Homework I-KEY

1. Consider the following set of points: {(-2, -1) , (1, 1) , (3 , 2)}
   a) Find the least square regression line for the given data points.
   b) Plot the given points and the regression line in the same rectangular system of axes.

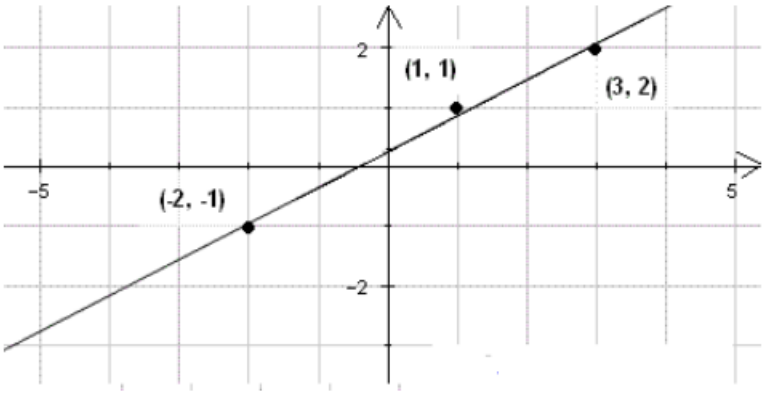
[5 Points]

a) Let us organize the data in a table.

x	y	x y	x^2
-2	-1	2	4
1	1	1	1
3	2	6	9
$\Sigma x = 2$	$\Sigma y = 2$	$\Sigma xy = 9$	$\Sigma x^2 = 14$

Find the least square regression line y = a x + b.

We now use the above formula to calculate a and b as follows
$a = (n\Sigma x\, y - \Sigma x \Sigma y) / (n\Sigma x^2 - (\Sigma x)^2) = (3*9 - 2*2) / (3*14 - 2^2) = 23/38$
$b = (1/n)(\Sigma y - a\, \Sigma x) = (1/3)(2 - (23/38)*2) = 5/19$
b) We now graph the regression line given by y = a x + b and the given points.



2. The values of y and their corresponding values of y are shown in the table below

x	0	1	2	3	4
y	2	3	5	4	6

a) Find the least square regression line $y = a x + b$.
b) Estimate the value of y when x = 10.

[5 Points]

a) We use a table to calculate a and b.

x	y	x y	x^2
0	2	0	0
1	3	3	1
2	5	10	4
3	4	12	9
4	6	24	16
$\Sigma x = 10$	$\Sigma y = 20$	$\Sigma x\, y = 49$	$\Sigma x^2 = 30$

We now calculate a and b using the least square regression formulas for a and b.
$a = (n\Sigma x\, y - \Sigma x \Sigma y) / (n\Sigma x^2 - (\Sigma x)^2) = (5*49 - 10*20) / (5*30 - 10^2) = 0.9$
$b = (1/n)(\Sigma y - a\, \Sigma x) = (1/5)(20 - 0.9*10) = 2.2$
b) Now that we have the least square regression line y = 0.9 x + 2.2, substitute x by 10 to find the value of the corresponding y.
y = 0.9 * 10 + 2.2 = 11.2

3. Explain why the size of the hypothesis space in the *EnjoySport* learning task is 973. How would the number of possible instances and possible hypotheses increase with the addition of the attribute *WaterCurrent*, which can take on the values *Light*, *Moderate* or *Strong*? More generally, how does the number of possible instances and hypotheses grow with the addition of a new attribute A that takes on *k* possible values?

[5 Points]

Each hypothesis specifies either a required attribute value or "don't care", yielding 4*3*3*3*3*3 = 972 (+ 1 for the null hypothesis) possible combinations.

Adding Water-Current as an attribute, which can take on 3 values, multiplies the number of possible instances by 3 (to get 288) and the number of possible hypotheses by 4 (to get 3888 + 1 for null hypothesis).

In general, adding an attribute with k possible values, multiplies the number of possible instances by k and the number of possible hypotheses by k+1.

4. Consider the following sequence of positive and negative training examples describing the concept "pairs of people who live in the same house". Each training example describes an ordered pair of people, with each person described by their sex, hair color (black, brown or blonde), height (tall medium or short), and nationality (US, French, German, Irish, Indian, Japanese or Portuguese). **[15 Points]**

+ <<male brown tall US> <female black short US>

+ <<male brown short French> <female black short US>>

- <<female brown tall German> <female black short Indian>>

+ <<male brown tall Irish> <female brown short Irish>>

Consider a hypothesis space defined over these instances, in which each hypothesis is represented by a pair of 4-tuples, and where each attribute constraints may be a specific value , "*?*" or "*∅*", just as in the *EnjoySport* hypothesis representation. For example, the hypothesis

<<male ? tall ?> < female ? ? Japanese>>

represents the set of all pairs of people where the first is a tall male (of any nationality and hair color), and the second is s Japanese female (of any hair color and height).

a. Provide a hand trace of the Candidate-Elimination algorithm learning from the above training examples and hypothesis language. In particular, show the specific and general boundaries of the version space after it has processed the first training example, then the second training example, etc.
b. How many distinct hypotheses from the given hypothesis space are consistent with the following single positive training example?
<<male black short Portuguese> <female blonde tall Indian>>
c. Assume the learner has encountered only the positive example from part (b), and that it is now allowed to query the trainer by generating any instance and asking the trainer to classify it. Give a specific sequence of queries that assures the learner will converge to the single correct hypothesis, whatever it may be (assuming that the target concept is describable within the given hypothesis language). Give the shortest sequence of

queries you can find. How does the length of this sequence relate to your answer to question (b)?

d. Note that this hypothesis language cannot express all concepts that can be defined over instances (i.e., we can define sets of positive and negative examples for which there is no corresponding describable hypothesis). If we were to enrich the language so that it could express all concepts that can be defined over the instance language, then how would your answer to © change?

a.

S1: {<<male brown tall US><female black short US>>}

G1: {<< ? ? ? ?>< ? ? ? ?>>}

S2: {<<male brown ? ?><female black short US>>}

G2: {<< ? ? ? ?>< ? ? ? ?>>}

S3: {<<male brown ? ?><female black short US>>}

G3: {<<male ? ? ?>< ? ? ? ?>>,

<< ? ? ? ?>< ? ? ? US>>}

S4: {<<male brown ? ?><female ? short ?>>}

G4: {<<male ? ? ?>< ? ? ? ?>>}

b.

Given the single positive example, the consistent hypotheses are those which for each attribute either require the same value in the example, or don't care. This gives two possibilities for each attribute, for a total of 2^8 = 256 consistent hypotheses.

c.

Once you have a single positive instance, for each attribute a single query may be generated to determine whether or not the specific value for the attribute in the original positive instance is of relevance to the target concept. To generate such a query, it is sufficient to copy the values from the original positive instance, but change the value of the attribute-in-question to something different. Such a series of queries will be of length equal to the number of attributes, and will assure that the

learner will converge to a single, correct hypothesis, given that the hypothesis is expressible in the target language.

For example:

<<female black short Portuguese><female blonde tall Indian>>

^^^^^^

<<male brown short Portuguese><female blonde tall Indian>>

^^^^^

<<male black tall Portuguese><female blonde tall Indian>>

^^^^

<<male black short French ><female blonde tall Indian>>

^^^^^^

<<male black short Portuguese><male blonde tall Indian>>

^^^^

<<male black short Portuguese><female black tall Indian>>

^^^^^

<<male black short Portuguese><female blonde medium Indian>>

^^^^^^

<<male black short Portuguese><female blonde tall US >>

^^

Once a single positive example has been seen, each query will halve the remaining version space. The number of queries (8) necessary to converge on a single hypothesis is thus log-base-two of the size of the hypothesis space (256, as in part b).

d.

Once you start working with a complete hypothesis space in the version space framework, you lose all inductive bias, and hence the ability to generalize to unseen examples. Therefore, to converge on a single hypothesis, you'd have to generate a query for each possible instance, other than the one already seen, in order to determine whether or not it would be covered by the target concept. Such a sequence would be of length (2*3*3*7)*(2*3*3*7) = 15876.

5. Consider learning a Boolean valued function : $X \rightarrow Y$, where $X = \langle X_1, X_2, \ldots., X_N \rangle$, where *Y* and the X_i are all Boolean valued variables. You decide to consider a hypothesis space *H* where each hypothesis is of the form
   a. $if\left[(X_i = a) \wedge \left(X_j = b\right)\right] then\ Y = 1\ else\ Y = 0$

   where $i \neq j$, and where *a* and *b* can be either 0 or 1. Notice each hypothesis constrains exactly two of the features of X.

   How many distinct hypotheses are there in *H*?

[5 Points]

4nC2