DSE 220: Machine learning

Worksheet 1 — Solutions

1. *Casting an image into vector form.* A 10×10 greyscale image has 100 coordinates with 1 pixel per coordinate. Thus the corresponding vector has dimension $d = 100$.

2. *The length of a vector.* Say $x \in \mathbb{R}^d$ where $x_i = 1$ for $i = 1, \ldots, d$. Then by our Euclidean distance formula

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

3. *Euclidean distance.* $\sqrt{8}$.

4. *Euclidean distance.* $\sqrt{d}$.

5. *Accuracy of a random classifier.*
   (a) If there are four labels, then no matter what the correct label is, a random classifier has exactly a 25% chance of choosing it. Therefore it has an error rate of 75%.
   (b) The best constant classifier is the one that always returns label A. It is wrong whenever the label isn't A, which occurs 50% of the time. Thus the classifier that always returns label A has error rate 50%.

6. *Decision boundary of the nearest neighbor classifier.*
   (a) The label of (0.5, 0.5) is 2.
   (b) Let us call $x_1 = (0.5, 0.5)$ and $x_2 = (0.5, 1.5)$. If $x = (1.5, 0.5)$, we have

$$\begin{aligned}\|x - x_1\| &= \sqrt{(1.5 - 0.5)^2 + (0.5 - 0.5)^2} = 1 \\ \|x - x_2\| &= \sqrt{(1.5 - 0.5)^2 + (0.5 - 1.5)^2} = \sqrt{2}\end{aligned}$$

   Then x is closer to x_1, and the nearest neighbor will give the label of x_1, which is 2.
   (c) Now let $x = (2, 2)$. Then

$$\begin{aligned}\|x - x_1\| &= \sqrt{(2 - 0.5)^2 + (2 - 0.5)^2} = \sqrt{\frac{9}{2}} \\ \|x - x_2\| &= \sqrt{(2 - 0.5)^2 + (2 - 1.5)^2} = \sqrt{\frac{5}{2}}\end{aligned}$$

   Therefore, x is closer to x_2, and the nearest neighbor will give the label of x_2, which is 1.
   (d) This classifier will never predict label 3, since it has no points in that region.

(e) Consider a general point $x = (a, b)$. When will this point be closer to x_1 than x_2? This happens precisely when $\|x - x_1\| < \|x - x_2\|$ or, equivalently, when $\|x - x_2\|^2 - \|x - x_1\|^2 > 0$. Writing out

$$\begin{aligned}\|x - x_2\|^2 - \|x - x_1\|^2 &= ((a - 0.5)^2 + (b - 1.5)^2) - ((a - 0.5)^2 + (b - 0.5)^2) \\ &= (b - 1.5)^2 - (b - 0.5)^2 \\ &= \left(b^2 - 3b + \frac{9}{4}\right) - \left(b^2 - b + \frac{1}{4}\right) \\ &= 2 - 2b\end{aligned}$$

Now we see that the above is greater than 0 if and only if $b < 1$. Thus our 1-NN classifier classifies (a, b) as 1 if $b > 1$ and 2 if $b < 1$. Note that when $a < 1$, these predictions are correct, but when $a \geq 1$, they are incorrect. Therefore, if $X = (A, B)$ is drawn from the uniform distribution over the square, we have

$$\Pr(\text{1-NN is incorrect on } X) = \Pr(A \geq 1) = 0.5$$

Thus the error rate of the 1-NN is 50%.

7. *Programming exercise.*

8. We can work out the distances from the query to all the points.

Training point	Distance to query	label
(2,2)	$\sqrt{8.5}$	`star`
(2,4)	$\sqrt{2.5}$	`square`
(2,6)	$\sqrt{4.5}$	`star`
(4,2)	$\sqrt{6.5}$	`square`
(4,4)	$\sqrt{0.5}$	`star`
(4,6)	$\sqrt{2.5}$	`square`
(6,2)	$\sqrt{12.5}$	`square`
(6,4)	$\sqrt{6.5}$	`square`
(6,6)	$\sqrt{8.5}$	`star`

(a) The closest point to the query is (4,4). So the point will be classified as `star`.

(b) The 3 closest points to the query are (4,4), (2,4), and (4,6). So the point will be classified as `square`.

(c) The 4 closest points to the query are (4,4), (2,4), (4,6), and (2,6). These are split 50/50 between `star` and `square`. The next closest point is a tie between (4,2) and (6,4). However, since both of these have the same label (`square`), the 5-NN classifier will label the query `square` no matter how it breaks ties.

9. In 4-fold cross-validation, we evenly divide our data set into 4 subsets. We hold out one subset and train on the rest. In our case, this means each time we train we will do so with 7,500 data points.

10. For 1-NN, the LOOCV procedure will misclassify the two right points. Thus the LOOCV error for 1-NN will be 50%.

    For 3-NN, the LOOCV procedure will always label the test point +. Thus the LOOCV error for 3-NN will be 25%.

11. *Programming assignment*

(a) Error rate with l_1 distance = 17.74%
Error rate with l_2 distance = 22.58%

(b) Confusion matrix for l_1 distance:

	NO	DH	SL
NO	20	8	2
DH	0	0	0
SL	1	0	31

Confusion matrix for l_2 distance:

	NO	DH	SL
NO	19	9	2
DH	0	0	0
SL	3	0	29