James Zhao
HW2
A15939512

1 1

1.1 a

$\|(1,2,3,4)\|_1 = |1| + |2| + |3| + |4| = 10$

1.2 b

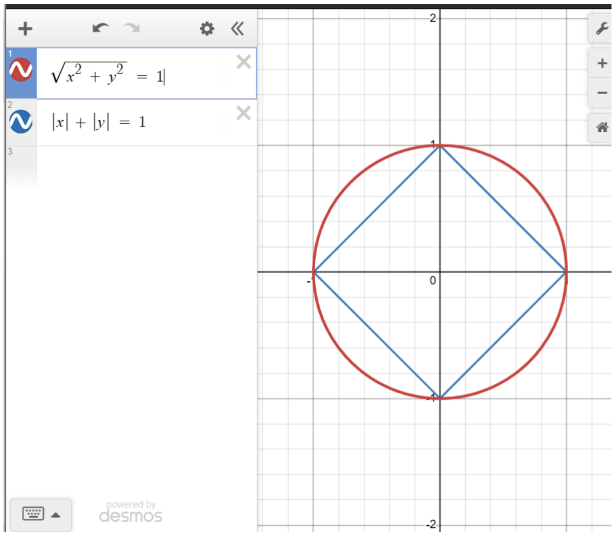
$\|(1,2,3,4)\|_2 = \sqrt{1^2+2^2+3^2+4^2} = \sqrt{1+4+9+16} = \sqrt{30}$

1.3 c

$\|(1,2,3,4,5)\|_\infty = max(abs([1,2,3,4])) = 4$

2 2

From lecture, we have seen that the unit 1-norm in R^2 looks like a ”diamond”, with corners on $(1,0), (-1,0), (0,1), (0,-1)$, and the unit 2-norm look slike a circle with radius 1. From the below drawing, it can be trivially inferred that the only intersections are the corners of the diamond, since at all points, the sides of the diamond are within the unit circle.



1 $\sqrt{x^2 + y^2} = 1$

2 $|x| + |y| = 1$

3

Thus, the intersection points are $(0,1), (0,-1), (1,0), (-1,0)$.

3 3

3.1 a

$d(x, y) = x - y$
Not a metric. Violates Property 1: Non-negativity.

$$d(1, 2) = 1 - 2 = -1 \ngeq 0$$

3.2 b

$Ham(x, y)$
The hamming distance is a metric.

- (a) $d(x, y) \geq 0$:
  Trivial because you can't differ by a negative number of positions
- (b) $d(x, y) = 0 iff x = y$:
  Forward: If x and y differ by 0 positions, then it is trivial to say that they are exactly the same.
  Backwards: If x and y are the same, then the two would not differ by any values. Thus, they would differ by 0 positions ($d(x, y) = 0$)
- (c) $d(x, y) = d(y, x)$:
  Since differing by positions is not order dependent (e.x. if x and y differ in position i, then y and x differ in position i), the hamming distance must be symmetric.
- (d) $d(x, z) \leq d(x, y) + d(y, z)$:
  If the hamming-distance is defined as the number of positions that x and z differ, then it is physically impossible to perform two separate hamming distance calculations that somehow enumerate a smaller number of differences.

3.3 c

Squared Euclidean Distance
Not a metric. Property (4) is violated:
$d(x, z) \leq d(x, y) + d(y, z)$
Choose $x = 1, z = -1, y = 0$: $d(x, z) = (-1 - 1)^2 = 4$
$d(x, y) = (-1 - 0)^2 = 1$
$d(y, z) = (y - -1)^2 = 1$
$d(x, z) = 4 \nleq 2 = d(x, y) + d(y, z)$

4 4

$d(p, q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$
$\frac{p(x)}{q(x)} = (\frac{1/2}{1/4}, \frac{1/4}{1/4}, \frac{1/8}{1/6}, \frac{1/16}{1/6}, \frac{1/16}{1/6}) = (2, 1, 3/4, 3/8, 3/8)$
$\log \frac{p(x)}{q(x)} = (1, 0, \log(3) - 2, \log(3) - 3, \log(3) - 3)$
$p(x) = (1/2, 1/4, 1/8, 1/16, 1/16)$
$d(p, q) = \frac{1}{2} + 0 + \frac{1}{8}(\log(3) - 2) + \frac{1}{16}(\log(3) - 3) + \frac{1}{16}(\log(3) - 3)$
$d(p, q) = \frac{1}{2} + \frac{1}{8}(\log(3) - 2) + \frac{1}{8}(\log(3) - 3))$
$d(p, q) = \frac{1}{2} + \frac{1}{8}\log(3) - \frac{1}{4} + \frac{1}{8}\log(3) - \frac{3}{8}$
$d(p, q) = -\frac{1}{8} + \frac{1}{4}\log(3)$

5 5

5.1 a

Classification. This is because the output space of {walking, sitting, running} is discrete and there is no significant natural ordering to the outputs.

5.2 b

Regression. The output space is continuous, and even we were predicting whole-number speeds, there is a natural ordering to the predictions.

5.3 c

Regression. The output space is continuous, and even we were predicting binned GPA's, there is a natural ordering to the predictions.

5.4 d

Classification. The output space is binary.

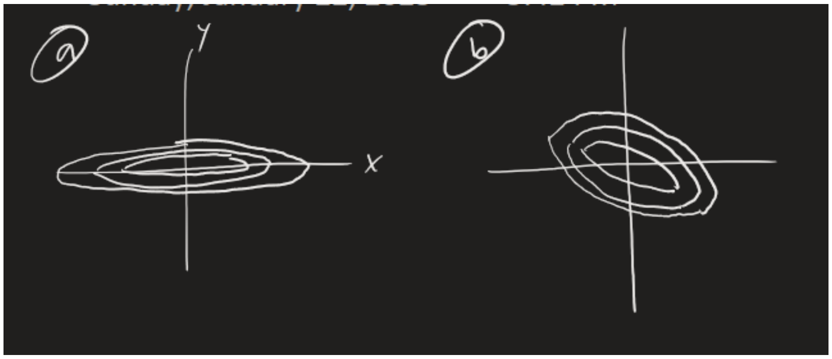
6 6

- 1. The (-) label points are very tightly distributed relative to the (+) points (and both distributions have the same mean), but the probability priors of (+) are much higher as to still maintain a larger value value of $\pi_j P_j(x)$. Thus in places in which (-) has a higher pdf value (such as at the mean), the value of $\pi_j P_j(x)$ is still lower, and the classifier assigns a label of (+).
- 2. In a degenerate case, the dataset could only have data points of (+)

7 7

Alch.	$\pi_1 P_1(x)$	$\pi_2 P_2(x)$	$\pi_3 P_3(x)$	Classification
12.0	0.33*small	**0.39 * 0.7**	0.28 * small	2
12.5	0.33*small	**0.39 * 0.6**	0.28 * 0.4	2
13.0	0.33 * 0.3	0.39 * 0.3	**0.28 * 0.8**	3
13.5	**0.33 * 0.8**	0.39 * small	0.28 * 0.6	1
14.0	**0.33 * 0.7**	0.39 * small	0.28 * 0.2	1

8 8



9 9

- (a)
  $q < 0$

- (b)
  $q = 0$

- (c)
  Let y be a linear function of x. That is $y = ax + b$, where $p = \text{Var}(x), r = \text{Var}(y), q = \text{Cov}(x, y)$.
  $\text{Var}(y) = \text{Var}(ax + b) = a^2\text{Var}(x)$
  $r = a^2p$
  $\text{Cov}(x, y) = \text{Cov}(x, ax + b) = a\text{Cov}(x, x) = a\text{Var}(x)$ $q = ap$
  Answer:
  Thus, we must have for some constant a, that $r = aq = a^2p$.
  One can also use some clever algebraic manipulation to realize that:
  $rp = a^2p^2 = q^2$
  $\pm\sqrt{rp} = q$

- (d)
  If $p = \text{Var}(x), q = \text{Cov}(x, y), r = \text{Var}(y)$ and y is a random variable with a constant value, then:
  $\text{Var}(y) = 0 = r$
  $\text{Cov}(x, ay + b) = \text{Cov}(x, 0y) = 0\text{Cov}(x, y) = 0$
  Thus, $r = 0, q = 0$

10 10

- **Case** $x < 0$:
  $P(X = x, Y = 1) = \pi_1 P_1(x) = \frac{7}{24}$
  $P(X = x, Y = 2) = \pi_2 P_2(x) = 0$
  $P(X = x, Y = 3) = \pi_3 P_3(x) = \frac{6}{24}$

- **Case** $x > 0$:
  $P(X = x, Y = 1) = \pi_1 P_1(x) = \frac{1}{24}$
  $P(X = x, Y = 2) = \pi_2 P_2(x) = \frac{4}{24}$
  $P(X = x, Y = 3) = \pi_3 P_3(x) = \frac{6}{24}$

Optimal Classifer:

$$h^*(x) : [-1, 1] \to \{1, 2, 3\},\ h^*(x) = \begin{cases} 1 & -1 \leq x < 0 \\ 3 & 0 \leq x \leq 1 \end{cases}$$

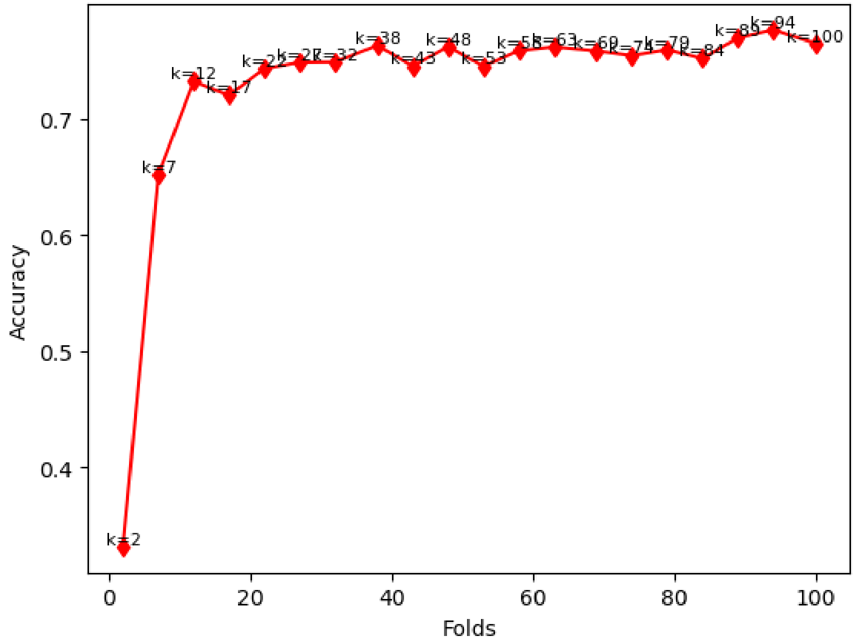
11 11

11.1 a

Accuracy: 0.770
Confusion Matrix:

	Label 1	Label 2	Label 3
Pred 1	52	5	3
Pred 2	3	54	14
Pred 3	4	12	31

11.2 b



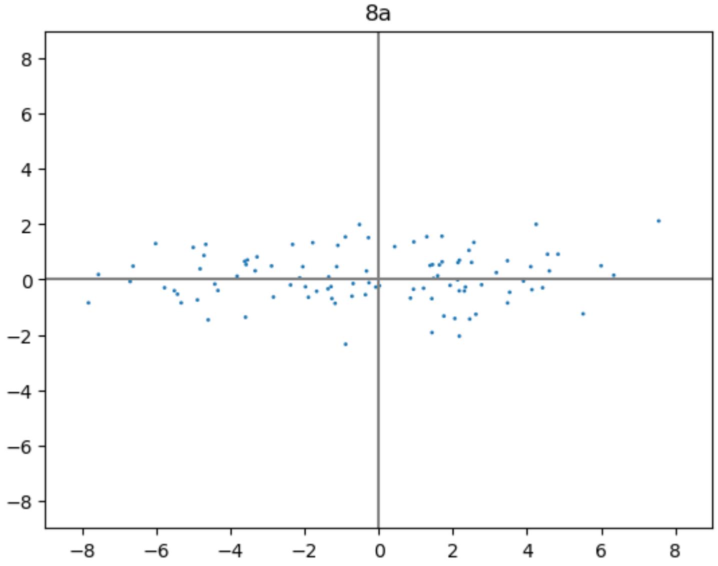
11.3 c

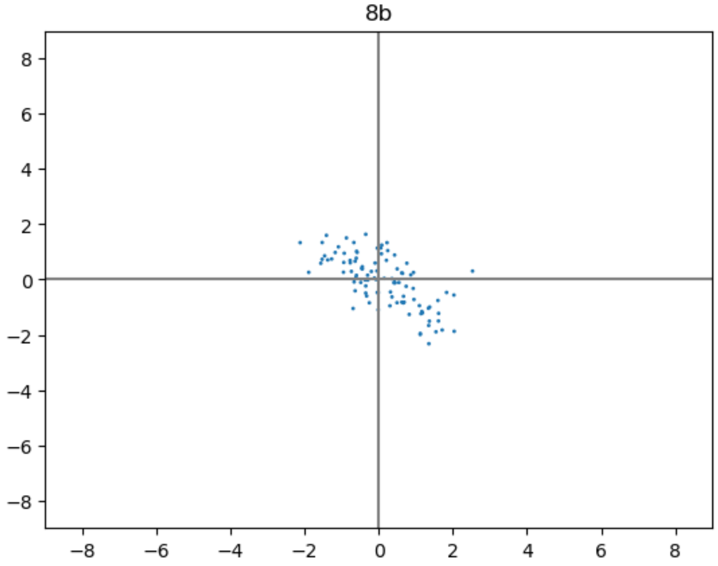
Accuracy: 0.949

	Label 1	Label 2	Label 3
Pred 1	59	5	0
Pred 2	0	62	0
Pred 3	0	4	48

The Normaliation helped a significant amount, increasing accuracy from 0.770 to 0.949.

12 12


8a
8
6
4
2
0
−2
−4
−6
−8
−8
−6
−4
−2
0
2
4
6
8


8b
8
6
4
2
0
−2
−4
−6
−8
−8
−6
−4
−2
0
2
4
6
8