CSE 151A: Machine learning

Homework 2

Instructions:

- You may discuss problems with your study group, but ultimately all your work (mathematical problems, code, experimental details) must be individual.
- Your solutions must be `typed up` and uploaded to Gradescope by 11.59PM on Thursday April 17. No late homeworks will be accepted under any circumstances, so you are encouraged to upload early.
- A subset of the problems will be graded.

Conceptual and mathematical problems

1. For the point $x = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$ in $\mathbb{R}^4$, compute the following.
   (a) $\|x\|_1$
   (b) $\|x\|_2$
   (c) $\|x\|_\infty$
2. *Comparing the ℓ_1, ℓ_2, and ℓ_∞ norms.*
   (a) Of all points $x \in \mathbb{R}^d$ with $\|x\|_\infty = 1$, which has the largest ℓ_1 norm? The largest ℓ_2 norm?
   (b) Of all points $x \in \mathbb{R}^d$ with $\|x\|_2 = 1$, which has the largest ℓ_1 norm? The largest ℓ_∞ norm?

   Here are some useful relationships between these three norms: for any $x \in \mathbb{R}^d$,

$$\|x\|_1 \;\geq\; \|x\|_2 \;\geq\; \|x\|_\infty$$
$$\|x\|_1 \;\leq\; \|x\|_2 \cdot \sqrt{d} \;\leq\; \|x\|_\infty \cdot d$$

   Something to think about if you have time (not for turning in): why do these inequalities hold? It should be possible to derive the first using algebra alone. For the second, one useful fact is the Cauchy-Schwarz inequality: that is, $|a \cdot b| \leq \|a\|_2 \|b\|_2$ for any vectors a, b.
3. The following table specifies a distance function on the space $\mathcal{X} = \{A, B, C, D\}$. Is this a metric? Justify your answer.

	A	B	C	D
A	0	2	1	5
B	2	0	4	3
C	1	4	0	2
D	5	3	2	0

4. Which of these distance functions is a *metric*? If it is a metric, just say so. 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 $m = 1$.)

5. The following vectors p and q specify probability distributions over a set of five outcomes.

$$p = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}\right)$$
$$q = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$$

   What is the KL divergence between them, $K(p, q)$?

6. 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.

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

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

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

7. *Covariance and correlation.* Random variables X, Y take on values in the range $\{-1, 0, 1\}$ and have the following joint distribution.

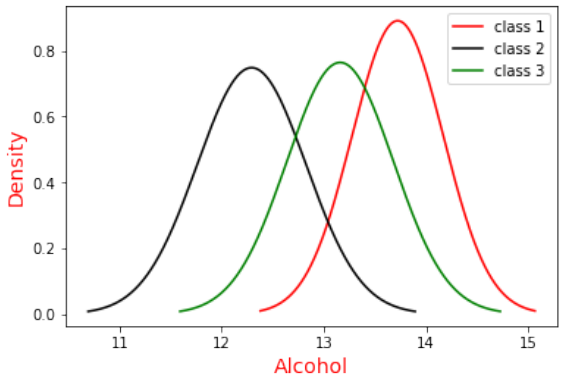
		Y		
		-1	0	1
X	-1	0	0	$1/3$
	0	0	$1/3$	0
	1	$1/3$	0	0

   (a) What is the covariance between X and Y?

   (b) What is the correlation between X and Y?

8. A generative approach is used for a binary classification problem (with classes $+$, $-$) and it turns out that the resulting classifier predicts $+$ at **all** points x in the input space. Why might this be?

9. *Winery classification.* For the winery example from lecture, the densities obtained are reproduced here:



The class probabilities are $\pi_1 = 0.33, \pi_2 = 0.39, \pi_3 = 0.28$. What labels would be assigned to the following points?

(a) 12.0

(b) 12.5

(c) 13.0

(d) 13.5

(e) 14.0

Programming problems

10. *Cross-validation for nearest neighbor classification.*

Download the `wine` data set from

`https://archive.ics.uci.edu/ml/datasets/wine`

This small data set has 178 observations. Each data point x consists of 13 features that capture visual and chemical properties of a bottle of wine. The label $y \in \{1, 2, 3\}$ indicates which of three wineries the bottle came from. The goal is to use the data to learn a classifier that can predict y from x.

Suppose we use the entire data set of 178 points for 1-NN classification with Euclidean distance. We would like to estimate the quality of this classifier.

(a) Use **leave-one-out cross-validation** (LOOCV) to estimate the **accuracy** of the classifier and also to estimate the 3×3 **confusion matrix**.

(b) Estimate the accuracy of the 1-NN classifier using k-fold cross-validation using 20 different choices of k that are fairly well spread out across the range 2 to 100. Plot these estimates: put k on the horizontal axis and accuracy estimate on the vertical axis.

(c) The various features in this data set have different ranges. Perhaps it would be better to normalize them so as to equalize their contributions to the distance function. There are many ways to do this; one option is to linearly rescale each coordinate so that the values lie in $[0, 1]$ (i.e. the minimum value on that coordinate maps to 0 and the maximum value maps to 1). Do this, and then re-estimate the accuracy and confusion matrix using LOOCV. Did the normalization help performance?