## **KL- Divergence and Cross Entropy**

How we deal with true and predicted distributions

1. Consider the following data:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **X** | **True Distribution: y** | **True IC(X)** | **Predicted Distribution: y** | **Predicted IC(X)** |
| A | y1 | -logy1 | ŷ1 | -logŷ1 |
| B | y2 | -logy2 | ŷ2 | -logŷ2 |
| C | y3 | -logy3 | ŷ3 | -logŷ3 |
| D | y4 | -logy4 | ŷ4 | -logŷ4 |

1. Initially, we do not know the values of the True distribution and thereby the True Information Content
2. Hence, we generate a Predicted distribution and use that to compute the predicted information content.
3. But, the actual message will come from the True distribution y.
4. So therefore, the No. of bits will **not be** but **instead**
5. This is because the value associated with each of these messages comes from the predicted distribution but the messages themselves comes from the True distribution
6. Now, we have formed the basis to talk about KL-Divergence:
   1. is called the entropy
   2. is called the cross entropy
   3. Now we want to find the difference/distance between the predicted case and the true case, using something more efficient than the squared error
   4. So y||ŷ =
   5. y||ŷ =
   6. This is called the KL-Divergence
7. Thus, we now have **KLD(y||ŷ) =**
8. Now, we have a way of computing the difference between the two distributions.