2.4.8 Kullback-Leibler Divergence

To measure the difference between two probability distributions over the same variable x, a measure, called the *Kullback-Leibler divergence*, or simply, the *KL divergence*, has been popularly used in the data mining literature. The concept was originated in probability theory and information theory.

The KL divergence, which is closely related to *relative entropy*, *information divergence*, and *information for discrimination*, is a non-symmetric measure of the difference between two probability distributions $p(x)$ and $q(x)$. Specifically, the Kullback-Leibler (KL) divergence of $q(x)$ from $p(x)$, denoted $D_{KL}(p(x), q(x))$, is a measure of the information lost when $q(x)$ is used to approximate $p(x)$.

Let $p(x)$ and $q(x)$ are two probability distributions of a discrete random variable x. That is, both $p(x)$ and $q(x)$ sum up to 1, and $p(x) > 0$ and $q(x) > 0$ for any x in X. $D_{KL}(p(x), q(x))$ is defined in Equation (2.1).

divergence of q(x) from p(x).

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)} \tag{2.1}$$

The KL divergence measures the expected number of extra bits required to code samples from $p(x)$ when using a code based on $q(x)$, rather than using a code based on $p(x)$. Typically $p(x)$ represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution. The measure $q(x)$ typically represents a theory, model, description, or approximation of $p(x)$.

The continuous version of the KL divergence is

$$D_{KL}(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx \tag{2.2}$$

Although the KL divergence measures the "distance" between two distributions, it is not a distance measure. This is because that the KL divergence is not a metric measure. It is not symmetric: the KL from $p(x)$ to $q(x)$ is generally not the same as the KL from $q(x)$ to $p(x)$. Furthermore, it need not satisfy triangular inequality. Nevertheless, $D_{KL}(P||Q)$ is a non-negative measure. $D_{KL}(P||Q) \geq 0$ and $D_{KL}(P||Q) = 0$ if and only if $P = Q$.

Notice that attention should be paid when computing the KL divergence. We know $lim_{p \to 0}\ p \log p = 0$. However, when $p \neq 0$ but $q = 0$, $D_{KL}(p||q)$ is defined as ∞. This means that if one event e is possible (*i.e.*, $p(e) > 0$), and the other predicts it is absolutely impossible (*i.e.*, $q(e) = 0$), then the two distributions are absolutely different. However, in practice, two distributions P and Q are derived from observations and sample counting, that is, from frequency distributions. It is unreasonable to predict in the derived probability distribution that an event is completely impossible since we must take into account the possibility of unseen events. A *smoothing* method can be used to derive the probability distribution from an observed frequency distribution, as illustrate in the following example.

Example 2.24. Computing the KL Divergence by Smoothing. Suppose there are two sample distributions P and Q as follows: $P : (a : 3/5, b :$

$1/5, c : 1/5)$ and $Q : (a : 5/9, b : 3/9, d : 1/9)$. To compute the KL divergence $D_{KL}(P||Q)$, we introduce a small constant ϵ, for example $\epsilon = 10^{-3}$, and define a smoothed version of P and Q, P' and Q', as follows.

The sample set observed in P, $SP = \{a, b, c\}$. Similarly, $SQ = \{a, b, d\}$. The union set is $SU = \{a, b, c, d\}$. By smoothing, the missing symbols can be added to each distribution accordingly, with the small probability ϵ. Thus, we have $P' : (a : 3/5 - \epsilon/3, b : 1/5 - \epsilon/3, c : 1/5 - \epsilon/3, d : \epsilon)$ and $Q' : (a : 5/9 - \epsilon/3, b : 3/9 - \epsilon/3, c : \epsilon, d : 1/9 - \epsilon/3)$. $D_{KL}(P', Q')$ can be computed easily.