Data Split Example

After collecting your data and sampling where needed, the next step is to split your data into **training sets** (/machine-learning/crash-course/glossary#training_set), **validation sets** (/machine-learning/crash-course/glossary#validation_set), and **testing sets** (/machine-learning/crash-course/glossary#test_set).

When Random Splitting isn't the Best Approach

While random splitting is the best approach for many ML problems, it isn't always the right solution. For example, consider data sets in which the examples are naturally clustered into similar examples.

Suppose you want your model to classify the topic from the text of a news article. Why would a random split be problematic?

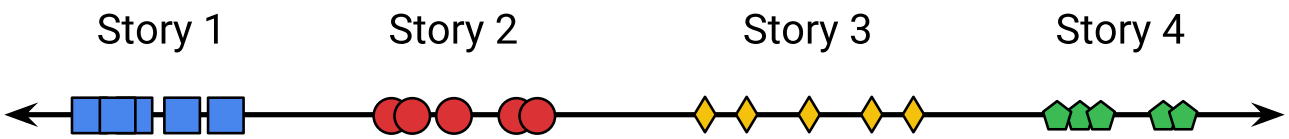

Figure 1. News Stories are Clustered.

News stories appear in clusters: multiple stories about the same topic are published around the same time. If we split the data randomly, therefore, the test set and the training set will likely contain the same stories. In reality, it wouldn't work this way because all the stories will come in at the same time, so doing the split like this would cause skew.

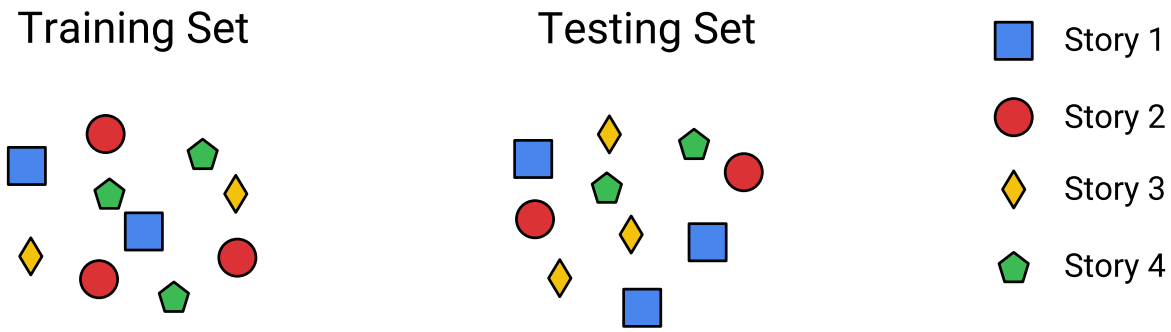

Figure 2. A random split will split a cluster across sets, causing skew.

A simple approach to fixing this problem would be to split our data based on when the story was published, perhaps by day the story was published. This results in stories from the

same day being placed in the same split.

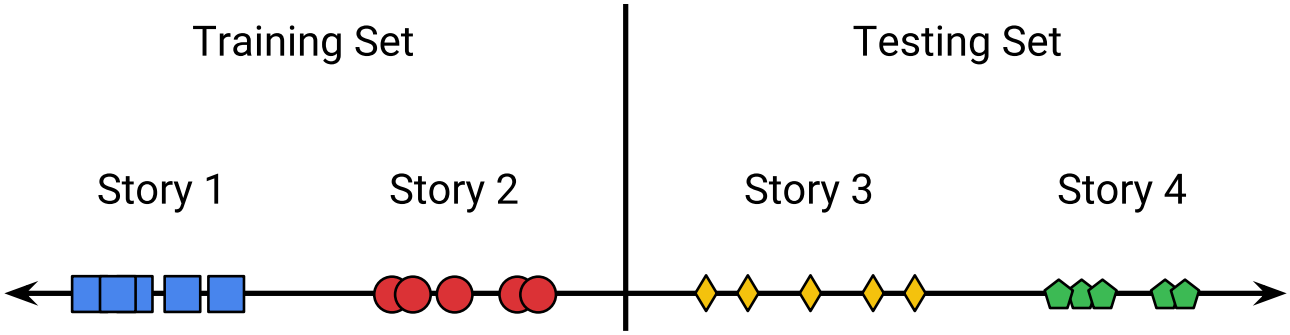

Figure 3. Splitting on time allows the clusters to mostly end up in the same set.

With tens of thousands or more news stories, a percentage may get divided across the days. That's okay, though; in reality these stories were split across two days of the news cycle. Alternatively, you could throw out data within a certain distance of your cutoff to ensure you don't have any overlap. For example, you could train on stories for the month of April, and then use the second week of May as the test set, with the week gap preventing overlap.

Previous
← Imbalanced Data (/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data)

Next
Splitting Your Data (/machine-learning/data-prep/construct/sampling-splitting/split) →