Dr. Mohit Kumar (CSE) E- 14/12/22

(265)

National Institute of Technology Hamirpur

End Semester Examination, Odd Sem, AY 2022-23

Machine Learning (CS-652)

Time: 2.30 PM to 5.30 PM **Date:** 14-12-2022

Duration: 3 hours **Max. Marks = 50**

Note 1: Attempt all questions from 1 to 5.

Note 2: If required to solve a question, make & and state your assumptions clearly.

---

1. **[4 + 6 = 10]**

(a) What do you understand by Instance based learning methods? Give an example of such a method. How do Instance based learning methods differ from model-based learning methods such as Logistic Regression?

(b) What is the main limitation of kNN? Explain using an example. What do you understand by k-fold cross validation? How can we use k-fold cross validation to find the optimal value of k in kNN model?

Variable	Definition
survival	Survival
pclass	Ticket class
sex	Sex
Age	Age in years
sibsp	# of siblings / spouses aboard the Titanic
parch	# of parents / children aboard the Titanic
ticket	Ticket number
fare	Passenger fare
cabin	Cabin number
embarked	Port of Embarkation

Figure 1: Description of Titanic dataset

2. **[6 + 4 = 10]**

(a) What do you understand by the term entropy in machine learning? Write the expression for calculating entropy and give its interpretation. What is Information Gain (IG)? What is the importance of IG in relation to Decision Trees (DT)? Why is sometimes Gini Impurity preferred to estimate IG instead of entropy?

(b) When constructing a DT for the Titanic dataset (Fig. 1), what problems could arise? How would you resolve those problems?

3. **[5 + 5 = 10]**

(a) What is the Bagging strategy for creating ensembles? Describe all the major considerations while building a Bagging model. Use appropriate illustrations wherever required.

(b) How does Boosting strategy for ensembling differ from the Bagging strategy? Explain briefly. What is the Cascading strategy in ensembling? Give two use cases where cascading is useful.

4. $[5 + 5 = 10]$

(a) What do you understand by a well posed learning problem? What is the need of machine learning?

(b) Explain the Newton method for solving the likelihood expression of Logistic Regression. To solve for the optimal parameters in Linear Regression, we have a deterministic method called the Normal Equation method. However, in practice, Gradient Descent is almost always the preferred choice. Explain the reasoning behind this behaviour.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Figure 2: PlayTennis data

5. $[6 + 4 = 10]$

(a) Formulate a Naive Bayes model for the dataset in Fig. 2. Explain your reasoning behind each step in the model formation. Show an example test case and how the outcome will be estimated given the model.

(b) The distance between the positive and negative hyperplanes in SVM formulation is taken to be 1. Show that any arbitrary distance would be valid in the same formulation. What are string kernels and graph kernels? Explain using examples and their applications.

***