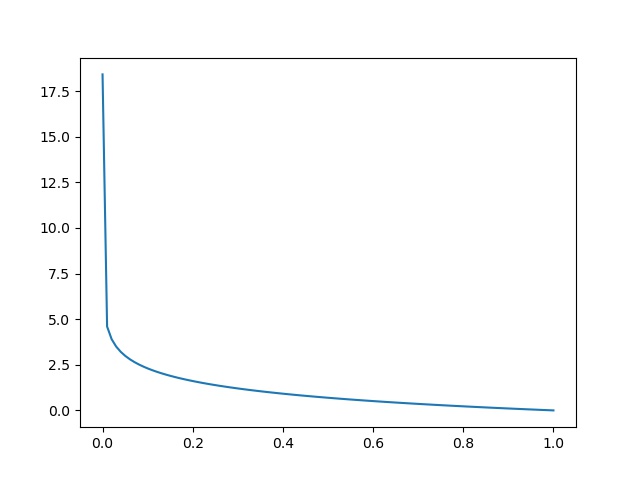
信息量

信息量用于定量某一事件中所包含信息的大小，由事件发生的概率计算而来。其计算公式如下所示：

函数图像如图下所示：



从图中可知，信息量与事件发生的概率成反比。不难理解，这是因为常见则不疑。如果某一件不可能或极少可能的事件发生了，那么相比于频繁发生的事件来说，前者的信息更大。

熵，其定义了一个事件组的信息量的均值。其计算公式如下所示：

同时，熵也可以表示一个体系的混乱程度。熵越大，则说明体系越混乱。熵越小，则说明体系越稳定，则该体系的概率分布呈驼峰形状。

KL divergence，用于计算两个分布的相似程度，具体说是用某一个分布Q来模仿分布P的近似程度。它不同于一般的距离，因为它满足距离计算的对称性和直递性（三角形性质）。其计算公式如下所示：

在分类的过程中，因为其目标概率分布是确定，却为one-hot形式，所以其熵为0。