Apollo Solutions Machine Learning Developer Test
Interpretation

Lilian Iazzai de Souza Oliveira

1 - In Figure 1 we have a data distribution, the dots represent the sparse data for the axis X and Y, and the lines represent the fit of a hypothetical classification model. Based on the distributions of Figure 1:

- Which distribution has the best balance between bias and variance?
- Describe your thoughts about your selection.

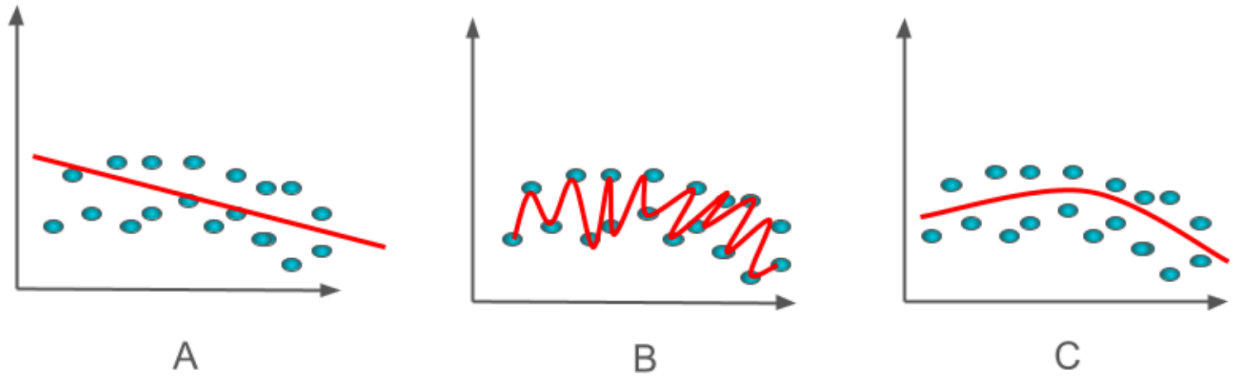


Figure 1 - Data distribution samples

The C Distribution is more balanced. It adjusts well to the training data, and also works well with unseen data.

In A, the model has high bias, which tends to make strong assumptions about the structure of the data, limiting their ability to capture the complexity of the problem. This often leads to systematic errors, or underfitting, where the model does not fit the data well.

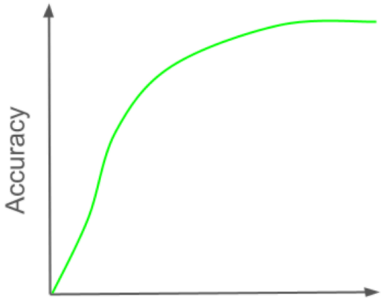
In B, the model has high variance, which makes it highly adaptive to different training data sets, possibly causing overfitting, where the model adapts so specifically to the training set that it loses generalization to new data. In this case, the model pays too much attention to the particular dataset it was trained on, causing it to perform poorly on unseen data.

2 - Figure 2 presents a simple graph with 2 curves and 1 line. In model selection and evaluation:
● What is the purpose of this graph and its name?
● What kind of model result does the dashed line represent?
● Which curve represents a better fit, the red or the green? Why?
● Describe your thoughts about your selection.

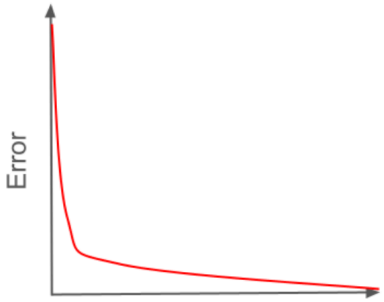
- The graphic represents a ROC (Receiver Operating Characteristic) curve. This curve relates the False Positive Rate (FPR) on the X-axis to the True Positive Rate (TPR) on the Y-axis. The curve illustrates the tradeoff between accurately predicting positive classes (true positives) and incorrectly predicting negative classes (false positives). A higher Area Under the Curve (AUC) indicates better model performance, with 1 representing perfect accuracy and 0.5 indicating random chance.
- Along the dashed line, AUC is 0.5, which means that the model would not be able to distinguish very well which record belongs to which class, it would do this randomly.
- The green curve represents a better fit, as the more the curve is above the dashed line (which is the random classifier), the more the AUC tends towards 1, improving the model's ability to distinguish classes. Red's AUC tends to 0, causing the model to lose almost all classifications.

3 - Figure 3 presents a classification model training and the evaluation. This model classifies 3 classes (A, B, C). Graph A represents the training accuracy over the epochs, Graph B represents the training loss over the epochs, and the table represents the evaluation of the model using some test samples, we used a confusion matrix to evaluate the classes trained.

- Can we say that the model has a good performance in the test evaluation?
- What phenomenon happened during the test evaluation?
- Describe your thoughts about your selection.



A



B

	A	B	C
A	0.5	0.25	0.25
B	0.15	0.45	0.35
C	0.1	0.4	0.5

Figure 3 - Model train and evaluation pipeline

Accuracy = (0.5 + 0.45 + 0.5) / 3 = 0.483
Precision for A = 0.5 / (0.5 + 0.15 + 0.1) = 0.66
Precision for B = 0.45 / (0.45 + 0.25 + 0.4) = 0.40
Precision for C = 0.5 / (0.5 + 0.25 + 0.35) = 0.45
Specificity for A = (0.45 + 0.35 + 0.4 + 0.5) / (0.15 + 0.1 + 0.45 + 0.35 + 0.4 + 0.5) = 0.87
Specificity for B = (0.5 + 0.25 + 0.1 + 0.5) / (0.25 + 0.4 + 0.5 + 0.25 + 0.1 + 0.5) = 0.84
Specificity for C = (0.5 + 0.25 + 0.15 + 0.45) / (0.25 + 0.35 + 0.5 + 0.25 + 0.15 + 0.45) = 0.69

This is not a particularly good model. This model has low accuracy (48.3%) and moderate to low precision for most classes, indicating it struggles to make correct predictions consistently. The specificity is relatively high for A and B but lower for C, suggesting that it can correctly identify negatives better than positives.
The confusion matrix reveals that the model struggles with distinguishing some classes (like B and C), which could result from limited data for those classes or suboptimal decision boundaries. The phenomenon might be class imbalance or unequal learning across classes.
But if the accuracy reaches 1 in the training model, the model is probably overfitting