National Graduate Programme in Computer Science

Data Mining and Machine Learning

Mid-Semester Examination, I Semester, 2018–2019

Date : 28 September, 2018
Duration : Two hours

Marks : 30
Weightage : 20%

1. In the market-basket analysis problem, suppose the set of items I has size 10^6, the number of transactions T is 10^9 and each transaction $t \in T$ contains at most 10 distinct items. Compute upper bounds for F_1 and F_2, the number of frequent itemset of size 1 and 2, respectively, for a support value of 0.1%. *(4 marks)*

2. Suppose we build a decision tree to do binary classification on a given set of training data, without any pruning, and we discover a leaf node that is not pure—it has representatives of both classes. What can we infer us about the attributes being used for classification? *(4 marks)*

3. We want to build a classifier to assign topics to a corpus of documents. Each document is modelled as a bag of words. However, we have additional structural information about each document. The words that appear in the title are listed separately from those that appear in the body. For a given topic, the generative model first generates words in the title with some distribution, and then generates words in the body, with a possibly different distribution. When assigning a topic to a document, we would like to give weightage w_t to the title and w_b to the body, where $w_t + w_b = 1$. Explain how to modify the standard naïve Bayes classifier to achieve this. *(4 marks)*

4. Explain why precision and recall are difficult to achieve simultaneously in a classifier. Describe an example where high precision is preferrable to high recall and another example where the converse is true. *(4 marks)*

5. The uniform convergence theorem tells us that for a hypothesis class $\mathcal{H}$ and thresholds ϵ and δ greater than zero, if a training set S of size $n \geq (1/2\epsilon^2)(\ln|\mathcal{H}|\ln(2/\delta))$ is drawn from a distribution D, then with probability greater than or equal to $1-\delta$, every $h \in \mathcal{H}$ satisfies $|err_S(h) - err_D(h)| \leq \epsilon$.

   Suppose we have a binary classification problem in which each data item is described as a vector $(a_1, a_2, \ldots, a_k)$ over k boolean attributes $(A_1, A_2, \ldots, A_k)$. Classifiers are expressed as disjunctions of positive or negative propositions, where each attribute is treated as a propositional variable. For instance, the classifier $A_2 \vee \neg A_6 \vee A_8$ will select all items $(a_1, a_2, \ldots, a_k)$ where a_2 or a_8 are true or a_6 is false. What sample size S should we choose to avoid overfitting with high probability, given thresholds ϵ and δ? *(5 marks)*

6. We have seen that the VC dimension of axis-parallel rectangles is exactly 4. Suppose we consider rectangles that can be rotated arbitrarily. Show that the VC dimention is at least 7. (Hint: Consider points arranged as a regular heptagon.) *(5 marks)*

7. Explain the difference between batch gradient descent and stochastic gradient descent and provide advantages and disadvantages for each. *(4 marks)*