Homework Assignment No. 7:

**HW No 7: Information Theory and Statistical Signficance**

submitted to:

Professor Joseph Picone

ECE 8527: Introduction to Pattern Recognition and Machine Learning

Temple University

College of Engineering

1947 North 12th Street

Philadelphia, Pennsylvania 19122

March 13th , 2022

prepared by:

Joe Arthur   
Email: joe.arthur@temple.edu

|  |  |  |  |
| --- | --- | --- | --- |
|  | Train | Dev | Randomly Generated |
| H(X1) | 4.6898 | 4.6889 | 3.6114 |
| X(X2) | 4.7121 | 4.731 | 3.5892 |
| H(X1,X2) | 9.1194 | 8.7503 | 8.2502 |
| H(X1|X2) | 4.4296 | 4.0614 | 3.9949 |
| H(X2|X1) | 4.4073 | 4.0193 | 3.9507 |
| I(X1:X2) | 0.2825 | 0.6696 | 0.2604 |

For this assignment we were asked to use both the train and dev component of dataset 8, to calculate the respective entropy, joint entropy, conditional entropy, and mutual information across each dataset. We were first asked to normalize our dataset using the formula:

This is a form of minmax normalization. The idea behind this is that we are scaling each datapoint by the range of its respective vector column. This compresses our range of values to 0 – 128. ‘Normalizing’ the data makes it easier to calculate probability density, which as we will see for our entropy calculation is important. Shannon entropy, which is used in information theory defines the entropy of a random variable as the average level of uncertainty or information which the random variable carries relative to its possible outcomes. It is defined as:

Chart, bubble chart

Description automatically generatedMap

Description automatically generatedIn order to calculate this is in python we can make use of the in built python function ‘entropy’ in the scipy.stats library. In order to run this function we need to provide a probability of each element occurring i.e. a probability distribution. Which means we need to know the frequency for which each value between 0 – 128 occurs. We can make use of the pandas ‘value.counts’ function to calculate this. An alternative way is to make use of numpy’s histogram function which I also used to verify my answers. However, before going further though it may be helpful to visualize both our datasets.

We can see that both datasets look very similar except that the ‘train’ dataset contains 30,000 elements and the ‘dev’ dataset contains 15,000 therefore the shape of the data is more certain for ‘train’. So having read, stored, visualized, and normalized the data I can now show (in the code below) how I calculated entropy.

Text

Description automatically generatedText

Description automatically generated

For joint entropy, it was simply a matter of creating a joint probability distribution of the 2 feature vectors for each dataset and then again using the built in function ‘entropy’ to calculate the joint entropy. I did this using the crosstab function and then flattened this to create a probability distribution.

Text

Description automatically generated

For both conditional entropy and mutual entropy we can use a few different statistical techniques to calculate them. If we know the joint entropy of 2 random variables (in this case H(X1, X2)) then if we also know the entropy of X1, we can use H(X1, X2) – H(X1) to describe the conditional entropy of X2 given X1. In other words:

H(X1, X2) – H(X1) = H(X2|X1)

By manipulating this formula and using the bayes rule we can calculate H(X1|X2). Finally for mutual information, we can say that this is the information which X1 and X2 share when the conditional information between the 2 variables is taken into account. Which we can write as:

I(X1;X2) = H(X1, X2) – H(X2|X1) – H(X1|X2)

Based on these formulas I calculated CE and MI and have added them to the table above.

Text

Description automatically generatedChart, scatter chart

Description automatically generatedFor the next part of the assignment was to repeat these calculations for and generated from statistically independent uniform random number generators. Below shows the code used to generate the data and a plot of the data.

After creating the data, it was a simple process of repeating the steps used above to calculate entropy, joint entropy, conditional entropy and mutual information. Comparing the values across the experimental datasets and the randomly generated data we can see that there are several differences. Firstly, both vector columns have lower entropy values. When we use some intuition along with comparing the plots this makes sense. The randomly generated dataset is distributed normally with the same mean and same variance for each class (0,1). Whilst the assignment of classes is random with each class being assigned 5000 elements each we can say that this will be normally distributed due to central limit theorem, therefore each class a should have ~ N(0,1) distribution also. Compared to our 2 dataset plots which we saw that whilst they may have the same mean value (0) each class has a different variance. Therefore, there is more uncertainty in both of our datasets due to the differences in variances between classes. This consequently impacts the probability distribution of each element, which results in the higher value of entropy. The joint entropy for all values should be less than the combined value of H(X1) and H(X2) which they are. With regards to mutual information, if our goal is to minimize uncertainty in our prediction of class assignments then we want to minimize conditional entropy, which in turn means we want to maximize mutual information. We can see that the train dataset has almost the same mutual information as randomly generated dataset. This is because despite each class having different variance in ‘train’ with 30,000 samples, this will tend towards a normal distribution which our randomly generated dataset follows. Dev has slightly higher mutual information and this is likely due to the fact that it has only been run with 15,000 samples and as we saw on the visualization gave a slightly different image than ‘train’. Overall this assignment has looked at the role which random process play in uncertainty and how this applies to information theory, whilst looking at data we can easily visualize. It is important to remember the intent of all of this (maximizing mutual information) is still a form of discriminative training with the purpose being trying to create a predictors which give us the lowest error rate.

(2) You conduct a set of machine learning experiments in which you measure performance on a data set of files (). The baseline system gives an error rate of .

(a) Your new system delivers an error rate of . Is it statistically significant at a confidence level of ? Explain.

(b) What is the minimum decrease in error rate that will be statistically significant?

(c) Repeat (a) and (b) for , and , for confidence levels of and .

Present your results in a single, nicely formatted table. Explain their significance.

You might find this spreadsheet useful:

*https://isip.piconepress.com/courses/temple/ece\_8527/resources/statistical\_significance/*

.