

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: \hat{y}	Predicted IC(X)
A	y_1	$-\log y_1$	\hat{y}_1	$-\log \hat{y}_1$
B	y_2	$-\log y_2$	\hat{y}_2	$-\log \hat{y}_2$
C	y_3	$-\log y_3$	\hat{y}_3	$-\log \hat{y}_3$
D	y_4	$-\log y_4$	\hat{y}_4	$-\log \hat{y}_4$

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