Boston University
ENG EC 414 Introduction to Machine Learning
Exam 2 Solution
Released on Wednesday, 11 November, 2020 (120 minutes, 41 points + 2 bonus points)
Submit to Gradescope

- *There are* 6 *problems plus 1 bonus one.*
- *Unless explicitly written, for each part you must clearly outline the key steps and provide proper justification for your calculations in order to receive full credit.*
- *You can use any material from the class (slides, discussions, homework solutions, etc.), but you cannot look for solutions on the internet. Also, be aware of the limited time.*

Problem 2.1 [10pts] Consider the following 6 training feature vectors: $\mathbf{x}_A = (2,0)^\top, \mathbf{x}_B = (3,2)^\top, \mathbf{x}_C = (4,1)^\top, \mathbf{x}_D = (1,2)^\top, \mathbf{x}_E = (2,4)^\top, \mathbf{x}_F = (0,3)^\top$ with *class* labels $+1, +1, +1, -1, -1, -1$ respectively.

(a) [2pts] Hand-plot the training set and properly label axes and key points.

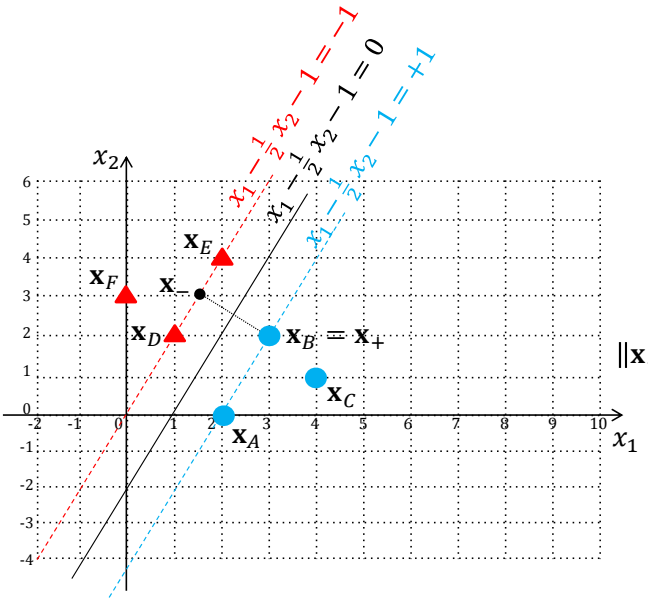
(b) [3pts] Compute the coordinates of the point in the convex hull of negative training feature vectors which is closest (in Euclidean distance) to the point $\mathbf{x}_B$.

(c) [3pts] Hand-compute the parameters of the hard-margin SVM hyperplane for this training set in *canonical* form. Sketch the SVM hyperplane.

(d) [2pts] Compute the size of the margin of the hard-margin SVM.

Solution:

(a) [2pts] The training set is plotted in the figure below.



$$\mathbf{w}_{\text{SVM}} = \left(1, -\tfrac{1}{2}\right)^{\mathrm{T}}$$
$$b_{\text{SVM}} = -1$$
$$\mathbf{x}_+ = \mathbf{x}_B$$
$$\mathbf{x}_- = \left(1\tfrac{2}{5}, 2\tfrac{4}{5}\right)^{\mathrm{T}}$$
$$\|\mathbf{x}_+ - \mathbf{x}_-\| = \tfrac{4}{\sqrt{5}}$$

(b) [3pts] $\boxed{(7/5, 14/5)^\top}$ From the figure, the point in the convex hull of negative feature vectors that is closest to $\mathbf{x}_B$ is $\mathbf{x}_-$ which lies on the line joining $\mathbf{x}_D$ and $\mathbf{x}_E$ which also passes through the origin. So let $\mathbf{x}_- = \gamma \mathbf{x}_D$. By the orthogonality principle, $(\mathbf{x}_B - \mathbf{x}_-) \perp \mathbf{x}_D \Rightarrow \mathbf{x}_D^\top(\mathbf{x}_B - \gamma \mathbf{x}_D) = 0 \Rightarrow \gamma = \mathbf{x}_D^\top \mathbf{x}_B / \|\mathbf{x}_D\|^2 = 7/5$. Thus $\mathbf{x}_- = (7/5, 14/5)^\top$.

(c) [3pts] $\boxed{\mathbf{w}_{SVM} = (1, -1/2)^\top, b_{SVM} = -1}$ Since $\mathbf{x}_B = \mathbf{x}_+ = (3, 2)^\top$ and $\mathbf{x}_- = (7/5, 14/5)^\top$,

$$\mathbf{w}_{SVM} = 2(\mathbf{x}_+ - \mathbf{x}_-)/\|\mathbf{x}_+ - \mathbf{x}_-\|^2 = (1, -1/2)^\top, \quad b_{SVM} = 1 - \mathbf{w}_{SVM}^\top \mathbf{x}_B = -1.$$

(d) [2pts] $\boxed{\text{Margin} = \frac{4}{\sqrt{5}}}$ It can be derived as $\|\mathbf{x}_+ - \mathbf{x}_-\|$ or equivalentely as $\frac{2}{\|\mathbf{w}_{SVM}\|}$.

Problem 2.2 [4pts] (True/False) For each statement, say if it is true or false. No justification is necessary here.

(i) [2pts] (SVM)

(a) The objective function of hard-margin SVM is convex.

(b) The hyperplane of an SVM $\mathbf{w}_{SVM}$ is a linear combination of training samples in both hard- and soft-margin SVMs.

(c) If we run a soft-margin SVM on a linearly separable dataset, we always get training error zero.

(d) SVMs cannot be kernelized due to the nonlinear hinge loss term.

(ii) [2pts] In polynomial regression with squared loss, as the degree increases, the *training* error:

(a) is non-decreasing

(b) is non-increasing

(c) increases first and then decreases

(d) decreases first and then increases

Solution:

(i) [2pts] (SVM)

(a) True: the objective function is convex and the constraints linear.

(b) True: we said it in class (without proof).

(c) False: it depends on the trade-off parameter C, so we might get a solution with some misclassified samples but with bigger margin than the solution with 0 errors.

(d) False: SVM can be kernelized!

(ii) [2pts] In polynomial regression with squared loss, as the degree increases, the *training* error:

(a) False

(b) True: we are considering strictly larger hypothesis classes.

(c) False

(d) False

Problem 2.3 [4pts] Say if each statement is true or false and **explain your choices to get full credit**.
Let $K(\mathbf{u}, \mathbf{v})$ a kernel, then:

(a) K is symmetric, that is $K(\mathbf{u}, \mathbf{v}) = K(\mathbf{v}, \mathbf{u}), \forall \mathbf{u}, \mathbf{v}$

(b) K is non-negative, that is $K(\mathbf{u}, \mathbf{v}) \geq 0, \ \forall \mathbf{u}, \mathbf{v}$

(c) There exists a unique function ϕ such that $K(\mathbf{u}, \mathbf{v}) = \phi(\mathbf{u})^\top \phi(\mathbf{v})$

(d) $K(\mathbf{u}, \mathbf{u}) \geq 0, \ \forall \mathbf{u}$.

Solution: Let $K(\mathbf{u}, \mathbf{v})$ a kernel, then:

(a) True, because the kernel is an inner product and inner products are always symmetric

(b) False, the linear kernel can return negative values too.

(c) False, the uniqueness is the critical point. Indeed, I can always construct another transformation adding a bunch of coordinates with 0 values, that is something like $\psi(\mathbf{u}) = [\psi(\mathbf{u}); 0]$

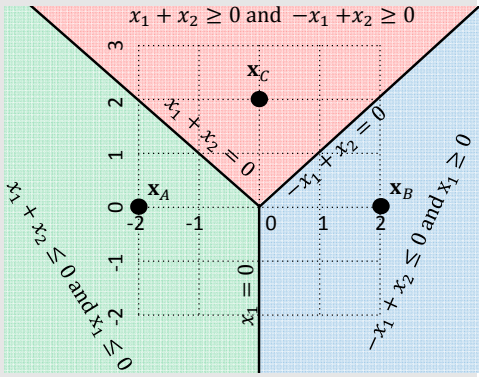
(d) True, because $K(\mathbf{u}, \mathbf{u}) = \phi(\mathbf{u})^\top \phi(\mathbf{u}) = \|\phi(\mathbf{u})\|^2 \geq 0$.

Problem 2.4 [6pts] Let $\mathbf{x}_A = (2, 0)^\top$, $\mathbf{x}_B = (-2, 0)^\top$, and $\mathbf{x}_C = (0, 2)^\top$ be three points in $\mathbb{R}^2$. Compute and sketch the Voronoi-tessellation of $\mathbb{R}^2$ induced by the three points for **Euclidean** distance.

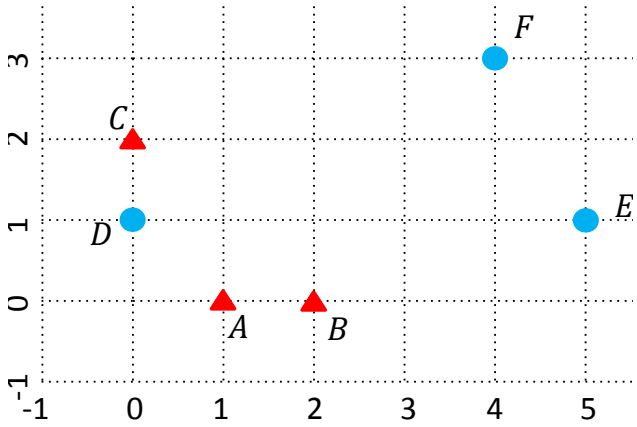
Solution:

Equation of points $\mathbf{x}$ at equal Euclidean distance from $\mathbf{u}$ and $\mathbf{v}$ is given by $\|\mathbf{x} - \mathbf{u}\|^2 = \|\mathbf{x} - \mathbf{v}\|^2 \Leftrightarrow 2(\mathbf{u} - \mathbf{v})^\top \mathbf{x} = \|\mathbf{u}\|^2 - \|\mathbf{v}\|^2 \Rightarrow$ $\mathcal{V}(\mathbf{x}_A) = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_A\| \leq \|\mathbf{x} - \mathbf{x}_B\|\} \cap \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_A\| \leq \|\mathbf{x} - \mathbf{x}_C\|\} = \left\{(x_1, x_2)^\top : x_1 + x_2 \leq 0\right\} \cap \left\{(x_1, x_2)^\top : x_1 \leq 0\right\}$. Similarly for $\mathbf{x}_B$ and $\mathbf{x}_C$.

$$\mathcal{V}(\mathbf{x}_A) = \left\{(x_1, x_2)^\top : x_1 + x_2 \leq 0\right\} \cap \left\{(x_1, x_2)^\top : x_1 \leq 0\right\}$$
$$\mathcal{V}(\mathbf{x}_B) = \left\{(x_1, x_2)^\top : -x_1 + x_2 \leq 0\right\} \cap \left\{(x_1, x_2)^\top : x_1 \geq 0\right\}$$
$$\mathcal{V}(\mathbf{x}_C) = \left\{(x_1, x_2)^\top : x_1 + x_2 \geq 0\right\} \cap \left\{(x_1, x_2)^\top : -x_1 + x_2 \geq 0\right\}$$



Problem 2.5 [7pts] A training set with points A, B, C in one class and D, E, F in another is shown in



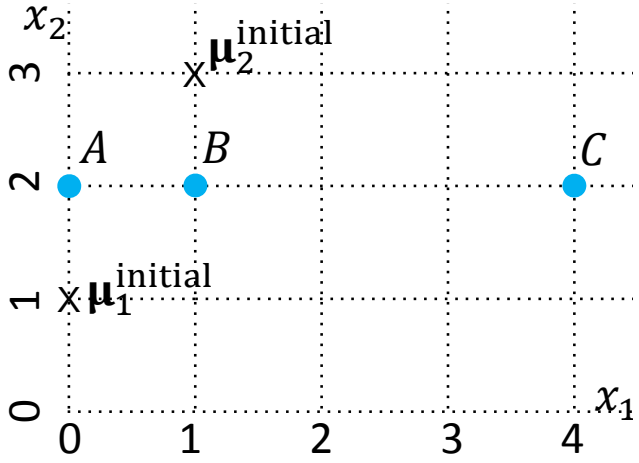
the figure. To select a value of k for k-NN classification using **Euclidean** distance, we perform leave one out cross-validation (LOOCV). (i) For each training point, list all its 5 nearest neighbors in the order of increasing Euclidean distance. (ii) For $k = 1, 3, 5$, list the validation examples that will be misclassified and the corresponding LOOCV error. Identify the best value of k among these choices.

Solution:

cross-validation point	nearest neighbor #				
	#1	#2	#3	#4	#5
A	B	D	C	E	F
B	A	D	C	E	F
C	D	A	B	F	E
D	C	A	B	F	E
E	F	B	A	D	C
F	E	B	C	A	D

k	misclassified points	LOOCV Error
1	C, D	$2/6 = 1/3$
3	D, E, F	$3/6 = 1/2$
5	A, B, C, D, E, F	$6/6 = 1$
Best choice: $k = 1$.		

Problem 2.6 [10pts] For the dataset of 3 points A, B, C and initial mean vectors $\boldsymbol{\mu}_1^{\text{initial}}, \boldsymbol{\mu}_2^{\text{initial}}$ shown



in the figure, compute the clusters and mean vectors found by running the 2-means algorithm (Euclidean distance) until convergence. In particular, complete the following table:

Iteration number	**Clusters**	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$
(Initialization)		$(0, 1)^\top$	$(1, 3)^\top$
...	...	...	...

Solution:

Iteration number	**Clusters**	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$
(Initialization)		$(0, 1)^\top$	$(1, 3)^\top$
1	$\{A\}, \{B, C\}$	$(0, 2)^\top$	$(2.5, 2)^\top$
2	$\{A, B\}, \{C\}$	$(0.5, 2)^\top$	$(4, 2)^\top$
(converged)			

Problem 2.7 [Bonus, 2pts] Consider x, y integers where $x, y \leq 100$. Show that $K(x, y) = \min(x, y)$ is a valid kernel finding the corresponding transformation ϕ. Hint: Consider $\phi : \mathbb{R} \to \mathbb{R}^{100}$.

Solution: In order to solve this problem, it is essentialy to make use of all the hidden hints in the text. For example, the fact that the numbers are integers, less or equal then 100, and the dimension of the feature space is also 100 are very important. The transformation is the following: Associate to each integer x a vector that contains 1 in the first x coordinates and 0 in the others. So, for example, $\phi(3) = [1, 1, 1, 0, \ldots, 0] \in \mathbb{R}^{100}$ and $\phi(5) = [1, 1, 1, 1, 1, 0, \ldots, 0] \in \mathbb{R}^{100}$. In this way, the inner product in this transformed space corresponds exactly to the min between the two numbers.