Machine Learning in Economics and Finance 1

Homework 3

August 8, 2020

- This homework assignment covers Lecture 3 – "Clustering"
- This homework is due 11 PM, Sunday, 16 August, 2020. The Google Form for submission will be sent out later.

Problem 1. (Adapted from Cover & Thomas and Yurdakul) *(60 points)*
If $p(x)$ and $q(x)$ are two probability mass functions (pmf's), the relative entropy or Kullback-Leibler divergence (or KL divergence), between $p(x)$ and $q(x)$, is defined to be

$$D(p||q) = \sum_{x \in X} p(x) log \frac{p(x)}{q(x)} = E_p log \frac{p(x)}{q(x)} \tag{1}$$

where we can have different bases for the log function (the most popular bases being e and 2), and use the convention that $0 \ ln\frac{0}{q} = 0$ and $p \ ln\frac{p}{0} = \infty$. X is the set of all possible values of x. The KL divergence is not considered to be a true distance between distributions as it is not symmetric and does not satisfy the triangle inequality.

(a) Prove that the KL divergence defined in Equation (1) is always non-negative and is zero if and only if $p = q$.

Let $X = \{0, 1\}$ and consider two Bernoulli distributions p and q on X. Let $p(0) = 1 - r$, $p(1) = r$, $q(0) = 1 - s$, $q(1) = s$.

(b) Derive $D(p||q)$ and $D(q||p)$ as functions of r and s.

(c) Verify that if r=s then $D(p||q) = D(q||p) = 0$.

(d) If $r = 1/2$ and $s = 1/4$, calculate $D(p||q)$ and $D(q||p)$ using base-2 logarithm in Equation (1).

(e) Consider three Bernoulli distributions $v = (0.5, 0.5)$, $w = (0.25, 0.75)$, and $u = (0.1, 0.9)$. Verify whether the KL divergence satisfies the triangle inequality with these three distributions.

In practice, when we have two populations $\hat{p}$ and $\hat{q}$, we normally group values of x into bins and write Equation (1) as

$$D(\hat{p}||\hat{q}) = \sum_{i=1}^{B} \hat{p}_i(x) log \frac{\hat{p}_i(x)}{\hat{q}_i(x)} \tag{2}$$

where B is the number of bins, and $\hat{p}_i$ and $\hat{q}_i$ are proportions of populations $\hat{p}$ and $\hat{q}$, respectively, in bin i. The *Population Stability Index (PSI)* is then defined as

$$PSI(\hat{p}, \hat{q}) = D(\hat{p}||\hat{q}) + D(\hat{q}||\hat{p}) \tag{3}$$

PSI is widely used to measure the difference between

- feature distributions of the training samples and samples being used for the model (the current samples) in a Machine Learning model;
- feature distributions between different points in time;

- outcome distributions.

In practice, there is a general rule of thumb: if PSI between the training samples and current samples is

- less than 10%, the model is considered appropriate;
- between 10% and 25%, we have to investigate the current samples to see why the PSI is so high;
- beyond 25%, we should retrain the model or develop a new model using more recent samples.

(e) Prove that

$$PSI(\hat{p}, \hat{q}) = \sum_{i=1}^{B} [\hat{p}_i(x) - \hat{q}_i(x)] [log(\hat{p}_i(x)) - log(\hat{q}_i(x))] \tag{4}$$

Problem 2 [Mini-project – K means clustering for the iris dataset]. *(20 points)*

In this problem we will use the iris dataset from *scikit-learn* library. More details on this dataset can be found in [2].

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

from sklearn.datasets import load_iris
iris = load_iris()
print("Features: ", iris.feature_names)
print("Labels: ", iris.target_names)
```

Using *scikit-learn's Kmeans* model to cluster the datapoints and calculate the cost function for *n_clusters* = 2, 3, 4, and 5. Which value of *n_clusters* yields the minumum value of the cost function?

Problem 3. [Mini-project – K means clustering for breast cancer diagnosis] *(20 points)*

In this problem you will logistic regression for a binary classification problem with the cancer dataset available in *scikit-learn* library. The image of a fine needle aspirate (FNA) of a breast mass is used to compute the 30 features in this dataset. For more details, see https://scikit-learn.org/stable/datasets/index.html#breast-cancer-dataset. There are two types of cancer classes: malignant (harmful) and benign (not harmful). The dataset can be loaded as follows

```python
# Import scikit-learn dataset library
from sklearn import datasets
# Load dataset
cancer = datasets.load_breast_cancer()
```

The feature names and label names can be printed as follows

```python
# print the names of the 30 features
print("Features: ", cancer.feature_names)
# print the label type of cancer('malignant' 'benign')
print("Label names: ", cancer.target_names)
```

The labels themselves can be printed as follows

```python
# print the labels
print("Labels:\n ", cancer.target)
```

The feature data and label data are in *cancer.data* and *cancer.target*.
(a) Run K-means clustering with K=2 on this dataset.
(b) As shown above, *cancer.target* stores the labels (0: 'malignant', 1: 'benign'). Calculate the accuracy, precision, recall, F1 score of your algorithm.

References

1. sklearn.datasets.load_iris https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html

2. Breast cancer wisconsin (diagnostic) dataset https://scikit-learn.org/stable/datasets/index.html#breast-cancer-dataset.

3. Thomas M. Cover and Joy A. Thomas. 2006. Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, New York, NY, USA.

4. Bilal Yurdakul, Statistical Properties of Population Stability Index (PSI), PhD Dissertation, Western Michigan University, 2018