Lab 8

Deep Learning, SVMs, Kernels

Using Zoom for Lectures

Sign in using:

your name

Please mute both:

your video cameras for the entire lecture

your audio/mics unless asking or answering a question

Asking/answering a question, option 1:

click on Participants

use the hand icon to raise your hand

I will call on you and ask you to unmute yourself

Asking/answering a question, option 2:

click on Chat

type your question, and I will answer it

Questions & Answers

1)The difference between deep learning and machine learning algorithms is that there is no need of feature engineering in machine learning algorithms, whereas, it is recommended to do feature engineering first and then apply deep learning.

A) True

B) False

Solution: (B).

Deep learning itself does feature engineering whereas machine learning requires manual feature engineering.

2) Which of the following is a representation learning algorithm?

A) Neural network

B) Random Forest

C) k-Nearest neighbor

D) None of the above

Solution: (A)

Neural network converts data in such a form that it would be better to solve the desired problem. This is called representation learning.

3) Increase in size of a convolutional kernel would necessarily increase the performance of a convolutional neural network.

A) TRUE

B) FALSE

Solution: (B)

Kernel size is a hyperparameter and therefore by changing it we can increase or decrease performance.

4) Suppose we have a 5-layer neural network which takes 3 hours to train on a GPU with 4GB VRAM. At test time, it takes 2 seconds for single data point.
Now we change the architecture such that we add dropout after 2nd and 4th layer with rates 0.2 and 0.3 respectively.
What would be the testing time for this new architecture?

A) Less than 2 secs
B) Exactly 2 secs
C) Greater than 2 secs
D) Can’t Say

Solution: (B)

The changes is architecture when we add dropout only changes in the training, and not at test time.

5) Which of the following options can be used to reduce overfitting in deep learning models?

1. **Add more data**
2. **Use data augmentation**
3. **Use architecture that generalizes well**
4. **Add regularization**
5. **Reduce architectural complexity**

A) 1, 2, 3

B) 1, 4, 5

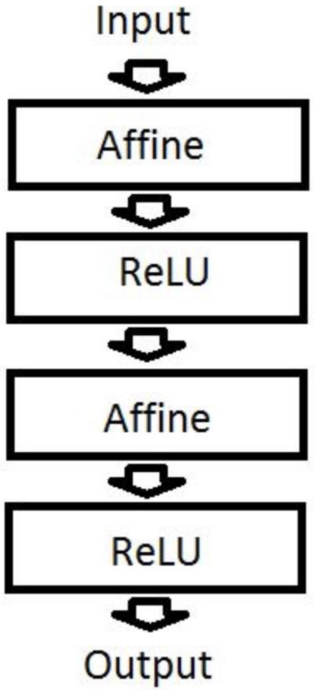
C) 1, 3, 4, 5

D) All of these

Solution: (D)

All of the above techniques can be used to reduce overfitting.

6)Suppose there is a neural network with the below configuration.
If we remove the ReLU layers, we can still use this neural network to model non-linear functions.



A) TRUE

B) FALSE

Solution: (B)

7)**Deep learning can be applied to which of the following Natural language Processing (NLP) tasks?**

A) Machine translation
B) Sentiment analysis
C) Question Answering system
D) All of the above

Solution: (D)

Deep learning can be applied to all of the above-mentioned NLP tasks.

8) Given an n-character word, we want to predict which character would be the (n+1)-th character in the sequence. For example, our input is “prediction” (which is a 9 character word) and we have to predict what would be the 10th character.

Which neural network architecture would be suitable to complete this task?

A) Fully-Connected Neural Network

B) Convolutional Neural Network

C) Recurrent Neural Network

D) Restricted Boltzmann Machine

Solution: (C)

Recurrent neural network works best for sequential data. Therefore, it would be best for the task.

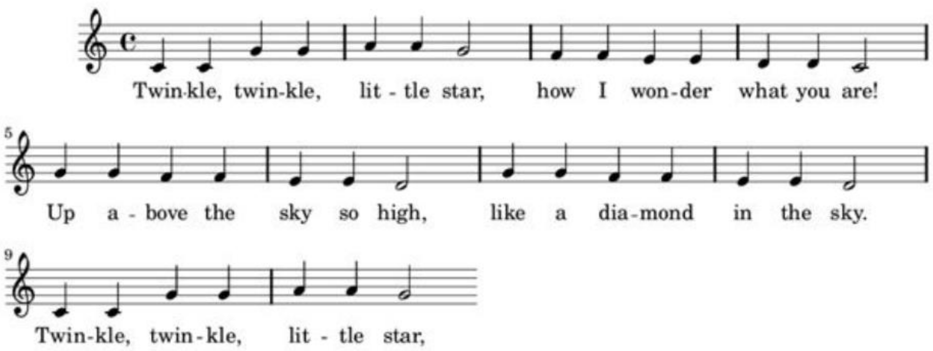
9) Exploding gradient problem is an issue in training deep networks where the gradient gets so large that the loss goes to an infinitely high value and then explodes.
What is the probable approach when dealing with “Exploding Gradient” problem in RNNs?

A) Use modified architectures like LSTM and GRUs
B) Gradient clipping
C) Dropout
D) None of these

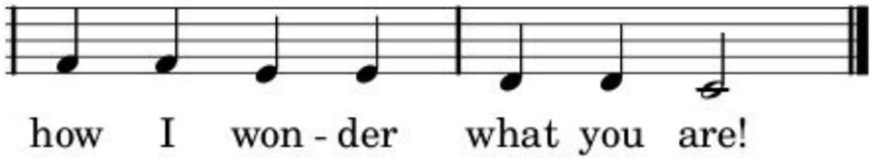
Solution: (B)

To deal with exploding gradient problem, it’s best to threshold the gradient values at a specific point. This is called gradient clipping.

10) Suppose your task is to predict the next few notes of song when you are given the preceding segment of the song.
For example: The input given to you is an image depicting the music symbols as given below,


Twin-kle, twin-kle, lit - tle star, how I won-der what you are!
5
Up a - bove the sky so high, like a dia-mond in the sky.
9
Twin-kle, twin-kle, lit - tle star,

Your required output is an image of succeeding symbols.



Which architecture of neural network would be better suited to solve the problem?

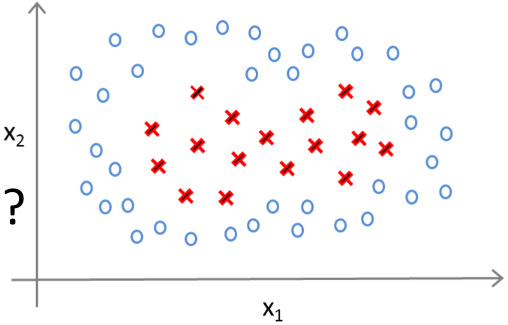
A) End-to-End fully connected neural network
B) Convolutional neural network followed by recurrent units
C) Neural Turing Machine
D) None of these

Solution: (B)

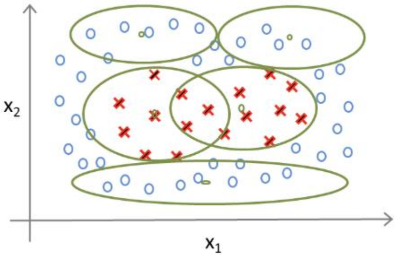
CNN work best on image recognition problems, whereas RNN works best on sequence prediction. Here you would have to use best of both worlds!

SVMs & Kernels

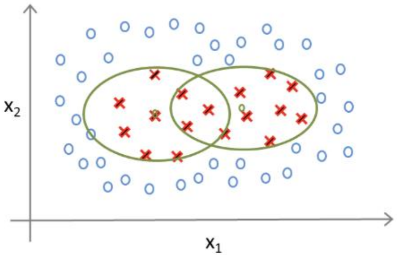
Consider the following dataset of 2-dimensional datapoints:



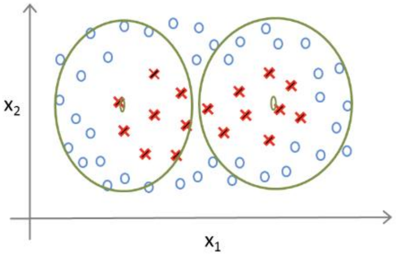
a)Which placement of Gaussian basis functions corresponds to a kernel feature representation for the dataset? Explain your answer in one sentence below.



(A)



(B)



(C)

(D) none of the above

Answer: (D)

to compute a kernel representation, we place a separate Gaussian distribution centered at **each** data point.

b) How does increasing the variance of the Gaussian affect the bias and variance of the resulting Gaussian Kernel SVM classifier? Explain.

Answer: It makes the boundary smoother, so the classifier has higher bias and lower variance.

c)For $k(x_1,x_2)$ to be a valid kernel, there must be a feature basis function $\phi(.)$ such that we can write $k(x_1,x_2)=\phi(x_1)^T\phi(x_2)$. Suppose $k_1(x_1,x_2)$ and $k_2(x_1,x_2)$ are valid kernels. Prove that the following is also a valid kernel:

$$k(x_1,x_2)=k_1(x_1,x_2)+k_2(x_1,x_2)$$

Answer:

$k(x_1,x_2)=k_1(x_1,x_2)+k_2(x_1,x_2)=\phi_1(x_1)^T\phi_1(x_2)+\phi_2(x_1)^T\phi_2(x_2)=\phi(x_1)^T(x_2)$ where $\phi(x)=[\phi_1(x), \phi_2(x)]$

d)Both SVMs with Gaussian kernels and Neural Networks with at least one hidden layer can be used to learn non-linear decision boundaries, such as the boundary between positive and negative examples in the dataset above. Describe the main similarity and the main difference in how these two approaches achieve this.

Answer: The similarity is that both can be thought of as mapping the input features x into a new feature space: the SVM kernel maps x to a new feature vector x, and the hidden layer of the neural network maps x to the activations of the last hidden layer, a. The difference is in how the mapping is done, where SVM uses training data as landmarks, but neural network learns the feature mapping through layer parameters.

e) Explain what slack variables are used for when training SVMs.

Answer: Slack variables are assigned to solve a problem that is not linearly separable by adding ‘slack’ values to the constraints.