National Graduate Programme in Computer Science

Data Mining and Machine Learning

Final Examination, I Semester, 2018–2019

Date : 30 November, 2018
Duration : Two hours

Marks : 40
Weightage : 40%

1. Consider logistic regression and decision trees as classification models.

   (a) Which of these models exhibit bias and which exhibit variance? Explain your answer.

   (b) In these models, what measures can we take to counteract bias and variance?

(6 marks)

2. How can an inverted index on individual words be adapted to answer queries based on phrases (sequences of words)?

(4 marks)

3. PageRank requires the stochastic matrix representing transition probabilities between web documents to be irreducible and aperiodic.

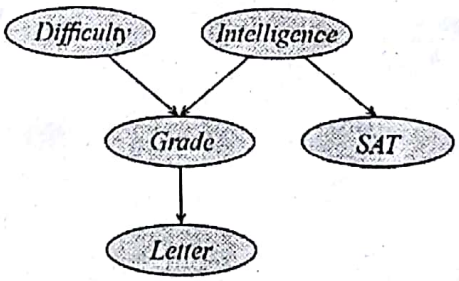
   (a) What do these properties mean, graph theoretically?

   (b) Why are these properties required?

   (c) How does PageRank ensure that these conditions are met?

(5 marks)

4. The Bayesian network to the right depicts the following situation. A student has taken a course a long time ago and asks the instructor for a reference letter. The instructor has completely forgotten the student but gives a reference letter based on the student's grade in the course. The student has also taken a standardized test. The nodes represent the following.



- *Difficulty* : the difficulty level of the course
- *Intelligence* : the student's basic intelligence
- *Grade* : the student's grade
- *SAT* : the student's score in the standardized test
- *Letter* : the quality of the reference letter

(a) The student now applies for a job using the reference letter and the standardized test scores. Expand the network to include the following variables. Explain your choices of dependencies.

- *Clarity* : the clarity of the instructor's teaching
- *Job* : whether the student gets the job
- *Happy* : whether the student is happy with his/her situation

Scanned by CamScanner

(b) In your model:

(i) Is *Clarity* independent of *Intelligence* given *Job*?

(ii) Is *Clarity* independent of *Intelligence* given *SAT*?

Justify your answers intuitively and show how they can be formally inferred from the structure of the network.

(8 marks)

5. Explain the difference between filtering and smoothing in a hidden Markov model. Given hidden states $X_0, X_1, , \ldots, X_t$ and evidence $e_1, e_2, \ldots, e_t$, derive a recursive expression to evaluate $P(X_k \mid e_{1:t})$, where $k < t$ and, as usual, for a sequence of values $y_0, y_1, \ldots, y_n$, $y_{i:j}$ denotes the segment $y_i, y_{i+1}, \ldots, y_j$.

(6 marks)

6. Consider the set $B_k = \{0,1\}^k$ of k-dimensional bit vectors, for a fixed $k > 0$. Let $C \subseteq B_k$ denote a concept class. When can such a concept C be represented by a multilayer perceptron network? Explain your answer.

(5 marks)

7. (a) For $z = wx + b$, how does the shape of the sigmoid function $\sigma(z) = (1 + e^{-z})^{-1}$ vary with w and b?

(b) Explain how to construct a neural network to approximate a "rectangular box" function $g(x)$ with height h between $x = \ell$ and $x = r$ for $0 \leq \ell < r \leq 1$. In other words, the function to be approximated is the following.

$$g(x) = \begin{cases} h & \text{if } \ell \leq x \leq r \\ 0 & \text{otherwise} \end{cases}$$

(6 marks)

Scanned by CamScanner