ECE M146
Introduction to Machine Learning
Instructor: Lara Dolecek

Spring 2021 Exam 2
May 17, 2021
Maximum score: 100 points

You have 2 hours to submit your work **directly on Gradescope under the Exam 2 submission link**.
Please read and carefully follow all the instructions.

Instructions

- The exam is accessible from 10 am PST on May 17th to 10 am PST on May 18th. Once you open the exam, you will have 2 hours to upload your work (therefore open the exam at least 2 hours before the closing time).
- This exam is open book, open notes. You are allowed to consult your own class notes (homework, discussion, lecture notes, textbook). You are not allowed to consult with each other or solicit external sources for help (e.g., an online forum).
- For each question, start a new sheet of paper. Therefore, the number of pages of your scan should be at least the number of questions. It is ok to write multiple parts of a question on one sheet. Properly erase or cross out any scratch work that is not part of the answer.
- Please submit your exam through the corresponding submission link on Gradescope.
- Make sure to include your **full name** and **UID** in your submitted file.
- Make sure to **show all your work**. Unjustified answers will be at a risk of losing points.
- Calculators are allowed for matrix inversion, entropy calculation and etc.
- **Policy on the Academic Integrity**
  "During this exam, you are **disallowed** to contact with a fellow student or with anyone outside the class who can offer a solution e.g., web forum."
  Please write the following statement on the first page of your answer sheet. You will **lose 10 points** if we can not find this statement.

I $\underline{\quad YourName \quad}$ with UID ____ have read and understood the policy on academic integrity.

1. (20 pts) k**-Nearest Neighbors**

In the following questions, you will consider a k-nearest neighbor classifier using L2 (Euclidean) distance or L1 (Manhattan) distance as the distance metric on a binary classification task. We assign the class of the test point to be the class of the majority of the k nearest neighbors. In the following dataset, the red dots and blue dots represent 2D features for data belonging to two classes. Note that the data is designed such that there will be no tie affecting your decision in the following problems.

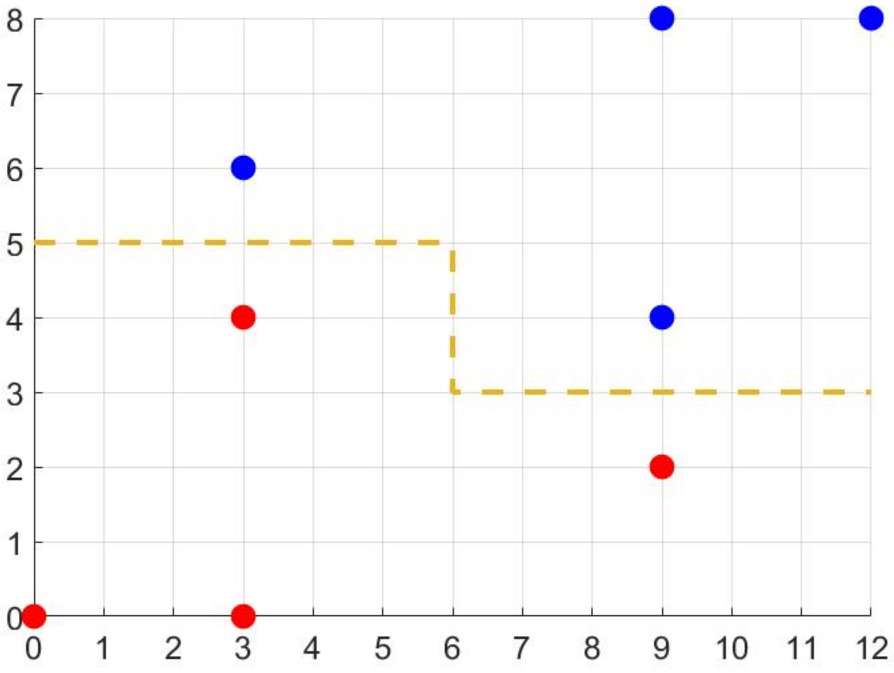


Figure 1: k-Nearest Neighbors

(a) (6 pts) The L2 distance between two vectors u and v is defined as the L2 norm of their difference, i.e., $\|u - v\|_2$. Using the L2 distance as your distance metric, in the above figure, sketch the 1-nearest neighbor decision boundary.

Solution:

The decision boundaries are shown above.

(b) (7 pts) Using the L2 distance as your distance metric, find the leave-one-out cross-validation accuracy for $k = 1$ and $k = 3$.

Solution:

The leave-one-out cross validation accuracy for $k = 1$ is $\frac{1}{2}$. The four points in the middle are classified erroneously. With $k = 3$, the point at $(3, 4)$ and $(9, 4)$ are now classified correctly, resulting in a validation accuracy of $\frac{3}{4}$.

(c) (7 pts) The L1 distance between two N dimensional vectors u and v is defined as $\sum_{i=1}^{N} |u_i - v_i|$. Using the L1 distance as your distance metric, find the leave-one-out cross-validation accuracy for $k = 1$ and $k = 3$.

Solution:

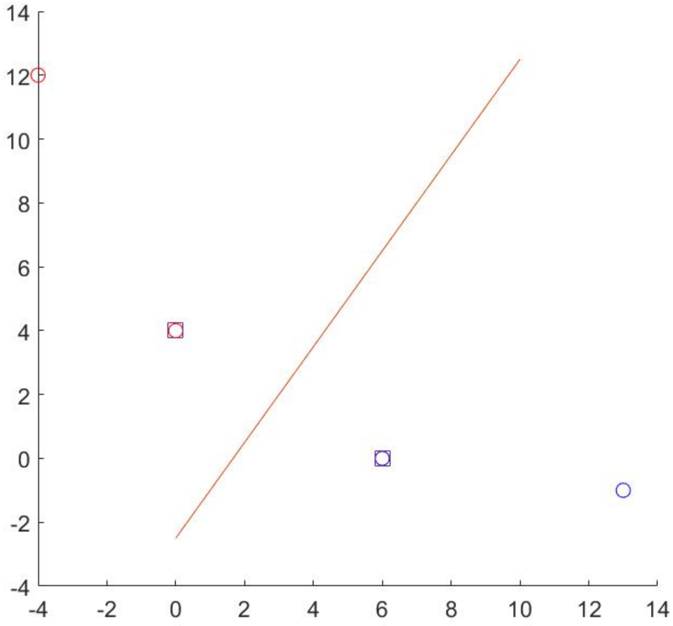
The leave-one-out cross validation accuracy for $k = 1$ is $\frac{1}{2}$. The four points in the middle are classified erroneously. With $k = 3$, the point at $(3, 4)$ and $(9, 4)$ are still classified erroneously resulting in a validation accuracy of $\frac{1}{2}$. Take the point $(3, 4)$ as an example, when using the L1 distance, its 3 nearest neighbors are $(3, 6)$, $(9, 4)$ and $(3, 0)$ instead of $(3, 6)$, $(0, 0)$ and $(3, 0)$ when using the L2 distance.

2. (20 pts) **Support Vector Machine**

You are given the following data set which is comprised of $x^{(i)} \in \mathbb{R}^2$ and $y^{(i)} \in \{-1, 1\}$.

i	$x_1^{(i)}$	$x_2^{(i)}$	y_i
1	-4	12	1
2	0	4	1
3	6	0	-1
4	13	-1	-1

(a) (3 pts) Plot the data by hand. Is the data linearly separable?
Solution: Yes, data is linearly separable. Plot is shown below.



(b) (5 pts) Suppose you are asked to find the maximum margin separating hyperplane of the form $[w_1, w_2][x_1, x_2]^T + b = 0$. Write down the (primal) optimization problem **explicitly** using only w_1, w_2 and b.
Solution:
The optimization problem is as follows:

$$\begin{aligned} \min_{w_1, w_2, b} \quad & w_1^2 + w_2^2 \\ s.t. \quad & -4w_1 + 12w_2 + b \geq 1, \\ & 4w_2 + b \geq 1, \\ & -6w_1 - b \geq 1, \\ & -13w_1 + w_2 - b \geq 1. \end{aligned}$$

(c) (6 pts) Look at the data and mark the support vectors by inspection. Find and plot the maximum margin separating hyperplane using basic geometry.

Solution:
The two support vectors are $[0,4]^T$ and $[6,0]^T$. The line that has normal vector $[-6,4]$ and also passes through the midpoint of support vectors $([3,2]^T)$ is $-6x_1 + 4x_2 + 10 = 0$.

(d) (6 pts) Solve the dual problem for the Lagrange multipliers α_i and use your dual solution to find the w and b of the primal problem.
Solution:
Since we only have two support vectors, only the Lagrange multiplier corresponding to the support vectors are non-zero. Let α_2 denote the Lagrange multiplier for $x^{(2)}$ and similarly α_3 for $x^{(3)}$. From the condition $\sum_{i=1}^{4} \alpha_i y_i = 0$, we get $\alpha_2 = \alpha_3 = \alpha_0$. Write down the objective of the dual problem of the SVM

$$\begin{aligned} W(\alpha) &= \sum_{i=1}^{4} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{4} y_i y_j \alpha_i \alpha_j x^{(i)T} x^{(j)} \\ &= 2\alpha_0 - \frac{1}{2}\alpha_0^2 x^{(2)T} x^{(2)} + \alpha_0^2 x^{(2)T} x^{(3)} - \frac{1}{2}\alpha_0^2 x^{(3)T} x^{(3)} \\ &= 2\alpha_0 - 26\alpha_0^2. \end{aligned}$$

Maximizing $W(\alpha)$ over α_0, we get $\alpha_3 = \alpha_2 = \alpha_0 = \frac{1}{26}$. Using $w = \sum_{m \in \mathcal{S}} \alpha_m y^{(m)} x^{(m)}$, we get $w = \frac{1}{26}[-6,4]^T$. To find b, recall that

$$y^{(i)} \left(w^T x^{(i)} + b \right) = 1$$

for any support vectors $x^{(i)}$. Using any support vector, we get $b = \frac{5}{13}$. The w and b we find by solving the dual problem is a scaled version of $[w_1, w_2]^T$ and w_0 in part (c). These solutions therefore give the same separating hyperplane.

3. (20 pts) **Kernels**

Suppose you are given 6 **one-dimensional** points: 3 points with label -1: $x_1 = -1, x_2 = 0, x_3 = 1$ and 3 points with label $+1$: $x_4 = -3, x_5 = -2, x_6 = 3$. In this question, we compare the performance of a linear classifier with and without kernel.

(a) (3 pts) Is the data linearly separable?
Solution: No. If we draw those points on a line, we can not separate points with label -1 and points with label $+1$ with a vertical line.

(b) (5 pts) Consider a linear classifier of the form $f(x) = \text{sign}(w_1 x + w_0)$. Write down the optimal value of w and its classification accuracy on the above 6 points. There might be more than one optimal solution, writing down one of them is enough.

Solution: One optimal solution is $w = [-1, -3/2]^T$ which gives the accuracy of 5/6.

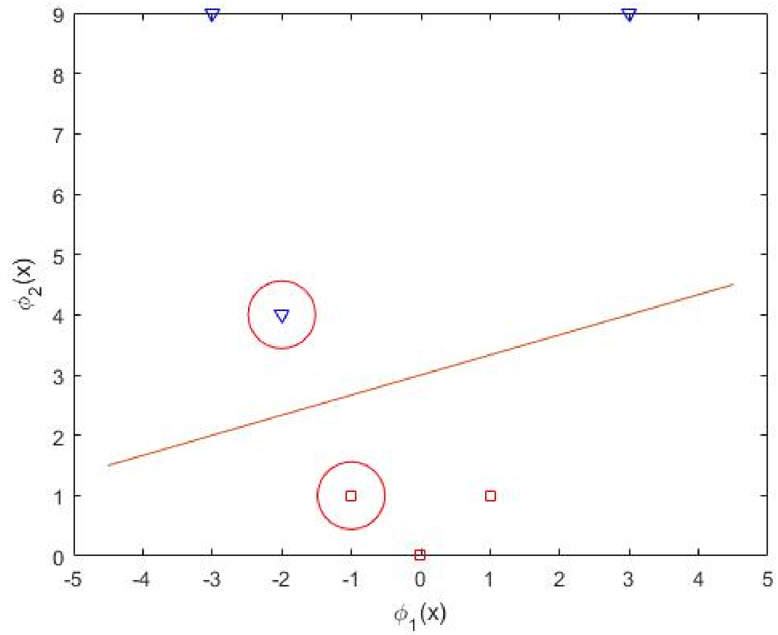
(c) (6 pts) Given two samples u and v in $\mathbb{R}$, define the kernel $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ as

$$K(u, v) = uv(1 + uv).$$

Find the corresponding feature map $\phi(x)$. Hint: This kernel maps the data to features in $\mathbb{R}^2$.

Solution: $\phi(x) = [x, x^2]^T$. One can verify that $\phi(u)^T \phi(v) = uv(1 + uv)$.

(d) (6 pts) Apply $\phi(x)$ to the data and plot the points in the induced feature space $\mathbb{R}^2$. Are these points linearly separable now? What is the highest classification accuracy that we can achieve with a linear classifier of the form $f(\phi(x)) = \text{sign}(w^T \phi(x) + b)$?
Solution: Yes. Since the data is linearly separable, we can get 100% classification accuracy with a linear classifier. The line below shows the optimal margin linear classifier solved by SVM.



4. (20 pts) **Naïve Bayes Classifier**

In the previous exam, we used the decision tree for medical diagnostics. Now let us build a Naïve Bayes Classifier based on the same data. There are 8 patients who have been asked to answer yes (1) or no (0) for several symptoms. Their answer and whether they are diagnosed with $XDisease$ are summarized in the following table:

Patient #	Fatigue	Fever	Cough	Headache	XDisease
1	0	0	0	1	0
2	0	1	0	1	1
3	1	0	0	1	0
4	1	1	1	1	1
5	0	1	1	0	1
6	1	1	0	0	1
7	0	0	0	0	0
8	1	0	1	1	1

(a) (15 pts) Train a Naïve Bayes classifier by calculating the maximum likelihood estimate of the class priors and the class conditional distributions. Namely, calculate the maximum likelihood estimate of the prior distribution $P(\text{XDisease})$; and calculate the maximum likelihood estimate of $P(X|\text{XDisease})$ for $X \in \{\text{Fatigue, Fever, Cough, Headache}\}$, $\text{XDisease} \in \{0, 1\}$.

Solution: The maximum likelihood of class priors are just the relative frequency of each class. We therefore have:

$$P(\text{XDisease} = 0) = \frac{3}{8}, P(\text{XDisease} = 1) = \frac{5}{8}.$$

The class conditional distribution can be estimated similarly by calculating the relative frequency of the features conditional on the class. We get:

$$P(\text{Fatigue} = 0|\text{XDisease} = 0) = \frac{2}{3}, P(\text{Fatigue} = 0|\text{XDisease} = 1) = \frac{2}{5};$$
$$P(\text{Fever} = 0|\text{XDisease} = 0) = 1, P(\text{Fever} = 0|\text{XDisease} = 1) = \frac{1}{5};$$
$$P(\text{Cough} = 0|\text{XDisease} = 0) = 1, P(\text{Cough} = 0|\text{XDisease} = 1) = \frac{2}{5};$$
$$P(\text{Headache} = 0|\text{XDisease} = 0) = \frac{1}{3}, P(\text{Headache} = 0|\text{XDisease} = 1) = \frac{2}{5}.$$

(b) (5 pts) Predict whether patient #9 has XDisease using the trained Naive Bayes classifier.

Solution: For XDisease = 0,

$$P(\text{XDisease} = 0)P(\text{Fatigue=1,Fever=0, Cough=1, Headache=0}|\text{XDisease} = 0)$$
$$=\frac{3}{8} \times \frac{1}{3} \times 1 \times 0 \times \frac{1}{3} = 0.$$

For XDisease = 1,

$$P(\text{XDisease = 1})P(\text{Fatigue=1,Fever=0, Cough=1, Headache=0}|\text{XDisease = 1})$$
$$=\frac{5}{8}\times\frac{3}{5}\times\frac{1}{5}\times\frac{3}{5}\times\frac{2}{5} > 0.$$

We therefore decide that patient #9 is diagnosed with XDisease.

5. (20 pts) **Gaussian Discriminative Analysis** (You may use results from the class without proof in this question.)
Suppose we have the following data set:

$$
\begin{aligned}
x_1 &= \left[1, 3\right]^T, y_1 = 1;\\
x_2 &= \left[3, 1\right]^T, y_2 = 1;\\
x_3 &= \left[0, 1\right]^T, y_3 = 0;\\
x_4 &= \left[-2, 1\right]^T, y_4 = 0.
\end{aligned}
$$

We consider the Gaussian Discriminant Analysis (GDA) model, which models $P(x|y)$ using multivariate Gaussians. Writing out the model, we have:

$$
\begin{aligned}
P(y = 1) &= \phi = 1 - P(y = 0),\\
P(x|y = 0) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)\right),\\
P(x|y = 1) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\right).
\end{aligned}
$$

(a) (10 pts) Find the maximum likelihood estimate of the parameters: ϕ, μ_0, μ_1, and Σ.

Solution:

$$
\begin{aligned}
\phi &= \frac{N_1}{N} = \frac{2}{4} = \frac{1}{2},\\
\mu_0 &= \frac{1}{2}\begin{bmatrix} 0 \\ 1 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} -2 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix},\\
\mu_1 &= \frac{1}{2}\begin{bmatrix} 1 \\ 3 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix},
\end{aligned}
$$

$$
\Sigma = \frac{1}{4}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \end{bmatrix} + \frac{1}{4}\begin{bmatrix} -1 \\ 0 \end{bmatrix}\begin{bmatrix} -1 & 0 \end{bmatrix} + \frac{1}{4}\begin{bmatrix} -1 \\ 1 \end{bmatrix}\begin{bmatrix} -1 & 1 \end{bmatrix} + \frac{1}{4}\begin{bmatrix} 1 \\ -1 \end{bmatrix}\begin{bmatrix} 1 & -1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}
$$

(b) (10 pts) The decision boundary is of the form $w^T x + b = 0$. For every point on the decision boundary, we have $P(y = 0|x) = P(y = 1|x)$. Formally,

$$
P(y = 0|x) = P(y = 1|x), \forall x \in \{x | w^T x + b = 0\}.
$$

Find w and b.

Solution: From $P(y = 0|x) = P(y = 1|x)$, we have $P(y = 0)P(x|y = 0) = P(y = 1)P(x|y = 1)$. Use the model and plug in $P(y = 0)$ and $P(y = 1)$. We get:

$$
-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0) = -\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1).
$$

Therefore:

$$
\begin{aligned}
w &= \Sigma^{-1}(\mu_0 - \mu_1) = \begin{bmatrix} 2 & 2 \\ 2 & 4 \end{bmatrix}\begin{bmatrix} -3 \\ -1 \end{bmatrix} = \begin{bmatrix} -8 \\ -10 \end{bmatrix},\\
b &= \frac{-\mu_0^T \Sigma^{-1}\mu_0 + \mu_1^T \Sigma^{-1}\mu_1}{2} = -1 + 20 = 19.
\end{aligned}
$$