ECE101: Exploring Digital Information Technologies **Fall 2023**
Week 8 Exercises (with solutions in red) **based on L14,L15; 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.

Rather than using this version directly, we suggest that you use the version without solutions to solve the problems before looking at the answers. 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: signals observed by accelerometer, heartbeat tracker and multiple other sensors on the device
   Output: activities and calories burned
   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: the content of the email so far, the content of related emails (for responses)
   Output: the next word or the phrase
   C. A new brand of electric toothbrush now measures the movement of the brush and predicts how well you brush your teeth. The results are then shown in a phone app connected with Bluetooth.
   Input: signals observed by the sensors (accelerometer, gyroscope, and so forth) inside the brush
   Output: amount of time spent on each tooth, positioning of brush head
   D. (Not everyday life example, but an interesting one)
   A new Machine Learning algorithm developed by DeepMind can accurately predict 3D models of protein structures and is accelerating research in nearly every field of biology. You can read more here:
   https://www.deepmind.com/research/highlighted-research/alphafold
   Input: the chemical structure of a protein
   Output: how the protein "folds" into a three-dimensional structure, which tells scientists how likely it is to interact with other biological systems
   E. (Not everyday life example, but an interesting one)
   Historically, water containers and pipes were typically made of lead, a soft metal that is easily worked into water-tight shapes, but contaminates the water and can cause health problems. In Chicago, for example, many older buildings still have lead pipes and connections in their plumbing. Finding such lead pipes and connectors is extremely difficult, so researchers of UChicago developed Machine Learning methods that can identify children with the highest risk of lead poisoning. Read more here:
   https://www.cmu.edu/news/stories/archives/2020/september/lead-poisoning-risk.html
   Input: blood test records, public investigations, housing records, and sociodemographic data
   Output: populations in children who are most at risk of lead poisoning, who are thus most likely to benefit from limited resources for investigating and fixing living conditions

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.
   supervised learning; classification with two classes, 'normal' and 'contains tumor tissue'

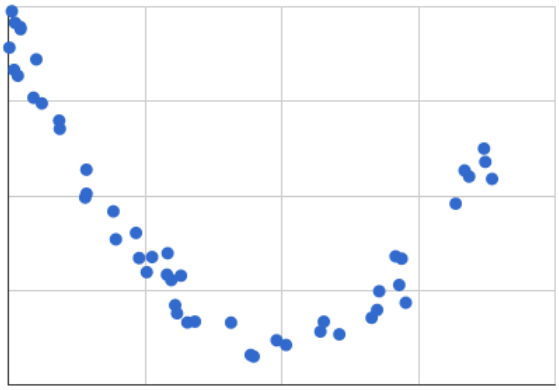
   B. We have the lyrics of thousands of songs on Spotify. The task is to group these songs in a way that songs with similar meaning are grouped together.
   unsupervised learning (clustering of songs into groups based on lyrics)

   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.
   unsupervised learning; use the existing photos to find a feature space with much smaller dimension than the original space of pixels, then use the reduced feature space to find photos that are "close" to our query image; this approach is similar to content-based recommendation strategies, except that the feature space is extracted from data using unsupervised learning (in practice, many recommendation engines also use such techniques in combination with human-specified features)

   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.
   supervised learning; regression in which many previous years' of data are used to predict future year(s)

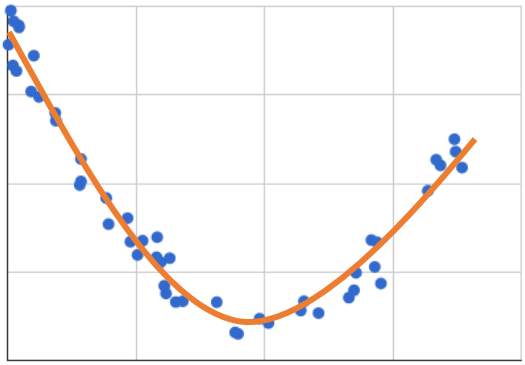
   E. We have images of leaves from hundreds of different plant species. Each leaf image is labeled with the corresponding plant name. The task is to identify the plant species of new leaf samples collected from Busey Woods.
   supervised learning; classification with hundreds of classes (plant species—each is one class)

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.)*

The best fit would be obtained using a quadratic function (degree 2). The data are obviously not linear, and while higher-degree polynomials would (always) give a smaller sum of differences, they would most likely be overfitting the data points shown. To illustrate, we also provide a fit line below (not required for your answer).

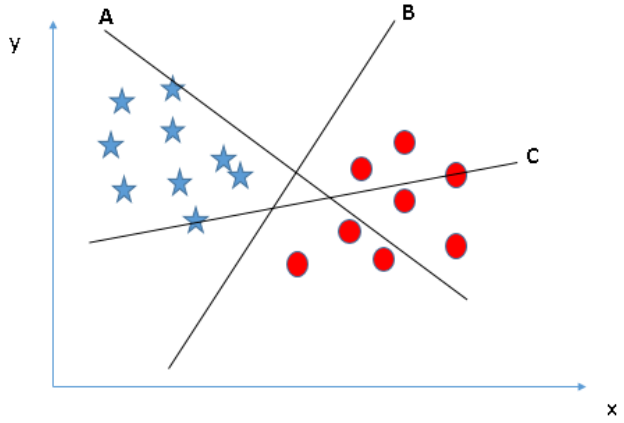


B. Can you comment on the relation shown between **V** and **E**? Does it match with what you learned in Physics class (if you took one)?

E and **V** are related quadratically: **E** is proportional to the square of **V**. This result matches with the standard equation for kinetic energy: $\mathbf{E = \frac{1}{2} mV^2}$.

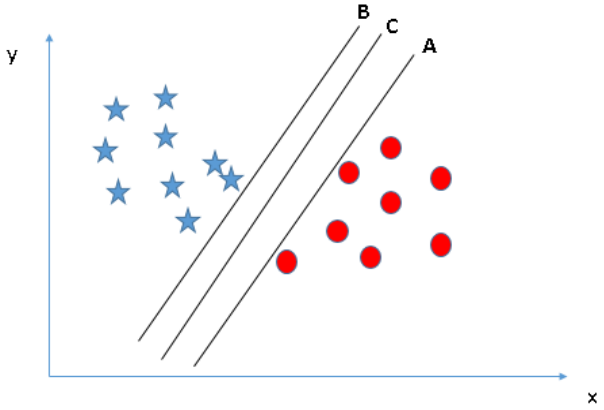
4. [L15] (parts of this exercise are taken from analyticsvidya.com's tutorial on SVM)
Let's look at a simple linear classification problem and try to use an SVM (support vector machine) to solve it. Our aim is to find a linear decision boundary that separates the blue stars from the red circles in the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

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



Line B is best since it completely separates the classes.

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

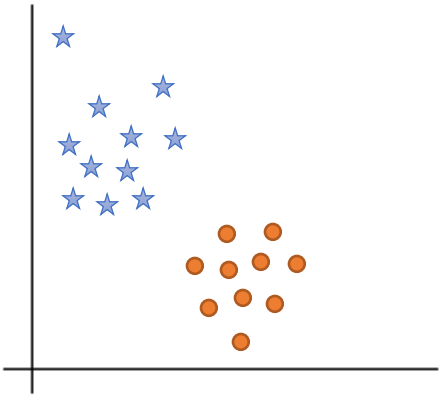


Line C is best since it maintains a larger distance from the datapoints of both classes. Line A is too close to the circles, while line B is too close to the stars.

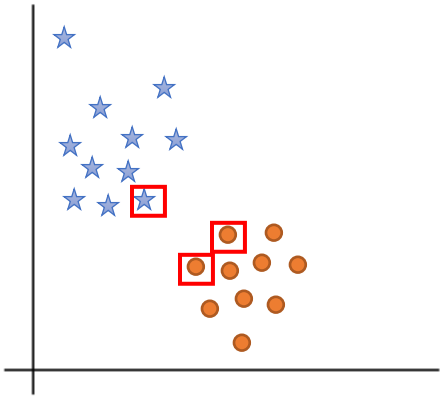
C. From **Parts A** and **B**, we can conclude that the best decision boundary is the one that separates all (or most) datapoints clearly, and lies at the maximum distance from the datapoints of both the classes.

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?

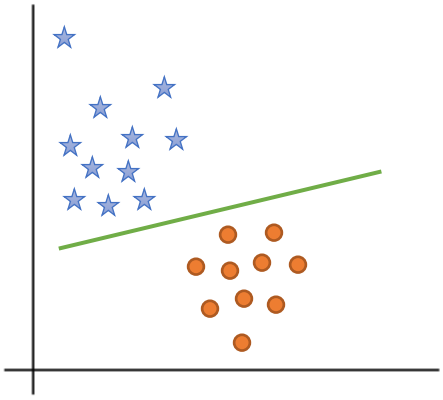


Support vectors are the data points on the decision boundary. In this case, support vectors are highlighted with red in the plot below:

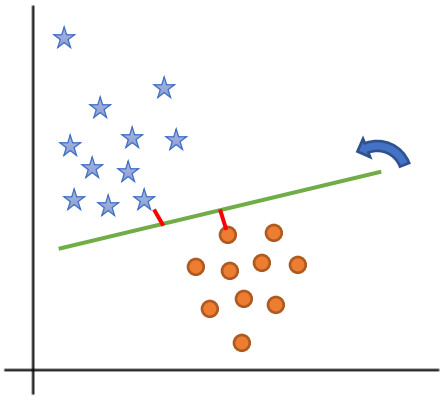


D. With the support vectors that you identified in **Part C**, you now need to position your decision boundary by making sure that it is at a maximum possible distance from the support vectors.

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 counter-clockwise) should the decision boundary tilted in order to make it accurate?



The line shown id close to the highlighted support vectors. Tilting the line **counter-clockwise** would help to increase the distance from these support vectors. Remember that our aim is to maximize the distance from the support vectors.



In the plot below, we provide a more accurate decision boundary, which is at a maximum possible distance from **all** support vectors.