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/

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

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

Solution 1 (Cluster quality evaluation):

1. First, let us find an upper bound on the clustering purity.

$$\begin{aligned}\text{purity} &= \frac{1}{|X|}\sum_{V\in\mathcal{V}}\max_{U\in\mathcal{U}}|U\cap V| \\ &= \frac{1}{N}\sum_{j=1}^{C}\max_{i=1,\dots,R}|U_i\cap V_j| \le \frac{1}{N}\sum_{j=1}^{C}\underbrace{\sum_{i=1}^{R}|U_i\cap V_j|}_{|V_j|} = 1.\end{aligned}$$

Now, let us consider a pair of identical partitions of X into singleton sets: $\mathcal{U} = \mathcal{V} = \{\{x_1\},\dots\{x_N\}\}$ where $R = C = N$. Then

$$|U_i\cap V_j| = 1\{i=j\} \equiv \begin{cases}1 & \text{if } i=j\\ 0 & \text{otherwise}\end{cases}$$

and

$$\max_{i=1,\dots,R}|U_i\cap V_j| = 1.$$

Thus

$$\text{purity} = \frac{1}{N}\sum_{j=1}^{N}\max_{i=1,\dots,N}|U_i\cap V_j| = \frac{1}{N}\sum_{j=1}^{N}1 = 1$$

2. Now, we define probability distributions on clusters

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

and use the *mutual information* as an evaluation measure:

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

Now, we will show that

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

where the entropy of clusters in $\mathcal{U}$ is defined as $H(\mathcal{U}) = -\sum_{i=1}^{R}p_U(i)\log_2 p_U(i)$.

$$\begin{aligned}
I(\mathcal{U},\mathcal{V}) &= -\sum_{i=1}^{R}\sum_{j=1}^{C} p_{UV}(i,j)\log_2 \frac{p_U(i)p_V(j)}{p_{UV}(i,j)} \\
&\geq -\log_2 \left(\sum_{i=1}^{R}\sum_{j=1}^{C} p_{UV}(i,j)\frac{p_U(i)p_V(j)}{p_{UV}(i,j)}\right) && \text{(Jensen's inequality)} \\
&= -\log_2 \left(\sum_{i=1}^{R}\sum_{j=1}^{C} p_U(i)p_V(j)\right) \\
&= -\log_2(1) \\
&= 0.
\end{aligned}$$

$$\begin{aligned}
I(\mathcal{U},\mathcal{V}) &= \sum_{i=1}^{R}\sum_{j=1}^{C} p_{UV}(i,j)\log_2 \frac{p_{UV}(i,j)}{p_U(i)p_V(j)} \\
&= \sum_{i=1}^{R}\sum_{j=1}^{C} p_{UV}(i,j)\frac{p_V(j)}{p_V(j)}\log_2 \frac{p_{UV}(i,j)}{p_U(i)p_V(j)} \\
&= \sum_{i=1}^{R}\sum_{j=1}^{C} p_V(j)\frac{p_{UV}(i,j)}{p_V(j)}\left(\log_2 \frac{p_{UV}(i,j)}{p_V(j)} - \log_2 p_U(i)\right) \\
&= \sum_{j=1}^{C} p_V(j)\sum_{i=1}^{R}\frac{p_{UV}(i,j)}{p_V(j)}\left(\log_2 \frac{p_{UV}(i,j)}{p_V(j)} - \log_2 p_U(i)\right) \\
&= \sum_{j=1}^{C} p_V(j)\left(-H(\mathcal{U}\mid\mathcal{V}) - \sum_{i=1}^{R}\frac{p_{UV}(i,j)}{p_V(j)}\log_2 p_U(i)\right) \\
&= -H(\mathcal{U}\mid\mathcal{V}) - \sum_{j=1}^{C}\sum_{i=1}^{R} p_{UV}(i,j)\log_2 p_U(i) \\
&= -H(\mathcal{U}\mid\mathcal{V}) - \sum_{i=1}^{R} p_U(i)\log_2 p_U(i) \\
&= H(\mathcal{U}) - H(\mathcal{U}\mid\mathcal{V})
\end{aligned}$$

Similarly, $0 \leq I(\mathcal{U},\mathcal{V}) = H(\mathcal{V}) - H(\mathcal{V}\mid\mathcal{U})$. Then since $H(\mathcal{U}) \geq H(\mathcal{U}\mid\mathcal{V})$ and $H(\mathcal{V}) \geq H(\mathcal{V}\mid\mathcal{U})$, the result follows.

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 in terms of the upper bound? How can mutual information be modified to account for this?

   Suppose again that $\mathcal{U} = \mathcal{V} = \{\{x_1\},\dots\{x_N\}\}$. Then $I(\mathcal{U},\mathcal{V}) = H(\mathcal{U}) = H(\mathcal{V}) = \log_2 N$. Since this is the maximal entropy value for a discrete distribution over support X, this is also the maximum value for I.

   These measures tend to prefer clusterings composed of many small clusters. One modification to correct for this is the *normalized mutual information*

$$\frac{2I(\mathcal{U},\mathcal{V})}{H(\mathcal{U})+H(\mathcal{V})}.$$

Solution 2 (Dirichlet process):

For any continuous base distribution F_0, the probability of drawing an X_k matching exactly one of the finite number of samples already drawn $(X_1, \dots, X_{k-1})$ is zero. Thus, the expected number of distinct samples $S(n)$ among $X_1, \dots, X_n$ is equal to the expected number of times we draw the next X_k from the base distribution. That is,

$$S(n) = E\left[\sum_{k=1}^{n} 1\{X_k \text{ is drawn from } F_0\}\right] = \sum_{k=1}^{n} P(X_k \text{ is drawn from } F_0) = \sum_{k=1}^{n} \frac{\alpha}{\alpha + k - 1}.$$

To find the asymptotics of $S(n)$, let us consider the integral $I(n) = \int_1^{n+1} \frac{\alpha}{\alpha+x-1} dx$. After looking at the graph of function $\frac{\alpha}{\alpha+x-1}$, it is easy to see that the integral $I(n)$ is bounded by the sum $\sum_{k=1}^{n} \frac{\alpha}{\alpha+k-1} = S(n)$ from above and by $\sum_{k=2}^{n+1} \frac{\alpha}{\alpha+k-1} = S(n) - 1 + \frac{\alpha}{\alpha+n}$ from below, that is,

$$S(n) < I(n) < S(n) - 1 + \frac{\alpha}{\alpha + n} \quad \forall n > 1,$$

which can be reformulated equivalently as

$$I(n) + 1 - \frac{\alpha}{\alpha + n} < S(n) < I(n) \quad \forall n > 1.$$

After computing the value of the integral

$$I(n) = \int_1^{n+1} \frac{\alpha}{\alpha + x - 1} dx = \alpha(\ln(\alpha + n) - \ln(\alpha)),$$

we see that both bounds $I(n) + 1 - \frac{\alpha}{\alpha+n}$ and $I(n)$ on $S(n)$ are asymptotically equivalent to $\alpha \ln(n)$ as $n \to \infty$, which concludes the solution:

$$S(n) \sim \alpha \ln(n) \quad \text{as} \quad n \to \infty.$$