Problem Set 1

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 ✔ **Answer:** 100

100

Answer

Correct:

Each coordinate of the vector corresponds to one pixel of the image. Thus the total number of coordinates is just the overall number of pixels, 100.

? **Hint (1 of 1):** How many pixels are there in an image, total? Next Hint

Submit

Answers are displayed within the problem

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

Answers are displayed within the problem

Problem 3

1/1 point (graded)

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

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

2.82

Answer: 2.828

2.82

Answer

Correct: The answer is $\sqrt{(1-3)^2+(2-2)^2+(3-1)^2}=\sqrt{8}.$

Submit

Answers are displayed within the problem

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

✔

Explanation

Plugging into the formula for L_2 distance, we get $\sqrt{1^2 + 1^2 + \cdots + 1^2} = \sqrt{d}$.

Submit

Answers are displayed within the problem

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.

Explanation

The points at L_2 distance exactly 1 lie on the surface of a sphere of radius 1. The points at distance ≤ 1 lie on or within this sphere. The other two cases (cube, diamond) correspond to L_∞ and L_1 distance, respectively.

Submit

Answers are displayed within the problem

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

Explanation

The two points differ by exactly 2 on each individual coordinate. Plugging into the formula for L_2 distance, we get $\sqrt{2^2 + 2^2 + \cdots + 2^2} = \sqrt{4d}$.

Submit

Answers are displayed within the problem

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 $1/4$? Give your answer as a number in the range $[0,1]$.

0.75

Answer: 0.75

0.75

Answer

Correct:

Whatever the correct label might be, the probability of randomly guessing it correctly is $1/4$ = 0.25. Thus the probability of being wrong is 0.75.

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

a

Answer: A

Answer

Correct: Yes, because this is the most frequently-occurring label.

c) What is the error rate of the classifier from b)? Give your answer as a number in the range $[0, 1]$.

0.5

✔ **Answer:** 0.5

0.5

Answer
Correct:
The classifier from b) is wrong whenever the label happens to be something other than its prediction. We can look at the table of frequencies to see how often this happens.

Submit

Answers are displayed within the problem

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?

✔

Explanation

In general, the error rate of a classifier on the training set tends to be an under-estimate of its true error in practice. A much better estimate is given by the error on a separate test set.

Submit

Answers are displayed within the problem

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 $(1, 1)$? Your answer should be 1, 2, or 3.

2

✔ **Answer:** 2

2

Answer
Correct: This point lies squarely in the region of class 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

✔ **Answer:** 2

2

Answer
Correct: The nearest neighbor to point $(3,1)$ is $(1,1)$, which has label 2.

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

1

✔ **Answer:** 1

1

Answer
Correct: The nearest neighbor to point $(4,4)$ is $(1,3)$, which has label 1.

d) Which label will this classifier never predict?

3

✔ **Answer:** 3

3

Answer
Correct:
There are only two data points in the training set, with labels 1 and 2. Thus no point will ever be assigned label 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? Give your answer as a number in the range $[0, 1]$.

0.5 ✔ **Answer:** 0.5

0.5

Answer
Correct:
The 1-NN classifier correctly classifies all points in class 1 or 2, and incorrectly classifies all points in class 3. Since class 3 occurs half the time, the error rate of the classifier is 0.5.

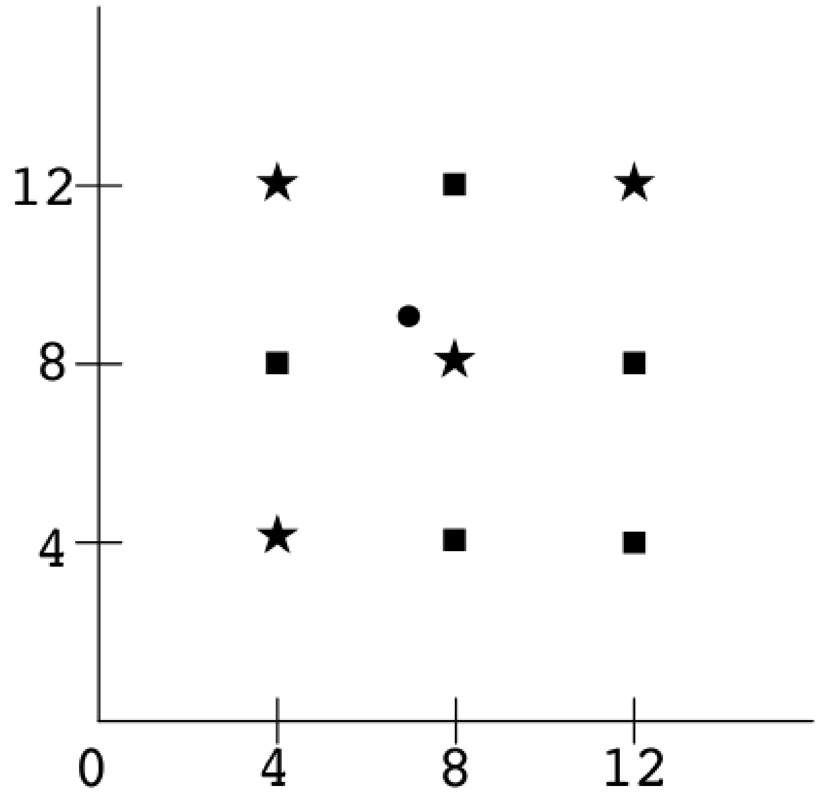
Submit

Answers are displayed within the problem

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

✔ **Answer:** star

b) By 3-NN?

square

✔ **Answer:** square

c) By 5-NN?

square

✔ **Answer:** square

Explanation

The nearest neighbors of this point are, in order: $(8, 8)$; $(4, 8)$ and $(8, 12)$; $(4, 12)$; $(8, 4)$ and $(12, 8)$.

Submit

Answers are displayed within the problem

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

Answer: 7500

7500

Answer

Correct:

We divide the training set into four equal-sized chunks, and take turns using three of these chunks for training and one for testing. Thus the chunks are of size 2500 and the training sets are of size 7500.

Submit

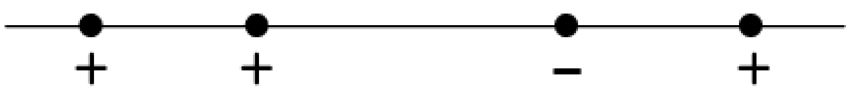
Answers are displayed within the problem

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? Your answer should be a number in the range $[0, 1]$.

2/4

Answer: 0.5

$\frac{2}{4}$

Answer

Correct:

The two points on the left are correctly classified by doing 1-NN on the remaining points, while the two on the right are incorrectly classified.

b) What is the LOOCV for 3-NN?

1/4 ✔ **Answer:** 0.25

$\frac{1}{4}$

Answer

Correct:

When doing 3-NN, every point is classified as +. Thus only one of them is misclassified.

? **Hint (1 of 1):** This is a small problem that you can solve by hand. No need for a computer!

Next Hint

Submit

Answers are displayed within the problem

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.

Explanation
We need to compute distances from the query point to each of the n data points, which takes time proportional to n.

Submit

Answers are displayed within the problem

Problem 15

0 points possible (ungraded)
(*This problem is ungraded and meant to be a thought exercise*. You are not expected to enter an answer. Feel free to discuss on the forums.)

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?

Explanation
One thing to consider is that the different features have very different ranges of values. Nearest neighbor using regular Euclidean distance might be dominated by the largest-valued features: how can this be fixed?

Submit

Answers are displayed within the problem

Problem 16

0 points possible (ungraded)
(*This problem is ungraded and meant to be a thought exercise*. You are not expected to enter an answer. Feel free to discuss on the forums.)

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?

Explanation
Here's one idea; there are many other possibilities. Suppose we represent each user by a list of movies he/she has watched and the ratings assigned to these movies. We then come up with a distance function between users: what might this be? We can then make predictions for a user based on his/her nearest neighbor.

Submit

Answers are displayed within the problem

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?

2.82

Answer: 2.828

2.82

Answer

Correct: It is $\sqrt{(-1-1)^2 + (1-1)^2 + (-1-1)^2 + (1-1)^2} = \sqrt{8}$.

What is the L_1 distance between them?

4

Answer: 4

4

Answer

Correct: It is $|-1-1| + |1-1| + |-1-1| + |1-1| = 4$.

What is the L_∞ distance between them?

2

Answer: 2

2

Answer

Correct: It is $\max\left(|-1-1|, |1-1|, |-1-1|, |1-1|\right) = 2$.

Submit

Answers are displayed within the problem

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

✔ **Answer:** 10

10

Answer

Correct: This is $|1| + |2| + |3| + |4|$.

b) $\|x\|_2$

5.47

✔ **Answer:** 5.477

5.47

Answer

Correct: This is $\sqrt{1^2 + 2^2 + 3^2 + 4^2}$.

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

4

✔ **Answer:** 4

4

Answer

Correct: This is $\max(1, 2, 3, 4)$.

Submit

Answers are displayed within the problem

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

Answer: ball

b) ℓ_1

diamond

Answer: diamond

c) ℓ_∞

box

Answer: box

Explanation
This is directly from one of the lecture slides.

Submit

Answers are displayed within the problem

Problem 20

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

4

Answer: 4

4

Explanation
The points with L_1 length 1 form a diamond. The points with L_2 length 1 form a circle. This circle and diamond intersect in exactly four points: the four corners of the diamond, $(1,0),(0,1),(-1,0),(0,-1)$.

Hint (1 of 1): Can you draw the set of points with L_1 length 1 and the set of points with L_2 length 1? We did this in lecture.

Next Hint

Submit

Answers are displayed within the problem

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,

$d\left(x,y\right)=\sum_{i=1}^{m}\left(x_i-y_i\right)^2$.

(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

✔

Explanation

For part (c), consider the three points $x=-1, y=0, z=1$. Then $d\left(x,z\right)=4$ whereas $d\left(x,y\right)=d\left(y,z\right)=1$, so $d\left(x,z\right)>d\left(x,y\right)+d\left(y,z\right)$.

Submit

Answers are displayed within the problem

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. $d(x, z) \leq d(x, y) + d(y, z)$)

✔

Explanation
To see that $d(x, y)$ is a metric, you need to systematically check that it satisfies all four properties. This involves algebraic manipulation.

Submit

Answers are displayed within the problem

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 *classification* problem or a *regression* problem.

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

- (selected) classification
- regression

✔

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

- classification
- (selected) regression

✔

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

- classification
- (selected) regression

✔

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

- (selected) classification
- regression

✔

Submit

Answers are displayed within the problem