Boston University
Department of Electrical and Computer Engineering

ENG EC 414 Introduction to Machine Learning

HW 7 Solution

Issued: Fri 23 Oct 2020 **Due:** 10:00am Fri 30 Oct 2020 in Gradescope (non-code) + Blackboard

Important: Before you proceed, please read the documents pertaining to *Homework formatting and submission guidelines* in the Homeworks section of Blackboard. **In particular, for computer assignments you are prohibited from using any online code or built-in MATLAB functions except as indicated in the problem or skeleton code (when provided).**

Important: To obtain full grade, please clearly motivate all your answers.

Note: Problem difficulty = number of coffee cups ☕

Problem 7.1 [23pts] *(Voronoi cells)* Let $\mathbf{x}_A = (0,0)^\top$, $\mathbf{x}_B = (2,2)^\top$, and $\mathbf{x}_C = (0,4)^\top$ be three points in $\mathbb{R}^2$.

(a) [4pts] Hand-compute and sketch the equation of the set of all points that are equidistant from $\mathbf{x}_A$ and $\mathbf{x}_B$ when distance is measured by Euclidean distance.

(b) [9pts] Hand-compute and sketch the Voronoi-tessellation of $\mathbb{R}^2$ induced by the three points $\mathbf{x}_A, \mathbf{x}_B, \mathbf{x}_C$ when distance is measured by Euclidean distance.

(c) [10pts] ☕ Repeat part (a) for Manhattan (i.e., ℓ_1) distance.

Solution:

(a) [4pts] $\boxed{x_1 + x_2 - 2 = 0 \text{ and see Figure 1}}$ To determine the equation of all points $\mathbf{x}$ that are equidistant from $\mathbf{x}_A$ and $\mathbf{x}_B$ according to Euclidean distance we set:

$$\|\mathbf{x} - \mathbf{x}_A\|^2 - \|\mathbf{x} - \mathbf{x}_B\|^2 = 0$$

where $\|\cdot\|$ denotes Euclidean distance. Using the following identity for squared Euclidean distance:

$$\|\mathbf{u} - \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\mathbf{u}^\top\mathbf{v}$$

we get

$$\begin{aligned}\left(\|\mathbf{x}\|^2 + \|\mathbf{x}_A\|^2 - 2\mathbf{x}_A^\top\mathbf{x}\right) - \left(\|\mathbf{x}\|^2 + \|\mathbf{x}_B\|^2 - 2\mathbf{x}_B^\top\mathbf{x}\right) &= 0\\ \Rightarrow 2(\mathbf{x}_B - \mathbf{x}_A)^\top\mathbf{x} + \|\mathbf{x}_A\|^2 - \|\mathbf{x}_B\|^2 &= 0\end{aligned}$$

Substituting $\mathbf{x}_A = (0,0)^\top, \mathbf{x}_B = (2,2)^\top$ we get

$$2 \cdot (2-0, 2-0) \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + (0^2 + 0^2) - (2^2 + 2^2) = 0$$

which simplifies to: $x_1 + x_2 - 2 = 0$. Although this is not needed, we note in passing that the equation $2(\mathbf{x}_B - \mathbf{x}_A)^\top \mathbf{x} + \|\mathbf{x}_A\|^2 - \|\mathbf{x}_B\|^2 = 0$ can also be re-written as $(\mathbf{x}_B - \mathbf{x}_A)^\top \left(\mathbf{x} - \frac{1}{2}(\mathbf{x}_A + \mathbf{x}_B)\right) = 0$ which reveals the geometric fact that points equidistant from $\mathbf{x}_A$ and $\mathbf{x}_B$ (according to Euclidean distance) are perpendicular to $(\mathbf{x}_B - \mathbf{x}_A)$ (the vector joining the two points) relative to the mid-point of $\mathbf{x}_A$ and $\mathbf{x}_B$ as the origin.

(b) [9pts]

$$\mathcal{V}(\mathbf{x}_A) = \left\{(x_1, x_2)^\top : x_1 + x_2 \leq 2\right\} \cap \left\{(x_1, x_2)^\top : x_2 \leq 2\right\}$$
$$\mathcal{V}(\mathbf{x}_B) = \left\{(x_1, x_2)^\top : x_1 + x_2 \geq 2\right\} \cap \left\{(x_1, x_2)^\top : -x_1 + x_2 \leq 2\right\}$$
$$\mathcal{V}(\mathbf{x}_C) = \left\{(x_1, x_2)^\top : x_2 \geq 2\right\} \cap \left\{(x_1, x_2)^\top : -x_1 + x_2 \geq 2\right\}$$
(See Figure 1)

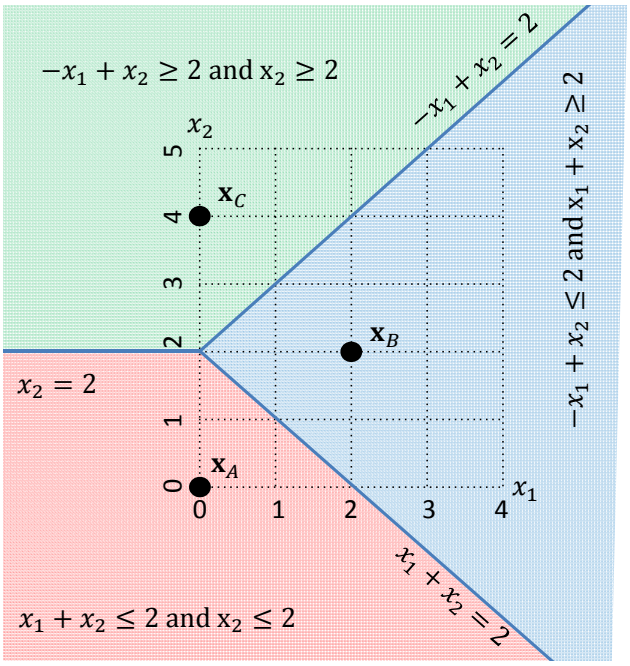


Figure 1: *Euclidean-distance Voronoi-tessellation of* $\mathbb{R}^2$ *induced by the points* $\mathbf{x}_A, \mathbf{x}_B, \mathbf{x}_C$.

The Voronoi cell of a point, say $\mathbf{x}_A$, is the set of all points $\mathbf{x}$ which are closer to $\mathbf{x}_A$ than to any other point. This can be expressed as

$$\begin{aligned}
\mathcal{V}(\mathbf{x}_A) &= \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_A\| \leq \|\mathbf{x} - \mathbf{x}_B\| \text{ and } \|\mathbf{x} - \mathbf{x}_A\| \leq \|\mathbf{x} - \mathbf{x}_C\|\} \\
&= \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_A\| \leq \|\mathbf{x} - \mathbf{x}_B\|\} \cap \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_A\| \leq \|\mathbf{x} - \mathbf{x}_C\|\} \\
&= \left\{\mathbf{x} : 2(\mathbf{x}_B - \mathbf{x}_A)^\top \mathbf{x} + \|\mathbf{x}_A\|^2 - \|\mathbf{x}_B\|^2 \leq 0\right\} \cap \left\{\mathbf{x} : 2(\mathbf{x}_C - \mathbf{x}_A)^\top \mathbf{x} + \|\mathbf{x}_A\|^2 - \|\mathbf{x}_C\|^2 \leq 0\right\} \\
&= \left\{(x_1, x_2)^\top : x_1 + x_2 \leq 2\right\} \cap \left\{(x_1, x_2)^\top : x_2 \leq 2\right\}
\end{aligned}$$

where we used the expression that we derived from part (a) for $\|\mathbf{x} - \mathbf{x}_A\|^2 - \|\mathbf{x} - \mathbf{x}_B\|^2$ (a similar one holds for the pair $(\mathbf{x}_A, \mathbf{x}_C)$), we substituted numerical values for the coordinates of points $\mathbf{x}_A, \mathbf{x}_B$, and $\mathbf{x}_C$, and simplified the answer. Thus we see that the Voronoi cell for point $\mathbf{x}_A$ is obtained by intersecting the regions for each pair of points $(\mathbf{x}_A, \mathbf{x}_B)$ and $(\mathbf{x}_A, \mathbf{x}_C)$, where the region for each pair is the set of points

that are closer to $\mathbf{x}_A$ than to the other point in the pair. The Voronoi regions for the other two points can be computed in the same way and is depicted in Figure 1.

(c) [10pts]

$$\{(x_1, x_2) : x_1 + x_2 = 2, x_1, x_2 \in [0, 2]\}$$
$$\cup \{(x_1, x_2) : 0 \le x_1, 2 \le x_2\}$$
$$\cup \{(x_1, x_2) : 2 \le x_1, x_2 \le 0\}$$
(See Figure 2)

To determine the equation of all points $\mathbf{x}$ that are equidistant from $\mathbf{x}_A$ and $\mathbf{x}_B$ according to Manhattan

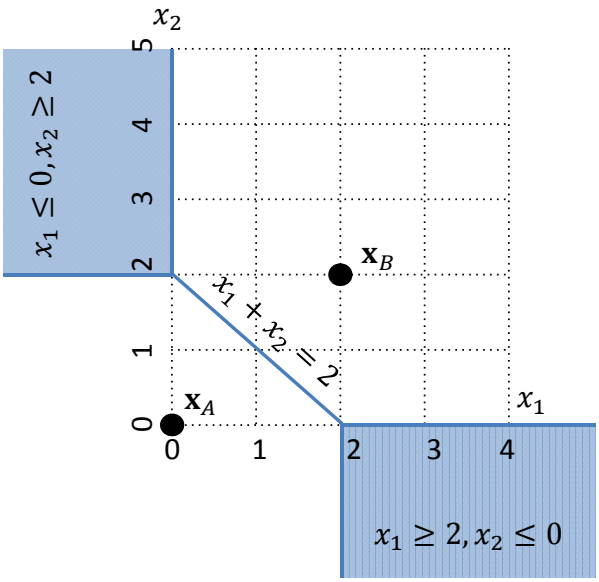


Figure 2: *Points that are equidistant from $\mathbf{x}_A, \mathbf{x}_B$ according to Manhattan-distance are depicted in blue color.*

distance we set:

$$|x_1 - 0| + |x_2 - 0| = |x_1 - 2| + |x_2 - 2|$$

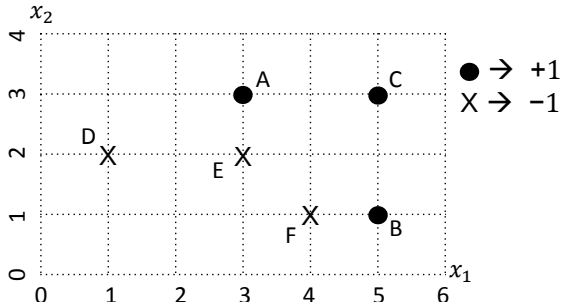
and solve for coordinates (x_1, x_2). To do this we note that $|a - b| = (b - a)$ if $a \le b$ and $|a - b| = (a - b)$ if $b \le a$. Thus we break up the plane $\mathbb{R}^2$ into 9 regions corresponding to whether the coordinates x_1 and x_2 lie to the left or right of the first and second coordinates of points $\mathbf{x}_A$ and $\mathbf{x}_B$ respectively and then solve the above equation. Arranging our calculations into a 3×3 table we have

	$x_1 < 0$	$0 \le x_1 \le 2$	$2 < x_1$
$2 < x_2$	$-x_1 + x_2 = 2 - x_1 + x_2 - 2$ $\Rightarrow 0 = 0$ Solution: $x_1 < 0,\ 2 < x_2$	$x_1 + x_2 = 2 - x_1 + x_2 - 2$ $\Rightarrow x_1 = 0$ Solution: $x_1 = 0,\ 2 < x_2$	$x_1 + x_2 = x_1 - 2 + x_2 - 2$ $\Rightarrow 0 = -4$ No solution
$0 \le x_2 \le 2$	$-x_1 + x_2 = 2 - x_1 + 2 - x_2$ $\Rightarrow x_2 = 2$ Solution: $x_1 < 0,\ x_2 = 2$	$x_1 + x_2 = 2 - x_1 + 2 - x_2$ $\Rightarrow$ Solution: $x_1 + x_2 = 2$	$x_1 + x_2 = x_1 - 2 + 2 - x_2$ $\Rightarrow x_2 = 0$ Solution: $2 < x_1,\ x_2 = 0$
$x_2 < 0$	$-x_1 - x_2 = 2 - x_1 + 2 - x_2$ $\Rightarrow 0 = 2$ No solution	$x_1 - x_2 = 2 - x_1 + 2 - x_2$ $\Rightarrow x_1 = 2$ Solution: $x_1 = 2,\ x_2 < 0$	$x_1 - x_2 = x_1 - 2 + 2 - x_2$ $\Rightarrow 0 = 0$ Solution: $2 < x_1,\ x_2 < 0$

The final solution can be compactly expressed as the union of 3 regions as described analytically in the boxed equations at the beginning of the solution and can be visualized graphically in Figure 2.

Problem 7.2 [7pts] *(Choosing the best k in kNN classification)* Consider a training set for binary classification with positive examples (label +1) at $A = (3,3)^\top, B = (5,1)^\top, C = (5,3)^\top$ and negative examples (label −1) at $D = (1,2)^\top, E = (3,2)^\top, F = (4,1)^\top$. To select a value of k for k-NN classification (using Euclidean distance), we perform leave-one-out cross-validation (LOOCV). For $k = 1, 3, 5$, list the validation examples that will be misclassified and the corresponding LOOCV error (0/1 loss). What is the best value of k (among these choices)?

Solution:



fold number	cross-validation point (label)	nearest neighbor # in training set (label)					majority label		
		#1	#2	#3	#4	#5	1NN	3NN	5NN
1	A (+)	E (-)	C (+)	D/F (-)	F/D (-)	B (+)	(-)	(-)	(-)
2	B (+)	F (-)	C (+)	E (-)	A (+)	D (-)	(-)	(-)	(-)
3	C (+)	A/B (+)	B/A (+)	E/F (-)	F/E (-)	D (-)	(+)	(+)	(-)
4	D (-)	E (-)	A (+)	F (-)	B/C (+)	C/B (+)	(-)	(-)	(+)
5	E (-)	A (+)	F (-)	D (-)	B/C (+)	C/B (+)	(+)	(-)	(+)
6	F (-)	B (+)	E (-)	A/C (+)	C/A (+)	D (-)	(+)	(+)	(+)
						Number of misclassified points:	4	3	6

k	misclassified validation points	LOOCV error
1	A, B, E, F	4/6
3	A, B, F	3/6
5	A, B, C, D, E, F	6/6

Best choice: $k = 3$.

Note: Validation points A, C, D, F have multiple pairs of points in the training set that are equidistant from them. However, in each such pair both points have the same label. Thus, even though there is ambiguity in the choice of the k-th nearest neighbor when there are two equidistant choices, either choice results in the same majority label decision and does not affect the LOOCV error calculations. Randomly choosing among equidistant choices with equal probability will also give the same LOOCV error.

Problem 7.3 [10pts] In this problem, we will implement k-NN and use it on the cats vs dogs dataset (see Lecture 12). Warning: the training/validation/test split in the code is random, so you'll get different results every time you run it.

(a) [5pts] As said in class, k-NN does not require training. So, implement directly the k-NN prediction algorithm we saw in class in a Matlab function with prototype

```
[yhat] = predict_knn(X, y, Xtest, k)
```

where `yhat` is a column vector of predictions for the n testing samples, `X` is the matrix of the training samples $\in \mathbb{R}^{m \times d}$, `y` is the vector of labels $\in \mathbb{R}^{m}$, `Xtest` is the matrix of the testing samples $\in \mathbb{R}^{n \times d}$, and k the number of neighbors to consider. There is no skeleton code. Hint: to take the majority vote of n numbers in $\{-1, 1\}$ it is enough to sum them and take the sign. Also, `sort` in Matlab sorts vector of numbers.

Solution: See code.

(b) [3pts] Let's now use the function in the point above to select the best k using a validation set. You decide a good range of value of k to try. Plot the 0/1 loss on the validation set with respect k.

Solution:

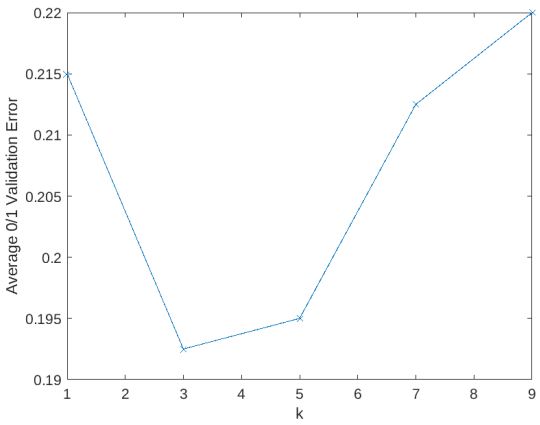


Figure 3: Average 0/1 validation loss for different choices of k in k-NN.

First, we decide to use $k = 1, 3, 5, 7, 9$ and possibly change the range of values based on the validation results. Note that you should not consider even values of k otherwise you cannot compute a reliable majority vote. Figure 3 shows how the average 0/1 validation error depends on k in k-NN. Given the behaviour of the validation error, there is no need to consider other values of k. Note that your figure might be different because the train/validation/test split is random.

(c) [2pts] Test on the test set using the best k found using the validation set and report the 0/1 error.

Solution: The value of k that gives the best validation error is $k = 3$. On the test set, this values gives a 0/1 average test loss of 0.2325. As we said in class, even if the data is very high-dimensional, the decent performance is probably due to the fact the intrinsic dimension of the cats and dogs images from this dataset is rather small.

Code-submission via Blackboard: You must prepare 1 files: `predict_knn.m` for Problem 7.3(a). Place them in a **single** directory which should be zipped and uploaded into Blackboard. Your directory must be named as follows: <`yourBUemailID`>_`hwX` where `X` is the homework number. For example, if your BU email address is `charles500@bu.edu` then for homework number 7 you would submit a single directory named: `charles500_hw7.zip` which contains all the MATLAB code (and only the code).

Three corresponding skeleton code files are provided for your reference. Reach-out to the TAs via Piazza and their office/discussion hours for questions related to coding.