Elements of information theory

A machine learning problem can also be analyzed in terms of information transfer or exchange. Our dataset is composed of **n** features, which are considered independent (for simplicity, even if it's often a realistic assumption) drawn from **n** different statistical distributions. Therefore, there are **n** probability density functions $p_i(x)$ which must be approximated through other $nq_i(x)$ functions. In any machine learning task, it's very important to understand how two corresponding distributions diverge and what is the amount of information we lose when approximating the original dataset.

The most useful measure is called **entropy**:

$$H(X) = -\sum_{x \in X} p(x) log_2 p(x)$$

This value is proportional to the uncertainty of **X** and it's measured in **bits** (if the logarithm has another base, this unit can change too). For many purposes, a high entropy is preferable, because it means that a certain feature contains more information. For example, in tossing a coin (two possible outcomes), **H(X)** = 1 bit, but if the number...

Previous Section (/book/big_data_and_business_intelligence/9781785889622/2/ch02lvl1s

Next Section (/book/big_data_and_business_intelligence/9781785889622/2/ch02lvl1sec1