Machine Learning - Unit 1 Question Bank

1. What do you understand from the term ‘Machine Learning’? How different is ML different from traditional programing?
2. What is machine learning? What is its significance?
3. Give examples of ***any five*** ML Applications. Justify why you think they fall under the category of ML Applications by highlighting what the machine learns in each of these applications.
4. What is the difference between Supervised and Unsupervised ML Algorithms? Explain with one case study example the difference between the two.
5. With one case study example each give a brief overview of the four types of ML Algorithms.
6. Name ***any two*** Supervised Learning Algorithms you have learnt. With examples justify why they belong to the category of ‘Supervised’ Learning?
7. With an e.g. each explain difference between ‘Classification’ & ‘Regression’ ML Algorithms.
8. With relevant examples briefly explain the difference between Artificial Intelligence, Machine Learning and Deep Learning.

(**Solution**: Refer ML introduction_PB.pdf. Also look at other solutions.)

9. In a Medical Classification Model to diagnose if a patient is diabetic or not 50,000 patients were examined. The number of TP = 8000, FP = 5000,TN = 35,000, FN = 2000. Do the following for this Medical Classification Model:
   - **a)** Draw the Confusion Matrix
   - **b)** Calculate the Accuracy value
   - **c)** Assess the Sensitivity Value
   - **d)** Estimate the Specificity Value
   - **e)** Measure the Error Rate
   - **f)** Determine the Precision Value
   - **g)** Compute the Recall Value
   - **h)** Evaluate the F-Measure
   - **i)** Is the data sample given to you a Balanced or an Imbalanced one? Justify your answer
   - **j)** Is this Model a ‘High Precision’ or ‘High Recall’ model? Justify your answer

(**Solution**: Refer 02_Accuracy_Precision_Recall_Confusion Matrix_MKN.pdf for a similar problem. Values are changed. Recalculate yourself.)

10. In a Spam Mail Filter Classification Model to filter out spam 5000 email samples were examined. The number of TP = 800, FP = 500,TN = 3500, FN = 200. Do the following for this Spam Mail Filter Classification Model:
    - **a)** Draw the Confusion Matrix
    - **b)** Calculate the Accuracy value
    - **c)** Assess the Sensitivity Value

- **d)** Estimate the Specificity Value
- **e)** Measure the Error Rate
- **f)** Determine the Precision Value
- **g)** Compute the Recall Value
- **h)** Evaluate the F-Measure
- **i)** Is the data sample given to you a Balanced or an Imbalanced one? Justify your answer
- **j)** Is this Model a 'High Precision' or 'High Recall' model? Justify your answer

(**Solution**: Refer 02_Accuracy_Precision_Recall_Confusion Matrix_MKN.pdf for a similar problem. Values are changed. Recalculate yourself.)

11. What is the difference between the measures 'Precision' and 'Recall'? Give one case study example of when you would use 'Precision' and when 'Recall'.

(**Solution**: Refer 02_Accuracy_Precision_Recall_Confusion Matrix_MKN)

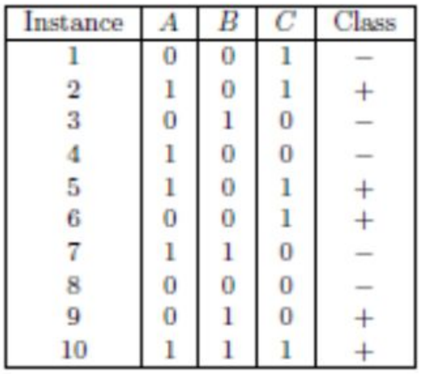
12. Give three computer applications EACH for which machine learning approaches seem appropriate and inappropriate. Give a brief justification why you think so.

13. Consider the following case studies. Determine which measure among 'Precision' and 'Recall' would you use for each of them. Justify your answer.

- **a)** People are innocent until proven guilty as per Indian Law. We want to avoid false convictions, even at the cost of criminals running free. (**Answer**: High Precision Model)
- **b)** In a Spam Mail Filter ML Model, we need to correctly classify so that no proper mails are incorrectly identified as spam. (**Answer**: High Precision Model)
- **c)** In an Examination Results ML Model, we need to correctly identify students who have failed. The model should never send a failed student home by classifying them as pass. (**Answer**: High Precision Model)
- **d)** In a Medical Diagnosis ML model we need to correctly identify the patients who are sick. The model should never send a sick patient home by classifying them as healthy.(**Answer**: High Recall Model)
- **e)** In Planet Utopia, the law demands that the number of criminals running free is brought as minimum as possible, even at the cost of wrongly convicting innocent people. (**Answer**: High Recall Model)
- **f)** In a Credit Card Fraud detection ML Model, we need to correctly identify fraudulent cards. The model should never classify a fraud card as safe and clean.(**Answer**: High Recall Model)

14. Consider the data set shown in Table below:

Dataset for a 2 class problem:



Instance	A	B	C	Class
1	0	0	1	−
2	1	0	1	+
3	0	1	0	−
4	1	0	0	−
5	1	0	1	+
6	0	0	1	+
7	1	1	0	−
8	0	0	0	−
9	0	1	0	+
10	1	1	1	+

(i) Estimate the conditional probabilities for $P(A = 1|+)$, $P(B = 1|+)$, $P(C = 1|+)$, $P(A = 1|-)$, $P(B = 1|-)$, *and* $P(C = 1|-)$

(ii) Use the conditional probabilities in part (i) to predict the class label for a test sample $(A = 1, B = 1, C = 1)$ using the **Naive Bayes** approach.

15. What do you mean by Imbalanced Datasets? Why does Accuracy fail as a performance metrics for these datasets? (**Solution**: Refer 03_F-Measures_Precision_Recall_MKN.pdf, 02_Accuracy_Precision_Recall_Confusion Matrix_MKN,

16. Discuss the applications of machine learning, classification and regression in machine learning with examples.

17. What do you understand by 'Matthew's Correlation Coefficient'? In which situation is it applied? Give an example. (**Solution**: Refer ML Terminologies_PB.pdf)

18. Discuss the following:

(i) Gibbs algorithm. Compare it with Bayes optimal classifier.

(ii) Working of Naïve Bayes classifier with an example.

(iii) Optimal classifier using Bayes method.

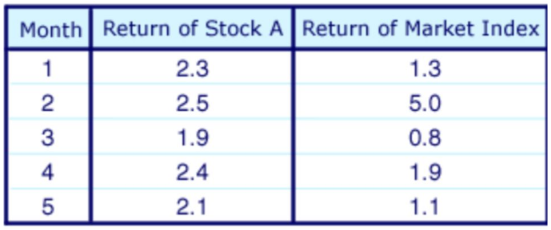
19. What do you understand by ROC Curves? Where is it used? Give one example.

20. What is the purpose of AUC in ROC Curves?

21. How are probabilities estimated? What is the role of ***m-estimate*** in the estimation of probabilities?

22. What is model overfitting? How is overfitting detected?

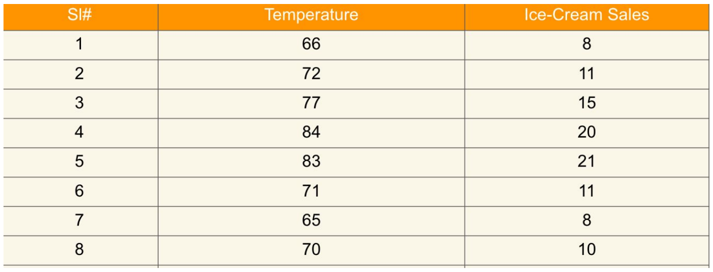
23. Given the following return information what is the covariance between return of Stock A and the return of Market Index?



Month	Return of Stock A	Return of Market Index
1	2.3	1.3
2	2.5	5.0
3	1.9	0.8
4	2.4	1.9
5	2.1	1.1

(**Solution**: http://ci.columbia.edu/ci/premba_test/c0331/s7/s7_5.html. Look at other problems too)

24. How 'correlated' are the parameters 'Temperature' and 'Ice-Cream Sales' for the following information given below:



Sl#	Temperature	Ice-Cream Sales
1	66	8
2	72	11
3	77	15
4	84	20
5	83	21
6	71	11
7	65	8
8	70	10

(**Solution**: Refer 07_Covariance_Correlation Coeff_MKN.pdf)

25. Explain with an example each the difference between the following:

a) Sensitivity and Specificity as Performance Measures of an ML Model
b) Precision and Recall as Performance Measures of an ML Model
c) Classification and Regression ML Algorithms
d) Balanced and Imbalanced Datasets
e) ROC and AUC Curves
f) Bias and Variance of the ML Model
g) Variance and Covariance of the ML Model
h) Overfitting and Underfitting of the ML Model
i) Train, Validation and Test Dataset

(**Solution**: Refer ML Terminologies_PB.pdf , 03_F-Measures_Precision_Recall_MKN.pdf, 02_Accuracy_Precision_Recall_Confusion Matrix_MKN, 06_Bias Variance_MKN .pdf, 08_ROC_AUC_MKN.pdf, 07_Covariance_Correlation Coeff_MKN.pdf)

26. What do you understand by the following terms? Give relevant examples / case studies

a) Bayesian Rule
b) Prior Probability, Joint Probability
c) Conditional Probability
d) Posterior Probability
e) Maximum A Posterior (MAP) Classification Rule
f) Probabilistic Classification Principle - Generative and Discriminative Model

(**Solution**: Refer The naïve Bayes' classifier_PB.pdf. Also look at other solutions.)

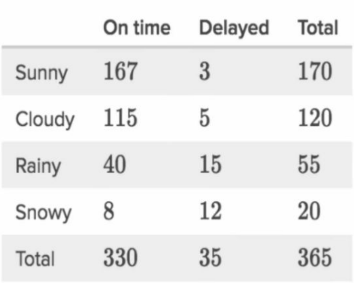
27. A hospital is testing patients for a certain disease. If a patient has the disease, the test is designed to return a "positive" result. If a patient does not have the disease, the test should return a "negative" result. No test is perfect though.

- 99% percent of patients who have the disease will test positive.
- 5% percent of patients who don't have the disease will also test positive.
- 10% percent of the population in question has the disease.

If a random patient tests positive, what is the probability that they have the disease?

(**Solution**:https://www.khanacademy.org/math/ap-statistics/probability-ap/stats-conditional-probability/a/tree-diagrams-conditional-probability)

28. James is interested in weather conditions to find out if the downtown train he sometimes takes runs on time. For a year, James records if each day is 'Sunny', 'Cloudy', 'Rainy' or 'Snowy' as well as if the train arrives late or on time on these weather conditions. His results are displayed in the table below. You are to use Conditional Probability to find out if the events "delayed" and "snowy" are independent of each other.



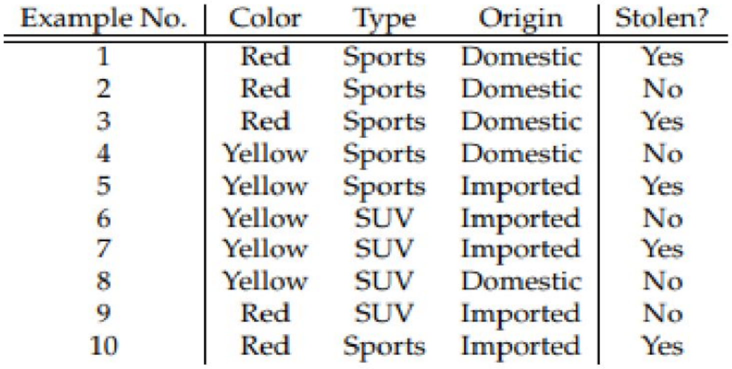
	On time	Delayed	Total
Sunny	167	3	170
Cloudy	115	5	120
Rainy	40	15	55
Snowy	8	12	20
Total	330	35	365

(**Solution**: Refer 04_Conditional Probability_Bayes Theorm_MKN.pdf. Look at other e.g. also.)

29. With an appropriate example explain Bayesian Rule. Derive Bayesian Rule from Conditional Probabilities. (**Solution**: Refer 05_Naive Bayes_MKN .pdf.)

30. Describe the Learning Phase and Testing Phase Algorithm for Naive Bayes Classifier.
(**Solution**: Refer The naïve Bayes' classifier_PB.pdf.)

31. Apply Naive Bayes Classifier to the below dataset and classify if the red, domestic, SUV is stolen?

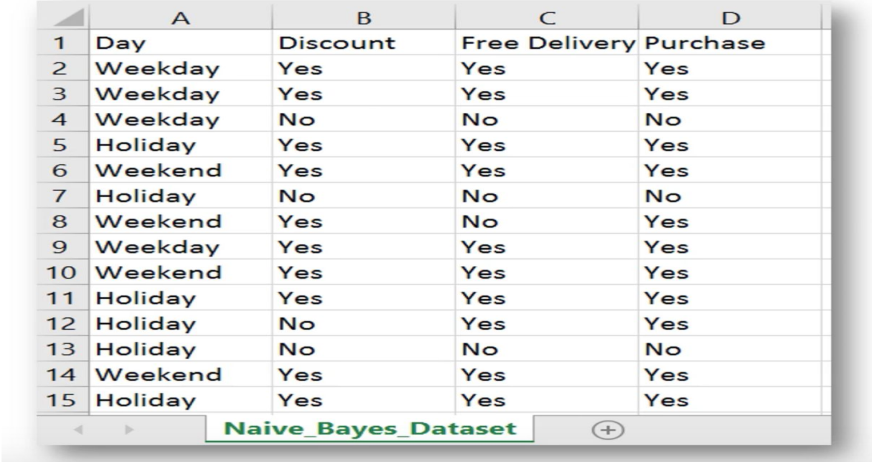


Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

(**Solution**: Refer The naïve Bayes' classifier_PB.pdf. Also look at other solutions in 05_Naive

Bayes_MKN.pdf and Stephen Marsland Section 2.3.2)

32. Apply Naive Bayes Classifier to the below dataset and classify if a customer will buy a product on a holiday, with discount and free delivery.



	A	B	C	D
1	Day	Discount	Free Delivery	Purchase
2	Weekday	Yes	Yes	Yes
3	Weekday	Yes	Yes	Yes
4	Weekday	No	No	No
5	Holiday	Yes	Yes	Yes
6	Weekend	Yes	Yes	Yes
7	Holiday	No	No	No
8	Weekend	Yes	No	Yes
9	Weekday	Yes	Yes	Yes
10	Weekend	Yes	Yes	Yes
11	Holiday	Yes	Yes	Yes
12	Holiday	No	Yes	Yes
13	Holiday	No	No	No
14	Weekend	Yes	Yes	Yes
15	Holiday	Yes	Yes	Yes

(**Solution**: Refer 05_Naive Bayes_MKN .pdf. Also look at other solutions in The naïve Bayes' classifier_PB.pdf and Stephen Marsland Section 2.3.2)