ECE101: Exploring Digital Information Technologies **Spring 2024**
Week 8 Exercises **Tested on Exam II**

Machine Learning

These exercises are intended to help you master and remember the material discussed in lectures and explored in labs. In future semesters, we may make some or all of these exercises required, but for now they remain optional. We suggest that you do them as we go over the material, but you may also want to use them to review concepts before the exam.

We suggest that you use this version rather than the version without solutions to solve the problems before looking at the version with solutions. Many studies have shown that people often trick themselves into believing that they know how to solve a problem if they are presented with the answer before they try to solve the problem themselves.

1. [L14] Machine Learning is already used widely in most people's daily lives. Make a list of Machine-Learning-based services and software that you use regularly. To help you start, a couple of examples have been provided below. Try to come up with at least three examples, and try to find examples that may not be obvious.

   For each case, also try to identify the inputs and outputs of the abstract ML box.

   A. Most fitness trackers or wearable devices track the signals from your movements around the clock and use Machine Learning to predict the activities that you are performing.
   Input: *(your answer here)*
   Output:

   B. A new service in Gmail and Outlook email apps use machine learning to try to autocomplete sentences by suggesting words and phrases as we write our emails.
   Input:
   Output:

2. [L14] For each of the Machine Learning tasks below, state whether the learning is supervised or unsupervised. For supervised learning tasks, also state whether the learning uses classification or regression.

   A. We have a data of thousands of MRI images of the brain. Each image is labeled as either 'normal' or 'contains tumor tissue'. The task is to build a model that, given an MRI image of a brain, can determine whether or not that brain contains a tumor.

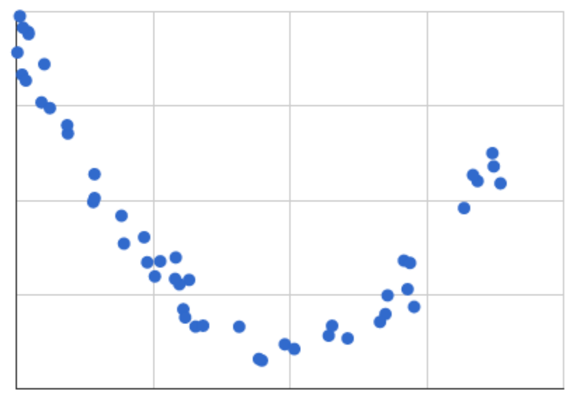
   B. We have the lyrics of thousands of songs on Spotify. The task is to group these songs in a way that songs with similar meaning are grouped together.

   C. We have millions of images from the Internet. The task is to support search based on a query image – in other words, to identify images from the Internet that are similar to the query image.

   D. We have historical data of the rainfall in the Rocky Mountains. The task is to predict how much rain will fall there during the next rainy season.

   E. We have images of leaves from hundreds of different plant species. Each leaf image is labeled with the corresponding plant name. The task is to identify the plant species of new leaf samples collected from Busey Woods.

3. [L14] Students in Physics 101 are performing an experiment in the lab to measure the relation between the velocity and kinetic energy of an object. They measured the kinetic energy of the object at different velocities. Their data are plotted below with velocity **V** on the x-axis and kinetic energy **E** on the y-axis.

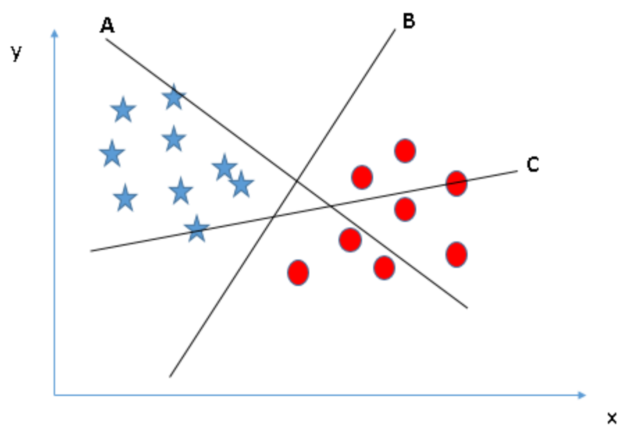


A. To find a relation between **V** and **E**, we need to **fit the curve** using an appropriate polynomial function. What degree of polynomial function do you think would be most suitable? *(We are not asking you to fit the data to an exact expression, just to give the best degree for the polynomial.)*

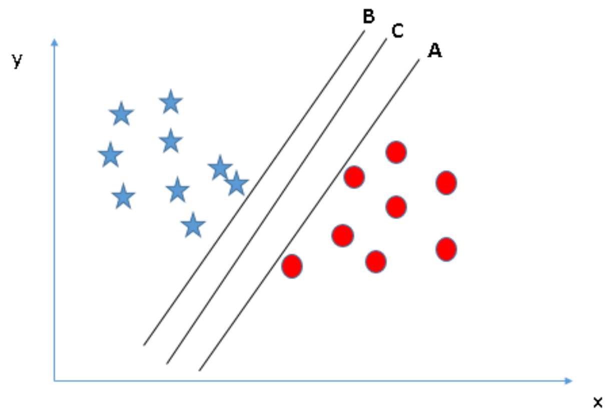
B. Can you comment on the relation shown between **V** and **E**? Does it match with what you learned in Physics class (if you took one)?

4. [L15] (parts of this exercise are taken from analyticsvidya.com's tutorial on SVM)
Let's look at a simple linear classification problem and try to use an SVM (support vector machine) to solve it. Our aim is to find a linear decision boundary that separates the blue stars from the red circles in the plot below.

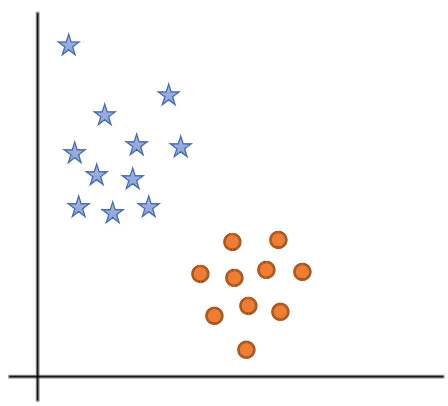
A. For the plot below, which line can serve best as a decision boundary?



B. Now consider another scenario. Which of the three lines works the best as a decision boundary in this case?



C. From **Parts A** and **B**, we can conclude that the best decision boundary is the one that separates all (or most) datapoints clearly, and lies at the maximum distance from the datapoints of both the classes.
Overall, our aim is to find a line that maximizes the minimum distance from any class. SVM does exactly that! To achieve this goal, we first need to identify the datapoints of both the classes that are at minimum distance from the decision boundary. These are the points lying **closest** to the decision boundary. These are called support vectors.
Can you identify **support vectors** in the plot below to separate the blue and orange points?



D. With the support vectors that you identified in **Part C**, you now need to position your decision boundary by making sure that it is at a maximum possible distance from the support vectors.
In the plot below, one possible decision boundary is shown. Do you think it is the best one? If not, in what direction (clockwise or counterclockwise) should the decision boundary tilted in order to make it accurate?