Roni's 10-601 Practice Midterm Questions

1. [**xx pts**] Consider a credit card fraud detection problem, where the goal is to learn the concept *fraudulent transaction*. The input instances are credit card transactions, all of which are represented by six fields (attributes), each of which assumes one of a small set of possible values as follows:

    - cardholder AGE: $\{< 30; 30+\}$
    - cardholder annual household INCOME: {<20K ; $20K+$}
    - cardholder credit LINE: {<1000 ; 1000+}
    - transaction AMOUNT: {<100 ; 100+}
    - transaction TIME: {day (6AM-6PM), night (6PM-6AM)}
    - transaction LOCATION: {urban, suburban }

    (a) [**xx pts**] What is the size of the input space, "X"?

    (b) [**xx pts**] What is the size of the concept space (i.e. the space of all possible concepts)?

    (c) [**xx pts**] Consider the set of hypotheses that can be expressed by specifying the values of exactly three different attributes (e.g. "TIME=night AND AGE= 30+ AND AMOUNT < 100").
    What is the size of this hypothesis space, H? Does it have a bias? If so, is it hard (restriction bias) or soft (preference/search bias)? If not, why not?

    (d) [**xx pts**] What is the size of the version space $VS(H, D)$ for the H in part (c) above and the empty training set D (i.e. a set containing no instances)?

2. An experiment consists of simultaneously flipping a fair dime and a fair penny. You are asked to predict whether the penny ends up heads or tails.

   (a) What is the irreducible entropy of the random variable corresponding to the outcome of flipping the penny?

   (b) Say that you (rationally) predicted (denote these predictions Q) that the penny would come up heads with probability .5 and tails with probability .5. What is the cross-entropy from the truth to your prediction.

   (c) You are now told the total number of "tails" in the experiment (this could be 0, 1 or 2). Does this help your prediction, on average? What is the average entropy of your prediction now? Show your work, or explain.

   (d) As in (b), but you are only told whether the total number of "tails" in the experiment was odd or even. Does this help your prediction, on average? What is the average entropy of your prediction now? Show your work, or explain.

3. Let X; Y be some jointly distributed numerical random variables. In each of the following, fill in the blank with exactly one of: $\{<, \leq, =, \geq, >, ?\}$, where ' ?' means that none of the other relations necessarily holds:

   $H(X) + H(Y)$ ____ $H(X, Y) + I(X; Y)$

   $H(X) + H(Y)$ ____ $H(X + Y)$ (Notice this is $H(X + Y)$, not $H(X, Y)$)

   $H(cos(Y))$ ____ $H(Y)$

   $H(X^3)$ ____ $H(X)$

   $H(X^4)$ ____ $H(X)$

   $H(Y)$ ____ $H(Y|X = x)$ for some given x

   $H(Y)$ ____ $H(Y|X)$

   $H(Y)$ ____ $P(X = x) \cdot H(Y|X = x)$ for some given x

4. Let X, Y be jointly distributed as follows:

		Y: 1	Y: 3	Y: 5	Y: 7
X	2	0	0	0	0.25
X	3	0	0	0.25	0
X	4	0	0.25	0	0
X	5	0.25	0	0	0

Please Calculate:

(a) The marginal distribution $P(X)$

(b) The conditional probability distribution $P(X|Y = 3)$

(c) Correlation coefficient:

(d) Mutual Information (in bits):

5. Suppose that we work at Netflix and our boss wants us to build a predictive model to estimate the number of viewers that a new movie will have. We gather data about 100 previous movies. Our attributes include the size of the cast, the average viewing statistics for series by those cast members and by the director, the rating (a categorical feature taking one of 4 values), the primary language of the series, a number of binary features indicating for which languages closed captioning is available etc. In all, imagine that we have 80 features. Because we have only been studying machine learning for 2 months, we only really have one option for which regression model to use: linear regression.

   (a) You hold out 20 movies to evaluate your model, leaving you with 80 remaining for training. You train a model using all 80 features of the 80 movies. Assuming that the features are linearly independent of each other, how well do you expect your model to fit your training data?

   (b) Do you expect to perform better, as well, or worse on the hold-out data?

   (c) Your boss wants you to use only 10 features in your mode (perhaps the board room executives want all the features to fit on one power-point slide). Trying out all combinations of features requires training over 10^{12} models). Suggest a computationally feasible alternative.

6. Say that you train a linear regression model with an ample number of examples relative to features (say 10000 examples and 50 features) and find yourself in the opposite situation: Instead of getting zero error and overfitting, you *get high error on both the training and test data.* Which of the following approaches might improve your model?

   - Go out and acquire more features.
   - Try removing some features.
   - Include polynomial features $x_1^2, x_1 x_2, ...$
   - Try multiplying your features by constants $\alpha_1 \cdot x_1, \alpha_2 \cdot x_2, ...$
   - Add some new features formed by taking linear combinations of your existing features, e.g. $x_i + 2x_j - 4.3x_k$

Do not remove this page! Use this page for scratch work.