Roni's 10-601 Practice Midterm Questions
Solutions

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"?

   2^6
   Note in the exam you won't have a calculator so we won't expect anything more than this.

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

   $2^{|X|} = 2^{2^6} = 2^{64}$

   (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?

   $|H| = C_3^6 \cdot 2^3 = 20 \cdot 8 = 160$
   Yes, H has a hard bias.

   (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)?

   $VS(H; \emptyset) = H$, so $|VS(H; \emptyset)| = |H| = 160$

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?

   $H(P) = H(0.5, 0.5) = 1\text{bit}$
   You should know easy values of entropy like this one.

   (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.

   $CH(P, Q) = .5 \log 2 + .5 \log 2 = 1 \text{ bit} = H(P)$
   Because your prediction and the true probabilities are the same, the cross entropy from P to Q is equal to the entropy of P.

   Because your predictions are different from the true probabilities, the cross entropy from P to Q will be greater than $H(P)$. Note that while you will get qualitative questions on the test, because you won't have a calculator, you will not be expected to perform precise calculations of quantities like $\log 4/3$.

   (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.

   $H(P|\#T) = P(\#T = 0)H(P|\#T = 0) + P(\#T = 1)H(P|\#T = 1) + P(\#T = 2)H(P|\#T = 2)$
   $= 0.25 \cdot 0 + 0.5 \cdot H(0.5; 0.5) + 0.25 \cdot 0$
   $= 0.5\text{bits}$

   (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.

   $H(P|T) = P(T = odd)H(P|T = odd) + P(T = even)H(P|T = even)$
   $= 0.5 \cdot H(0.5, 0.5) + 0.5 \cdot H(0.5, 0.5)$
   $= 0.5 \cdot 1 + 0.5 \cdot 1 = 1 \text{ bit}$

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

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

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

   $H(cos(Y)) \leq H(Y)$

   $H(X^3) = H(X)$

   $H(X^4) \leq H(X)$

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

   $H(Y) \geq H(Y|X)$

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

   Note: Roni loves questions like this, so make sure you fully understand entropy for a varied type of distributions. Additionally, you might want to understand these kind of rule for expectation/variance/etc

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

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

Please Calculate:

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

$P(X) = 2 : .25, 3 : .25, 4 : .25, 5 : .25$

We calculate the marginal probability for each value of X by summing the values across each row.

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

$P(X|Y = 3) = 2 : 0, 3 : 0, 4 : 1, 5 : 0$

(c) Correlation coefficient:

$\rho(X, Y) = -1$ Explanation: all points lie along a straight line, whose slope is negative.

(d) Mutual Information (in bits):

$I(X; Y) = 2$

Explanation: Given X, Y is fully known, so

$$I(X; Y) = H(Y) = H(0 : 25; 0 : 25; 0 : 25; 0 : 25) = 2bits$$

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?

   Because you are solving a linear system $Xw = y$, and because the matrix X has full rank, you will exactly fit the training data (0 error on the training set).

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

   Worse. Because your model is so flexible (with as many parameters as training points) it's easily capable of overfitting the training data, but this doesn't assure that we'll do well on previously unseen 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.

   There may be multiple reasonable answers two that come to mind: (i) the most obvious is to A simple (and popular) approach would be to add features one at a time greedily choosing at each step the feature that explains the most variance in the label (i.e. that leads to the biggest reduction in squared error).
   (ii) While we haven't yet discussed regularization at length, another reasonable approach is to optimize our linear model using a sparsifying regularizer like L1. For suitably strong regularization strength, this model will produce parameter values of 0 corresponding to all but the desired number of parameters.

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 $x_1^2, x_1x_2, ...$
   - 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. $x_i + 2x_j - 4.3x_k$

- If the new features have predictive value, they should help.
- If you're not overfitting, you will not improve performance by removing features.
- Nonlinear features might help if the label is linear in these transformed features but not linear in the raw features.
- Constant factors do not make a difference your model could always learn constant multipliers just by changing the weights in a corresponding fashion.
- The original model was already capable of learning any linear combination of the original features. Adding new features consisting of linear combinations will not provide any predictive value.

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