is 6, second roll should be 4.

$$\Pr(\text{sum} = 10 \mid \text{first roll} = 6) = \Pr(\text{second roll} = 4) = \frac{1}{6}$$

   (b) Using the formula for conditional probabilities,

$$\begin{aligned} \Pr(\text{sum} = 10 \mid \text{first roll is even}) &= \frac{\Pr(\text{sum} = 10 \text{ AND first roll is even})}{\Pr(\text{first roll is even})} \\ &= \frac{\Pr(\{(4,6),(6,4)\})}{1/2} = \frac{2/36}{1/2} = \frac{1}{9} \end{aligned}$$

   (c) Let A be the event when both rolls have same value: $A = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$.

$$\Pr(A) = \frac{6}{36} = \frac{1}{6}$$

9. *Bayes rule.* The information we are given is:

   - $\Pr(\text{male}) = \Pr(\text{female}) = 0.5$
   - $\Pr(\text{disease} \mid \text{male}) = 0.05$
   - $\Pr(\text{disease} \mid \text{female}) = 0.01$

Applying Bayes' rule,

$$\Pr(\text{male} \mid \text{disease}) = \Pr(\text{male}) \times \frac{\Pr(\text{disease} \mid \text{male})}{\Pr(\text{disease})} = \frac{1}{2} \times \frac{0.05}{0.05 * 0.5 + 0.01 * 0.5} = \frac{5}{6}$$

10. *Programming assignment*

    (a) Error rate with l_1 distance = 21.667%
    Error rate with l_2 distance = 23.33%

    (b) Confusion matrix for l_1 distance:

	NO	DH	SL
NO	14	0	2
DH	9	9	0
SL	1	1	24

    Confusion matrix for l_2 distance:

	NO	DH	SL
NO	12	1	3
DH	9	9	0
SL	1	0	25