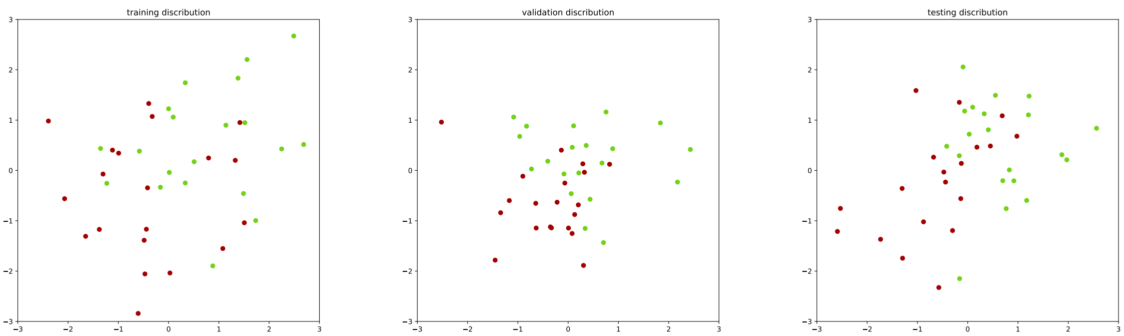
Exercise 5: More Overfitting

Name: .... UTID:................

This far we talked about overfitting to a training set. In this exercise we'll see how you can overfit to a validation set too.

Consider a binary classificaiton task and a dataset that is sampled from a two dimensional normal distribution. We sample positive samples from a $x, y \sim \mathcal{N}(0.3, 1)$, negative samples from $x, y \sim \mathcal{N}(-0.3, 1)$. The training, validation and testing datasets are listed below:



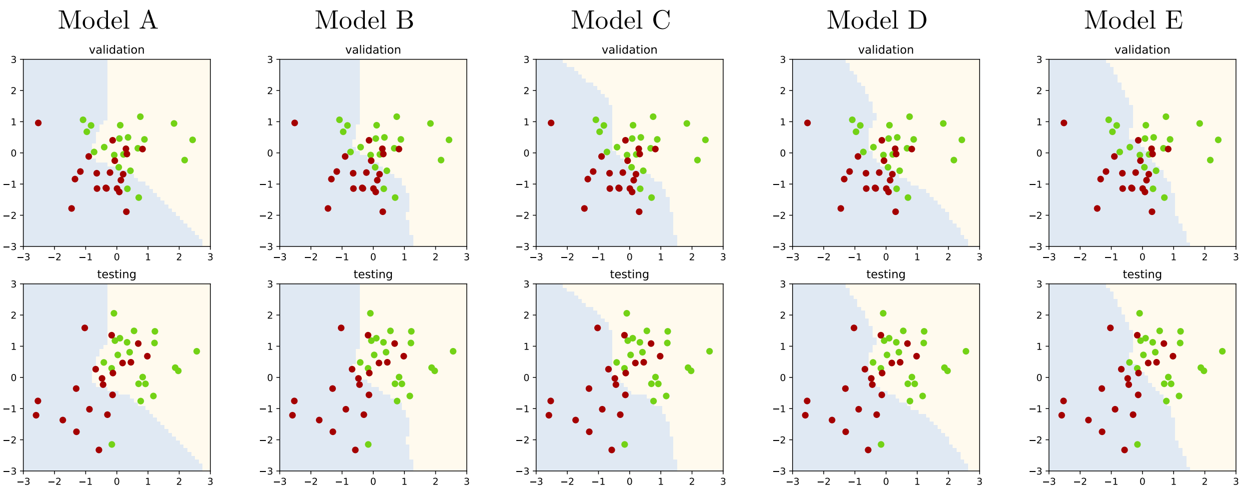
1) The training, validation and testing distributions look quite different.

a) Explain why?

b) How would you address this?

c) How would this scale if you used higher dimensional input data (e.g. images). Would the distribution look more even, or skewed? Explain your answer.

You now go and train a fully connected network (two fully connected layers and a ReLU) to fit this data. However, instead of just fitting one model you dedicde to train 5 networks with different random initalizations, but the same hyper parameters. Their decision boundary is shown below. For each model we overlay the validation and testing samples (these are real models).



The training and validation accuracies are given below

	Model A	Model B	Model C	Model D	Model E
training accuracy	0.775	0.750	0.775	0.775	0.700
validation accuracy	0.750	0.700	0.725	0.725	0.650

a) Which model do you select? What do you expects its test performance to be? What is the variance of that estimate?

b) You go and select the best performing model A. However it does not perform best during testing. Look at the decision boundaries of the models above. Can you think of reasons why?