**ECE 8527: Introduction to
Machine Learning and Pattern Recognition**

# HW No. 07: Information Theory and StatisticAL SIGNIFICANCE

(1) Data set no. 8 is located on the course web site in [*/resources/data/set\_08/*](https://isip.piconepress.com/courses/temple/ece_8527/resources/data/set_08/). Only use training data (train.csv) and the dev data (dev.csv). This set consists of a two-dimensional feature vectors: $[x\_{1},x\_{2}]$. Quantize each element of each vector to a set of 128 discrete values by dividing each value by the range:

$y=round\left(\left({x}/{(x\_{max}-x\_{min⁡)}}\right)\*128\right)$ .

Compute the following quantities: $H(x\_{1}), H\left(x\_{2}\right), H\left(x\_{1},x\_{2}\right), H\left(x\_{1}\left|x\_{2}\right.\right), H\left(x\_{2}\left|x\_{1}\right.\right),$ and $I\left(x\_{1};x\_{2}\right)$. Do this for both the train and dev sets. Repeat these calculations for $x\_{1}$ and $x\_{2}$ generated from statistically independent uniform random number generators. Present your results in a table whose columns represent the different data sets. Compare the quantities calculated experimentally to the theoretical values for the last case. Discuss what you have learned from this analysis and what it reveals about the random process uses to generate set #8.

(2) You conduct a set of machine learning experiments in which you measure performance on a data set of $1,000$ files ($N=1,000$). The baseline system gives an error rate of $20.0\%$.

(a) Your new system delivers an error rate of $19.0\%$. Is it statistically significant at a confidence level of $80\%$? Explain.

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

(c) Repeat (a) and (b) for $N=100, 500, 2000, 5000 $, and $10,000$, for confidence levels of $85\%, 90\%$ and $95\%$.

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/*