CS6720 Data Mining
Take Home Examination 1

Time: Due in or before Class on Monday **Max. marks:** 70

Honour Code: Please write out the following, sign and return it along with your exam. If you are submitting the exam online, still leave a sheet of paper in my pigeon hole.

I have not discussed the problems with any one else. I have not requested for solutions in forums. I have used only my class notes, other materials provided on Moodle, and whatever is already available on the web.
(Hint: If the answer to any question is *It depends*, explain clearly the factors it depends on, and the nature of the dependence.)

1. **(5 marks)** A regularizer implicitly or explicitly penalizes complex models, which is one of the roles of a prior in the Bayesian setting. Consider a Bayesian setting for parameter estimation, what will the parametric form of the prior distribution in order to perform l_2 regularization?

2. **(10 marks)** In spectral clustering, suppose we are interested in partitioning the graph into 3 sets. Let v_2 denote the range normalized *Fiedler vector*, i.e., the elements of the vector lie between 0 and 1.[1] Can the scheme given below be used to find the cluster assignments? Explain your answer.

   `If` $v_2^{(i)}$ `< 1/3 then assign node` i `to cluster '1'`
   `else if 1/3` $\leq$ $v_2^{(i)}$ `< 2/3 then assign to cluster'2'`
   `else assign to cluster '3'`
   where $v_2^{(i)}$ denotes the i-th element of v_2.

3. **(5 marks)** Consider a scenario where the notion of shingling is applied at the word level rather than at the level of characters, i.e., consider n consecutive words as a shingle. Would this be equivalent to a n-gram model of the text? What are the advantages and difficulties in using such a representation? (Hint: Think about the size of the representation, indexing, etc.)

[1] The most negative number will be 0 and the largest number will be 1. Look up *min-max* normalization, if you want more explanation.

4. **(10 marks)** In the case of kNN classifiers expectation at a point is approximated by voting over a region. Suppose you have an efficient data structure and storage mechanism (based on LSH) that can answer *Near* neighbour queries efficiently but not nearest neighbour queries. How would kNN perform if this mechanism is used to perform the nearest-neighbour search? Would it be a good idea to still perform kNN? How would 1NN perform in this scenario?

5. **(5 marks)** When we discussed gradient boosting, we talked about estimating the residual gradients of the loss function and fitting the next iteration of the base learner to predict the residuals. Regression tree is one of the most popular base learners used in gradient boosting. Given the number of parameters of a regression tree and the complex dependency between the parameters and the prediction error, how would you estimate the residual gradients in the case of regression trees?

6. **(10 marks)** We typically use a 0-1 loss function for modeling classification problems. In some cases when it is hard to optimise the 0-1 loss function, we resort to surrogate loss functions that produce the same solutions as 0-1 loss, but are easier to optimise. A different loss function that we encountered in class is the exponential loss function used in AdaBoost. Can it be used as a surrogate to the 0-1 loss? If yes, give arguments to support your claim. If no, provide a counter example in which exponential loss has a minima different from that of the 0-1 loss.

7. Consider a Bayesian setting for parameter estimation. Provide the parametric form of distributions for the scenarios given below :

   (a) **(5 marks)** When the likelihood is a probability density function over a continuous random variable X, what could be the prior distribution?

   (b) **(5 marks)** A document is represented by shingles and we have a strong domain knowledge that shorter documents are more likely, what could be the prior distribution?

   (c) **(5 marks)** Consider a modified Jaccard distance as defined below:

$$d(x, y) = 1 - \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)} \tag{1}$$

Thus if x and y are equal, then the distance is 0. If we are interested in modeling the distribution over distances given by the above measure, what would be a suitable family of distributions? (Hint: Think about the range of the distance measure. Note that this is the regular Jaccard distance when the features are Boolean.)

8. **(10 marks)** Consider the modified Jaccard distance specified in Eqn. 1. Would the *minhashing* scheme as described in class work with this distance? Specifically can you give the same LSH guarantees that minhashing provides for the regular Jaccard distance in this case. Can you give any LSH guarantees for this distance? Justify your answers.

All the Best