Homework 5

Posted: May 1, 2019
Due: May 15, 2019

1. Write an **R** function that starts with a contingency table for two partitions and generates a vector containing the true positive, the false positive, the true negative, and the false negative counts. Use this function to compute the F-index of the partitions.
2. Starting from the `iris` database (limited to its first four columns, because you need to drop the `species` attribute) apply the k-means algorithm for $k = 2, 3$ and $k = 4$. Which algorithm produces a clustering that has the highest F-value? Why?
3. Let $\pi = \{B_1, \ldots, B_m\}$ be a partition of a set S with $|S| = n$ and let $|B_i| = b_i$ for $1 \leqslant i \leqslant m$. Write and test an **R** function that computes the β-entropy:

$$\mathcal{H}_\beta = \frac{1}{1 - 2^{1-\beta}} \left(1 - \sum_{i=1}^{k} \left(\frac{|B_i|}{|S|} \right)^\beta \right).$$

   The argument of the function should be the vector $\mathbf{v} = (b_1, \ldots, b_n)$. Test the function for several values of β (e.g., 1.01, 2, and 3).
4. Let $\mathbf{p} = (p_1, \ldots, p_m)$ and $\mathbf{q} = (q_1, \ldots, q_m)$ be two probability distributions with $p_i \neq 0$ and $q_i \neq 0$ for $1 \leqslant i \leqslant m$.
   (a) Prove that for $x > 0$ we have $\ln x \geqslant 1 - \frac{1}{x}$.
   (b) Define the Kullback-Leibler divergence of $\mathbf{p}$ and $\mathbf{q}$ as

$$\mathrm{KL}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}.$$

   Prove that $\mathrm{KL}(\mathbf{p}, \mathbf{q}) \geqslant 0$. What is the relationship between $\mathbf{p}$ and $\mathbf{q}$ when $\mathrm{KL}(\mathbf{p}, \mathbf{q}) = 0$?