Problem Set 1

Bookmark this page

Problems 1-9 correspond to "Nearest neighbor classification"

Problem 1

1/1 point (graded)

A 10×10 greyscale image is mapped to a d-dimensional vector, with one pixel per coordinate. What is d?

100 ✔

100

Submit

Problem 2

1/1 point (graded)

Which of these is the correct notation for 4-dimensional Euclidean space?

- $4\mathbb{R}$

- $4^{\mathbb{R}}$
- $\mathbb{R}^4$ ✔
- $\mathbb{R}4$

Submit

Problem 3

1/1 point (graded)

What is the Euclidean (also known as L_2) distance between the following two points in $\mathbb{R}^3$?

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

2.8284 ✔

2.8284

Submit

Problem 4

1/1 point (graded)

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?

- 1
- $\sqrt{d}$ ✔
- d
- d^2

Submit

Problem 5

1/1 point (graded)

Which of the following accurately describes the set of all points in $\mathbb{R}^3$ whose (Euclidean) length is ≤ 1?

- A ball centered at the origin. ✔
- A cube centered at the origin.
- A diamond centered at the origin.

Submit

Problem 6

1/1 point (graded)

What is the Euclidean distance between the following two points $x, x' \in \mathbb{R}^d$?

- x has all coordinates equal to 1.
- x' has all coordinates equal to -1.

- $\sqrt{d}$
- d
- $2\sqrt{d}$ ✔
- $\sqrt{2}d$

Submit

Problem 7

3/3 points (graded)

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 $1/4$?

0.75 ✔

0.75

b) One very simple type of classifier just returns the same label, always. What label should it return?

A ✔

c) What is the error rate of the classifier from b)?

0.5 ✔

0.5

Submit

Problem 8

2/2 points (graded)

A nearest neighbor classifier is built using a large training set, and then its performance is also evaluated on a separate test set.

- Which is likely to be smaller:

- training error ✔
- test error?

- Which is likely to be a better predictor of future performance:

- training error
- test error? ✔

Submit

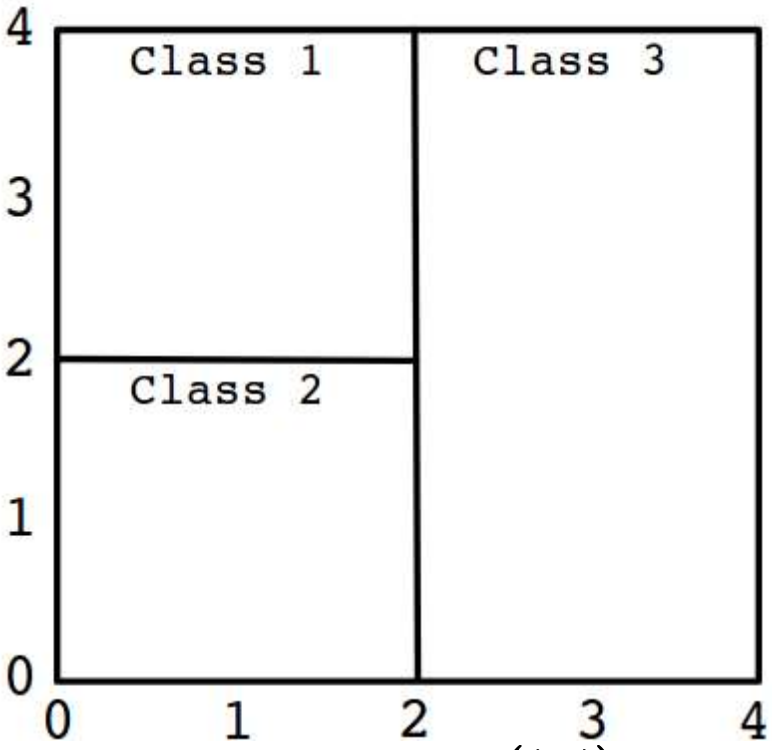
Problem 9

5/5 points (graded)

In this problem,

- The data space is $X = [0, 4]^2$: each point has two coordinates, and they lie between 0 and 4.
- The labels are $Y = \{1, 2, 3\}$.

The correct labels in different parts of X are as shown below.



a) What is the label of point $(1, 1)$?

2

2

For parts (b) through (e), assume you have a training set consisting of just two points, located at

$$(1, 1),\ (1, 3).$$

b) What label will the nearest neighbor classifier assign to point $(3, 1)$?

2

2

c) What label will the nearest neighbor classifier assign to point $(4, 4)$?

1 ✔

1

d) Which label will this classifier never predict?

3 ✔

3

e) Now suppose that when the classifier is used, the test points are uniformly distributed over the square X. What is the error rate of the 1-NN classifier?

1/2 ✔

$\frac{1}{2}$

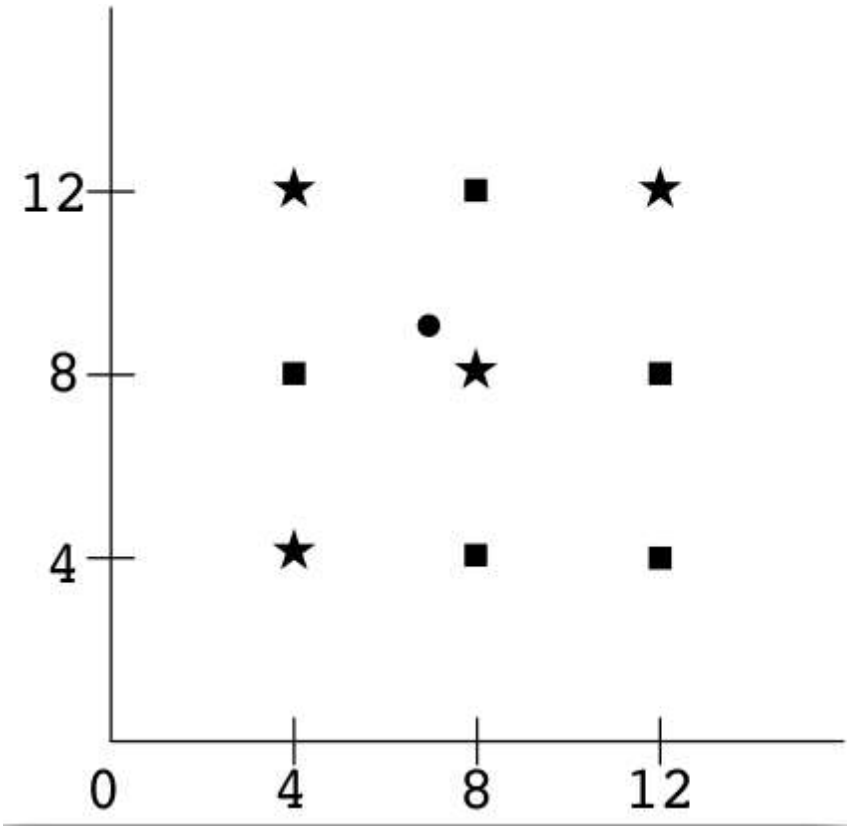
Submit

Problems 10-16 correspond to "Improving nearest neighbor"

Problem 10

3/3 points (graded)

In the picture below, there are nine training points, each with label either **square** or **star**. These will be used to predict the label of a query point at $(7, 9)$, indicated by a circle.



Suppose Euclidean (L_2) distance is used.

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

star ✔

b) By 3-NN?

square ✔

c) By 5-NN?

square ✔

Submit

Problem 11

1/1 point (graded)

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?

7500 ✔

$\mathbf{7500}$

Submit

Problem 12

2/2 points (graded)

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:

a) What is the LOOCV for 1-NN?

1/2 ✔

$\frac{1}{2}$

b) What is the LOOCV for 3-NN?

1/4 ✔

$\frac{1}{4}$

Submit

Problem 13

2/2 points (graded)

An emergency room wishes to build a classifier that will use basic information about entering patients to decide which ones are at high risk and need to be prioritized. As soon as a patient enters the facility, the following information is collected:

- age
- temperature
- heart rate
- nine-digit identification number

Suppose a nearest neighbor classifier is used, with L_2 distance.

a) Which of these four features is least relevant to the classification problem?

- age
- temperature
- heart rate

- nine-digit identification number ✔

b) Which of these four features is likely to have the greatest influence on the Euclidean distance function?

- age
- temperature
- heart rate
- nine-digit identification number ✔

Submit

Problem 14

1/1 point (graded)

Suppose we do nearest neighbor classification using a training set of n data points, and we do not use any special data structures to speed up the classifier. Which of the following correctly describes the running time for classifying a single test point?

- It does not depend on n.
- It is proportional to $\log n$.
- It is proportional to n. ✔

- It is proportional to n^2.

Submit

Problem 15

0 points possible (ungraded)

A bank decides to use nearest neighbor classification to decide which clients to offer a certain investment option. It has a database of clients that were already offered this product, along with information about whether these clients accepted or declined. This is the training set. It also has a long list of other clients who have not yet been offered this product; it wants to choose clients that are reasonably likely to accept, and will do so by using nearest neighbor using the training set.

Suppose the following information is available on each client:

- age
- annual income
- amount in bank
- zip code
- driver license number

Which of these features do you think would be most relevant to the classification problem? Would it make sense to use Euclidean distance, or would something else be better?

Submit

Problem 16

0 points possible (ungraded)

How might nearest neighbor be used in a recommender system? Suppose a movie streaming service keeps track of which movies its users watch and what their ratings are. Is there a way to use this information to make movie recommendations to users? What would the data space be, and what kind of distance function would be suitable?

Submit

Problems 17-22 correspond to "Useful distance functions for machine learning"

Problem 17

3/3 points (graded)

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

What is the L_2 distance between them?

2.8284 ✔

2.8284

What is the L_1 distance between them?

4 ✔

4

What is the L_∞ distance between them?

2

2

Submit

✔ Correct (3/3 points)

Problem 18

3/3 points (graded)

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

a) $\|x\|_1$

10

10

b) $\|x\|_2$

5.4772

5.4772

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

4

4

Submit

✔ Correct (3/3 points)

Problem 19

3/3 points (graded)

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

ball ✔

b) ℓ_1

diamond ✔

c) ℓ_∞

box ✔

Submit

✔ Correct (3/3 points)

Problem 20

1/1 point (graded)

How many points in $\mathbb{R}^2$ have $\|x\|_1 = \|x\|_2 = 1$?

4 ✔

4

Submit

✔ Correct (1/1 point)

Problem 21

3/3 points (graded)

Which of these distance functions is a ***metric***? If it is not a metric, select which of the four metric properties it violates (possibly more than one of them).

a) Let $X = \mathbb{R}$ and define $d(x, y) = x - y$.

- [ ] this function is a metric
- [x] not a metric; violates non-negativity (i.e. $d(x, y) \geq 0$)
- [x] not a metric; violates symmetry (i.e. $d(x, y) = d(y, x)$)
- [ ] not a metric; violates identity (i.e. $d(x, y) = 0$ iff $x = y$)
- [ ] not a metric; violates triangle inequality (i.e. $d(x, z) \leq d(x, y) + d(y, z)$)

✔

b) Let Σ be a finite set and $X = \Sigma^m$. The $Hamming\ distance$ on X is

$d(x, y) =$ # of positions on which x and y differ.

- [x] this function is a metric
- [ ] not a metric; violates non-negativity
- [ ] not a metric; violates symmetry
- [ ] not a metric; violates identity
- [ ] not a metric; violates triangle inequality

✔

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 $m = 1$.)

- [ ] this function is a metric
- [ ] not a metric; violates non-negativity
- [ ] not a metric; violates symmetry
- [ ] not a metric; violates identity
- [x] not a metric; violates triangle inequality

✔

Submit

✔ Correct (3/3 points)

Problem 22

1/1 point (graded)

Suppose d_1 and d_2 are two metrics on a space X. Define d to be their sum:

$$d(x,y) = d_1(x,y) + d_2(x,y).$$

Is d necessarily a metric? If not, which of the four metric properties might it violate?

- [x] this function is a metric
- [ ] not a metric; violates non-negativity (i.e. $d(x,y) \geq 0$)
- [ ] not a metric; violates symmetry (i.e. $d(x,y) = d(y,x)$)
- [ ] not a metric; violates identity (i.e. $d(x,y) = 0$ iff $x = y$)
- [ ] not a metric; violates triangle inequality (i.e. $d(x,z) \leq d(x,y) + d(y,z)$)

✔

Submit

✔ Correct (1/1 point)

Problem 23 corresponds to "A host of prediction problems"

Problem 23

4/4 points (graded)

For each of the following prediction tasks, state whether it is best thought of as a $\boldsymbol{classification}$ problem or a $\boldsymbol{regression}$ problem.

a) Based on sensors in a person's cell phone, predict whether they are walking, sitting, or running.

- (•) classification ✔
- ( ) regression

b) Based on sensors in a moving car, predict the speed of the car directly in front.

- ( ) classification
- (•) regression ✔

c) Based on a student's high-school SAT score, predict their GPA during freshman year of college.

- classification
- regression ✔

d) Based on a student's high-school SAT score, predict whether or not they will complete college.

- classification ✔
- regression

Submit

✔ Correct (4/4 points)

Learn About Verified Certificates


Download on the
App Store

ANDROID APP ON
Google play