### **Interpretation**

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?

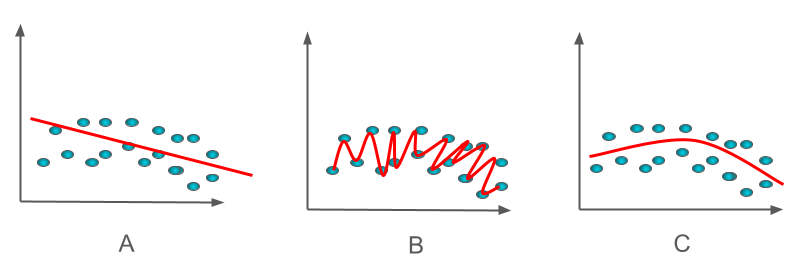
*Answer*: The distribution that presents the best balance between bias and variance is Figure C.

* Describe your thoughts about your selection.

*Answer*: Figure C shows a closer relationship between the data points and the model’s line, capturing the overall trend without being overly simplistic (as in Figure A) or overly complex (as in Figure B). In Figure A, the line is too simple and fails to capture the relationship between the data, indicating high bias. On the other hand, Figure B shows a line that is excessively fitted, following almost every data point perfectly, which indicates high variance and can lead to overfitting.

Figure C, however, balances these factors by capturing the general trend of the data without being overly complex, reducing the risk of both underfitting and overfitting. This balance makes the model in Figure C more robust and effective for generalizing to new observations, which is a fundamental goal in predictive modeling.

Therefore, selecting Figure C reflects a good trade-off between bias and variance, optimizing the model’s performance in real-world scenarios.



**Figure 1 - Data distribution samples**

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?

*Answer*: The graph is known as the ROC Curve (Receiver Operating Characteristic), and it is used to evaluate the performance of binary classification models. It shows the relationship between the true positive rate (on the Y-axis) and the false positive rate (on the X-axis) for different decision thresholds of the model. The goal is to analyze the model’s ability to distinguish between positive and negative classes, with a performance metric called the area under the ROC curve (AUC).

* What kind of model result does the dashed line represent?

*Answer*: The dashed line represents the performance of a random classifier, meaning a model that makes predictions completely randomly, with no real ability to distinguish between the classes. This line indicates that for every increase in the true positive rate, there is a proportional increase in the false positive rate. The AUC value for this line is 0.5, which is considered the minimum acceptable performance for a model.

* Which curve represents a better fit, the red or the green? Why?

*Answer*: The green curve represents a better fit compared to the red curve. This is because it is closer to the top-left corner of the graph, which indicates a high true positive rate and a low false positive rate. An ideal model’s performance would have a curve that directly reaches the top-left corner, indicating perfect separation between the classes.

* Describe your thoughts about your selection.

*Answer*: The green curve reflects a more effective model for classification because it maximizes the true positive rate while minimizing the false positive rate. This results in a higher area under the ROC curve (AUC), which is indicative of better overall performance. Conversely, the red curve is closer to the dashed line, indicating that the model has little capacity to distinguish between the classes.

Selecting a model with an ROC curve closer to the ideal (like the green curve) is critical in scenarios where minimizing false positives or maximizing true positives is essential to the system’s success. 

**Figure 2 - Simple graph**

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?

*Answer*: We cannot say that the model demonstrated good performance in the test evaluation. Although the accuracy graph shows a significant increase over epochs (Graph A) and the training error decreases (Graph B), the confusion matrix indicates that the model struggles to distinguish between classes, with low to moderate accuracy when classifying classes A, B, and C. For example:

* Only 50% of the samples for class A were correctly classified as A.
* For class B, 45% were classified correctly as B, but 15% were incorrectly classified as A and 35% as C.
* For class C, only 50% of the samples were classified correctly, while 40% were mistakenly assigned to class B.

These results indicate that the model did not generalize well and that its performance on the test data is unsatisfactory.

* What phenomenon happened during the test evaluation?

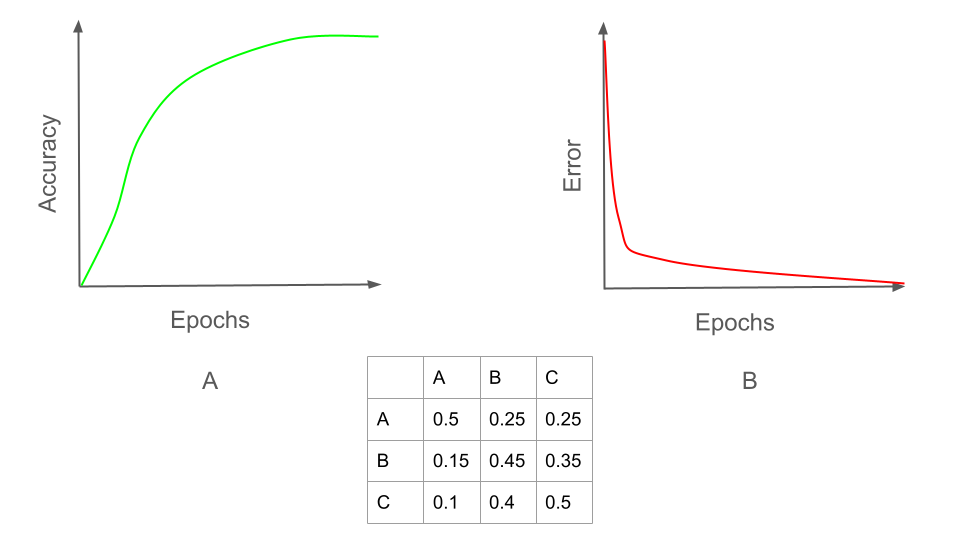
*Answer*: The phenomenon observed is known as overfitting. During training, the model achieved high accuracy and low error on the training set (as indicated by the graphs), but it performed poorly when evaluated on the test set. This indicates that the model memorized patterns from the training set rather than learning to generalize to new data. Overfitting may have occurred due to an excessive number of training epochs or a lack of regularization in the model.

* Describe your thoughts about your selection.

*Answer*: The model demonstrates good fitting to the training data, as evidenced by the increasing accuracy and decreasing error on the training set. However, the results in the confusion matrix reveal significant generalization issues. The low accuracy across several classes and frequent misclassification between the classes indicate that the model failed to capture the overall patterns of the problem.

To mitigate overfitting and improve performance on the test set, several strategies can be considered, such as:

* Using regularization techniques like L1/L2 regularization or dropout.
* Implementing early stopping to halt training before the model begins to excessively memorize the training data.
* Expanding the training dataset with data augmentation techniques or collecting additional data.
* Adjusting hyperparameters, such as the learning rate, to avoid overfitting.



**Figure 3 - Model train and evaluation pipeline**