CS 4780/5780: Introduction to Machine Learning **(Due: 09/15/20)**

Assignment #1: Version Spaces, k-NN, Decision Trees

Instructor: Nika Haghtalab, Thorsten Joachims *Name:* Student name(s), *Netid:* NetId(s)

Course Policy: Read all the instructions below carefully before you start working on the assignment, and before you make a submission.

- Please include your name and NetIDs on the first page. We recommend typesetting your submission in LaTeX, and an Overleaf template is linked on the canvas module. When submitting, remember to mark which page has which response.
- Assignments are due at 5 pm on the due date in PDF form on Gradescope.
- Late assignments can be submitted on Gradescope up to Sunday, Sept 20 at 5pm EST. This is also when the solutions will be released.
- You can do this assignment in groups of 1-2. Please submit no more than one submission per group. Collaboration across groups is not permitted.
- All sources of material outside the course must be cited. The University Academic Code of Conduct will be strictly enforced.

Problem 1: Version Spaces

(5+5+5+5+7+8=35 points)

Table T shows the last five dishes Alice had along with various attributes and whether they liked the dish or not. The attributes are

- Spiciness S: A dish can be *mild*, *medium* or *hot*.
- Meat M: A dish can have *no meat*, *chicken* or *beef*.
- Cilantro C: A binary variable indicating whether or not a dish has cilantro.

Dish	Spiciness	Meat	Cilantro	Like?
spicy chicken	hot	chicken	yes	yes
pepper steak	medium	beef	no	yes
beef stew	mild	beef	no	no
lentil curry	hot	no meat	yes	no
chicken soup	medium	chicken	no	no

Bob would like to learn Alice's eating preferences based on her feedback on dishes she's eaten before. Formally, Bob wants to learn a function $S \times M \times C \rightarrow \{yes, no\}$.

(a) What is the size of instance space $\mathcal{X}$?

(b) Assuming that the hypothesis space H consists of all possible functions on $S \times M \times C \rightarrow \{yes, no\}$, what is the size of H?

(c) Initially, Bob thinks that *Alice would like any hot dish.* Is this hypothesis h consistent with table T above? Why or why not? Justification should mention entries from table T.

(d) For the above training set T and hypothesis space H defined in (b), what is the size of the version space $VS_{(H,T)}$?

Realizing from (b) that the size of H is too large, Bob maps each attribute value to a numerical value as follows:

- Spiciness S: *mild* = -1, *medium* = 0 and *hot* = 1.
- Meat M: *no meat* = -1, *chicken* = 0 and *beef* = 1.
- Cilantro C: *no* = -1, and *yes* = 1.

They apply the same mapping to table T and get a new table T'. The new hypothesis space H' now contains all functions that only use the sum of any two attribute values to make a prediction. In particular, for every two attributes A and B ($A \neq B$), Bob constructs two hypotheses g and g':

$$g(A,B) = \begin{cases} yes, & \text{if } A + B > 0 \\ no, & \text{otherwise} \end{cases}$$

$$g'(A,B) = \begin{cases} no, & \text{if } A + B > 0 \\ yes, & \text{otherwise} \end{cases}$$

(e) Write down all hypotheses $h \in H'$ using the shorthand notation above. What is the size of H'? For instance, we have $h_1 \in H'$, where

$$h_1 = g(S,M) = \begin{cases} yes, & \text{if } S + M > 0 \\ no, & \text{otherwise} \end{cases}$$

(f) For T' and H' defined above, what is the size of the version space $VS_{(H',T')}$? To show your work, first complete the following table by enumerating functions in H' and filling in their evaluations. We have filled in the value of h_1 as the first column.

Dish	S	M	C	h_1	$h_2 \ldots$
spicy chicken	1	0	1	yes	
pepper steak	0	1	-1	yes	
beef stew	-1	1	-1	no	
lentil curry	1	-1	1	no	
chicken soup	0	0	-1	no	

Problem 2: k-Nearest Neighbor (14+5+10=29 points)

In this problem, you are going to look at a small dataset to develop a better understanding of various properties of k-NN. Suppose that there exists a set of points S in $\mathbb{R}^2$, and that each point belongs to one of two different classes $\{red, blue\}$. Below are the coordinates (x, y) of all points in S, along with their class labels:

- Points in class *Red* : $(3, 3), (4, 4), (6, 1)$
- Points in class *Blue* : $(1, 1), (4, 2), (6, 3), (7, 2)$

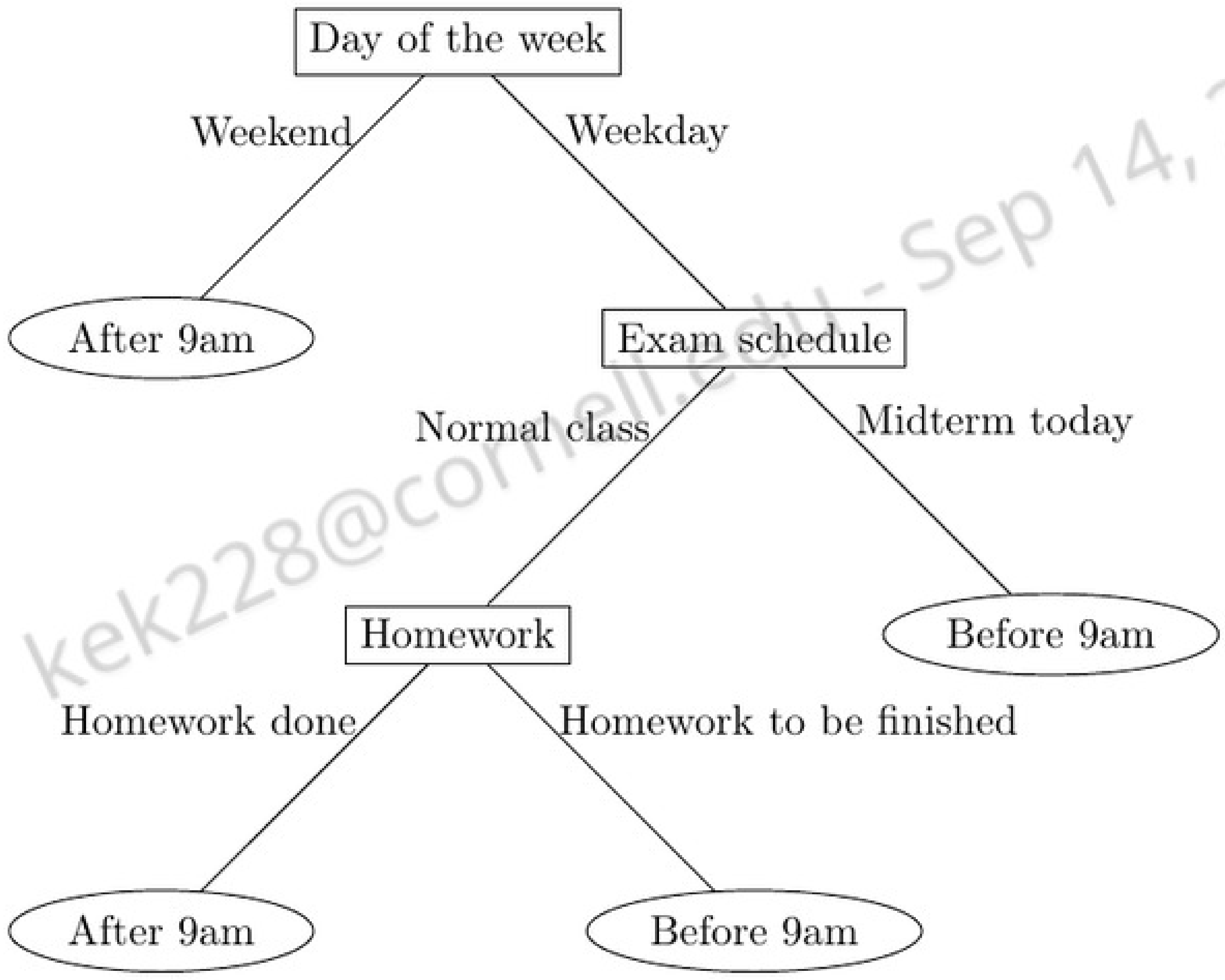
(a) Draw the k-nearest-neighbor decision boundary for $k = 1$. Remember that the decision boundary is defined as the line (or set of disjoint lines) where the classification of a test point changes. Use the standard Euclidean distance between points to determine the nearest neighbors. Start by plotting the points as a two-dimensional graph. Please use the corresponding colors for points of each class (e.g. *blue* and *red*). If you draw the plots by hand, make sure they are legible. Note the Overleaf template has some plotting assistance commented out.

(b) If the x-coordinate of each point was multiplied by 5, what would happen to the $k = 1$ boundary (Draw another picture)? Explain briefly how this effect might cause problems when working with real data, and propose at least one strategy to mitigate this problem.

(c) Suppose now we have a test point at $(3, 4)$. How would it be classified under 3-NN, under the setting of part (a) (i.e. with the original x-coordinates)? Given that you can modify the 3-NN decision boundary by adding points to the training set in the diagram, what is the minimum number of points that you need to add to change the classification at $(3, 4)$? List the coordinates (x, y) of each of these points, and also show them on a plot with the data points and test point.

Problem 3: Decision Trees (5+5+6+9=25 points)

We have constructed a decision tree which classifies if student A will get up before 9am in the morning on a certain day. The instance space is $\{Weekend, Weekday\} \times \{Normal\ class, Midterm\ today\} \times \{Homework\ done, Homework\ to\ be\ finished\}$.



(a) Use the decision tree to classify whether student A will get up before 9am on the following days.

- On a weekend day where there is normal class and homework needs to be finished, does student A get up before 9am?
- On a weekday where there is a midterm and the homework is done, does student A get up before 9am?
- On a weekday where there is normal class and the homework is done, does student A get up before 9am?

Use the given decision tree to answer the following question: If student A woke up after 9am today and it is a weekday, was there a midterm today?

(b) We have provided in Table 1 a sample of data on Student B's wake up schedule for seven randomly chosen days. Included are information on the day of week, whether homework has been finished, and if there is a midterm on that day. The same instance space is used from part a).

Day	Weekend?	Homework done?	Midterm?	Wake up before 9am?
day 1	no	yes	yes	no
day 2	yes	no	yes	no
day 3	yes	yes	yes	yes
day 4	no	yes	no	yes
day 5	yes	yes	yes	yes
day 6	no	yes	yes	no
day 7	no	no	no	yes

Table 1: Sample of Student B's wake up schedule for seven days.

Construct three decision trees using the data given in Table 1. Draw them in a similar manner to the tree given in part a). The Overleaf template includes basic tree drawing code commented out. If you draw the trees by hand, make sure they are legible.

We will use three different methods to choose on which feature we split.

1. For the first decision tree (DT 1), we select the feature that splits the dataset into groups as **evenly as possible**. More formally, in the binary case where each feature A can take one of two values $\{a_0, a_1\}$ and corresponding sample sets $S_i = \{(x, y) \in S : \text{feature } A \text{ of input vector } x \text{ has value } a_i\}$ for $i = \{0, 1\}$, pick the feature which **minimizes** $||S_1| - |S_0||$.

2. For the second decision tree (DT 2), we select the feature that minimizes **error**. Recall from lecture that the error of a set $\mathrm{Err}(S)$ is the size of the minority label and that the error $\mathrm{Err}(S \mid A) = \sum_{a_i} \mathrm{Err}(S_i)$. We would select A that maximizes $\mathrm{Err}(S) - \mathrm{Err}(S \mid A)$.

3. For the third decision tree (DT 3), we select the feature based on **entropy**. This means we pick the feature A that **maximizes the information gain**. Recall from lecture `03-c-growing-DTs.pdf`, slide "Which Split?" that information gain is defined as follows:

$$\mathrm{IG(S)} = H(S) - \mathrm{H}(S \mid A)$$

(Hint: Refer to this slide for the definition of entropy as well). You do not need to show all your calculations for the selection process. Instead, please write the expression and the evaluation of $IG(S)$ for your **chosen feature for the root of the decision tree** down to 5 significant digits.

Build each decision tree to the fullest extent, or until each leaf node is pure.