40 Machine Learning Questions to test a data scientist

Created	@May 11, 2020 8:39 PM
Created By	Chandana Brdwj
Last Edited By	Chandana Brdwj
Last Edited Time	@May 19, 2020 9:34 PM
Stakeholders	
Status	
Type	
URL	https://www.analyticsvidhya.com/blog/2017/04/40-questions-test-data-scientist-machine-learning-solution-skillpower-machine-learning-datafest-2017/

Questions & Solutions

Question Context

A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college.

1) Which of the following statement is true in following case?

A) Feature F1 is an example of nominal variable.

B) Feature F1 is an example of ordinal variable.

C) It doesn't belong to any of the above category.

D) Both of these

Solution: (B)

Ordinal variables are the variables which has some order in their categories. For example, grade A should be considered as high grade than grade B.

---

2) Which of the following is an example of a deterministic algorithm?

A) PCA

B) K-Means

C) None of the above

Solution: (A)

A deterministic algorithm is that in which output does not change on different runs.

PCA would give the same result if we run again, but not k-means.

---

3) [True or False] A Pearson correlation between two variables is zero but, their values can still be related to each other.

A) TRUE

B) FALSE

Solution: (A)

$Y=X2$. Note that, they are not only associated, but one is a function of the other and Pearson correlation between them is 0.

---

4) Which of the following statement(s) is / are true for Gradient Descent (GD) and Stochastic Gradient Descent (SGD)?

1. **In GD and SGD, you update a set of parameters in an iterative manner to minimize the error function.**
2. **In SGD, you have to run through all the samples in your training set for a single update of a parameter in each iteration.**
3. **In GD, you either use the entire data or a subset of training data to update a parameter in each iteration.**

A) Only 1

B) Only 2

C) Only 3

D) 1 and 2

E) 2 and 3

F) 1,2 and 3

Solution: (A)

In SGD for each iteration you choose the batch which generally contains the random sample of data. But in case of GD each iteration contains all of the training observations.

---

5) Which of the following hyper parameter(s), when increased may cause random forest to overfit the data?

1. **Number of Trees**
2. **Depth of Tree**
3. **Learning Rate**

A) Only 1

B) Only 2

C) Only 3

D) 1 and 2

E) 2 and 3

F) 1,2 and 3

Solution: (B)

Usually, if we increase the depth of tree it will cause overfitting. Learning rate is not an hyperparameter in random forest. Increase in the number of tree will cause under fitting.

---

6) Imagine, you are working with "Analytics Vidhya" and you want to develop a machine learning algorithm which predicts the number of views on the articles.

Your analysis is based on features like author name, number of articles written by the same author on Analytics Vidhya in past and a few other features. Which of the following evaluation metric would you choose in that case?

1. **Mean Square Error**
2. **Accuracy**

3. **F1 Score**

A) Only 1

B) Only 2

C) Only 3

D) 1 and 3

E) 2 and 3

F) 1 and 2

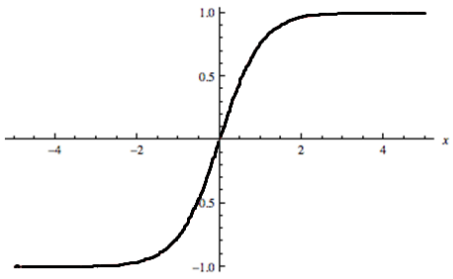
Solution:(A)

You can think that the number of views of articles is the continuous target variable which fall under the regression problem. So, mean squared error will be used as an evaluation metric.

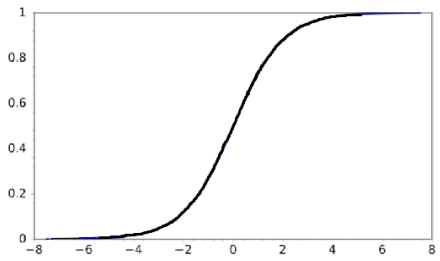
---

7) Given below are three images (1,2,3). Which of the following option is correct for these images?

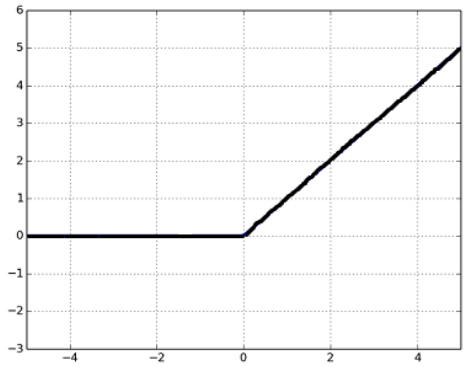
A)



B)



C)



A) 1 is tanh, 2 is ReLU and 3 is SIGMOID activation functions.

B) 1 is SIGMOID, 2 is ReLU and 3 is tanh activation functions.

C) 1 is ReLU, 2 is tanh and 3 is SIGMOID activation functions.

D) 1 is tanh, 2 is SIGMOID and 3 is ReLU activation functions.

Solution: (D)

The range of SIGMOID function is [0,1].

The range of the tanh function is [-1,1].

The range of the RELU function is [0, infinity].

So Option D is the right answer.

---

8) Below are the 8 actual values of target variable in the train file.

[0,0,0,1,1,1,1,1]

What is the entropy of the target variable?

A) -(5/8 log(5/8) + 3/8 log(3/8))

B) 5/8 log(5/8) + 3/8 log(3/8)

C) 3/8 log(5/8) + 5/8 log(3/8)

D) 5/8 log(3/8) – 3/8 log(5/8)

Solution: (A)

The answer is A.

---

9) Let's say, you are working with categorical feature(s) and you have not looked at the distribution of the categorical variable in the test data.

You want to apply one hot encoding (OHE) on the categorical feature(s). What challenges you may face if you have applied OHE on a categorical

variable of train dataset?

A) All categories of categorical variable are not present in the test dataset.

B) Frequency distribution of categories is different in train as compared to the test dataset.

C) Train and Test always have same distribution.

D) Both A and B

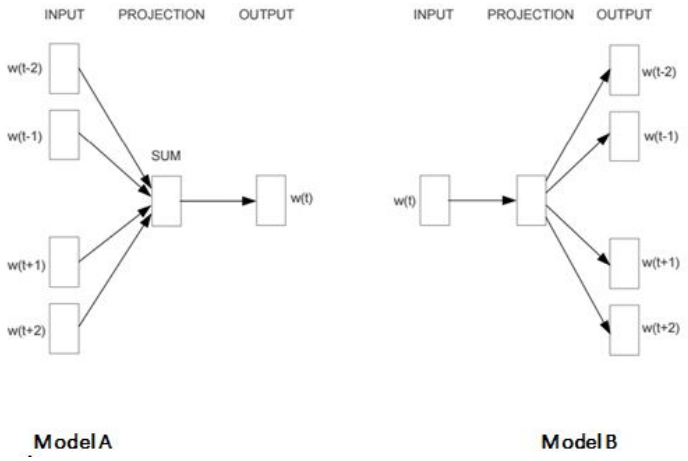
E) None of these

Solution: (D)

Both are true, the OHE will fail to encode the categories which is present in test but not in train so it could be one of the main challenges while applying OHE. The challenge given in option B is also true, you need to be more careful while applying OHE if the frequency distribution is not same in train and test.

---

10) Skip gram model is one of the best models used in Word2vec algorithm for words embedding. Which one of the following models depict the skip gram model?



A) A

B) B

C) Both A and B

D) None of these

Solution: (B)

Both models (modelA and modelB) are used in Word2vec algorithm. The modelA represents a CBOW model whereas ModelB represents the Skip gram

model.

---

11) Let's say, you are using activation function X in hidden layers of neural network. At a particular neuron for any given input, you get the output as "-0.0001". Which of the following activation function could X represent?

A) ReLU

B) tanh

C) SIGMOID

D) None of these

Solution: (B)

The function is a tanh because the this function output range is between (-1,-1).

---

12) [True or False] LogLoss evaluation metric can have negative values.

A) TRUE B) FALSE

Solution: (B)

Log loss cannot have negative values.

13) Which of the following statements is/are true about "Type-1" and "Type-2" errors?

1. **Type1 is known as false positive and Type2 is known as false negative.**
2. **Type1 is known as false negative and Type2 is known as false positive.**
3. **Type1 error occurs when we reject a null hypothesis when it is actually true.**

A) Only 1

B) Only 2

C) Only 3

D) 1 and 2

E) 1 and 3

F) 2 and 3

Solution: (E)

In statistical hypothesis testing, a type I error is the incorrect rejection of a true null hypothesis (a "false positive"), while a type II error is incorrectly

retaining a false null hypothesis (a “false negative”).

---

14) Which of the following is/are one of the important step(s) to pre-process the text in NLP based projects?

1. **Stemming**
2. **Stop word removal**
3. **Object Standardization**

A) 1 and 2

B) 1 and 3

C) 2 and 3

D) 1,2 and 3

Solution: (D)

Stemming is a rudimentary rule-based process of stripping the suffixes (“ing”, “ly”, “es”, “s” etc) from a word.

Stop words are those words which will have not relevant to the context of the data for example is/am/are.

Object Standardization is also one of the good way to pre-process the text.

---

15) Suppose you want to project high dimensional data into lower dimensions. The two most famous dimensionality reduction algorithms used here are PCA and t-SNE. Let’s say you have applied both algorithms respectively on data “X” and you got the datasets “X_projected_PCA” , “X_projected_tSNE”.

Which of the following statements is true for “X_projected_PCA” & “X_projected_tSNE” ?

A) X_projected_PCA will have interpretation in the nearest neighbour space.

B) X_projected_tSNE will have interpretation in the nearest neighbour space.

C) Both will have interpretation in the nearest neighbour space.

D) None of them will have interpretation in the nearest neighbour space.

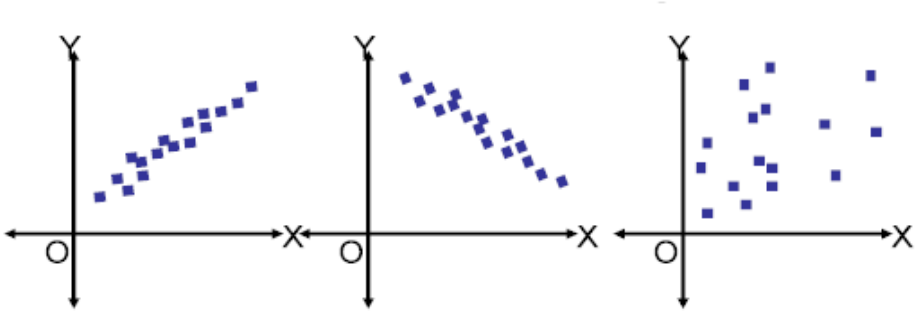
Solution: (B)

t-SNE algorithm considers nearest neighbour points to reduce the dimensionality of the data. So, after using t-SNE we can think that reduced

dimensions will also have interpretation in nearest neighbour space. But in case of PCA it doesn't.

Context: 16-17

Given below are three scatter plots for two features (Image 1, 2 & 3 from left to right).



16) In the above images, which of the following is/are example of multi-collinear features?

A) Features in Image 1

B) Features in Image 2

C) Features in Image 3

D) Features in Image 1 & 2

E) Features in Image 2 & 3

F) Features in Image 3 & 1

Solution: (D)

In Image 1, features have high positive correlation whereas in Image 2 has high negative correlation between the features so in both images, pair of features are the example of multicollinear features.

---

17) In previous question, suppose you have identified multi-collinear features. Which of the following action(s) would you perform next?

1. **Remove both collinear variables.**
2. **Instead of removing both variables, we can remove only one variable.**
3. **Removing correlated variables might lead to loss of information. In order to retain those variables, we can use penalized regression models like**

ridge or lasso regression.

A) Only 1

B)Only 2

C) Only 3

D) Either 1 or 3

E) Either 2 or 3

Solution: (E)

You cannot remove both the features. If you remove both the features you may lose the data. So, you should either remove 1 feature or use the regularization algorithms like L1 and L2.

---

18) Adding a non-important feature to a linear regression model may result in.

1. **Increase in R-square**
2. **Decrease in R-square**

A) Only 1 is correct

B) Only 2 is correct

C) Either 1 or 2

D) None of these

Solution: (A)

After adding a feature in feature space, whether that feature is important or not, the R-squared always increases.

---

19) Suppose, you are given three variables X, Y and Z. The Pearson correlation coefficients for (X, Y), (Y, Z) and (X, Z) are C1, C2 & C3 respectively.

Now, you have added 2 in all values of X (i.e., new values become X+2), subtracted 2 from all values of Y (i.e. new values are Y-2) and Z remains the same. The new coefficients for (X,Y), (Y,Z) and (X,Z) are given by D1, D2 & D3 respectively. How do the values of D1, D2 & D3 relate to C1, C2 & C3?

A) D1= C1, D2 < C2, D3 > C3

B) D1 = C1, D2 > C2, D3 > C3

C) D1 = C1, D2 > C2, D3 < C3

D) D1 = C1, D2 < C2, D3 < C3

E) D1 = C1, D2 = C2, D3 = C3

F) Cannot be determined

Solution: (E)

Correlation between the features won't change if you add or subtract a value in the features.

20) Imagine, you are solving a classification problems with highly imbalanced class. The majority class is observed 99% of times in the training data.

Your model has 99% accuracy after taking the predictions on test data. Which of the following is true in such a case?

1. Accuracy metric is not a good idea for imbalanced class problems.
2. Accuracy metric is a good idea for imbalanced class problems.
3. Precision and recall metrics are good for imbalanced class problems.
4. Precision and recall metrics aren't good for imbalanced class problems.

A) 1 and 3

B) 1 and 4

C) 2 and 3

D) 2 and 4

Solution: (A)

---

21) In ensemble learning, you aggregate the predictions for weak learners, so that an ensemble of these models will give a better prediction than prediction of individual models.

Which of the following statements is / are true for weak learners used in ensemble model?

1. **They don't usually overfit.**
2. **They have high bias, so they cannot solve complex learning problems**
3. **They usually overfit.**

A) 1 and 2

B) 1 and 3

C) 2 and 3

D) Only 1

E) Only 2

F) None of the above

Solution: (A)

Weak learners are sure about particular part of a problem. So, they usually don't overfit which means that weak learners have low variance and high bias.

---

22) Which of the following options is/are true for K-fold cross-validation?

1. **Increase in K will result in higher time required to cross validate the result.**
2. **Higher values of K will result in higher confidence on the cross-validation result as compared to lower value of K.**
3. **If K=N, then it is called Leave one out cross validation, where N is the number of observations.**

A) 1 and 2

B) 2 and 3

C) 1 and 3

D) 1,2 and 3

Solution: (D)

Larger k value means less bias towards overestimating the true expected error (as training folds will be closer to the total dataset) and higher running time (as you are getting closer to the limit case: Leave-One-Out CV). We also need to consider the variance between the k folds accuracy while selecting the k.

Question Context 23-24

Cross-validation is an important step in machine learning for hyper parameter tuning. Let's say you are tuning a hyper-parameter "max_depth" for GBM by selecting it from 10 different depth values (values are greater than 2) for tree based model using 5-fold cross validation.

Time taken by an algorithm for training (on a model with max_depth 2) 4-fold is 10 seconds and for the prediction on remaining 1-fold is 2 seconds.

Note: Ignore hardware dependencies from the equation.

23) Which of the following option is true for overall execution time for 5-fold cross validation with 10 different values of "max_depth"?

A) Less than 100 seconds

B) 100 – 300 seconds

C) 300 – 600 seconds

D) More than or equal to 600 seconds

C) None of the above

D) Can't estimate

Solution: (D)

Each iteration for depth "2" in 5-fold cross validation will take 10 secs for training and 2 second for testing. So, 5 folds will take 12*5 = 60 seconds. Since we are searching over the 10 depth values so the algorithm would take 60*10 = 600 seconds. But training and testing a model on depth greater than 2 will take more time than depth "2" so overall timing would be greater than 600.

---

24) In previous question, if you train the same algorithm for tuning 2 hyper parameters say "max_depth" and "learning_rate".

You want to select the right value against "max_depth" (from given 10 depth values) and learning rate (from given 5 different learning rates). In such cases, which of the following will represent the overall time?

A) 1000-1500 second

B) 1500-3000 Second

C) More than or equal to 3000 Second

D) None of these

Solution: (D)

Same as question number 23.

---

25) Given below is a scenario for training error TE and Validation error VE for a machine learning algorithm M1. You want to choose a hyperparameter (H) based on TE and VE.

# 1	# 105	# 90	Aa Title
2	200	85	Untitled
3	250	96	Untitled
4	105	85	Untitled
5	300	100	Untitled

Which value of H will you choose based on the above table?

A) 1

B) 2

C) 3

D) 4

E) 5

Solution: (D)

Looking at the table, option D seems the best

---

26) What would you do in PCA to get the same projection as SVD?

A) Transform data to zero mean

B) Transform data to zero median

C) Not possible

D) None of these

Solution: (A)

When the data has a zero mean vector PCA will have same projections as SVD, otherwise you have to centre the data first before taking SVD.

Question Context 27-28

Assume there is a black box algorithm, which takes training data with multiple observations (t1, t2, t3,…….. tn) and a new observation (q1). The black box outputs the nearest neighbor of q1 (say ti) and its corresponding class label ci.

You can also think that this black box algorithm is same as 1-NN (1-nearest neighbor).

27) It is possible to construct a k-NN classification algorithm based on this black box alone.

Note: Where n (number of training observations) is very large compared to k.

A) TRUE

B) FALSE

Solution: (A)

In first step, you pass an observation (q1) in the black box algorithm so this algorithm would return a nearest observation and its class.

In second step, you through it out nearest observation from train data and again input the observation (q1). The black box algorithm will again return the a nearest observation and it's class.

You need to repeat this procedure k times

---

28) Instead of using 1-NN black box we want to use the j-NN (j>1) algorithm as black box. Which of the following option is correct for finding k-NN using j-NN?

1. **J must be a proper factor of k**
2. **J > k**
3. **Not possible**

A) 1

B) 2

C) 3

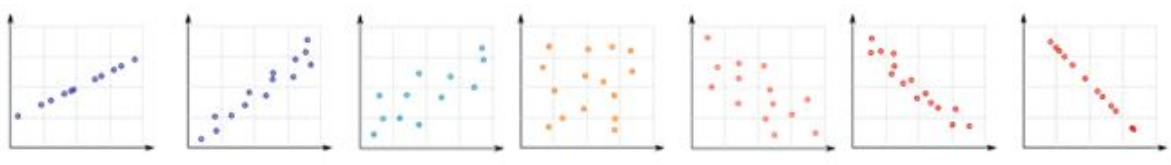
Solution: (A)

Same as question number 27

---

29) Suppose you are given 7 Scatter plots 1-7 (left to right) and you want to compare Pearson correlation coefficients between variables of each scatterplot.

Which of the following is in the right order?



1. 1<2<3<4
2. 1>2>3 > 4
3. 7<6<5<4
4. 7>6>5>4

A) 1 and 3

B) 2 and 3

C) 1 and 4

D) 2 and 4

Solution: (B)

from image 1to 4 correlation is decreasing (absolute value). But from image 4 to 7 correlation is increasing but values are negative (for example, 0, -0.3, -0.7, -0.99).

---

30) You can evaluate the performance of a binary class classification problem using different metrics such as accuracy, log-loss, F-Score. Let's say, you are using the log-loss function as evaluation metric.

Which of the following option is / are true for interpretation of log-loss as an evaluation metric?

1. If a classifier is confident about an incorrect classification, then log-loss will penalise it heavily.

$$logLoss = \frac{-1}{N}\sum_{i=1}^{N}(y_i(logp_i) + (1-y_i)log(1-p_i))$$

2. For a particular observation, the classifier assigns a very small probability for the correct class then the corresponding contribution to the log-loss will be very large.
3. Lower the log-loss, the better is the model.

A) 1 and 3

B) 2 and 3

C) 1 and 2

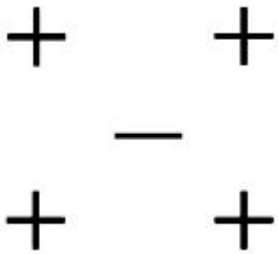
D) 1,2 and 3

Solution: (D)

Options are self-explanatory.

Question 31-32

Below are five samples given in the dataset.



Note: Visual distance between the points in the image represents the actual distance.

31) Which of the following is leave-one-out cross-validation accuracy for 3-NN (3-nearest neighbor)?

A) 0

D) 0.4

C) 0.8

D) 1

Solution: (C)

In Leave-One-Out cross validation, we will select (n-1) observations for training and 1 observation of validation. Consider each point as a cross validation point and then find the 3 nearest point to this point. So if you repeat this procedure for all points you will get the correct classification for all positive class given in the above figure but negative class will be misclassified. Hence you will get 80% accuracy.

---

32) Which of the following value of K will have least leave-one-out cross validation accuracy?

A) 1NN

B) 3NN

C) 4NN

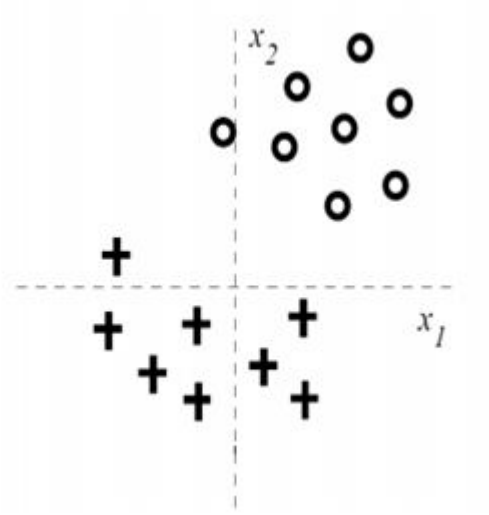
D) All have same leave one out error

Solution: (A)

Each point which will always be misclassified in 1-NN which means that you will get the 0% accuracy.

---

33) Suppose you are given the below data and you want to apply a logistic regression model for classifying it in two given classes.



You are using logistic regression with L1 regularization.

Where C is the regularization parameter and w1 & w2 are the coefficients of x1 and x2.

Which of the following option is correct when you increase the value of C from zero to a very large value?

A) First w2 becomes zero and then w1 becomes zero

B) First w1 becomes zero and then w2 becomes zero

C) Both becomes zero at the same time

D) Both cannot be zero even after very large value of C

Solution: (B)

By looking at the image, we see that even on just using x2, we can efficiently perform classification. So at first w1 will become 0. As regularization parameter increases more, w2 will come more and more closer to 0.

---

34) Suppose we have a dataset which can be trained with 100% accuracy with help of a decision tree of depth 6. Now consider the points below and choose the option based on these points.

Note: All other hyper parameters are same and other factors are not affected.

1. **Depth 4 will have high bias and low variance**
2. **Depth 4 will have low bias and low variance**

A) Only 1

B) Only 2

C) Both 1 and 2

D) None of the above

Solution: (A)

If you fit decision tree of depth 4 in such data, it will be more likely to underfit the data. So, in case of underfitting you will have high bias and low variance.

---

35) Which of the following options can be used to get global minima in k-Means Algorithm?

1. **Try to run algorithm for different centroid initialization**
2. **Adjust number of iterations**
3. **Find out the optimal number of clusters**

A) 2 and 3

B) 1 and 3

C) 1 and 2

D) All of above

Solution: (D)

All of the option can be tuned to find the global minima.

---

36) Imagine you are working on a project which is a binary classification problem. You trained a model on training dataset and get the below confusion matrix on validation dataset.

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

Based on the above confusion matrix, choose which option(s) below will give you correct predictions?

1. **Accuracy is ~0.91**
2. **Misclassification rate is ~ 0.91**
3. **False positive rate is ~0.95**
4. **True positive rate is ~**0.95

A) 1 and 3

B) 2 and 4

C) 1 and 4

D) 2 and 3

Solution: (C)

The Accuracy (correct classification) is (50+100)/165 which is nearly equal to 0.91.

The true Positive Rate is how many times you are predicting positive class correctly so true positive rate would be 100/105 = 0.95 also known as "Sensitivity" or "Recall"

---

37) For which of the following hyperparameters, higher value is better for decision tree algorithm?

1. **Number of samples used for split**
2. **Depth of tree**
3. **Samples for leaf**

A)1 and 2

B) 2 and 3

C) 1 and 3

D) 1, 2 and 3

E) Can't say

Solution: (E)

For all three options A, B and C, it is not necessary that if you increase the value of parameter the performance may increase. For example, if we have a very high value of depth of tree, the resulting tree may overfit the data, and would not generalize well. On the other hand, if we have a very low value, the tree may underfit the data. So, we can't say for sure that "higher is better".

Context 38-39

Imagine, you have a 28 * 28 image and you run a 3 * 3 convolution neural network on it with the input depth of 3 and output depth of 8.

Note: Stride is 1 and you are using same padding.

38) What is the dimension of output feature map when you are using the given parameters.

A) 28 width, 28 height and 8 depth

B) 13 width, 13 height and 8 depth

C) 28 width, 13 height and 8 depth

D) 13 width, 28 height and 8 depth

Solution: (A)

The formula for calculating output size is

output size = (N – F)/S + 1

where, N is input size, F is filter size and S is stride.

---

39) What is the dimensions of output feature map when you are using following parameters.

A) 28 width, 28 height and 8 depth

B) 13 width, 13 height and 8 depth

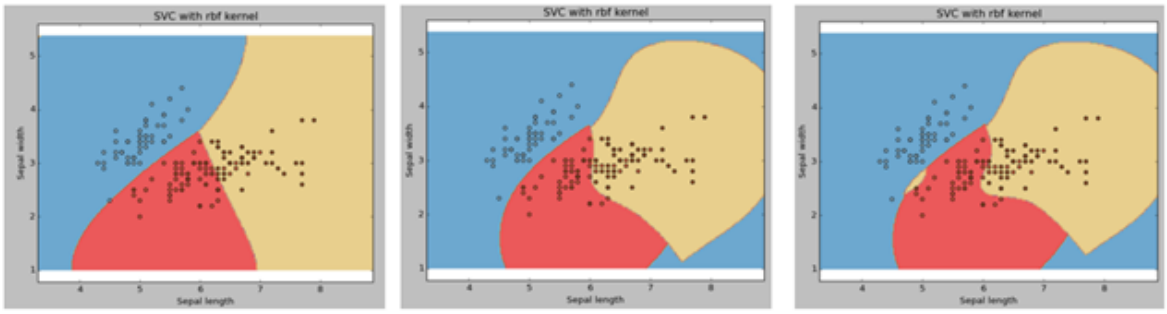
C) 28 width, 13 height and 8 depth

D) 13 width, 28 height and 8 depth

Solution: (B)

Same as above

40) Suppose, we were plotting the visualization for different values of C (Penalty parameter) in SVM algorithm. Due to some reason, we forgot to tag the C values with visualizations. In that case, which of the following option best explains the C values for the images below (1,2,3 left to right, so C values are C1 for image1, C2 for image2 and C3 for image3) in case of rbf kernel.



A) C1 = C2 = C3

B) C1 > C2 > C3

C) C1 < C2 < C3

D) None of these

Solution: (C)

Penalty parameter C of the error term. It also controls the trade-off between smooth decision boundary and classifying the training points correctly. For large values of C, the optimization will choose a smaller-margin hyperplane. Read more here.