Exercises
Advanced Machine Learning
Fall 2019

Machine Learning Laboratory
Dept. of Computer Science, ETH Zürich
Prof. Joachim M. Buhmann
Web https://ml2.inf.ethz.ch/courses/aml/

Series 7 10 Dec 2019 (Nonparametric Bayesian methods and Clustering)

Teaching assistant: **Mikhail Karasikov**
mikhaika@inf.ethz.ch

Problem 1 (Cluster quality evaluation):

Suppose we are given a data set $X = \{x_1, \ldots, x_N\}$ and two clusterings $\mathcal{U} = \{U_1, \ldots, U_R\}$ and $\mathcal{V} = \{V_1, \ldots, V_C\}$, where $\mathcal{U}$ is a reference clustering and $\mathcal{V}$ is computed by our favorite clustering algorithm. We want to evaluate how well our algorithm performs relative to the reference. For simplicity, assume that $\mathcal{U}$ and $\mathcal{V}$ are disjoint partitions of $X = \sqcup_i U_i = \sqcup_j V_j$, that is,

$$\bigcup_i U_i = X \quad \text{and} \quad U_k \cap U_l = \emptyset \ \forall k \neq l.$$

Now consider the following measures for evaluation of clusters.

1. One simple measure is cluster *purity*, which is defined as

$$\text{purity} = \frac{1}{|X|} \sum_{V \in \mathcal{V}} \max_{U \in \mathcal{U}} |U \cap V|.$$

   What is the maximal value of this measure? Provide a degenerate pair of clusterings of X that achieve this value.

2. Suppose we choose to use *mutual information* as a measure, where probabilities are defined as follows:

$$p_U(i) = \frac{|U_i|}{|X|}, \qquad p_V(j) = \frac{|V_j|}{|X|}, \qquad p_{UV}(i,j) = \frac{|U_i \cap V_j|}{|X|}.$$

   The mutual information between $\mathcal{U}$ and $\mathcal{V}$ is then defined as

$$I(\mathcal{U}, \mathcal{V}) := \sum_{i=1}^{R} \sum_{j=1}^{C} p_{UV}(i,j) \log_2 \frac{p_{UV}(i,j)}{p_U(i) p_V(j)}.$$

   Recall that the entropy of $\mathcal{U}$ is defined as

$$H(\mathcal{U}) = -\sum_{i=1}^{R} p_U(i) \log_2 p_U(i).$$

   Show that

$$0 \leq I(\mathcal{U}, \mathcal{V}) \leq \min(H(\mathcal{U}), H(\mathcal{V})).$$

3. How do the clusterings that maximize $I(\mathcal{U}, \mathcal{V})$ compare to those that maximize purity? How does this relate to the entropy terms in the upper bound? How can mutual information be modified to account for this?

Problem 2 (Dirichlet process):

Consider the following algorithm for sampling from the Dirichlet process with base distribution F_0 and concentration parameter α.

1. Draw the first sample $X_1 \sim F_0$.
2. For $i = 2, 3, \ldots$, draw

$$X_i | X_1, \ldots, X_{i-1} = \begin{cases} X \sim \hat{F}_{i-1}, & \text{with probability } p = \frac{i-1}{\alpha+i-1}, \\ X \sim F_0, & \text{with probability } p = \frac{\alpha}{\alpha+i-1}, \end{cases}$$

   where $\hat{F}_{i-1}$ is the empirical distribution of $X_1, \ldots, X_{i-1}$.

Find the asymptotics of the expected number of distinct samples drawn, as a function of the total number of samples drawn: $X_1, \ldots, X_n$. Or equivalently, the number of occupied tables in the Chinese restaurant process metaphor.