47717: Data Mining
Homework #4: Supervised Learning, Evaluation
Due: **July 24th, 11:59pm, on Moodle**

THE HEBREW UNIVERSITY OF JERUSALEM

Read the instructions **carefully** (that's a good idea in general).

- This is a short theoretical homework. It is *optional*, and will be used as *Magen* for another theoretical homework (as in, the worst theoretical part of HW 1-4 will not count towards your grade).
- This homework only includes a theoretical part (no programming). Each person submits their own answers, a single file in pdf format only (no docx or jpg) named **ex4t_ID.pdf** (ID is your ID).
- If you are submitting handwritten answers, make sure they are crystal clear.
- The meta question should be answered through a questionnaire. The link will be posted on the moodle.
- Points may be reduced for submissions that fail to comply.
- Make sure you follow the News forum for any updates.

Problem 1 (Decision Trees).

(a) Below is a dataset about when you choose to go and play tennis. Build a decision tree from the dataset. You should build a tree to predict PlayTennis, based on the other attributes (do not use the day in your tree!). Show all of your work, calculations, and decisions as you build the tree. What is the classification accuracy?

(b) Is it possible to find a subset of the days to use as training examples that will get the algorithm to include the attribute Temperature in the learned tree, even though the true target concept is independent of Temperature? if no, explain. If yes, give such a set.

(c) Build a tree using only examples D1-D7. What is the classification accuracy for the training set? what is the accuracy for the test set (examples D8-D14)? Compare to the tree of (a) and explain why you think these are the results.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Problem 2 (Inter-agreement).

Suppose we want to predict whether a blog post contains an argument that is liberal, conservative, or neutral. Two annotators label 900 blog posts independently and produce the following contingency table:

		$\mathcal{A}$			
		liberal	conservative	neutral	
$\mathcal{B}$	liberal	100	30	170	300
	conservative	50	130	20	200
	neutral	100	40	260	400
		250	200	450	900

(a) Compute the inter-annotator agreement in terms of percent agreement.

(b) Compute Cohen's Kappa agreement under the assumption that the annotators are unbiased and therefore each one independently labels an equal number of blog posts as liberal, conservative, and neutral (i.e., 300 each).

(c) Compute the Cohen's Kappa agreement under the assumption that the annotators are biased and follow their individual biases, which are reflected in the contingency matrix (like we did in class).