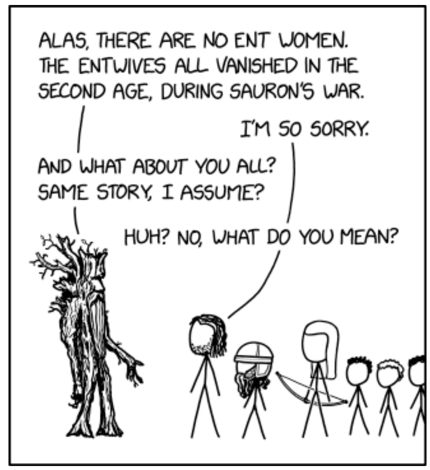
CSE 353: Homework 5 — Due Thursday November 9

1. **(1pt)** *Machine learning fundamentals.* Decide if the following ways of thinking fall under the umbrella of *supervised*, *semi-supervised*, or *unsupervised* learning. Justify your answer.

   There is only and exactly one correct answer (no "both" or "all 3", no "none").

   (a) I have just arrived to Middle Earth, and I am shocked by all the different creatures I see. Some are tall, some are short. Some have thorny heads, others have hair, or are bald. I decide to group everything I see into categories, and give them funny names (entpeople, olephants, etc.) These names are entirely my design, and capture feature similarities between clusters of species.

   (credit: xkcd.com)

   (b) I have now met an "entperson", who tells me that there are specific anatomical features that are unique to entmen and entwomen. For example, entmen tend to have really thorny heads, while entwomen have bushier feet. These are not hard and fast rules, in that there are entmen with bushy feat and entwomen with thorny heads, but they are rare. I therefore devise a classification scheme that decides the gender of entpeople based on these features. My entfriend then goes through a list of all the people in our neighborhood and helps me label them as men or women.

   (c) I have now lived in Middle Earth for 150 years, so I have a pretty good grasp of the "main" species, such as dwarfs, or giants, or elves. But every once in a while I see something completely new, and am not sure exactly which of these categories it belongs. So, I use a mixture of the clustering strategy I used when I first arrived to kind of lump this new species in some category, and then use the labels I acquired by talking to a lot of different species over the past 150 years, and guess an appropriate species name for this new entity.

2. *Which is it?* **(2pts)**

   (a) I need to go to the bank to deposit a check, but my schedule is tight and I want to predict exactly how long it will take. So I illegally download a bunch of bank security cameras and record how long each person waited in line to deposit a check, and record features like bank location, day of the weak, time of the day, etc. From this, I form a model that predicts how long it will take me to deposit this check, given these features.

   This is an example of a

   regression model or **classification model.**

   (b) I need to go to the bank to deposit a check, but my schedule is tight and I want to go to the ATM that I know has the shortest line. In my town there are 3 ATMs to choose from, and I have all the security footages from each one. So I form a model that predicts that, given the day of the week, the time of day, the rotation of the planets, which ATM will have the shortest line. I use this to predict which ATM I should visit on a warm Tuesday afternoon when Jupiter is in retrograde.

   This is an example of a

   regression model, or **classification model.**

   (c) A

   discriminative model or **generative model**

   forms a probability distribution based on training data and uses it to predict future samples using maximum likelihood estimation.

   (d) A

   discriminative model, **generative model**

   produces discriminative lines that well-separates different classes, but does not really focus on the probability distribution away from the class separation points.

   (e) I want to know if an image is a dog or a cat. So I go online, download all the image search results for cat, and for dog, and train a model over my downloaded images to correctly predict cat or dog.

   This is an example of

   supervised learning, or **unsupervised learning.**

   (f) I want to know if an image is a dog or a cat. So I adopt 100,000 animals that are all either cats or dogs, and then put them in separate rooms based on their physical similarities, and attribute self-made labels to each room.

   This is an example of

   supervised learning, or **unsupervised learning.**

3. *Probability of getting poisoned.* **(2pts)** Open the attached python notebook, and load the Mushrooms dataset. It contains 8124 samples of different mushrooms and 23 features, plus a target feature (poisonous or not). The details of how to read the loaded dataset is given below.

Attribute Name	Description
poisonous	edible=e, poisonous=p
cap-shape	bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
cap-surface	fibrous=f,grooves=g,scaly=y,smooth=s
cap-color	brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
bruises	bruises=t,no=f
odor	almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
gill-attachment	attached=a,descending=d,free=f,notched=n
gill-spacing	close=c,crowded=w,distant=d
gill-size	broad=b,narrow=n
gill-color	black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=
stalk-shape	enlarging=e,tapering=t false
stalk-root	bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?
stalk-surface-above-ring	fibrous=f,scaly=y,silky=k,smooth=s
stalk-surface-below-ring	fibrous=f,scaly=y,silky=k,smooth=s
stalk-color-above-ring	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
stalk-color-below-ring	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
veil-type	partial=p,universal=u
veil-color	brown=n,orange=o,white=w,yellow=y
ring-number	none=n,one=o,two=t
ring-type	cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
spore-print-color	black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
population	abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
habitat	grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

Assume that each sample is taken i.i.d. from the field. Follow the instructions in the python notebook, and report the following here.

(a) What are the prior, posterior, and likelihood probabilities that a mushroom is poisonous, with the feature that its cap is convex? Give your answer to at least 3 significant digits.

(b) Which feature(s) / feature value(s) pair best answers this question: "If a mushroom has this feature value, then it is definitely poisonous"? Is this a prior, posterior, or likelihood probability?

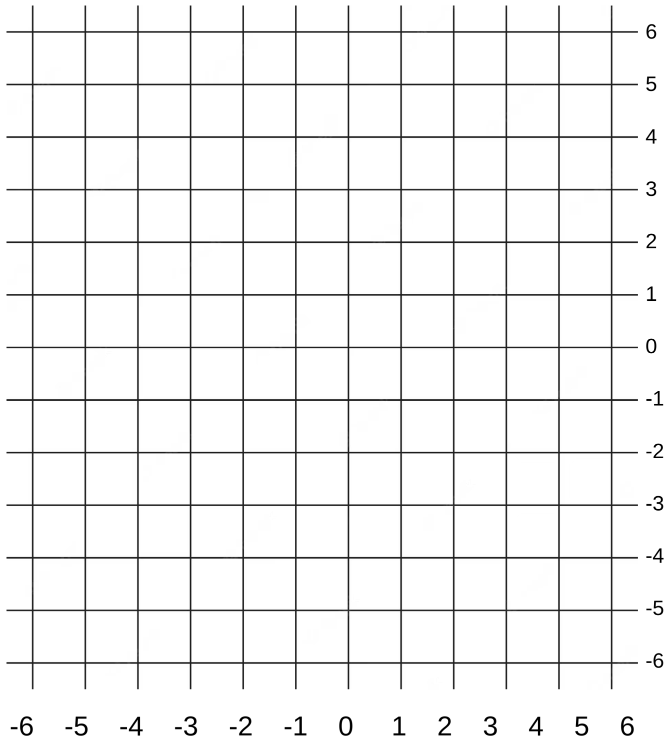
(c) Which feature(s) / feature value(s) pair best answers this question: "To identify a poisonous mushroom, I should look for this feature value"? Is this a prior, posterior, or likelihood probability?

4. **KNN and multiclass logistic regression(2pts)** Consider the following 2-D dataset

points	(2,3)	(4,4)	(3,6)	(-4,1)	(-3,-3)	(2,-4)	(1,-1)	(6,-5)	(1,-6)
label	red	red	red	green	green	green	blue	blue	blue

(a) **KNN**

i. Plot these points, in the color of the given label. This is your training set.



ii. Find the centroid of these points by label.

center of red points = , center of blue points = , center of green points =

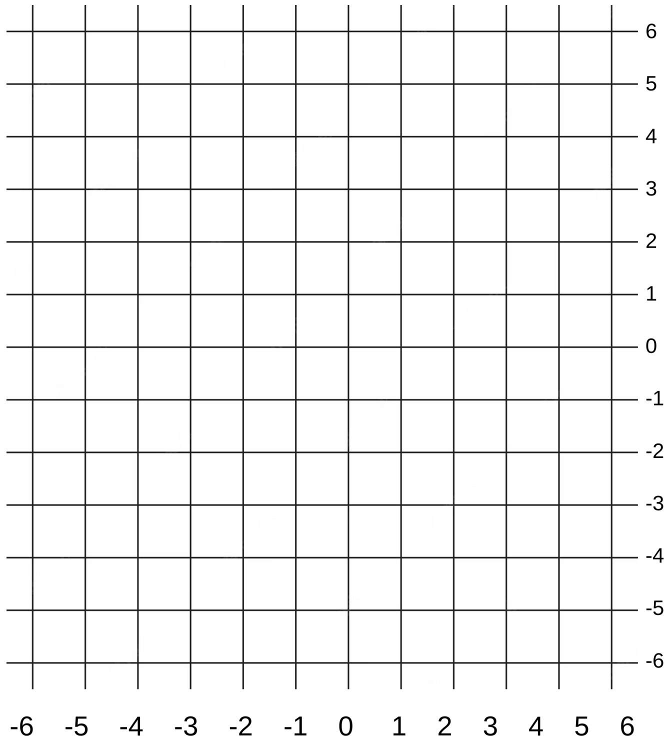
iii. Draw *Voronoi regions*; that is, the regions for which any new point in that region would be classified as a specific label, and color the region according to that label.

iv. Now consider this test set. Using a KNN classification strategy, give a predicted label to each point below. Write in 'R', 'G', or 'B' in the last row. Compute the classification accuracy of your guess.

points	(4,5)	(4,-3)	(7,-6)	(-4,0)	(3,3)
true label	red	blue	blue	green	red
pred. label					

(b) **Multiclass logistic regression**

i. Replot these points, in the color of the given label.



ii. Now, suppose that your multiclass logistic regression code spat out the following solution

$$\Theta = \begin{bmatrix} 1 & -1 & 500 \\ 4 & 1 & 1 \end{bmatrix}$$

First, compute a normalized Θ so that each *column* of Θ, represented as θ_i, has norm 1. That is, construct

$$\Theta_{\text{norm}} = \begin{bmatrix} \frac{\theta_1}{\|\theta_1\|_2} & \frac{\theta_2}{\|\theta_2\|_2} & \frac{\theta_3}{\|\theta_3\|_2} \end{bmatrix}$$

Report Θ_{norm} with rounded values including at least 3 significant digits.

iii. Let's define the three columns of Θ_{norm} as θ_1, θ_2, and θ_3 (in order). Recall that in multiclass logistic regression, our classification rule is

$$\hat{y} = \underset{i\in\{1,2,3\}}{\text{argmax}} \theta_i^T x.$$

So, if we are choosing between $\hat{y} = 1$ and $\hat{y} = 2$, then the discriminating line will be some function of θ_1 and θ_2. Write down an equation involving θ_1, θ_2, and x which implies that classifying x is exactly on the boundary between being labeled class 1 or class 2. Draw this set of x on the plot above. Repeat this for the discriminator between class 2 and 3, and between 3 and 1.

iv. In the test set, again assign a prediction to each label, by picking

$$\text{predicted label} = \underset{i\in\{1,2,3\}}{\text{argmax}}\ \theta_i^T x$$

for each datapoint x. Supposing that red = class 1, green = class 2, blue = class 3, give the labels again in terms of 'R', 'G', and 'B'. Report the classification accuracy over this test set.

points	(4,5)	(4,-3)	(7,-6)	(-4,0)	(3,3)
true label	red	blue	blue	green	red
pred. label					

5. **Coding.(3pts)** Complete the two linear layers problem on github.