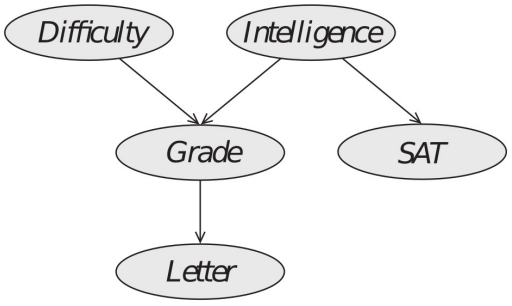
Data Mining and Machine Learning

Final Examination, II Semester, 2021–2022

Date : 11 May, 2022 | Marks : 40
Duration : 3 hours | Weightage : 40%

1. Explain how clustering can be used for semi-supervised learning. *(4 marks)*
2. DBscan and LOF are both density-based approaches for outlier detection. Explain the difference in how density is defined and used in the two approaches. *(6 marks)*
3. Explain how locally linear embeddings are computed. *(5 marks)*
4. There are two biased coins $\{c_1, c_2\}$ with probabilities of heads $\{p_1, p_2\}$, respectively. The following action is performed 50 times: one of the coins is chosen uniformly at random and tossed twice. You are given the resulting sequence of 100 coin tosses without any information about which coin was chosen at each step. Describe, in algorithmic pseudocode, an iterative procedure to estimate p_1 and p_2. Assume that you have prior information that $p_1 < 0.5$ and $p_2 > 0.5$. *(6 marks)*
5. Explain whether we can use kernels with linear separators that are computed using the perceptron algorithm. *(4 marks)*
6. Explain the distinction between parameters and hyperparameters in neural networks. Give two examples of hyperparameters. *(4 marks)*
7. Give examples of Markov chains that are (a) not irreducible and (b) not aperiodic. Explain why these can fail to have a stationary distribution. *(5 marks)*
8. 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 in the course
- *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 in the course
- *Job* : whether the student gets the job
- *Happy* : whether the student is happy with his/her situation

(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 explain how they can be formally inferred from the structure of the network.

(6 marks)

Data Mining and Machine Learning 2023

Final Exam

Chennai Mathematical Institute

26 April 2023, 09:30–12:30 (3 hours) Marks: 50. Weightage: 50%

1. Our task is to build a classifier to predict if a person is at risk from diabetes. The training set has patient data with age, gender, presence or absence of 1000 genes in their DNA, and whether or not they have diabetes.

   (a) Describe how to build and use a Naive Bayes classifier for this dataset. Mention any assumptions that you have made. *(5 marks)*

   (b) Suppose we are given additional patient data which does not include information about whether each individual has diabetes or not. Can we use this additional data to improve your classifier? If so, how? *(5 marks)*

2. We are studying a species of fish found in the Indian Ocean. This species has evolved into sub-species based on the region of the ocean that they live in. We have a dataset of latitudes and longitudes where this species of fish has been observed. How would we estimate the number of sub-species of this fish species? *(5 marks)*

3. We have a dataset of customers who have taken personal loans from a bank over the past 30 years. The data about each customer consists of their age, loan amount, address and whether or not they have paid back the loan in full. Additionally, for the past 5 years, we have data about whether the customer was a salaried employee or a businessman.

   The bank wishes to estimate the probability of full repayment of the loan for salaried employees. How can we use expectation-maximization to estimate this probability based on the entire data over 30 years? *(5 marks)*

4. We have a dataset of 10000 rice farms growing the same variety of rice. For each farm, we have data about how much fertilizer was used per acre and the yield per acre (in kgs). In addition, for 100 equally spaced days between planting and harvesting, there is data about the water-level (in cm), the average temperature and the average height of the rice plants.

   (a) Explain how to build a model that predicts the yield of rice, given all the other data about a farm. *(2 marks)*

   (b) We wish to run our ML model on a mobile phone (in an app). How can we reduce the computational cost without compromising too much on accuracy? *(3 marks)*

5. Suppose we have a neural network with four input features x_1, x_2, x_3, x_4 and a single output y. As usual, we assume that each pair of adjacent layers is completely connected and there is a single output layer. How many parameters do we have to estimate in the following situations?

   (a) A shallow network with 1 hidden layer consisting of 18 nodes.

   (b) A deep network with 3 hidden layers, where the first two layers have 3 nodes each and the third layer has 2 nodes. *(5 marks)*

6. We made the following assumptions about the loss (cost) function C for neural networks.

- For each input x, $C(x)$ is a function of only the output layer activation.
- The total cost across the training set is the average of the individual input costs.

Explain why these assumptions are important for effective learning of the parameters.

(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) Given two input features x_1, x_2, explain how to construct a neural network to approximate a "rectangular box" function $g(x_1, x_2)$ with height h for $\ell_1 \leq x_1 \leq r_1$ and $\ell_2 \leq x_2 \leq r_2$. In other words, the function to be approximated is the following.

$$g(x_1, x_2) = \begin{cases} h & \text{if } \ell_i \leq x_i \leq r_i, i \in \{1, 2\} \\ 0 & \text{otherwise} \end{cases}$$

(5 marks)

8. Consider a neural network that is layered and completely connected. Suppose we initialize two nodes n_1 and n_2 from the same layer with the same biases and same weights on incoming and outgoing edges. What can we say about the final weights and biases that will be learned for n_1 and n_2 through backpropagation? What can we conclude about initialization strategies for such networks? *(5 marks)*

9. In a nuclear power station, an alarm is triggered when a temperature gauge exceeds a given threshold. The gauge measures the temperature of the core of the reactor. Consider the boolean variables A (alarm sounds), F_A (alarm is faulty), and F_G (gauge is faulty) along with multivalued variables G (gauge reading) and T (actual core temperature).

   (a) Draw a Bayesian network for this scenario, given that the gauge is more likely to fail when the core temperature gets too high. Explain the structure of your network.

   (b) Suppose G and T each take just two values, normal and high. Assume that the gauge gives the correct temperature with probability x when it is working and with probability y when it is faulty. Describe the conditional probability table for G.

(5 marks)

Data Mining and Machine Learning

Final Examination, II Semester, 2023–2024

Date : 3 May, 2024 | Marks : 40
Duration : 3 hours | Weightage : 40%

1. There are three biased coins c_1, c_2, and c_3. You are given a sequence of 1000 coin tosses, where each outcome corresponds to tossing one of $\{c_1, c_2, c_3\}$, chosen uniformly at random. Let $\{p_1, p_2, p_3\}$ be the probabilities of heads for the coins $\{c_1, c_2, c_3\}$, respectively. You have prior information that p_1 is less than 0.5 and p_2 and p_3 are greater than 0.5. Describe, in algorithmic pseudocode, an iterative procedure to estimate $\{p_1, p_2, p_3\}$. *(5 marks)*

2. Explain how to cluster points using a mixture of Gaussians. Can this also be used to detect outliers? *(5 marks)*

3. Explain how clustering can be used for image segmentation — that is, to identify objects in an image. *(5 marks)*

4. The 0−1 loss function assigns a cost of 1 to every misclassified input and a cost of 0 to every correctly classified input. This loss function is minimized when the model makes no errors on the training data. Explain with respect to the perceptron algorithm why the 0−1 loss function is not always adequate to learn a good model. *(5 marks)*

5. (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) Given two input features x_1, x_2, explain how to construct a neural network to approximate a "rectangular box" function $g(x_1, x_2)$ with height h for $\ell_1 \leq x_1 \leq r_1$ and $\ell_2 \leq x_2 \leq r_2$. In other words, the function to be approximated is the following.

$$g(x_1, x_2) = \begin{cases} h & \text{if } \ell_i \leq x_i \leq r_i, i \in \{1, 2\} \\ 0 & \text{otherwise} \end{cases}$$

(5 marks)

6. Consider a neural network that is layered and completely connected. Suppose we initialize two nodes n_1 and n_2 from the same layer with the same biases and same weights on incoming and outgoing edges. What can you say about the final weights and biases that will be learned for n_1 and n_2 through backpropagation? What can you conclude about initialization strategies for such networks? *(5 marks)*

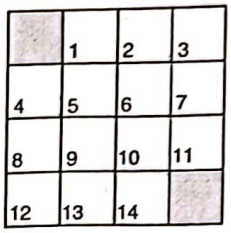
7. Two astronomers independently count stars in the same region of the sky using their telescopes. The region has N stars. The counts reported by the astronomers are M_1 and M_2, respectively. Each astronomer has a small probability of miscounting the stars by ± 1. It is also possible that their telescopes are faulty and do not focus properly, denoted by boolean events F_1 and F_2, respectively. With a faulty telescope, an astronomer may undercount by as many as 3 stars.

   (a) Draw a Bayesian network to represent the relationship between N, M_1, M_2, F_1 and F_2.

   (b) Suppose $M_1 = 12$ and $M_2 = 14$. What are the possible values of N for each of the different combinations of F_1 and F_2?

(5 marks)

8. Consider the 4×4 grid-world to the right. The non-terminal states are $\{1, 2, \ldots, 14\}$ and the terminal states are the shaded squares. There are four actions, $\{\text{up}, \text{down}, \text{left}, \text{right}\}$, which result in a deterministic move in the given direction. A move that would take the agent off the grid leaves the position unchanged. The reward is -2 for any transition that results in a change of position. A move off the grid that does not change the position has a reward of -1. Formally, $r(s, a, s') = -2$ if $s \neq s'$ and $r(s, a, s') = -1$ if $s = s'$.



(a) Consider the uniformly random policy π that chooses each of the four directions with equal probability. Assume we start with an initial value $v(s) = 0$ for each state s. Compute one iteration of v_π.

(b) Describe the new policy after applying policy improvement based on this one step computation of v_π.

(5 marks)

Data Mining and Machine Learning

Mid-Semester Examination, II Semester, 2024–2025

Sample questions

1. In the market-basket analysis problem, suppose the set of items I has size 10^7, the number of transactions T is 10^{10} 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%.

2. Recall that a class association rule has the class attribute as its target. To reduce overfitting, a class association rule can be generalized by dropping attributes from its left hand side and checking if the performance improves over random test data.

   Given a decision tree, explain how to interpret paths in the tree as class association rules. How can we apply the generalization strategy for association rules to generalize decision trees? In what way would this be different from generalization through the usual method of pruning?

3. Your team has computed the solution to a linear regression problem on n attributes as $\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$. Your partner argues that the relative importance of the attributes can be computed from the coefficients. The most significant attribute is the one with the largest coefficient (in magnitude), the second most significant attribute is the one with the second largest coefficient, and so on. Explain whether your partner's claim is justified.

4. You are building a decision tree on tabular data with attributes $\{A_1, A_2, \ldots, A_7\}$ where $\{A_3, A_5\}$ are numeric and the other five attributes are categorical. Attribute A_3 takes integer values in the range $[-100, 100]$ and attribute A_5 takes integer values in the range $[1, 10000]$. There are 2000 items in the training set. You adopt a pre-pruning strategy to build the tree where you do not split any node with 50 items or fewer. Across all possible decision trees that can be built on this training set, what is the maximum height of the resulting tree? Explain your answer.

5. We want to build a naïve Bayes classifier for junk email. Each message is modelled as a bag of words. However, an email message has some structure that can be exploited: we can separate out the sender's address, the subject line and the body of the message. We assume all three parts are constucted from a common vocabulary, but with different probability distributions. The corresponding generative model first generates the sender's address with some probability distribution, then the subject line, with a different distribution, and finally the body, with yet another distribution. When classifying an email as junk, we would like to give weightage w_1 to the sender's address, w_2 to the subject line and w_3 to the body, $w_1 + w_2 + w_3 = 1$. Explain how to modify the standard naïve Bayes classifier to achieve this.

6. How can we use a random forest classifier to rank input features in order of importance? Why is this calculation more effective for a random forest than for a single decision tree?

7. When we use bagging, we need not keep aside a separate test set to validate our model. Explain.

Data Mining and Machine Learning

Mid-Semester Examination, II Semester, 2021–2022

Date : 16 March, 2022	Marks : 30
Duration : 2 hours + 0.5 hours upload time	Weightage : 20%

1. A number of new vaccines are being deployed to treat a recently discovered disease. Reports are emerging of patients having side effects caused by vaccinations. Some side effects are vaccine-specific, some occur across vaccines.

   For each reported case, there is information available about the nature of the side effect, the vaccine used, demographic details about the patient (age, gender, race, ...) as well as information about prevailing health conditions of the patient (diabetes, hypertension, ...) that may create complications.

   Explain how market-basket analysis can help doctors determine risk factors associated with vaccinations, in general, and specific vaccines, in particular. *(5 marks)*

2. Consider the following situation when building a decision tree for binary classification. We have a node with 60 samples, with 40 belonging to the majority class and 20 to the minority class. We have only one attribute A available to query, which splits the node into two subsets S_1 and S_2. S_1 has 35 samples with 21 in the majority class and S_2 has 25 samples with 19 in the majority class.

   Compute the impurity gain using misclassification rate as a measure of impurity and contrast it with the impurity gain due to a nonlinear impurity measure such as Gini index or entropy. What can you conclude from this? *(5 marks)*

3. Explain why squared error is a natural loss function for normal regression while cross entropy is more suitable for logistic regression. *(5 marks)*

4. We want to build a naïve Bayes classifier for junk email. Each message is modelled as a bag of words. However, an email message has some structure that can be exploited: we can separate out the sender's address, the subject line and the body of the message. We assume all three parts are constucted from a common vocabulary, but with different probability distributions. The corresponding generative model first generates the sender's address with some probability distribution, then the subject line, with a different distribution, and finally the body, with yet another distribution. When classifying an email as junk, we would like to give weightage w_1 to the sender's address, w_2 to the subject line and w_3 to the body, $w_1 + w_2 + w_3 = 1$. Explain how to modify the standard naïve Bayes classifier to achieve this. *(5 marks)*

5. How can we use a decision tree to rank input features in order of importance? Compare the effectiveness of this calculation if we use a random forest rather than a single decision tree. *(5 marks)*

6. Suppose we apply gradient boosting to solve a regression problem, using a sequence of regression trees. Describe a strategy to estimate the optimum number of regression trees to use. *(5 marks)*

Data Mining and Machine Learning

Mid-Semester Examination, II Semester, 2023–2024

Date : 2 March, 2024 Marks : 30
Duration : 2 hours Weightage : 20%

1. In the market-basket analysis problem, suppose the set of items I has size 10^7, the number of transactions T is 10^{10} 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%. *(5 marks)*

2. Recall that a class association rule has the class attribute as its target. To reduce overfitting, a class association rule can be generalized by dropping attributes from its left hand side and checking if the performance improves over random test data.

   Given a decision tree, explain how to interpret paths in the tree as class association rules. How can we apply the generalization strategy for association rules to generalize decision trees? In what way would this be different from generalization through the usual method of pruning? *(5 marks)*

3. Your team has computed the solution to a linear regression problem on n attributes as $\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$. Your partner argues that the relative importance of the attributes can be computed from the coefficients. The most significant attribute is the one with the largest coefficient (in magnitude), the second most significant attribute is the one with the second largest coefficient, and so on. Explain whether your partner's claim is justified. *(5 marks)*

4. How can we use a random forest classifier to rank input features in order of importance? Why is this calculation more effective for a random forest than for a single decision tree? *(5 marks)*

5. We have a dataset $X = \{x_1, x_2, \ldots, x_N\}$ equipped with a symmetric distance function: $d(x_i, x_j) = d(x_j, x_i)$ is the distance between x_i and x_j. We construct an $N \times N$ matrix D such that $D[i, j] = d(x_i, x_j)$. We can cluster the N columns of D using the usual Euclidean distance in N dimensions, since each column is a vector of length N. Explain whether the clusters formed by the columns of D have any meaningful interpretation with respect to the original set X. *(5 marks)*

6. Explain how locally linear embeddings are computed. *(5 marks)*

Data Mining and Machine Learning

Practice Assignment 1, II Semester, 2024–2025

3 February, 2025

1. A number of new vaccines are being deployed to treat a recently discovered disease. Reports are emerging of patients having side effects caused by vaccinations. Some side effects are vaccine-specific, some occur across vaccines.

   For each reported case, there is information available about the nature of the side effect, the vaccine used, demographic details about the patient (age, gender, race, ...) as well as information about prevailing health conditions of the patient (diabetes, hypertension, ...) that may create complications.

   Explain how market-basket analysis can help doctors determine risk factors associated with vaccinations, in general, and specific vaccines, in particular.

2. In the market-basket analysis problem, suppose the set of items I has size 10^7, the number of transactions T is 10^{10} 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%.

3. You are building a decision tree on tabular data with attributes $\{A_1, A_2, \ldots, A_7\}$ where $\{A_3, A_5\}$ are numeric and the other five attributes are categorical. Attribute A_3 takes integer values in the range $[-100, 100]$ and attribute A_5 takes integer values in the range $[1, 10000]$. There are 2000 items in the training set. You adopt a pre-pruning strategy to build the tree where you do not split any node with 50 items or fewer. Across all possible decision trees that can be built on this training set, what is the maximum height of the resulting tree? Explain your answer.

4. The algorithm we described to build a decision tree is deterministic. However, we saw that the decision tree library implemented in Python's `sklearn` library uses a random seed. Why should a random seed be needed? (Hint: Consider the example from the iris dataset.)

5. An airport security system consists of a full body scanner followed by manual frisking. If the full body scanner beeps, the passenger is checked manually and then allowed to proceed if there is nothing amiss. If the full body scanner does not beep, no frisking is done.

   (a) In terms of the entries in the confusion matrix, what ratio should the full body scanner maximize to ensure that no suspicious person is let through unchecked?

   (b) Similarly, what ratio should manual frisking maximize?

Data Mining and Machine Learning 2023

Mid-Semester Exam

Chennai Mathematical Institute

23 February 2023, 09:30–11:30 (2 hours) Marks: 30, Weightage: 20%

The following questions carry 5 marks each.

1. A research paper contains the data of about 1000 university students with the following attributes: (1) the time they went the sleep previous night, (2) the time they woke up, (3) the meals they had during the day and (4) their concentration level during the day. Can you use this data to build a ML model that takes the first three attributes and predicts the concentration level with good accuracy for an *average person*? Explain your answer.

2. Consider the class of *decision stumps*, which are decision trees of height 1.

   (a) Consider the ensemble model A build by Bagging 100 decision stumps. Can A be represented as a decision stump?

   (b) Consider another ensemble model B built by AdaBoosting 100 decision stumps. Can B be represented as a decision stump?

   Justify your answers.

3. You are building a decision tree on tabular data with attributes $\{A_1, A_2, \ldots, A_7\}$ where $\{A_3, A_5\}$ are numeric and the other five attributes are categorical. Attribute A_3 takes integer values in the range $[-100, 100]$ and attribute A_5 takes integer values in the range $[1, 10000]$. There are 2000 items in the training set. You adopt a pre-pruning strategy to build the tree where you do not split any node with fewer than 50 items. Across all possible decision trees that can be built on this training set, what is the maximum height of the resulting tree? Explain your answer.

4. An airport security system consists of a full body scanner that all passengers pass through, followed by manual frisking. Passengers who pass through the scanner without setting off the alarm are let through without frisking. Passengers who set off the alarm are manually frisked. If nothing suspicious is found during manual frisking they are let through, otherwise they are detained.

   Think of the full body scanner and manual frisking as two classifiers to detect suspicious passengers. A false positive is an innocent passenger who is classified as suspicious and a false negative is a suspicious passenger who is classified as innocent.

   In terms of the confusion matrix, explain what metrics the two stages in the classifier should aim to optimize so that, to the extent possible, all suspicious passengers are detained, and all innocent passengers are let through with minimum hassle.

5. You often hear airlines announce that a departure is delayed due to the late arrival of the incoming flight. Since these delays cascade, it would be useful to know, for instance, if a delay in a morning flight from Amritsar to Delhi would have an impact on your afternoon flight from Hyderabad to Cochin. Suppose the airline provides you with daily data about delayed flights for the past year. Each day's data has the list of flights that were delayed on that day. Explain how you could use this data to answer questions of the form "Does a delay in flight A imply a delay in flight B?" Be as precise as possible about how you model this problem and the analysis that you need to perform on the model.

6. Suppose your team computes the solution to a linear regression problem on n attributes as $\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$. Your partner argues that the relative importance of the attributes to the final answer can be computed from the coefficients. The most significant attribute has the largest coefficient (in magnitude), the second most significant attribute has the second largest coefficient, and so on. Explain whether your partner's claim is justified.

Data Mining and Machine Learning

Mid-Semester Examination, II Semester, 2024–2025

Date : 6 March, 2025
Duration : 3 hours

Marks : 30
Weightage : 20%

1. A number of new vaccines are being deployed to treat a recently discovered disease. Reports are emerging of patients having side effects caused by vaccinations. Some side effects are vaccine-specific, some occur across vaccines.

   For each reported case, there is information available about the nature of the side effect, the vaccine used, demographic details about the patient (age, gender, race, ...) as well as information about prevailing health conditions of the patient (diabetes, hypertension, ...) that may create complications.

   Explain how market-basket analysis can help doctors determine risk factors associated with vaccinations, in general, and specific vaccines, in particular. *(4 marks)*

2. 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)*

3. Suppose we are building a naïve Bayesian classifier and some attribute values are missing in the training data. What problem can this cause with prediction and how can we mitigate the situation? *(3 marks)*

4. The algorithm we described to build a decision tree is deterministic. However, we saw that the decision tree library implemented in Python's `sklearn` library uses a random seed. Why should a random seed be needed? *(3 marks)*

5. (a) Explain whether the following statements are true.

   (i) Polynomial regression can always achieve 100% accuracy with respect to the training data.

   (ii) A decision tree can always achieve 100% accuracy with respect to the training data.

   (b) Explain whether 100% accuracy with respect to the training data is a desirable target to achieve.

   (4 marks)

6. Explain why cross entropy is a more suitable loss function than squared error for logistic regression. *(4 marks)*

7. How can we use a random forest classifier to rank input features in order of importance? Why is this calculation more effective for a random forest than for a single decision tree? *(4 marks)*

8. What is a validation set? Explain how a validation set can be used to determine the optimum number of models to use in boosting. *(4 marks)*
