

## Feedback — Lecture 10 Quiz

[Help Center](#)

You submitted this quiz on **Fri 1 Jul 2016 1:01 AM CEST**. You got a score of **7.00** out of **8.00**. However, you will not get credit for it, since it was submitted past the deadline.

### Question 1

When learning a mixture of experts, it is desirable that each expert specializes in a different area of the input space. But then at test time, how will we know which expert to use?




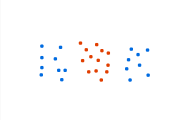
Your Answer	Score	Explanation
<input type="radio"/> We see which training case the test case is closest to (in input space) and use the model that was used for that training case.		
<input type="radio"/> We also learn a "manager" model that sees the input and assigns probabilities for picking each expert. We then get predictions from all the experts and take their weighted average using the probabilities.		
<input checked="" type="radio"/> We also learn a "manager" model that sees the input and assigns probabilities for picking each expert. We then choose the expert that has the highest probability and use it to make a prediction.	✖ 0.00	Averaging many models is usually better than picking the best one.
<input type="radio"/> We uniformly average the predictions of each expert.		
Total	0.00 / 1.00	

### Question 2

Which data set **can not** be classified perfectly with a linear classifier but **can** be classified perfectly with a mixture of two experts where each expert is a linear classifier ?

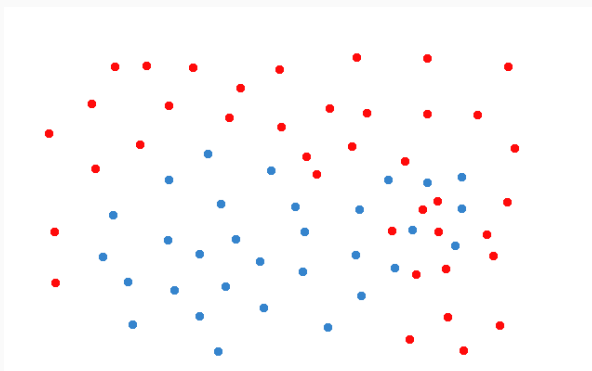
Assume that the manager is also linear, i.e., it can decide which expert to use based on a linear

criterion only. (In other words, the manager has a linear function  $f$  and given any input case  $x$ , it must decide to apply expert 1 with probability 1 if  $f(x) > 0$  and expert 2 with probability 1 if  $f(x) \leq 0$ .

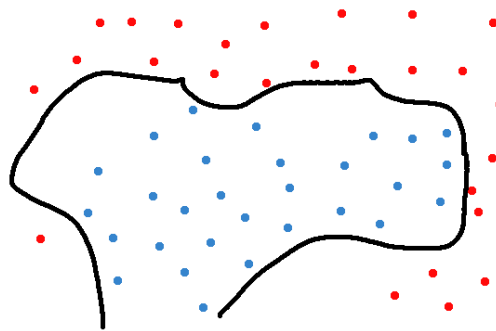
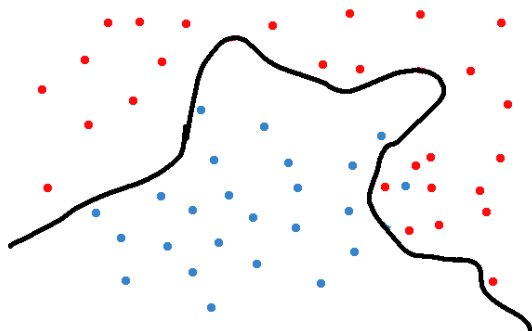
Your Answer	Score	Explanation
<input type="checkbox"/> 	<input checked="" type="checkbox"/> 0.50	This set is linearly separable already.
<input type="checkbox"/> 	<input checked="" type="checkbox"/> 0.50	This cannot be separated by two linear classifiers no matter how powerful the manager is.
<input checked="" type="checkbox"/> 	<input checked="" type="checkbox"/> 0.50	No linear classifier can separate this data set. However, we can train one linear classifier for the top part and one for the bottom. The manager can then choose the expert based on a linear criterion ('top' or 'bottom').
<input checked="" type="checkbox"/> 	<input checked="" type="checkbox"/> 0.50	No linear classifier can separate this data set. However, we can train one linear classifier for the left part set of blue points vs. red points and another classifier for right set of blue points vs. red points.
Total	2.00 / 2.00	

### Question 3

Andy has a dataset of points that he wishes to classify. This set is shown below.



Being knowledgeable about bagging, he samples this data set and creates 3 separate ones. He then uses a neural net to learn separate classifiers on each data set. The learned classifier boundaries are shown below. Note that each data set is a subset of the complete dataset.



Which of the following statements is true ?

**Your Answer**

**Score Explanation**

☐ All the learned models make a lot of errors ("high-bias"), so model averaging is likely to help in generalization.

☐ The learned models are different ("high-variance") and do well on their training sets, so model averaging is unlikely to help in generalization.

☐ All the learned models make a lot of errors ("high-bias"), so model averaging is unlikely to help in generalization.

☒ The learned models are different ("high-variance") and do well on their training sets, so model averaging is likely to help in generalization.

✓ 1.00

The classifiers are fairly different, but overfitted to their training sets. Averaging them could lead to a more generalizable model.

Total

1.00 /  
1.00

## Question 4

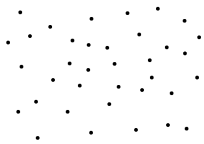
In Bayesian learning, we learn a probability distribution over parameters of the model. Then at test time, how should this distribution be used to get predictions with the highest possible accuracy?

Your Answer	Score	Explanation
<input type="radio"/> Sample the distribution once to get a parameter setting and use it to make a prediction.		
<input type="radio"/> Sample a lot of parameters using some sampling procedure (such as MCMC) and average the parameters. Then use the averaged parameter setting to obtain a prediction.		
<input type="radio"/> Pick the parameter setting that has maximum probability and use it to make a prediction.		
<input checked="" type="radio"/> Sample a lot of parameters using some sampling procedure (such as MCMC) and average the predictions obtained by using each parameter setting separately.	✓ 1.00	This method makes sure that we use a lot of models and also choose the models in proportion to how much we can trust them.
Total	1.00 / 1.00	

## Question 5

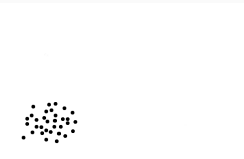
Amy is trying different MCMC samplers to sample from a probability distribution. Each option shows a few samples obtained by running a sampler. It is known that the distribution is multimodal and peaked. Which among the following is the best sampler ?

Your Answer	Score	Explanation
<input type="radio"/>		



1.00

This sampler can access multiple modes and does not give a lot of samples from low probability regions. This is what a good sampler is expected to do.



Total

1.00 /  
1.00

## Question 6

Brian wants to learn a classifier to predict if a movie is "good" or "bad" given its review. He has a lot of computational power and wants to use a very powerful model. However, he only has a small dataset of labelled movie reviews. He tried learning a massive neural net and found that it achieves zero training error very easily, but its test performance is much worse. He then trained a small neural net and found that it does not get zero training error, but still the test performance is no better than what the big model got. Neither a big nor a small model works for him! He is completely disappointed with neural nets now. He is willing to spend as much computational power as needed during training and testing. What suggestion can you give to help him?

### Your Answer

### Score Explanation

☐ Look for a better optimization algorithm to help the large neural net.



0.50

The big neural net already fits the training data perfectly. Having a better optimization algorithm will only make it overfit faster!

☒ Train the big neural net with dropout in the hidden units.



0.50

Adding drop out helps prevent overfitting in large nets.

<input type="checkbox"/> Train lots of small neural nets of the same architecture on the whole data and average their predictions.	✓	0.50	If the models are very alike and high-bias, averaging them is unlikely to benefit.
<input checked="" type="checkbox"/> Train many different models - neural nets, SVMs, decision trees - and average their predictions.	✓	0.50	Model averaging with different kinds of models is useful because each model is likely to be different and make different errors.
Total		2.00 / 2.00	