Introduction to Machine Learning

Week 1

Prof. B. Ravindran, IIT Madras

1. (2 Marks) Which of the following are supervised learning problems (Multiple Correct)?

   (a) Clustering Spotify users based on their listening history
   (b) Weather forecast using data collected by a satellite
   (c) Predicting tuberculosis using patient's chest X-Ray
   (d) Training a humanoid to walk using a reward system

   Sol. b and c

2. (2 Marks) Which of the following are regression tasks (Multiple Correct)?

   (a) Predicting the outcome of an election
   (b) Predicting the weight of a giraffe based on its weight
   (c) Predicting the emotion conveyed by a sentence
   (d) Identifying abnormal data points

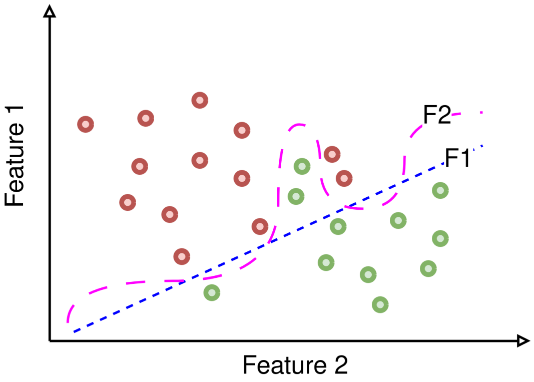
   Sol. b

3. (2 Marks) Which of the following are classification tasks (Multiple Correct)?

   (a) Predicting the outcome of an election
   (b) Predicting the weight of a giraffe based on its weight
   (c) Predicting the emotion conveyed by a sentence
   (d) Identifying abnormal data points

   Sol. a, and c

(Common Data for Questions 4 and 5)
Here is a 2-dimensional plot showing two functions that classify data points into two classes. The red points belong to one class, and the green points belong to another. The dotted blue line (F1) and dashed pink line (F2) represent the two trained functions.



4. (1 Mark) Which of the two functions overfit the training data?

   (a) Both functions F1 & F2

   (b) Function F1

   (c) Function F2

   (d) None of them

   Sol. c

5. (1 Mark) Which of the following 2 functions will yield higher training error?

   (a) Function F1

   (b) Function F2

   (c) Both functions F1 & F2 will have the same training error

   (d) Can not be determined

   Sol. a

6. (1 Mark) What does the term ‘policy’ refer to in reinforcement learning?

   (a) A set of rules governing the environment

   (b) The reward function

   (c) The initial state of the environment

   (d) The strategy the agent follows to choose actions

   Sol. d

7. (1 Mark) Given the following dataset, for $k = 3$, use KNN regression to find the prediction for a new data-point (2,3) *(Use Euclidean distance measure for finding closest points)*

X1	X2	Y
2	5	3.4
5	5	5
3	3	3
6	3	4.5
2	2	2
4	1	2.8

   (a) 2.0

   (b) 2.6

   (c) 2.8

   (d) 3.2

   Sol. c: The closest k points are (2,5), (3,3) and (2,2). Their corresponding labels averaged is $(3.4 + 3 + 2)$ / $3 = 2.8$

8. (1 Mark) For any given dataset, comment on the bias of K-nearest classifiers upon increasing the value of K.

(a) The bias of the classifier decreases

(b) The bias of the classifier does not change

(c) The bias of the classifier increases

(d) Can not be determined

Sol. c: Refer to lecture

9. (1 Mark) Bias and variance are given by:

(a) $\mathbb{E}[\hat{f}(x)] - f(x)$, $\mathbb{E}\big[(\mathbb{E}[\hat{f}(x)] - \hat{f}(x))^2\big]$

(b) $\mathbb{E}[\hat{f}(x)] - f(x)$, $\mathbb{E}\big[(\mathbb{E}[\hat{f}(x)] - \hat{f}(x))\big]^2$

(c) $(\mathbb{E}[\hat{f}(x)] - f(x))^2$, $\mathbb{E}\big[(\mathbb{E}[\hat{f}(x)] - \hat{f}(x))^2\big]$

(d) $(\mathbb{E}[\hat{f}(x)] - f(x))^2$, $\mathbb{E}\big[(\mathbb{E}[\hat{f}(x)] - \hat{f}(x))\big]^2$

Sol. a

10. (1 Mark) Which of the following statements are FALSE regarding bias and variance?

(a) Models which overfit have a high bias

(b) Models which overfit have a low bias

(c) Models which underfit have a high variance

(d) Models which underfit have a low variance

Sol. a and c

Introduction to Machine Learning

Week 2

Prof. B. Ravindran, IIT Madras

1. (1 Mark) **State True or False**: Typically, linear regression tend to underperform compared to k-nearest neighbor algorithms when dealing with high-dimensional input spaces.

   (a) True

   (b) False

   Sol. b

2. (2 Marks) Given the following dataset, find the uni-variate regression function that best fits the dataset.

X	Y
2	5.5
3	6.5
4	9
10	18.5

   (a) $f(x) = 1 \times x + 4$

   (b) $f(x) = 1 \times x + 5$

   (c) $f(x) = 1.5 \times x + 3$

   (d) $f(x) = 2 \times x + 1$

   Soln. C

3. (1 Mark) Given a training data set of 500 instances, with each input instance having 6 dimensions and each output being a scalar value, the dimensions of the design matrix used in applying linear regression to this data is

   (a) 500×6

   (b) 500×7

   (c) 500×6^2

   (d) None of the above

   Soln. B

4. (1 Mark) **Assertion A:** Binary encoding is usually preferred over One-hot encoding to represent categorical data (eg. colors, gender etc)
   Reason R: Binary encoding is more memory efficient when compared to One-hot encoding

   (a) Both A and R are true and R is the correct explanation of A

   (b) Both A and R are true but R is not the correct explanation of A

   (c) A is true but R is false

(d) A is false but R is true

Soln. D

5. (1 Mark) Select the **TRUE** statement.

   (a) Subset selection methods are more likely to improve test error by only focussing on the most important features and by reducing variance in the fit.

   (b) Subset selection methods are more likely to improve train error by only focussing on the most important features and by reducing variance in the fit.

   (c) Subset selection methods are more likely to improve both test and train error by focussing on the most important features and by reducing variance in the fit.

   (d) Subset selection methods don't help in performance gain in any way.

   Sol. a

6. (1 Mark) Rank the 3 subset selection methods in terms of computational efficiency:

   (a) Forward stepwise selection, best subset selection, and forward stagewise regression.

   (b) forward stepwise selection, forward stagewise regression and best subset selection.

   (c) Best subset selection, forward stagewise regression and forward stepwise selection.

   (d) Best subset selection, forward stepwise selection and forward stagewise regression.

   Sol. b

7. (1 Mark) Choose the **TRUE** statements from the following: (Multiple correct choice)

   (a) Ridge regression since it reduces the coefficients of all variables, makes the final fit a lot more interpretable.

   (b) Lasso regression since it doesn't deal with a squared power is easier to optimize than ridge regression.

   (c) Ridge regression has a more stable optimization than lasso regression.

   (d) Lasso regression is better suited for interpretability than ridge regression.

   Sol. c, d

8. (2 Marks) Which of the following statements are TRUE? Let x_i be the i-th datapoint in a dataset of N points. Let v represent the first principal component of the dataset. (Multiple answer questions)

   (a) $v = \arg\max \sum_{i=1}^{N}(v^T x_i)^2$ s.t. $|v| = 1$

   (b) $v = \arg\min \sum_{i=1}^{N}(v^T x_i)^2$ s.t. $|v| = 1$

   (c) Scaling at the start of performing PCA is done just for better numerical stability and computational benefits but plays no role in determining the final principal components of a dataset.

   (d) The resultant vectors obtained when performing PCA on a dataset can vary based on the scale of the dataset.

   Soln. A and D

Introduction to Machine Learning

Week 3

Prof. B. Ravindran, IIT Madras

1. (1 Mark) For a two-class problem using discriminant functions (δ_k - discriminant function for class k), where is the separating hyperplane located?

   (a) Where $\delta_1 > \delta_2$

   (b) Where $\delta_1 < \delta_2$

   (c) Where $\delta_1 = \delta_2$

   (d) Where $\delta_1 + \delta_2 = 1$

   Soln. C

2. (1 Mark) Given the following dataset consisting of two classes, A and B, calculate the prior probability of each class.

Feature 1	**Class**
2.3	A
1.8	A
3.2	A
2.7	B
3.0	A
2.1	A
1.9	B
2.4	B

   What are the prior probabilities of class A and class B?

   (a) $P(A) = 0.5, \quad P(B) = 0.5$

   (b) $P(A) = 0.625, \quad P(B) = 0.375$

   (c) $P(A) = 0.375, \quad P(B) = 0.625$

   (d) $P(A) = 0.6, \quad P(B) = 0.4$

   Soln. B

3. (1 Mark) In a 3-class classification problem using linear regression, the output vectors for three data points are [0.8, 0.3, -0.1], [0.2, 0.6, 0.2], and [-0.1, 0.4, 0.7]. To which classes would these points be assigned?

   (a) 1, 2, 1

   (b) 1, 2, 2

   (c) 1, 3, 2

   (d) 1, 2, 3

Soln. D

4. (1 Mark) If you have a 5-class classification problem and want to avoid masking using polynomial regression, what is the minimum degree of the polynomial you should use?

   (a) 3

   (b) 4

   (c) 5

   (d) 6

   Soln. B

5. (1 Mark) Consider a logistic regression model where the predicted probability for a given data point is 0.4. If the actual label for this data point is 1, what is the contribution of this data point to the log-likelihood?

   (a) -1.3219

   (b) -0.9163

   (c) +1.3219

   (d) +0.9163

   Soln. B

6. (1 Mark) What additional assumption does LDA make about the covariance matrix in comparison to the basic assumption of Gaussian class conditional density?

   (a) The covariance matrix is diagonal

   (b) The covariance matrix is identity

   (c) The covariance matrix is the same for all classes

   (d) The covariance matrix is different for each class

   Soln. C

7. (1 Mark) What is the shape of the decision boundary in LDA?

   (a) Quadratic

   (b) Linear

   (c) Circular

   (d) Can not be determined

   Soln. B

8. (1 Mark) For two classes C_1 and C_2 with within-class variances $\sigma_{w1}^2 = 1$ and $\sigma_{w2}^2 = 4$ respectively, if the projected means are $\mu_1' = 1$ and $\mu_2' = 3$, what is the Fisher criterion $J(w)$?

   (a) 0.5

   (b) 0.8

   (c) 1.25

(d) 1.5

Soln. B
$S_w = \sigma_{w1}^2 + \sigma_{w2}^2 = 1 + 4 = 5$ $S_b = (\mu_2' - \mu_1')^2 = (3-1)^2 = 4$ $J(w) = \frac{S_b}{S_w} = \frac{4}{5} = 0.8$

9. (2 Marks) Given two classes C_1 and C_2 with means $\mu_1 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ and $\mu_2 = \begin{bmatrix} 5 \\ 7 \end{bmatrix}$ respectively, what is the direction vector w for LDA when the within-class covariance matrix S_w is the identity matrix I?

   (a) $\begin{bmatrix} 4 \\ 3 \end{bmatrix}$

   (b) $\begin{bmatrix} 5 \\ 7 \end{bmatrix}$

   (c) $\begin{bmatrix} 0.7 \\ 0.7 \end{bmatrix}$

   (d) $\begin{bmatrix} 0.6 \\ 0.8 \end{bmatrix}$

   Soln. D
   $S_w \propto \mu_1 - \mu_2$

Introduction to Machine Learning

Week 4

Prof. B. Ravindran, IIT Madras

---

1. (1 Mark) In the context of the perceptron learning algorithm, what does the expression $\frac{f(x)}{||f'(x)||}$ represent?

   (a) The gradient of the hyperplane

   (b) The signed distance to the hyperplane

   (c) The normal vector to the hyperplane

   (d) The misclassification error

   Soln. B

2. (1 Mark) Why do we normalize by $\|\boldsymbol{\beta}\|$ (the magnitude of the weight vector) in the SVM objective function?

   (a) To ensure the margin is independent of the scale of $\boldsymbol{\beta}$

   (b) To minimize the computational complexity of the algorithm

   (c) To prevent overfitting

   (d) To ensure the bias term is always positive

   Soln. A

3. (1 Mark) Which of the following is NOT one of the KKT conditions for optimization problems with inequality constraints?

   (a) Stationarity: $\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^{p} \nu_j \nabla h_j(\mathbf{x}^*) = 0$

   (b) Primal feasibility: $g_i(\mathbf{x}^*) \leq 0$ for all i, and $h_j(\mathbf{x}^*) = 0$ for all j

   (c) Dual feasibility: $\lambda_i \geq 0$ for all i

   (d) Convexity: The objective function $f(\mathbf{x})$ must be convex

   Soln. D

4. (1 Mark) Consider the 1 dimensional dataset:

x	y
-1	1
0	-1
2	1

(Note: x is the feature and y is the output)

State true or false: The dataset becomes linearly separable after using basis expansion with the following basis function $\phi(x) = \begin{bmatrix} 1 \\ x^3 \end{bmatrix}$

(a) True

(b) False

Soln. B

After applying basis expansion, $x_1^{'} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, $x_2^{'} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, and $x_3^{'} = \begin{bmatrix} 1 \\ 8 \end{bmatrix}$. Despite the basis expansion, it still remains linearly inseparable.

5. (1 Mark) Consider a polynomial kernel of degree d operating on p-dimensional input vectors. What is the dimension of the feature space induced by this kernel?

   (a) $p \times d$

   (b) $(p + 1) \times d$

   (c) $\binom{p+d}{d}$

   (d) p^d

   Soln. C

6. (1 Mark) State True or False: For any given linearly separable data, for any initialization, both SVM and Perceptron will converge to the same solution.

   (a) True

   (b) False

   Soln. B

   For Q7,8: Kindly download the modified version of Iris dataset from this link.
   Available at: (https://goo.gl/vchhsd)
   The dataset contains 150 points, and each input point has 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. In the following questions, to report accuracy, use the test dataset. You can round off the accuracy value to the nearest 2-decimal point number. (Note: Do not change the order of data points.)

7. (2 marks) Train a Linear perceptron classifier on the modified iris dataset. We recommend using sklearn. Use only the first two features for your model and report the best classification accuracy for l1 and l2 penalty terms.

   (a) 0.91, 0.64

   (b) 0.88, 0.71

   (c) 0.71, 0.65

   (d) 0.78, 0.64

Sol. (d)

Following code will give the desired result.

```
>>clf = Perceptron(penalty=”l1”).fit(X[0:100,0:2],Y[0:100])
>>clf.score(X[100:,0:2], Y[100:])
>>clf = Perceptron(penalty=”l2”).fit(X[0:100,0:2],Y[0:100])
>>clf.score(X[100:,0:2], Y[100:])
```

8. (2 marks) Train a SVM classifier on the modified iris dataset. We recommend using sklearn. Use only the first three features. 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.

   (a) 0.98

   (b) 0.88

   (c) 0.99

   (d) 0.92

   Sol. (a)

   Following code will give the desired result.

```
>>clf = svm.SVC( C=1.0, kernel=’rbf’, decision_function_shape=’ovr’, gamma = 0.5)).fit(X[0:100,0:3],
Y[0:100])
>>clf.score(X[100:,0:3], Y[100:])
```

Introduction to Machine Learning

Week 5

Prof. B. Ravindran, IIT Madras

1. (1 Mark) Given a 3 layer neural network which takes in 10 inputs, has 5 hidden units and outputs 10 outputs, how many parameters are present in this network?

   (a) 115
   (b) 500
   (c) 25
   (d) 100

   Soln. A

2. (1 Mark) Recall the XOR(tabulated below) example from class where we did a transformation of features to make it linearly separable. Which of the following transformations can also work?

X_1	X_2	Y
-1	-1	-1
1	-1	1
-1	1	1
1	1	-1

   (a) Rotating x_1 and x_2 by a fixed angle.
   (b) Adding a third dimension $z = x * y$
   (c) Adding a third dimension $z = x^2 + y^2$
   (d) None of the above

   Sol. (b)

3. We use several techniques to ensure the weights of the neural network are small (such as random initialization around 0 or regularisation). What conclusions can we draw if weights of our ANN are high?

   (a) Model has overfitted.
   (b) It was initialized incorrectly.
   (c) At least one of (a) or (b).
   (d) None of the above.

   Sol. (d)
   Overfitting may be because of high weights but the two are not always associated.

4. (1 Mark) In a basic neural network, which of the following is generally considered a good initialization strategy for the weights?

   (a) Initialize all weights to zero
   (b) Initialize all weights to a constant non-zero value (e.g., 0.5)
   (c) Initialize weights randomly with small values close to zero
   (d) Initialize weights with large random values (e.g., between -10 and 10)

   Soln. C

5. (1 Mark) Which of the following is the primary reason for rescaling input features before passing them to a neural network?

   (a) To increase the complexity of the model
   (b) To ensure all input features contribute equally to the initial learning process
   (c) To reduce the number of parameters in the network
   (d) To eliminate the need for activation functions

   Soln. B

6. (1 Mark) In the Bayesian approach to machine learning, we often use the formula: $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$ Where θ represents the model parameters and D represents the observed data. Which of the following correctly identifies each term in this formula?

   (a) $P(\theta|D)$ is the likelihood, $P(D|\theta)$ is the posterior, $P(\theta)$ is the prior, $P(D)$ is the evidence
   (b) $P(\theta|D)$ is the posterior, $P(D|\theta)$ is the likelihood, $P(\theta)$ is the prior, $P(D)$ is the evidence
   (c) $P(\theta|D)$ is the evidence, $P(D|\theta)$ is the likelihood, $P(\theta)$ is the posterior, $P(D)$ is the prior
   (d) $P(\theta|D)$ is the prior, $P(D|\theta)$ is the evidence, $P(\theta)$ is the likelihood, $P(D)$ is the posterior

   Soln. B

7. (1 Mark) Why do we often use log-likelihood maximization instead of directly maximizing the likelihood in statistical learning?

   (a) Log-likelihood provides a different optimal solution than likelihood maximization
   (b) Log-likelihood is always faster to compute than likelihood
   (c) Log-likelihood turns products into sums, making computations easier and more numerically stable
   (d) Log-likelihood allows us to avoid using probability altogether

   Soln. C

8. (1 Mark) In machine learning, if you have an infinite amount of data, but your prior distribution is incorrect, will you still converge to the right solution?

   (a) Yes, with infinite data, the influence of the prior becomes negligible, and you will converge to the true underlying solution.
   (b) No, the incorrect prior will always affect the convergence, and you may not reach the true solution even with infinite data.

(c) It depends on the type of model used; some models may still converge to the right solution, while others might not.

(d) The convergence to the right solution is not influenced by the prior, as infinite data will always lead to the correct solution regardless of the prior.

Soln. A

9. Statement: Threshold function cannot be used as activation function for hidden layers.
Reason: Threshold functions do not introduce non-linearity.

   (a) Statement is true and reason is false.

   (b) Statement is false and reason is true.

   (c) Both are true and the reason explains the statement.

   (d) Both are true and the reason does not explain the statement.

   Sol. (a)
   The reason is that threshold function is non-differentiable so we will not be able to calculate gradient for backpropagation.

10. Choose the correct statement (multiple may be correct):

   (a) MLE is a special case of MAP when prior is a uniform distribution.

   (b) MLE acts as regularisation for MAP.

   (c) MLE is a special case of MAP when prior is a beta disrubution .

   (d) MAP acts as regularisation for MLE.

   Sol. (a), (d)
   Ref. lecture

Introduction to Machine Learning

Week 6

Prof. B. Ravindran, IIT Madras

1. (1 Mark) Entropy for a 90-10 split between two classes is:

   (a) 0.469

   (b) 0.195

   (c) 0.204

   (d) None of the above

   Sol. (a)

2. (2 Mark) Consider a dataset with only one attribute(categorical). Suppose, there are 8 unordered values in this attribute, how many possible combinations are needed to find the best split-point for building the decision tree classifier?

   (a) 511

   (b) 1023

   (c) 512

   (d) 127

   Sol. (d)
   Suppose we have q unordered values; the total possible splits would be $2^{q-1} - 1$. Thus, in our case, it will be $2^7 - 1 = 127$.

3. (2 mark) Having built a decision tree, we are using reduced error pruning to reduce the size of the tree. We select a node to collapse. For this particular node, on the left branch, there are three training data points with the following outputs: 5, 7, 9.6, and for the right branch, there are four training data points with the following outputs: 8.7, 9.8, 10.5, 11. The average value of the outputs of data points denotes the response of a branch. The original responses for data points along the two branches (left & right respectively) were *response_left* and, *response_right* and the new response after collapsing the node is *response_new.* What are the values for *response_left*, *response_right* and *response_new* (numbers in the option are given in the same order)?

   (a) 9.6, 11, 10.4

   (b) 7.2; 10; 8.8

   (c) 5, 10.5, 15

   (d) Depends on the tree height.

   Sol. (b)

4. (1 Mark) Which of the following is a good strategy for reducing the variance in a decision tree?

(a) If improvement of taking any split is very small, don't make a split. (Early Stopping)

(b) Stop splitting a leaf when the number of points is less than a set threshold K.

(c) Stop splitting all leaves in the decision tree when any one leaf has less than a set threshold K points.

(d) None of the Above.

Sol. (b)

5. (1 Mark) Which of the following statements about multiway splits in decision trees with categorical features is correct?

(a) They always result in deeper trees compared to binary splits

(b) They always provide better interpretability than binary splits

(c) They can lead to overfitting when dealing with high-cardinality categorical features

(d) They are computationally less expensive than binary splits for all categorical features

Sol. (c)

6. (1 Mark) Which of the following statements about imputation in data preprocessing is most accurate?

(a) Mean imputation is always the best method for handling missing numerical data

(b) Imputation should always be performed after splitting the data into training and test sets

(c) Missing data is best handled by simply removing all rows with any missing values

(d) Multiple imputation typically produces less biased estimates than single imputation methods

Sol. (d)

7. (2 Marks) Consider the following dataset:

feature1	*feature2*	*output*
18.3	187.6	a
14.7	184.9	a
19.4	193.3	a
17.9	180.5	a
19.1	189.1	a
17.6	191.9	b
19.9	190.2	b
17.3	198.6	b
18.7	182.6	b
15.2	187.3	b

Which among the following split-points for *feature2* would give the best split according to the misclassification error?

(a) 186.5

(b) 188.6

(c) 189.2

(d) 198.1

Sol. (c)

Introduction to Machine Learning

Week 7

Prof. B. Ravindran, IIT Madras

---

1. (1 Mark) Define active learning:

   (a) A learning approach where the algorithm passively receives all training data at once

   (b) A technique where the model learns from its own predictions without human intervention

   (c) An iterative learning process where the model selects the most informative data points for labeling

   (d) A method where the model randomly selects data points for training to reduce bias

   Sol. (c) - Refer to the lectures

2. (2 Mark) Given 100 distinct data points, if you sample 100 times with replacement, what is the expected number of distinct points you will obtain?

   (a) Approximately 50

   (b) Approximately 63

   (c) Exactly 100

   (d) Approximately 37

   Sol. (b) -
   Probability of not selecting a point in 100 tries = $\frac{99}{100}^{100}$
   Probability of selecting a point atleast once in 100 tries = 1 - $\frac{99}{100}^{100}$
   Expectation = $\sum x \times P(x) = 100(1 - (\frac{99}{100})^{100}) \approx 63$

3. (1 Mark) What is the key difference between bootstrapping and cross-validation?

   (a) Bootstrapping uses the entire dataset for training, while cross-validation splits the data into subsets

   (b) Cross-validation allows replacement, while bootstrapping does not

   (c) Bootstrapping creates multiple samples with replacement, while cross-validation creates subsets without replacement

   (d) Cross-validation is used for model selection, while bootstrapping is only used for uncertainty estimation

   Sol. (c) - Refer to the lectures

4. (2 Marks) Consider the following confusion matrix for a binary classification problem:

	Predicted Positive	Predicted Negative
Actual Positive	85	15
Actual Negative	20	80

   What are the precision, recall, and accuracy of this classifier?

(a) Precision: 0.81, Recall: 0.85, Accuracy: 0.83

(b) Precision: 0.85, Recall: 0.81, Accuracy: 0.85

(c) Precision: 0.80, Recall: 0.85, Accuracy: 0.82

(d) Precision: 0.85, Recall: 0.85, Accuracy: 0.80

Sol. (a) -

Precision $= \frac{TP}{TP+FP} = \frac{85}{85+20} = \frac{85}{105} \approx 0.81$
Recall $= \frac{TP}{TP+FN} = \frac{85}{85+15} = \frac{85}{100} = 0.85$
Accuracy $= \frac{TP+TN}{TP+TN+FP+FN} = \frac{85+80}{200} = \frac{165}{200} = 0.825 \approx 0.83$

5. (1 Mark) AUC for your newly trained model is 0.5. Is your model prediction completely random?

(a) Yes

(b) No

(c) ROC curve is needed to derive this conclusion

(d) Cannot be determined even with ROC

Sol. (c) - An AUC of 0.5 suggests that the model is either making random predictions, predicting all instances as a single class, or making systematically incorrect predictions that could be corrected by inverting its outputs.

6. (1 Mark) You are building a model to detect cancer. Which metric will you prefer for evaluating your model?

(a) Accuracy

(b) Sensitivity

(c) Specificity

(d) MSE

Sol. (b) - In medical application, FP is the most important (which sensitivity captures)

7. (1 Mark) You have 2 binary classifiers A and B. A has accuracy=0% and B has accuracy=50%. Which classifier is more useful?

(a) A

(b) B

(c) Both are good

(d) Cannot say

Sol. (a) - Flip the labels and get 100% accuracy!

8. (1 Mark) You have a special case where your data has 10 classes and is sorted according to target labels. You attempt 5-fold cross validation by selecting the folds sequentially. What can you say about your resulting model?

(a) It will have 100% accuracy.

(b) It will have 0% accuracy.

(c) It will have close to perfect accuracy.

(d) Accuracy will depend on the compute power available for training.

Sol. (b) - The training and test sets are partitioned in a way that some classes are only present in the test set. This means the classifier will never learn about these classes and therefore cannot predict them

Introduction to Machine Learning

Week 8

Prof. B. Ravindran, IIT Madras

1. (1 Mark) In Bagging technique, the reduction of variance is maximum if:

   (a) The correlation between the classifiers is minimum

   (b) Does not depend on the correlation between the classifiers

   (c) Similar features are used in all classifiers

   (d) The number of classifiers in the ensemble is minimized

   Soln. A - This ensures diverse predictions that effectively average out errors

2. (1 Mark) If using squared error loss in gradient boosting for a regression problem, what does the gradient correspond to?

   (a) The absolute error

   (b) The log-likelihood

   (c) The residual error

   (d) The exponential loss

   Soln. C - $\Delta(y - f(x; w))^2 = 2(y - f(x; w))\Delta f(x; w)$

3. (1 Mark) In a random forest, if T (number of features considered at each split) is set equal to P (total number of features), how does this compare to standard bagging with decision trees?

   (a) It's exactly the same as standard bagging

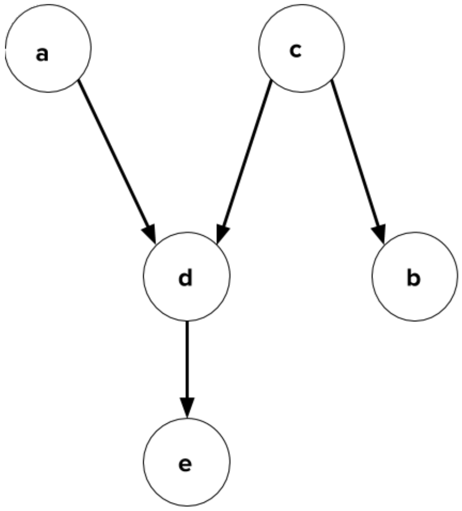
   (b) It will always perform better than standard bagging

   (c) It will always perform worse than standard bagging

   (d) Can not be determined

   Soln. A - Random Forests differ from standard bagging by sampling a sub-set of features at every node

4. (1 Mark) **Multiple Correct:** Consider the following graphical model, which of the following are true about the model? (multiple options may be correct)



(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

Soln. A, D - Refer to the lectures

5. (1 Mark) Consider the Bayesian network given in the previous question. Let “a”, “b”, “c”, “d” and “e” denote the random variables shown in the network. Which of the following can be inferred from the network structure?

(a) “a” causes “d”
(b) “e” causes “d”
(c) Both (a) and (b) are correct
(d) None of the above

Soln. D - Node “d” is dependendant on both “a” and “c” and “e” can not cause “d”

6. (2 Marks) A single box is randomly selected from a set of three. Two pens are then drawn from this container. These pens happen to be blue and green colored. What is the probability that the chosen box was Box A?

Box	Green	Blue	Yellow
A	3	2	1
B	2	1	2
C	4	2	3

(a) 37/18
(b) 15/56

(c) 18/37

(d) 56/15

Soln. C -

Probability of choosing one box $P(A) = P(B) = P(C) = 1/3$. Here the event (E) is choosing the green and blue balls from the random box. Therefore, P(E | A) = $\frac{{}^3C_1 * {}^2C_1}{{}^6C_2}$ = 6/15 = 2/5

$$\text{P(E | B)} = \frac{{}^2C_1 *^1 C_1}{{}^5C_2} = 2/10 = 1/5$$
$$\text{P(E | C)} = \frac{{}^4C_1 *^2 C_1}{{}^9C_2} = \frac{8}{72/2} = 2/9$$
$$P(A \mid E) = P(E \mid A)/[P(E \mid A) + P(E \mid B) + P(E \mid C)]$$
$$= \frac{2/5}{(2/5) + (1/5) + (2/9)}$$
$$= 18/37$$

7. (1 Mark) **State True or False:** The primary advantage of the tournament approach in multiclass classification is its effectiveness even when using weak classifiers.

(a) True

(b) False

Soln. B - Refer to the lectures

8. (1 Mark) A data scientist is using a Naive Bayes classifier to categorize emails as either “spam” or “not spam”. The features used for classification include:

- Number of recipients (To, Cc, Bcc)
- Presence of “spam” keywords (e.g., ”URGENT”, ”offer”, ”free”)
- Time of day the email was sent
- Length of the email in words

Which of the following scenarios, if true, is most likely to violate the key assumptions of Naive Bayes and potentially impact its performance?

(a) The length of the email follows a non-Gaussian distribution

(b) The time of day is discretized into categories (morning, afternoon, evening, night)

(c) The proportion of spam emails in the training data is lower than in real-world email traffic

(d) There’s a strong correlation between the presence of the word ”free” and the length of the email

Soln. D - This scenario violates the Naive Bayes assumption of feature independence, as it the features are dependent on each other.

9. Consider the two statements:
Statement 1: Bayesian Networks are inherently structured as Directed Acyclic Graphs (DAGs).
Statement 2: Each node in a bayesian network represents a random variable, and each edge represents conditional dependence.
Which of these are true?

   (a) Both the statements are True.

   (b) Statement 1 is true, and statement 2 is false.

   (c) Statement 1 is false, and statement 2 is true.

   (d) Both the statements are false.

   Soln. **A** - Bayesian Networks are structured as DAGs and each node represents a random variable, with edges indicating conditional dependencies.