Assignment 1

Introduction to Machine Learning

Prof. B. Ravindran

1. Which of the following is a supervised learning problem?

   (a) Grouping people in a social network.

   (b) Predicting credit approval based on historical data

   (c) Predicting rainfall based on historical data

   (d) all of the above

   Sol. (b) and (c)
   (a) does not have labels to indicate the groups. (b) and (c) have the correct answers for the examples in the dataset.

2. Which of the following are classification problems? (multiple options may be correct)

   (a) Predicting the amount of rain fall for a particular day.

   (b) Predicting whether it will rain or not on a particular day.

   (c) Given all the actors in a movie, predicting its genre.

   (d) Filtering of spam messages

   Sol. (b), (c), (d)
   In (a), the amount of rain fall is a continuous variable. So it is a regression task. In other 3 options, the output variable is a discrete class, so these are classification tasks.

3. Which of the following is a regression task?

   (a) Predicting whether a given document is related to sports or not

   (b) Predicting the gender of a human

   (c) Predicting the share price of a company

   (d) Finding clusters in a given data

   Sol. (c)
   In (c), the task is to predict a continuous variable. Tasks (a) and (b) involve the prediction of a discrete variable. Task (d) is an unsupervised learning task.

4. Which of the following is an unsupervised learning task?

   (a) Learning to ride a bicycle.

   (b) Grouping related documents from an unannotated corpus.

   (c) Grouping students into following groups- primary, high school, college.

   (d) both (a) and (c)

   Sol. (b)
   (a) is an RL problem. (c) is a classification problem. In (b), since the corpus is unannotated we do not have labels.

5. Which of the following is a categorical feature?

   (a) Weight of a person

   (b) Ethnicity of a person

   (c) Height of a person

   (d) Income of a person

   Sol. (b)
   Height, weight and income are continuous valued variables. Ethnicity on the other hand can take values from a limited set of values.

6. A new phone, E-Corp X1 has been announced and it is what you've been waiting for, all along. You decide to read the reviews before buying it. From past experiences, you've figured out that good reviews mean that the product is good 90% of the time and bad reviews mean that it is bad 70% of the time. Upon glancing through the reviews section, you find out that the X1 has been reviewed 1269 times and only 127 of them were bad reviews. What is the probability that, if you order the X1, it is a bad phone?

   (a) 0.1362

   (b) 0.160

   (c) 0.840

   (d) 0.773

   Sol. (b)
   For the solution, let's use the following abbreviations.

   - BP - Bad Phone
   - GP - Good Phone
   - GR - Good Review
   - BR - Bad Review

   From the given data, $Pr(BP|BR) = 0.7$ and $Pr(GP|GR) = 0.9$. Using this, Pr(BP—GR) = 1 - Pr(GP—GR) = 0.1.
   Hence,

$$\begin{aligned} Pr(BP) &= Pr(BP|BR) \cdot Pr(BR) + Pr(BP|GR) \cdot Pr(GR) \\ &= 0.7 \cdot \frac{127}{1269} + 0.1 \cdot \frac{1269 - 127}{1269} \\ &= 0.160 \end{aligned}$$

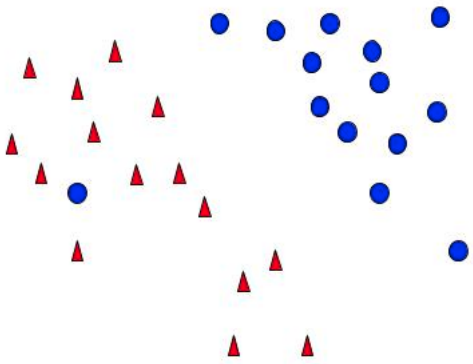
7. What would be the ideal complexity of the curve which can be used for separating the two classes shown in the image below.

   (a) Linear

   (b) Quadratic

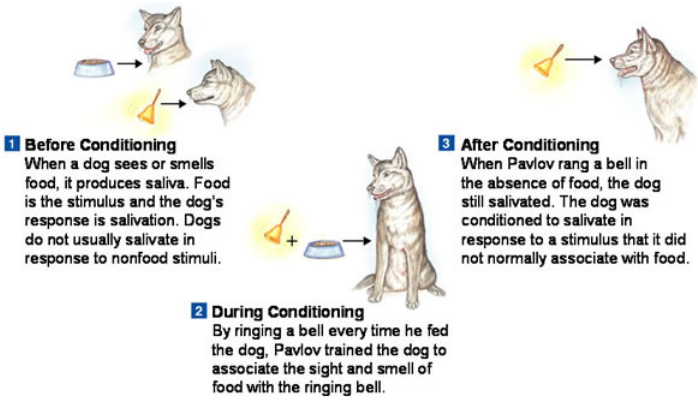
   (c) Cubic

   (d) insufficient data to draw conclusion



Sol. (a)
The blue point in the red region is an outlier (most likely noise). The rest of the data is linearly separable.

8. **Pavlov's Experiment**
Pavlov's experiment is a classic experiment conducted by the Russian physiologist Ivan Pavlov. He experiments on dogs shown in the image below.



(Image source - http://www.goldiesroom.org/)

Before conditioning dog responds with saliva only in presence of the food but after conditioning it starts salivating just with the bell. Select the correct option(s) about the experiment.

(a) In this experiment, the dog learns in a supervised setting

(b) In this experiment, the dog acts as a Reinforcement learning agent

(c) Comparing this experiment to Reinforcement learning theory, the various states are

- Presence of just the bell
- Presence of just food
- Presence of both food and bell

Sol. (b) and (c)
In this scenario, we see that as the dog interacts with the environment it is in, it starts to modify its behaviour based on rewards it receives from the environment (in this case, food). This is clearly a reinforcement learning situation. We also note the different states of the environment which differ as the experiment progresses - initially just the food is present resulting in the dog producing saliva, next the bell is introduced whenever the dog is given food as part of the conditioning process, and finally, the desired behaviour (i..e, salivation) can be produced when just the bell is rung (without the food).

9. One of the most common uses of Machine Learning today is in the domain of Robotics. Robotic tasks include a multitude of ML methods tailored towards navigation, robotic control and a number of other tasks. Robotic control includes controlling the actuators available to the robotic system. An example of this is control of a painting arm in automotive industries. The robotic arm must be able to paint every corner in the automotive parts while minimizing the quantity of paint wasted in the process. Which of the following learning paradigms would you select for training such a robotic arm?

   (a) Supervised learning

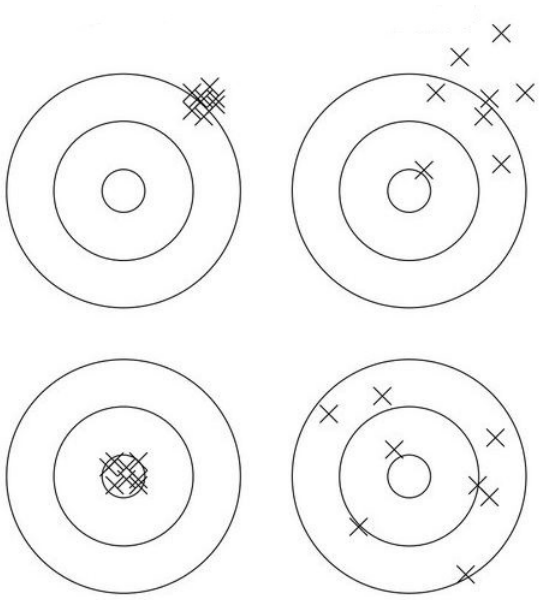
   (b) Unsupervised learning

   (c) Combination of supervised and unsupervised learning

   (d) Reinforcement learning

   Sol. (d)
   This kind of a learning problem warrants the use of Reinforcement Learning. We see that the robotic arm has to cover every corner, i.e. *maximize* the area covered and all the while *minimizing* the quantity of paint wasted in the process. One can design a primitive reward signal that takes into account the area covered and paint wasted (normalized to some extent) and use it to train a reinforcement learning agent.
   Supervised Learning cannot be used in this setting as there will not be any kind of training data available for the agent. Similarly, Unsupervised Learning cannot be used as it is not a pattern discovery problem.

10. Consider the following diagram where we assume that the target data points lie on the bullseye of the targets, and the predicted data points are represented by the X's.



Based on the above diagram, which of the following statements are true?

(a) the top left diagram represents a model with high bias and low variance

(b) the variance in the top right diagram is less than the variance in the bottom left diagram

(c) the variance in the top left diagram is less than the variance in the bottom right diagram

(d) to improve the model on the bottom right diagram, we should focus more on reducing variance rather than bias

Sol. (a), (c), (d)

Assignment 2

Introduction to Machine Learning

Prof. B. Ravindran

1. The parameters obtained in linear regression

   (a) can take any value in the real space

   (b) are strictly integers

   (c) always lie in the range [0,1]

   (d) can take only non zero values

   Sol. (a)

2. Given a set of n data points, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the best least squares fit $f(x)$ is obtained by minimization of:

   (a) $\sum_{i=1}^{n}[y_i - f(x_i)]$

   (b) $\min(y_i - f(x_i))$

   (c) $\sum_{i=1}^{n}[y_i - f(x_i)]^2$

   (d) $\max(y_i - f(x_i))$

   Sol. (c)

3. Consider forward selection, backward selection and best subset selection with respect to the same data set. Which of the following is true?

   (a) Best subset selection can be computationally more expensive than forward selection

   (b) forward selection and backward selection always lead to the same result

   (c) best subset selection can be computationally less expensive than backward selection

   (d) best subset selection and forward selection are computationally equally expensive

   (e) both (b) and (d)

   Sol. (a)
   Explanation Best subset selection has to explore all possible subsets which takes exponential time. It is not guaranteed that forward selection and backward selection take the same time. Forward selection and backward selection are computational much cheaper than best subset selection.

4. Adding interaction terms (such as products of two dimensions) along with original features in linear regression

   (a) reduces training error.

   (b) increases training error.

   (c) doesn't affect training error.

   Sol. (a)
   Adding interaction terms gives us additional modeling freedom. This can fit the training data better and hence leads to lower training error.

5. Consider the following five training examples

$$X = [2\ 3\ 4\ 5\ 6]$$

$$Y = [12.8978\ 17.7586\ 23.3192\ 28.3129\ 32.1351]$$

We want to learn a function $f(x)$ of the form $f(x) = ax + b$ which is parameterized by (a, b). Using squared error as the loss function, which of the following parameters would you use to model this function.

(a) (4 3)

(b) (5 3)

(c) (5 1)

(d) (1 5)

Sol. (b)

6. A study was conducted to understand the effect of number of hours the students spent studying to their performance in the final exams. You are given the following 8 samples from the study. What is the best linear fit on this dataset?

Table 1: Number of hours spent vs final score

Number of hours spent studying (x)	**Score in the final exam (0-100) (y)**
10	95
9	80
2	10
15	50
10	45
16	98
11	38
16	93

(a) y = -3.39x + 11.62

(b) y = 4.59x + 12.58

(c) y = 3.39x + 10.58

(d) y = 4.69x + 11.62

Sol. (b)

7. kNN regressor outputs the average of the k nearest neighbours of a query point. Consider a variant of kNN regressor where instead of returning the average we fit a linear regression model on the k neighbours. Which of the following **do not hold true** for this new variant?

(a) This method makes an assumption that the data is globally linear.

(b) This method makes an assumption that the data is locally linear.

(c) This method has lower bias compared to kNN

(d) This method has lower variance compared to kNN

Sol. (a) and (d)

8. Which of the following shrinkage methods leads to sparse solution?

   (a) Lasso regression

   (b) Ridge regression

   (c) Lasso and ridge regression both return sparse solutions

   Sol. (a)

9. Consider the design matrix X of dimension $N \times (p+1)$. Which of the following statements are true?

   (a) The rowspace of X is the same as the columnspace of $X^\top$

   (b) The rowspace of X is the same as the rowspace of $X^\top$

   (c) The eigenvectors of $XX^\top$ are the same as the eigenvectors of $X^\top X$

   (d) The eigenvalues of $XX^\top$ are the same as the eigenvalues of $X^\top X$

   Sol. (a) and (d)

10. Principal Component Regression (PCR) is an approach to find an orthogonal set of basis vectors which can then be used to reduce the dimension of the input. Which of the following matrices contains the principal component directions as its columns (follow notation from the lecture video)

    (a) X

    (b) S

    (c) X_c

    (d) V

    (e) U

    Sol. (d)

Assignment 3

Introduction to Machine Learning

Prof. B. Ravindran

1. Select the valid reasons for doing dimensionality reduction.

   (a) Variance can be controlled by controlling the number of features used in the model.

   (b) Lower number of features render the model to better interpretability.

   (c) Upon reducing the feature pool, we can sometimes increase the prediction accuracy, though not always.

   (d) Upon reducing the feature pool, we can reduce the time taken to run inference on the model at the cost of increased model building time.

   Sol. (a), (b), (c) and (d)

2. You work as a data analyst and your job is to analyze the growth of sale of a product. Suppose you decide to fit a linear regression model on a multivariate dataset. You find that a particular feature has a very high negative coefficient. What can you infer from this?

   (a) We can't comment on the importance of this feature without any additional information.

   (b) Since the feature has a large negative coefficient, so it is not an important feature. It is better to discard it.

   (c) Since the magnitude of the coefficient is very high, we should never discard that feature.

   Sol. (a)
   A high magnitude suggests that the feature is important. But if we haven't performed feature normalization we cannot comment on the feature importance by just looking at its magnitude. Even if we do feature normalization and a particular coefficient has a large magnitude, it may be the case that another feature is highly correlated with this feature and its coefficient also has a high magnitude with the opposite sign, in effect cancelling out the effect of the former. Thus, we cannot really remark on the importance of a feature just because its coefficient has a very large magnitude.

3. Which of the following dimensionality reduction methods are *hard* thresholding methods? (A method is *soft*-thresholding if it contains parameters that vary in a continuous fashion)

   (a) Forward Stepwise Regression

   (b) Backward Stepwise Regression

   (c) Forward Stagewise Regression

   (d) Ridge Regression

   (e) Least Absolute Shrinkage and Selection Operator (LASSO)

   (f) Principal Component Regression

   (g) Partial Least Squares Method

   Sol. (a), (b), (c), (f) and (g)

   (a) We either select a feature or we don't. We only have a finite set of solutions. Hence, hard thresholding.

(b) Same as (a).

(c) Same as (a).

(d) We have a continuous parameter λ that can be varied to produce an infinite set of solutions to the dimensionality reduction problem.

(e) Same as (d).

(f) There are only a finite set of Principal Components and in this method, we only select a finite subset of these. Hence, this is hard thresholding.

(g) Here, we only construct a finite set of features. We construct at most p features because the original data is fully specified by p features. No continuous parameters exist in this method and hence, this is hard thresholding.

4. Suppose that in applying linear regression, we are working with a data set where there are a large number of features, many of which we suspect to be redundant. We have discussed how using regularization we can constrain the magnitude of the weights associated with each feature. In fact, using regularization we can eliminate certain redundant features where the magnitude of the weights are found to be zero. Suppose we have a choice between two regularization schemes, L_2 regularization, where the additional penalty term is the sum of squares of the weights and L_1 regularization, where the penalty term is the sum of the absolute values of the weights. Which method do you think will result in eliminating more features (by reducing corresponding weights to zero)?

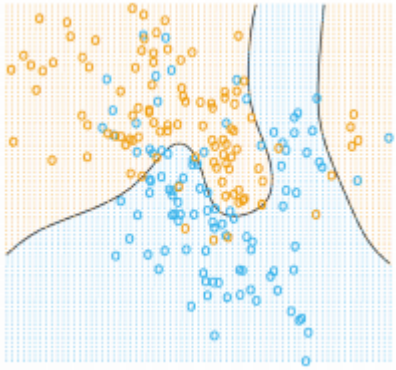
   (a) LASSO

   (b) Ridge

   (c) either of LASSO or Ridge regression can be used

   Sol. (a)
   LASSO leads to sparse solution. We can discard the features with zero coefficients. On the other hand ridge regression doesn't necessarily lead to sparse solutions.

5. Which of the following algorithm could have generated this decision boundary? (consider the situation where we do not allow for basis expansion)



   (a) Linear regression with indicator variable

   (b) Logistic regression

(c) 1-NN

(d) None of the above

Sol. (d)
Decision boundaries of (a), (b), (c) are linear. The image shows a non linear decision boundary.

6. Which of the following is true about a logistic regression based classifier?

   (a) The logistic function is non-linear in the weights

   (b) The logistic function is linear in the weights

   (c) The decision boundary is linear in the weights

   (d) The decision boundary is non-linear in the weights

   Sol. (a), (c) Refer to lecture videos.

7. For a two class classification problem, which among the following are true?

   (a) If both the covariance matrices are spherical and equal, the within class variance term has an effect on the LDA derived direction.

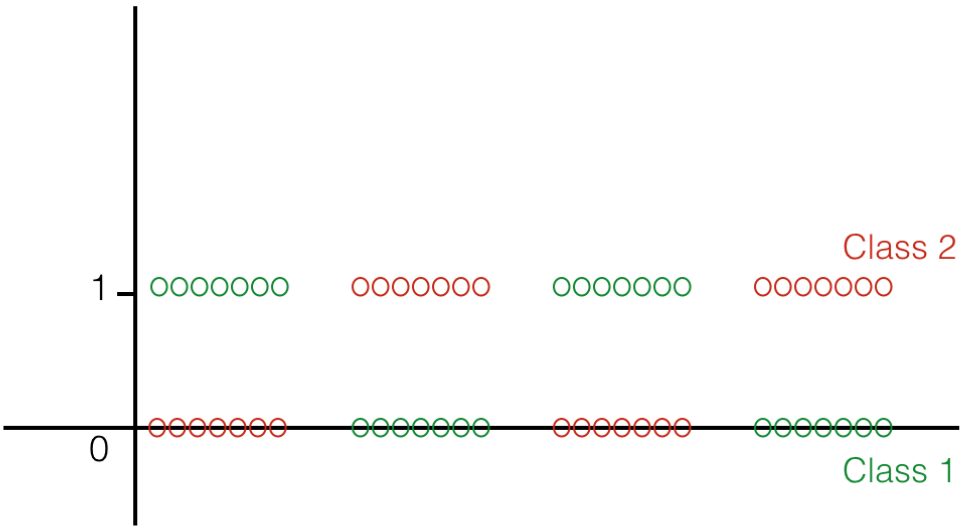
   (b) If both the covariance matrices are spherical and equal, the within class variance term has no effect on the LDA derived direction.

   (c) If both the covariance matrices are spherical but unequal, the within class variance term has an effect on the LDA derived direction.

   (d) If both the covariance matrices are spherical but unequal, the within class variance term has no effect on the LDA derived direction.

   Sol. (b), (d)
   Spherical covariance matrix represents perfect spherical classes in the space. For perfect spherical classes, there is no effect of changing the direction of projection. This can also be proved from the definition of covariance matrix and showing that rotations don't affect it.

8. In the following dataset, there are two classes arranged in the following manner:



Which of the following bases functions would you use to prevent any masking? Select all that apply.

(a) $1, x, x^2, x^3$

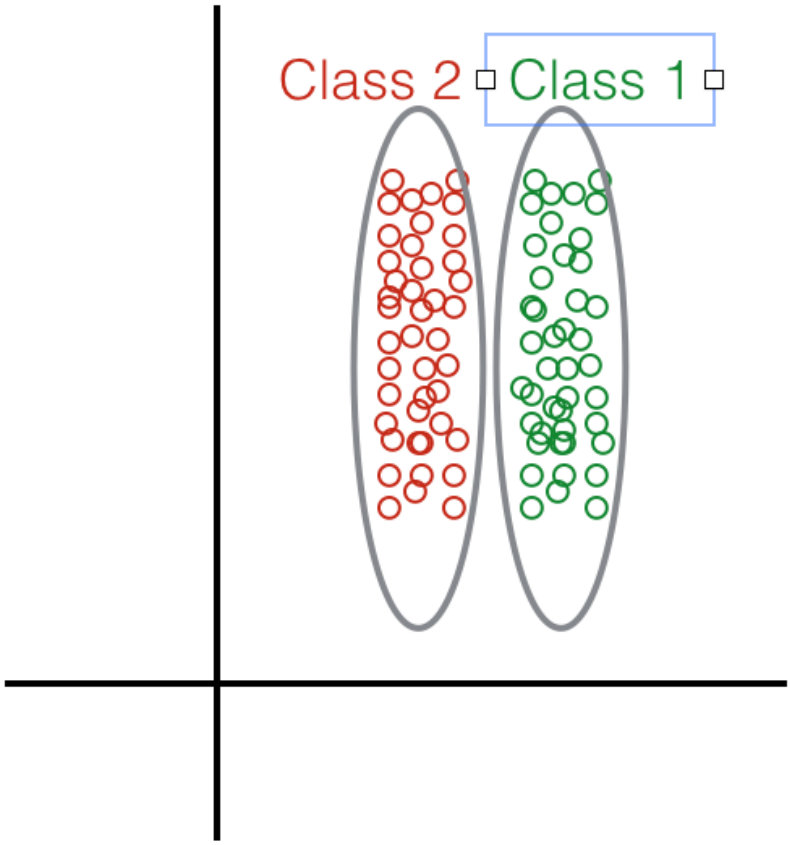
(b) $1, x, x^3, x^4$

(c) $1, x, x^2$

(d) $1, x, sin(x)$

Sol. The main concept here is that we need a curve with atleast 1 point of inflection in order to fit the points. Thus, a cubic curve and a biquadratic curve will be able to curve enough times to touch all 4 regions. A quadratic curve, on the other hand, will not work. Since the regions are approximately of equal size, a periodic function like $a.sin(X) + b$ can also be used to fit the given regions. Thus, options (a), (b) and (d) work.

9. Given the following distribution of data points:



What method would you choose to perform Dimensionality Reduction?

(a) Linear Discriminant Analysis

(b) Principal Component Analysis

Sol. (a)
PCA does not use class labels and will treat all the points as instances of the same pool. Thus the principal component will be the vertical axis, as the most variance is along that direction. However, projecting all the points into the vertical axis will mean that critical information is lost and both classes are mixed completely. LDA, on the other hand models each class with a gaussian. This will lead to a principal component along the horizontal axis which retains class information (the classes are still linearly separable)

10. In general, which of the following classification methods is the most resistant to gross outliers?

(a) Quadratic Discriminant Analysis (QDA)

(b) Linear Regression

(c) Logistic regression

(d) Linear Discriminant Analysis (LDA)

Sol. (c)
In general, a good way to tell if a method is sensitive to outliers is to look at the loss it incurs upon ignoring outliers.
Linear Regression uses a square loss and thus, outliers that are far away from the hyperplane contribute significantly to the loss.
LDA and QDA both use the L2-Norm and, for the same reason, sensitive to outliers.
Logistic Regression weights the points close to the boundary higher than points far away. This is an implication of the Logistic loss function (beyond the boundary, roughly linear instead of quadratic).

Assignment 4

Introduction to Machine Learning

Prof. B. Ravindran

1. Which of the following loss functions are convex?

   (a) 0-1 loss (sometimes referred as mis-classification loss)

   (b) Hinge loss

   (c) Logistic loss

   (d) Squared error loss

   Sol. (b), (c) and (d)

2. When using SVMs, what effect, in general, can you expect on the size of the margins when the C parameter is decreased?

   (a) the margins may become wider

   (b) the margins may become narrower

   (c) no relation between C and margin sizes

   Sol. (a)

   For Q3,4: Kindly download the synthetic dataset from this link. The dataset contains 1000 points and each input point contains 3 features.

3. Train a linear regression model (without regularization) on the above dataset. Report the coefficients of the best fit model. Report the coefficients in the following format: $\beta_0, \beta_1, \beta_2, \beta_3$

   (a) -1.2, 2.1, 2.2, 1

   (b) 1, 1.2, 2.1, 2.2

   (c) -1, 1.2, 2.1, 2.2

   (d) 1, -1.2, 2.1, 2.2

   (e) 1, 1.2, -2.1, -2.2

   Sol. (d)

4. Train a l2 regularized linear regression model on the above dataset. Vary the regularization parameter from 1 to 10. As you increase the regularization parameter, absolute value of the coefficients (excluding the intercept) of the model:

   (a) increase

   (b) first increase then decrease

   (c) decrease

   (d) first decrease then increase

   (e) does not change

Sol. (c)

For Q5,6: Kindly download the modified version of Iris dataset from this link. The dataset contains 150 points and each input point contains 4 features and belongs to one among three classes. Use the first 100 points as the training data and the remaining 50 as test data.

5. Train a $l2$ regularized logistic regression classifier on the modified iris dataset. We recommend using sklearn. Use only the first two features for your model. We encourage you to explore the impact of varying different hyperparameters of the model. Kindly note that the C parameter mentioned below is the inverse of the regularization parameter λ. As part of the assignment train a model with the following hyperparameters:
Model: logistic regression with one-vs-rest classifier, $C = 1e4$
For the above set of hyperparameters, report the best classification accuracy

   (a) 0.88

   (b) 0.86

   (c) 0.92

   (d) 0.68

   Sol. (b)

6. Train an SVM classifier on the modified iris dataset. We recommend using sklearn. Use only the first two features for your model. We encourage you to explore the impact of varying different hyperparameters of the model. Specifically try different kernels and the associated hyperparameters. As part of the assignment train models with the following set of hyperparameters
RBF-kernel, $gamma = 0.5$, one-vs-rest classifier, no-feature-normalization.
Try $C = 0.01, 1, 10$. For the above set of hyperparameters, report the best classification accuracy along with total number of support vectors on the test data.

   (a) 0.88, 69

   (b) 0.44, 69

   (c) 0.68, 44

   (d) 0.34, 44

   Sol. (a)

Assignment 5

Introduction to Machine Learning

Prof. B. Ravindran

1. Which of the following is/are true about the Perceptron classifier?

   (a) It can learn a OR function

   (b) It can learn a AND function

   (c) The obtained separating hyperplane depends on the order in which the points are presented in the training process.

   (d) For a linearly separable problem, there exists some initialization of the weights which might lead to non-convergent cases.

   Sol. (a), (b) and (c)
   OR is a linear function, hence can be learnt by perceptron.
   XOR is non linear function which cannot be learnt by a perceptron learning algorithm which can learn only linear functions.
   The perceptron learning algorithm dependent on the order on which the data is presented, there are multiple possible hyperplanes, and depending on the order we will converge to any one of them.
   We can also prove that the algorithm always converges to a separating hyperplane if it exists. Hence d is false.

2. You are given the following neural networks which take two binary valued inputs $x_1, x_2 \in \{0, 1\}$ and the activation function is the threshold function($h(x) = 1$ if $x > 0$; 0 otherwise). Which of the following logical functions does it compute?

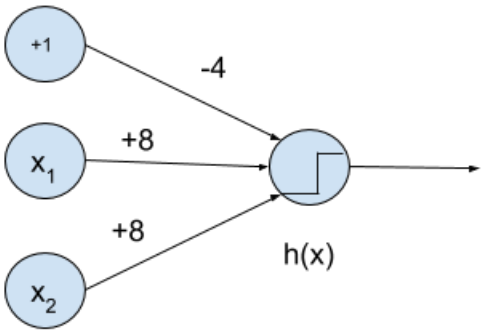


Figure 1: Q1

   (a) OR

   (b) AND

   (c) NAND

   (d) None of the above.

Sol. (a)

3. We have a function which takes a two-dimensional input $x = (x_1, x_2)$ and has two parameters $w = (w_1, w_2)$ given by $f(x, w) = \sigma(\sigma(x_1 w_1)w_2 + x_2)$ where $\sigma(x) = \frac{1}{1+e^{-x}}$. We use backpropagation to estimate the right parameter values. We start by setting both the parameters to 0. Assume that we are given a training point $x_1 = 0, x_2 = 1, y = 5$. Given this information answer the next two questions. What is the value of $\frac{\partial f}{\partial w_2}$?

   (a) 0.693

   (b) 0.098

   (c) 0.125

   (d) -0.531

   Sol. (b)
   Write $\sigma(x_1 w_1)w_2 + x_2$ as o_2 and $x_1 w_1$ as o_1

$$\frac{\partial f}{\partial w_2} = \frac{\partial f}{\partial o_2}\frac{\partial o_2}{\partial w_2}$$

$$\frac{\partial f}{\partial w_2} = \sigma(o_2)(1 - \sigma(o_2)) \times \sigma(o_1)$$

$$\frac{\partial f}{\partial w_2} = \sigma(1) \times (1 - \sigma(1)) \times \sigma(0) = 0.0983$$

4. If the learning rate is 0.5, what will be the value of w_2 after one update using backpropagation algorithm?

   (a) -0.5625

   (b) -0.4423

   (c) 0.5625

   (d) 0.4423

   Sol. (d) The update equation would be

$$w_2 = w_2 - \lambda \frac{\partial L}{\partial w_2}$$

   where L is the loss function, here $L = (y - f)^2$

$$w_2 = w_2 - \lambda \times 2(y - f) \times (-1) \times \frac{\partial f}{\partial w_2}$$

   Now putting in the given values we get the right answer.

5. We are given a 2-class classification problem with 0/1 output labels. We plan to use a neural network to implement the classifier. Which of the following functions is a suitable choice for the output neurons?

   (a) Hyperbolic Tangent Neuron - $tanh(\cdot)$

(b) Linear Neuron

(c) Arctangent Neuron - $arctan(\cdot)$

(d) Logistic Sigmoid Neuron - $\sigma(\cdot)$

Sol. (d)
Explanation The sigmoid neuron outputs a score between 0 and 1. This can be used as the probability of the input belonging to a given class.

6. Given N samples $x_1, x_2, \ldots, x_N$ drawn independently from a Gaussian distribution with variance σ_2 and unknown mean μ, find the MLE of the mean.

   (a) $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{\sigma^2}$

   (b) $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{2\sigma^2 N}$

   (c) $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$

   (d) $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N-1}$

   Sol. (c)

7. Which of the following statements is false:

   (a) The chances of overfitting decrease with Increasing the number of hidden nodes and increasing the number of hidden layers.

   (b) A neural network with one hidden layer can represent any Boolean function given sufficient number of hidden units and appropriate activation functions.

   (c) Two hidden layer neural networks can represent any continuous functions (within a tolerance) as long as the number of hidden units is sufficient and appropriate activation functions used.

   Sol. (a) By increasing the number of hidden nodes or hidden layers we are increasing the number of parameters. Increased set of parameters is more capable to memorize the training data. Hence it may result in overfitting.

8. Which of the following are true when comparing ANNs and SVMs?

   (a) ANN error surface has multiple local minima while SVM error surface has only one minima

   (b) After training, an ANN might land on a different minimum each time, when initialized with random weights during each run.

   (c) In training, ANN's error surface is navigated using a gradient descent technique while SVM's error surface is navigated using convex optimization solvers.

   (d) As shown for Perceptron, there are some classes of functions that cannot be learnt by an ANN. An SVM can learn a hyperplane for any kind of distribution.

   Sol. (a), (b) and (c)
   By universal approximate theorem, we can argue that option (d) is not true.

Assignment 6

Introduction to Machine Learning

Prof. B. Ravindran

1. What is specified at any non-leaf node in a decision tree?

   (a) Class of instance

   (b) Data value description

   (c) Test specification

   (d) Data process description

   Sol (c)

2. Which of the following statements are true with respect to the application of Cost-Complexity Pruning and Reduced Error Pruning with Cross-Validation?

   (a) In Reduced Error Pruning, the pruned tree error can never be less than the original tree on the training dataset.

   (b) In Cost Complexity Pruning, the pruned tree error can never be less than the original tree on the training dataset.

   (c) In Cost Complexity Pruning, the pruned tree error can never be less than the original tree on the validation dataset.

   (d) In Reduced Error Pruning, the pruned tree error can never be less than the original tree on the validation dataset.

   Sol. (a),(b)
   A pruned tree is always less accurate on the training dataset since a split is never made if it makes the tree worse at predicting the training set. Thus (a) and (b) are true.
   Nothing can be said about the validation set. Reduced Error Pruning always makes sure that the pruned tree error is less than the original tree on the validation dataset (which is the opposite of (d)).

3. Which of these classifiers do not require any additional modifications to their original descriptions (as seen in the lectures) to use them when we have more than 2 classes?

   (a) decision trees

   (b) logistic regression

   (c) support vector machines

   (d) k nearest neighbors

   Sol. (a), (d)
   Explanation Logistic regression and SVM's need to be tweaked to make them work for multiclass classification problems. Decision trees and kNN's on the other hand are agnostic to the number of classes.

4. Consider the following data set.

price	maintenance	capacity	airbag	profitable
low	low	2	no	yes
low	med	4	yes	yes
low	low	4	no	yes
low	med	4	no	no
low	high	4	no	no
med	med	4	no	no
med	med	4	yes	yes
med	high	2	yes	no
med	high	5	no	yes
high	med	4	yes	yes
high	med	2	yes	yes
high	high	2	yes	no
high	high	5	yes	yes

Considering 'profitable' as the binary valued attribute we are trying to predict, which of the attributes would you select as the root in a decision tree with multi-way splits using the cross-entropy impurity measure?

(a) price

(b) maintenance

(c) capacity

(d) airbag

Sol. (b)

5. In the above data set, what is the value of cross entropy when we consider capacity as the attribute to split on (multi-way splits)?

(a) 0.7973

(b) 0.8684

(c) 0.8382

(d) 0.7688

Sol. (c)

6. For the same data set, suppose we decide to construct a decision tree using binary splits and the gini index impurity measure. Which among the following feature and split point combinations would be the best to use as the root node assuming that we consider each of the input features to be unordered?

(a) price {med,low}|{high}

(b) capacity {2,4}|{5}

(c) maintenance {high}|{med,low}

(d) maintenance {high,med}|{low}

Sol. (c)

7. In the above question, what is the gini index value when we decide to split on the attribute price according to the following split: {med}|{low, high}?

   (a) 0.4505

   (b) 0.4573

   (c) 0.4196

   (d) 0.4615

   Sol. (d)

8. An important factor that influences the variance of decision trees is the average height of the tree. For the same dataset, if we limited the height of the trees to some H, how would the variance of the decision tree algorithm be affected?

   (a) Variance may increase with tree length H.

   (b) Variance may decrease with tree length H.

   (c) Variance is unaffected by tree length H.

   Sol. (a). Generally, a more complex classifier implies more variance and less bias.
   An intuitive way to imagine it is to think of a decision tree with $H = 1$ as a linear classifier. We know that a linear classifier has low variance and high bias. We also know that a general decision tree has high variance and low bias. Changing the average tree height H produces a spectrum of different classifiers ranging from low variance/high bias (a linear classifier) to high variance/low bias (a very tall tree).

9. In which of the following situations is it appropriate to introduce a new category 'Missing' for missing values?

   (a) When values are missing because the 108 emergency operator is sometimes attending a very urgent distress call.

   (b) When values are missing because the attendant spilled coffee on the papers from which the data was extracted.

   (c) When values are missing because the warehouse storing the paper records went up in flames and burnt parts of it.

   (d) When values are missing because the nurse/doctor finds the patient's situation too urgent.

   Sol. (a),(d). We typically introduce a 'Missing' value when the fact that a value is missing can also be a relevant feature. In the case of (a) is can imply that the call was so urgent that the operator couldn't note it down. This urgency could potentially be useful to determine the target.
   But a coffee spill corrupting the records is likely to be completely random and we glean no new information from it. In this case, a better method is to try to predict the missing data from the available data.

10. Which of the following properties are true in the context of decision trees?

    (a) High bias

(b) High variance

(c) Lack of smoothness of prediction surfaces

(d) Unbounded parameter set

Sol. (b), (c) & (d)
Decision trees are generally unstable considering that a small change in the data set can result in a very different set of splits. This is mainly due to the hierarchical nature of decision trees, since a change in split points in the initial stages will affect all the subsequent splits.

The decision surfaces that result from decision tree learning are generated by recursive splitting of the feature space using axis parallel hyper planes. They clearly do not produce smooth prediction surfaces such as the ones produced by, say, neural networks.

Decision trees do not make any assumptions about the distribution of the data. They are non-parametric methods where the number of parameters depends solely on the data set on which training is carried out

Assignment 7

Introduction to Machine Learning

Prof. B. Ravindran

1. Which of the following factors need to be taken into account while setting up an experiment?

   (a) Floor/Ceiling Effects

   (b) Order Effects

   (c) Sampling Bias

   Sol. (a), (b) and (c)

2. Select the correct equations.
   TP - True Positive, TN - True Negative, FP - False Positive, FN - False Negative

   (a) Precision = $\frac{TP}{TP+FP}$

   (b) Recall = $\frac{FP}{TP+FP}$

   (c) Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

   (d) Recall = $\frac{TP}{TP+FN}$

   Sol. (a), (c) and (d)

3. Which of the following measure best analyze the performance of a classifier?

   (a) Precision

   (b) Recall

   (c) Accuracy

   (d) Time complexity

   (e) Depends on the application

   Sol. (e)
   Explanation Different applications might need to optimize different performance measures. Applications of machine learning span over playing games to very critical domains(such as health and security). Measures like accuracy for instance cannot be reliable when we have a dataset with significant class imbalance. So there cannot be a single measure to analyze the effectiveness of a classifier in all environments.

4. For the ROC curve of True positive rate vs False positive rate, which of the following are true?

   (a) The curve is always concave (negative convex).

   (b) The curve is never concave.

   (c) The curve may or may not be concave

   Sol. (c)
   Explanation The nature of ROC curve is dependent on the classifier. Classifiers better than random classifier have a concave curve. Classifiers that perform worse than random classifier have a convex curve.

5. What are the quantities in the **Receiver Operating Characteristics** (ROC) curve along the x and y axes?

   (a) x - Precision, y - Recall

   (b) x - True Positive, y - True Negative

   (c) x - Specificity, y - Sensitivity

   (d) x - False Positive Rate, y - True Positive Rate

   Sol. (d)

6. In case of limited training data, which technique, bagging or stacking, would be preferred, and why?

   (a) Bagging, because we can combine as many classifier as we want by training each on a different sample of the training data

   (b) Bagging, because we use the same classification algorithms on all samples of the training data

   (c) Stacking, because each classifier is trained on all of the available data

   (d) Stacking, because we can use different classification algorithms on the training data

   Sol. (c)

7. How does bagging help in improving the classification performance?

   (a) If the parameters of the resultant classifiers are fully uncorrelated (independent), then bagging is inefficient.

   (b) It helps reduce bias

   (c) If the parameters of the resultant classifiers are fully correlated, then bagging is inefficient.

   (d) It helps reduce variance

   Sol. (c), (d)
   The lecture clearly states that correlated weights generally means that all the classifiers learn very similar functions. This means that bagging gives no extra stability.
   Having a lot of uncorrelated classifier helps to reduce variance since the resultant ensemble is more resistant to a single outlier (it likely only affects a small fraction of classifiers in the ensemble)

8. Which among the following prevents over-fitting when we perform bagging?

   (a) The use of sampling with replacement as the sampling technique

   (b) The use of weak classifiers

   (c) The use of classification algorithms which are not prone to overfitting

   (d) The practice of validation performed on every classifier trained

   Sol.(a)
   Bagging employs sampling with replacement to sample a subset of points from the original dataset. This means regions that are denser (and thus have more evidence) appear in every sample while outlier regions (very sparse) occur in only a fraction of classifiers, thus reducing their effect on the whole.

9. Which of the following statements are TRUE when comparing Committee Machines and Stacking

   (a) Committee Machines are, in general, special cases of 2-layer stacking where the second-layer classifier provides uniform weightage.

   (b) Both Committee Machines and Stacking have similar mechanisms, but Stacking uses different classifiers while Committee Machines use similar classifiers.

   (c) Committee Machines are more powerful than Stacking

   (d) Committee Machines are less powerful than Stacking

   Sol. (a), (d)
   Both Committee Machines and Stacked Classifiers use sets of different classifiers. Assigning constant weight to all first layer classifiers in a Stacked Classifier is simply the same as giving each one a single vote (Committee Machines).
   Since Committee Machines are a special case of Stacked Classifiers, they are less powerful than Stacking, which can assign an adaptive weight depending on the region.

10. Which of the following are true about using 5-fold cross validation with a data set of size n = 100 to select the value of k in the kNN algorithm. (More than one option may be correct)

    (a) Will always result in the same k since it does not involve any randomness.

    (b) Might give different answers depending on the splitting in 5 fold cross validation.

    (c) Does not make sense since n is larger than the number of folds.

    Sol. (b)

Assignment 8

Introduction to Machine Learning

Prof. B. Ravindran

1. Which of the following properties is false in the case of a Bayesian Network?

   (a) The edges are directed

   (b) Contains cycles

   (c) Represents conditional independence relations among random variables

   (d) All of the above

   Sol. (b)

2. A and B are Boolean random variables.
   Given: $P(A = True) = 0.3, P(A = False) = 0.7, P(B = True|A = True) = 0.4, P(B = False|A = True) = 0.6, P(B = True|A = False) = 0.6, P(B = False|A = False) = 0.4$.
   Calculate $P(A = True|B = False)$ by Bayes rule.

   (a) 0.49

   (b) 0.39

   (c) 0.37

   (d) 0.28

   Sol. (b)

   $$\frac{0.6 \times 0.3}{0.6 \times 0.3 + 0.7 \times 0.4} = 0.39$$

3. If you have a bad classifier, which of the following ensemble methods will give the worst performance when including the given classifier?

   (a) Gradient Boosting

   (b) AdaBoost

   (c) Bagging

   (d) Committee Machine

   Sol. (c)
   As mentioned in the lectures, Bagging tends to make a bad classifier even worse (partially due to the bootstrap mechanism providing fewer data points to an already bad classifier).

4. Which among Gradient Boosting and AdaBoost is less susceptible to outliers considering their respective loss functions?

   (a) AdaBoost

   (b) Gradient Boost

   (c) On average, both are equally susceptible.

Sol. (b)
Gradient Boosting (the one discussed in the lecture) uses a least squares loss function while AdaBoost uses an exponential loss function. AdaBoost penalizes outliers to an exponential amount whereas Gradient Boost penalizes them to a lesser extent and thus, cares less about them.

5. Which of the following method(s) is/are not inherently sequential? [Note: Multiple options may be correct]

   (a) Gradient Boosting
   (b) Committee machines
   (c) AdaBoost

   Sol. (b)
   Since Boosting is a strictly sequential process (the previous round residual errors are required for the next round to proceed), it is not as parallel as committee machines.

6. Boosting techniques typically give very high accuracy classifiers by sequentially training a collection of similar low-accuracy classifiers. Which of the following statements are true with respect to Boosting?

   (a) LogitBoost (like AdaBoost, but with Logistic Loss instead of Exponential Loss) is less susceptible to overfitting than AdaBoost.
   (b) Boosting techniques tend to have low bias and high variance
   (c) Boosting techniques tend to have low variance and high bias
   (d) For basic linear regression classifiers, there is no effect of using Gradient Boosting.

   Sol. (a),(b), (d) The logistic loss function is roughly linear for large values which means outliers matter less for the LogitBoost objective function when compared with AdaBoost function.
   Boosting essentially works by training recursively training a classifier on the residual of the previous set of classifiers. This means it effectively reduces the bias at each stage, but this also makes it sensitive to a slight shift in points (as some of the classifiers added towards the end focus on very few specific points)
   Gradient boosting is a linear combination of classifiers, which in the general case makes a more powerful classifier. However, for linear classifiers, a linear combination is still a primitive linear classifier.

7. While using Random Forests, if the input data is such that it contains a large number (¿80%) of irrelevant features (the target variable is independent of the these features), which of the following statements are TRUE?

   (a) Random Forests have reduced performance as the fraction of irrelevant features increases.
   (b) Random forests have increased performance as the fraction of irrelevant features increases.
   (c) The fraction of irrelevant features doesn't impact the performance of random forest.

   Sol. (a)
   Random Forests sample a subset of features for each decision tree. Assuming that the number of features in the sample is a fixed fraction of the total number of features, it is easy to see that the greater the *fraction* of irrelevant features, the more the fraction of useless classifiers (which dilutes the effect of the relevant classifiers).

8. Which of the following statements are true about ensemble classifiers?

   (a) The different learners in boosting based ensembles can be trained in parallel

   (b) The different learners in bagging based ensembles can be trained in parallel

   (c) Boosting based algorithms which iteratively re-weight training points, such as AdaBoost, are more sensitive to noise than bagging based methods.

   (d) Boosting methods generally use strong learners as individual classifiers

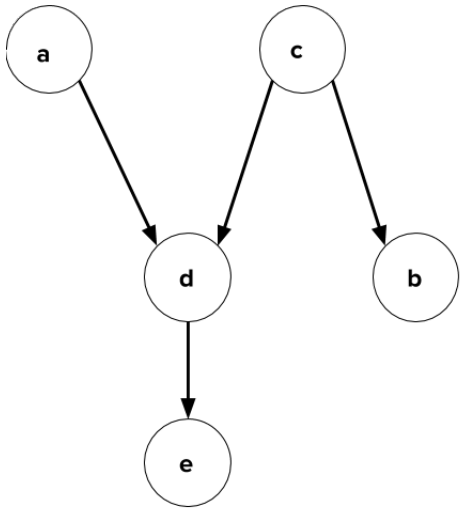
   (e) Boosting methods generally use weak learners as individual classifiers.

   (f) An individual classifier in a bagging based ensemble is trained with every point in the training set

   (g) An individual classifier in a boosting based ensemble is trained with every point in the training set.

   Sol. (b), (c), (e), (g)

9. Consider the following graphical model, which of the following are true about the model?



   (a) d is independent of b when c is known

   (b) a is independent of c when e is known

   (c) a is independent of b when e is known

   (d) a is independent of b when c is known

   Sol. (a), (d)

10. You are faced with a five class classification problem, with one class being the class of interest, i.e., you care more about correctly classifying data points belonging to that one class than the others. You are given a data set to use as training data. You analyze the data and observe the following properties. Which among these properties do you think are unfavourable from the point of view of applying general machine learning techniques?

(a) all data points come from the same distribution

(b) there is class imbalance with much more data points belonging to the class of interest than the other classes

(c) the given data set has some missing values

(d) each data point in the data set is independent of the other data points

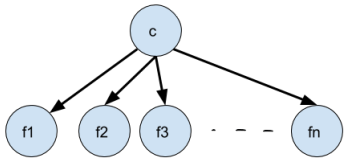
Sol. (c)
Option (a) is favourable, since it is an implicit assumption we make when we try applying supervised learning techniques. Option (b) indicates class imbalance which is usually an issue when it comes to classification. However, note that the imbalance is in favour of the class of interest. This means that there are a large number of examples for the class which we are interested in, which should allow us to model this class well. Option (c) is of course unfavourable as it would require us to handle missing data, and given the extent of the data that is missing, could severely affect the performance of any classifier we come up with. Finally, option (d) is favourable since, once again, it is an assumption about the data that we implicitly make.

Assignment 9

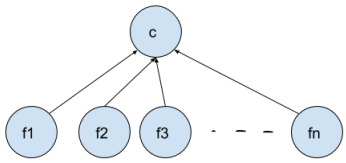
Introduction to Machine Learning

Prof. B. Ravindran

1. Which of the following graphical models capture the Naive Bayes assumption, where c represents the class label and f_i are the features?



(a)



(b)

(c) It cannot be captured by a graphical model.

(d) Graphical model can capture the assumption, but the given models don't do it.

Sol. (a)
The bayes assumption states that given the class label, the features are independent. This is captured when the class label is the parent node for all the feature nodes.

2. Select the correct pair of graphs and their implied independence results.
(Note 1: Arrows represent dependence as in normal Bayesian Networks.
Note 2: More than one options may be correct)

(a) $A \rightarrow C \leftarrow B$ implies A is independent of B given C

(b) $A \rightarrow C \leftarrow B$ implies A depends on B if C is known.

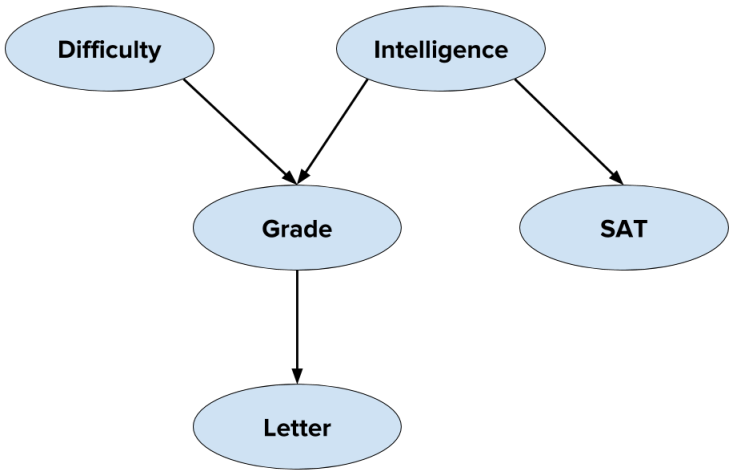
(c) $A \rightarrow C \rightarrow B$ implies B is independent of A if C is known.

(d) $A \leftarrow C \rightarrow B$ implies A is independent of B given C.

Sol. (b), (c) and (d)
Follow the lecture.

3. Here is a popular toy graphical model. It models the grades obtained by a student in a course and it's implications. Difficulty represents the difficulty of the course and intelligence is an indicator of how intelligent the student is, SAT represents the SAT scores of the student and

Letter presents the event of the student receiving a letter of recommendation from the faculty teaching the course.



Given this graphical model, which of the following statements are true?
(Note - More than one can be correct.)

(a) Given the grade, difficulty and letter are independent variables.

(b) Given grade, difficulty and intelligence are independent

(c) Without knowing any information, Difficulty and Intelligence are independent.

(d) Given the intelligence, SAT and grades are independent.

Sol. (a), (c) and (d)
To check independence between pairs of variables, first check all the paths between the pair of nodes. We have to ensure that all the paths should be blocked between the nodes. We call a path blocked in the following cases

- The nodes which occur on the path with head to tail or tail to tail should be known.
- The nodes which occur on the path with head to head shouldn't be known.

Using this strategy we will try to evaluate each of the option next. For option A, there is only one path between D and L, which passes through G. Since G is a node with head to tail, and G is known, hence the path is blocked which makes D, L independent. Similarly you can evaluate for all the options and reach the given solution.

4. The random variables given in the previous model are modeled as discrete variables and the corresponding CPDs are as below.

d^0	d^1
0.6	0.4

i^0	i^1
0.6	0.4

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.2	0.8
g^2	0.4	0.6
g^3	0.99	0.01

What is the probability of i^1, d^0, g^2, s^1, l^0 occurring?

(a) 0.004608
(b) 0.006144
(c) 0.003992
(d) 0.007309

Sol. (b)

$$P(i^1, d^0, g^2, s^1, l^0) = P(i^1)P(d^0)P(g^2|i^1, d^0)P(s^1|i^1)P(l^0|g^2)$$

5. Using the given example and CPD's compute the probability of following assignment, i^1, g^1, s^1, l^1 irrespective of the difficulty of the course?

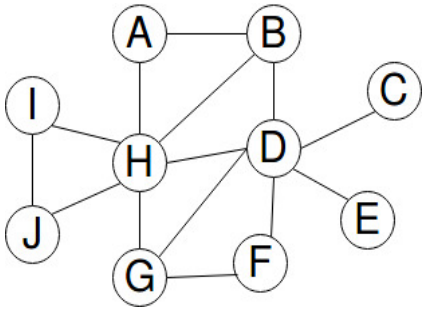
(a) 0.160
(b) 0.371
(c) 0.662
(d) 0.189

Sol. (d)

$$P(i^1, g^1, s^1, l^1) = P(i^1)P(s^1|i^1)P(l^1|g^1) \sum_{d}^{d=0,1} (P(d)P(g^1|i^1, d))$$

$$P(i^1, g^1, s^1, l^1) = 0.4 \times 0.8 \times 0.8 \times (0.9 \times 0.6 + 0.5 \times 0.4)$$

6. Consider the following Markov Random Field.



Which of the following nodes will have no effect on D given the Markov Blanket of D? (Note: more than one options may be correct)

(a) A

(b) B

(c) C

(d) E

(e) F

(f) G

(g) H

(h) I

(i) J

Sol. (a), (h) and (i)
In effect, the question requires you to select the random variables **not** in the Markov blanket of D. We see that the Markov blanket of D contains $B, C, E, F, G, \& H$. The only other variables, other than D are $A, I, \& J$. These three variables can have no effect on D once the Markov blanket is known/given.

7. Select the correct pairs of (Graphical Model, Inference Algorithm) (note: more than one option may be correct)

   (a) (Bayesian Networks, Variable Elimination)

   (b) (Viterbi Algorithm, Markov Random Fields)

   (c) (Viterbi Algorithm, Hidden Markov Models)

   (d) (Belief Propagation, Markov Random Fields)

   (e) (Variable Elimination, Markov Random Fields)

   Sol. (a), (c), (d) and (e)
   Viterbi Algorithm is for a sequence, while MRFs don't have a concept of sequence.

Assignment 10

Introduction to Machine Learning

Prof. B. Ravindran

1. In the CURE clustering algorithm, representative points of a cluster are moved a fraction of the distance between their original location and the centroid of the cluster. Would it make more sense to move them all a fixed distance towards the centroid instead? Why or why not?

   (a) Yes, because this approach will ensure that the original cluster shape is preserved.

   (b) No, because this approach will not be as effective against outliers as the original approach.

   Sol. (b)
   Moving representative points towards the cluster centroid helps in overcoming the effects of outliers. In the proposed approach, moving only a fixed distance towards the centroid would be less effective against outliers, since the distance between the outliers and the centroid may be much larger than the distance between the other points and the centroid.

2. If you are trying to find clusters in a dataset of roughly 5 billion data points (each data point taking roughly 128 bytes of data) and you have a machine with 2TB of RAM. Considering the large data size, which of the following algorithms would you use to efficiently find patterns in the data? (Note: More than one option may be correct)

   (a) BIRCH

   (b) CURE

   (c) K-means

   (d) Logistic Regression

   Sol.(a) or (b)
   Both K-means and Logistic Regression are not good at handling enormous amounts of data. (Especially with nearly 5 billion data points). You would also need to know the number of clusters in the final data which isn't provided here.
   BIRCH and CURE are tailor-made for operation in limited RAM and are efficient with large datasets.

3. In k-means clustering, globally minimizing the objective function for a known k is:

   (a) NP hard

   (b) impossible

   (c) possible in linear time

   (d) possible in polynomial time

   Sol. (a)
   Explanation For a known k, one has to analyze all possible divisions of items into k subsets. Which makes it NP hard.

4. What assumption does the CURE clustering algorithm make with regards to the shape of the clusters?

(a) No assumption

(b) Spherical

(c) Elliptical

Sol. (a)
Explanation CURE does not make any assumption on the shape of the clusters.

5. What would, in general, be the effect of increasing MinPts in DBSCAN while retaining the same Eps parameter? (Note that more than one statement may be correct)

   (a) Increase in the sizes of individual clusters

   (b) Decrease in the sizes of individual clusters

   (c) Increase in the number of clusters

   (d) Decrease in the number of clusters

   Sol. (b), (c)
   By increasing the MinPts, we are expecting large number of points in the neighborhood, to include them in cluster. In one sense, by increasing MinPts, we are looking for dense clusters. This can break not-so-dense clusters into more than one part, which can lead to reduce the cluster size and increase the number of clusters.

   In the following three questions we would like to understand the utility of different clustering algorithms on particular dataset. Kindly download DS1 and DS2. The first two columns in the dataset correspond to the co-ordinates of each data point. The third column corresponds two the actual cluster label.

6. Visualize the dataset DS1. Which of the following algorithms will be able to recover the true clusters.

   (a) K-means clustering

   (b) Single link hierarchical clustering

   (c) Complete link hierarchical clustering

   (d) None of the above

   Sol. (b)
   The dataset contains spiral clusters. Single link hierarchical clustering can recover spiral clusters with appropriate parameter settings.

7. Visualize the dataset DS2. Which of the following algorithms will be able to recover the true clusters.

   (a) K-means clustering

   (b) Single link hierarchical clustering

   (c) Complete link hierarchical clustering

   (d) None of the above

   Sol. (d)

8. For the dataset DS1, compute the Rand Index for the following methods: K-means, Single link hierarchical clustering and Complete link hierarchical clustering.
which of the method will return clustering with maximum Rand Index?

   (a) K-means clustering

   (b) Single link hierarchical clustering

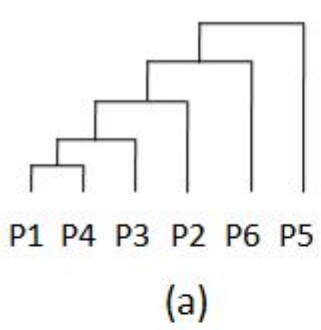
   (c) Complete link hierarchical clustering

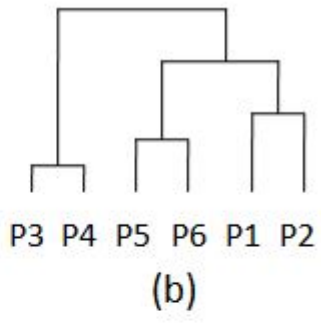
   (d) All the above will return the same clusters and hence have equal Rand Index

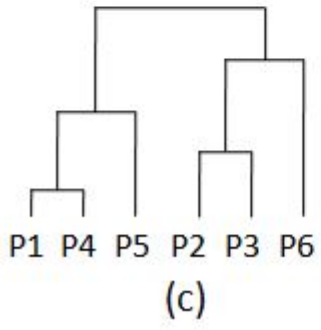
   Sol. (b) Single link hierarchical clustering recovers the exact clusters and hence will return the maximum Rand Index.

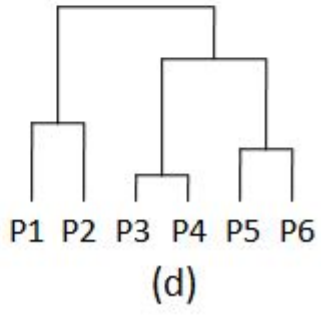
9. Consider the similarity matrix given below: Which of the following shows the hierarchy of clusters created by the single link clustering algorithm.

	P1	P2	P3	P4	P5	P6
P1	1.0000	0.7895	0.1579	0.0100	0.5292	0.3542
P2	0.7895	1.0000	0.3684	0.2105	0.7023	0.5480
P3	0.1579	0.3684	1.0000	0.8421	0.5292	0.6870
P4	0.0100	0.2105	0.8421	1.0000	0.3840	0.5573
P5	0.5292	0.7023	0.5292	0.3840	1.0000	0.8105
P6	0.3542	0.5480	0.6870	0.5573	0.8105	1.0000



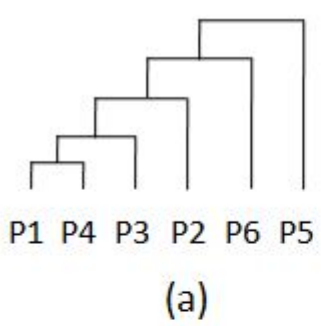




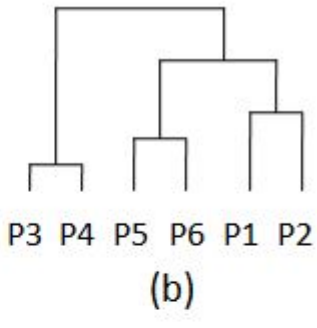


Sol. (b)

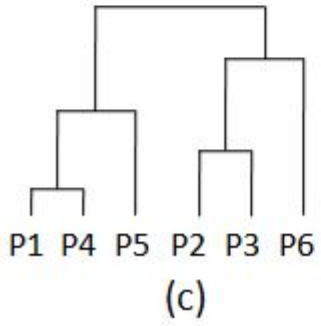
10. For the similarity matrix given in the previous question, which of the following shows the hierarchy of clusters created by the complete link clustering algorithm.



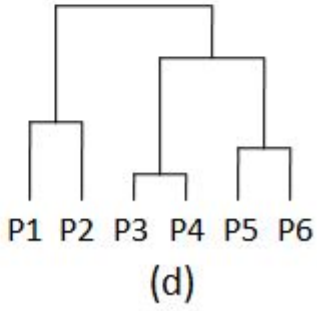
(a)



(b)



(c)



(d)

Sol. (d)

Assignment 11

Introduction to Machine Learning

Prof. B. Ravindran

1. Given N samples $x_1, x_2, \ldots, x_N$ drawn independently from a Gaussian distribution with variance σ and unknown mean μ, find the MLE of the mean. (1 mark)

   (a) $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{\sigma^2}$

   (b) $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{2\sigma^2 N}$

   (c) $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$

   (d) $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N-1}$

   Sol. (c)

2. Suppose we are trying to model a p dimensional Gaussian distribution. What is the actual number of independent parameters that need to be estimated? (2 marks)

   (a) 2

   (b) $2p$

   (c) $p(p+1)$

   (d) $p(p+3)/2$

   Sol. (d)
   Explanation Mean vector has p parameters. The covariance matrix is symmetric ($p \times p$) and hence has $p\frac{p+1}{2}$ independent parameters.

3. You are given n p-dimensional data points. The task is to learn a classifier to distinguish between k classes. You come to know that the dataset has missing values. Can you use EM algorithm to fill in the missing values ? (without making any further assumptions) (2 marks)

   (a) Yes

   (b) No

   Sol. (b)

4. During parameter estimation for a GMM model using data X, which of the following quantities are you minimizing (directly or indirectly)? (1 mark)

   (a) Log-likelihood

   (b) Negative Log-likelihood

   (c) Cross-entropy

   (d) Residual Sum of Squares (RSS)

   Sol. (b)

5. In Gaussian Mixture Models, π_i are the mixing coefficients. Select the correct conditions that the mixing coefficients need to satisfy for a valid GMM model. (2 marks)

   (a) $0 \leq \pi_i \leq 1 \forall i$

   (b) $-1 \leq \pi_i \leq 1 \forall i$

   (c) $\sum_i \pi_i = 1$

   (d) $\sum_i \pi_i$ need not be bounded

   Sol. (a), (c)

6. Expectation-Maximization, or the EM algorithm, consists of two steps - E step and the M-step. Using the following notation, select the correct set of equations used at each step of the algorithm.
   Notation.
   X Known/Given variables/data
   Z Hidden/Unknown variables
   θ Total set of parameters to be learned
   $\theta_{\mathbf{k}}$ Values of all the parameters after stage k
   $\mathbf{Q}(,)$ The Q-function as described in the lectures (2 marks)

   (a) $\mathbf{E} - \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta_{m-1}}[log(Pr(\mathbf{X},\mathbf{Z}|\theta))]$

   (b) $\mathbf{E} - \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta}[log(Pr(\mathbf{X},\mathbf{Z}|\theta_m))]$

   (c) $\mathbf{M} - argmax_\theta \sum_{\mathbf{Z}} Pr(\mathbf{Z}|\mathbf{X},\theta_{m-1}) \cdot log(Pr(\mathbf{X},\mathbf{Z}|\theta))$

   (d) $\mathbf{M} - argmax_\theta \mathbf{Q}(\theta,\theta_{m-1})$

   (e) $\mathbf{M} - argmax_\theta \mathbf{Q}(\theta,\theta_{m-2})$

   Sol. (a), (c), (d)

Assignment 12

Introduction to Machine Learning

Prof. B. Ravindran

1. In a tournament classifier with N classes. What is the complexity of the number of classifiers we require? (1 mark)

   (a) $\mathcal{O}(N)$

   (b) $\mathcal{O}(N^2)$

   (c) $\mathcal{O}(N.log(N))$

   (d) $\mathcal{O}(log(N))$

   Sol. (a) For a tournament classifier, we have $\frac{N}{2}$ classifiers in the first round that nominate $\frac{N}{2}$ candidates to the next round where there are $\frac{N}{4}$ classifiers. Thus, the total number of classifiers are $N + \frac{N}{2} + \frac{N}{4} + .. \approx N$.

2. In the context of Reinforcement Learning algorithms, which of the following definitions constitutes a valid Markov State? (2 marks)

   (a) For Chess: Positions of yours and the opponent's remaining pieces

   (b) For Tic-Tac-Toe: A snapshot of the game board (all Xs, Os and empty spaces)

   (c) For Chess: Positions of your pieces and the identities of the opponents defeated pieces.

   (d) For Tennis: Position and Velocity of the ball

   (e) For Tennis: Position of the ball

   Sol. (a) (b) (d)
   This is answered by considering if a given state has all the information needed to make the next move, or if previous states need to be accessed for more information.

3. You are designing a Reinforcement Learning agent for a racing game. Among the following reward schemes, which one leads to the best performance of the agent?

   (a) +5 for reaching the finish line, -1 for going off the road

   (b) +5 for reaching the finish line, -0.1 for every second that passes before the agent reaches the finish line

   (c) +5 for reaching the finish line, -0.1 for every second that passes before the agent reaches the finish line, +1 for the agent going off the road.

   (d) -5 for reaching the finish line, +0.1 for every second that passes before the agent reaches the finish line.

   Sol. (b) In order to obtain the most optimal solution, one must not attempt to give rewards to intermediate states that we believe might be part of the optimal solution. This discourages the agent from exploring other possibilities.
   Additionally, a *negative* reward for every second spent on the track incentivises the agent to find the fastest path to the goal (Otherwise, any path to the goal will fetch the same reward).

4. Recall the tic-tac-toe example from the reinforcement learning lecture. For each board position, we maintain the probability of winning from that board position. These board positions are updated using temporal difference learning. Assume that the probability values are all initialized to 0.5 and the opponent is an imperfect player. Suppose we always select the greedy action, i.e., the action which leads to the next state with highest probability. What problem(s) can you expect to encounter following this strategy? (2 marks)

   (a) It may happen that we never win even once.

   (b) No problems, this is an optimal strategy.

   (c) If a path exists in the game tree that always leads to victory, such a path can never be found using this strategy.

   (d) It is possible that we never fully explore the entire game tree.

   Sol. (d)

5. Suppose we want an RL agent to learn to play the game of golf. For training purposes, we make use of a golf simulator program. Assume that the original reward distribution gives a reward of +10 when the golf ball is hit into the hole and -1 for all other transitions. To aide the agents learning process, we propose to give an additional reward of +3 whenever the ball is within a 1 metre radius of the hole. Is this additional reward a good idea or not? Why? (1 mark)

   (a) Yes. The additional reward will help speed-up learning.

   (b) Yes. Getting the ball to within a metre of the hole is like a sub-goal and hence, should be rewarded.

   (c) No. The additional reward may actually hinder learning.

   (d) No. It violates the idea that a goal must be outside the agents direct control.

   Sol. (c)
   In this specific case, the additional reward will be detrimental to the learning process, as the agent will learn to accumulate rewards by keeping the ball within the 1 metre radius circle and not actually hitting the ball in the hole.

6. You want to toss a fair coin a number of times and obtain the probability of it falling on it's heads by taking a simple average. What is the estimated number of times you'll have to toss the coin to make sure that your estimated probability is within 10% of the actual probability, at least 90% of the time? (2 marks)

   (a) 400*ln(20)

   (b) 800*ln(20)

   (c) 200*ln(20)

   (d) 100*ln(40)

   Sol. (c) Since you're given that the coin is fair, $p = 0.5$ is the mean of your tosses if tossed infinitely. Hence, margin for error is 0.05. Now, using Chernoff-Hoeffding bounds, we can

obtain the required number of trails as follows.

$$\begin{aligned} Pr(|\bar{X} - 0.5| \geq 0.05) &\leq 2e^{-2\cdot(0.05)^2\cdot n} \\ \implies 0.1 &\leq 2e^{-2\cdot(0.05)^2\cdot n} \\ 0.05 &\leq e^{-2\cdot(0.05)^2\cdot n} \\ ln(\frac{1}{20}) &\leq -2\cdot(0.05)^2\cdot n \\ n &\geq \frac{1}{2\cdot(0.05)^2} ln(20) \\ n &\geq \frac{400}{2} ln(20) \\ \implies n &\geq 200 ln(20) \end{aligned}$$

7. You face a particularly challenging RL problem, where the reward distribution keeps changing with time. In order to gain maximum reward in this scenario, does it make sense to stop exploration or continue exploration? (1 mark)

   (a) Stop exploration

   (b) Continue exploration

   Sol. (b)
   Ideally, we would like to continue exploring, since this allows us to adapt to the changing reward distribution.