DSC 255: Machine learning

Week 1 — Worksheet

Nearest neighbor classification

1. *Casting an image into vector form.* A 10×10 greyscale image is mapped to a d-dimensional vector, with one pixel per coordinate. What is d?

2. *The length of a vector.* The Euclidean (or L_2) length of a vector $x \in \mathbb{R}^d$ is

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

   where x_i is the ith coordinate of x. This is the same as the Euclidean distance between x and the origin. What is the length of the vector which has a 1 in every coordinate? Your answer may be a function of d.

3. *Euclidean distance.* What is the Euclidean distance between the following two points in $\mathbb{R}^3$?

$$(1, 2, 3), \quad (3, 2, 1)$$

4. *Accuracy of a random classifier.* A particular data set has 4 possible labels, with the following frequencies:

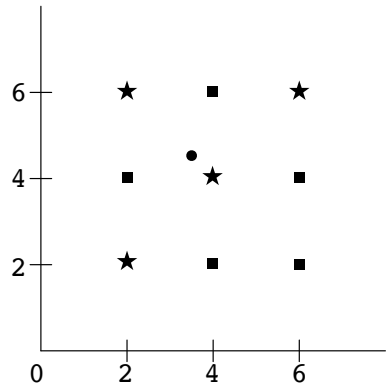
Label	Frequency
A	50%
B	20%
C	20%
D	10%

   (a) What is the error rate of a classifier that picks a label (A,B,C,D) at random, each with probability 1/4?

   (b) One very simple type of classifier just returns the same label, always.

   - What label should it return?
   - What will its error rate be?

5. In the picture below, there are nine training points, each with label either `square` or `star`. These will be used to guess the label of a query point at $(3.5, 4.5)$, indicated by a circle.



Suppose Euclidean distance is used.

(a) How will the point be classified by 1-NN? The options are `square`, `star`, or `ambiguous`.

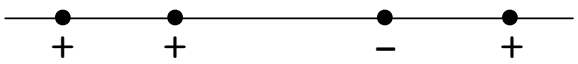
(b) By 3-NN?

(c) By 5-NN?

6. We decide to use 4-fold cross-validation to figure out the right value of k to choose when running k-nearest neighbor on a data set of size 10,000. When checking a particular value of k, we look at four different training sets. What is the size of each of these training sets?

7. An extremal type of cross-validation is *n-fold cross-validation* on a training set of size n. If we want to estimate the error of k-NN, this amounts to classifying each training point by running k-NN on the remaining $n-1$ points, and then looking at the fraction of mistakes made. It is commonly called *leave-one-out cross-validation* (LOOCV).

   Consider the following simple data set of just four points:



   What is the LOOCV error for 1-NN? For 3-NN?

Distance functions for ML

8. Consider the two points $x = (-1, 1, -1, 1)$ and $x' = (1, 1, 1, 1)$.

   (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?

9. For the point $x = (1, 2, 3, 4)$ in $\mathbb{R}^4$, compute the following.

   (a) $\|x\|_1$

   (b) $\|x\|_2$

   (c) $\|x\|_\infty$

10. For each of the following norms, consider the set of points with length ≤ 1. In each case, state whether this set is shaped like a *ball*, a *diamond*, or a *box*.
    (a) ℓ_2
    (b) ℓ_1
    (c) ℓ_∞
11. List all points in $\mathbb{R}^2$ with $\|x\|_1 = \|x\|_2 = 1$.
12. Which of these distance functions is a *metric*? If it is not a metric, state which of the four metric properties it violates.
    (a) Let $\mathcal{X} = \mathbb{R}$ and define $d(x, y) = x - y$.
    (b) Let Σ be a finite set and $\mathcal{X} = \Sigma^m$. The *Hamming distance* on $\mathcal{X}$ is $d(x, y) = \#$ of positions on which x and y differ.
    (c) Squared Euclidean distance on $\mathbb{R}^m$, that is, $d(x, y) = \sum_{i=1}^{m}(x_i - y_i)^2$. (It might be easiest to consider the case $m = 1$.)