Statistical Inference and Machine Learning
Homework 2

- This assignment can be solved in groups of 1 up to 5 students. You must mention the name of all the participants. Note that all the students in a group will get the same grade.
- Deadline: 25 November 2020, 23:59 (No late submissions will be accepted)
- Upload a single pdf file on Moodle containing your solution.

1 Feature Selection [60 pts]

Algorithm:
Given a dataset $S = \{(Y^i, X^i)\}_{i=1}^n$ of n instances, where features $X = (X_1, \ldots, X_d) \in R^d$, and labels $Y = \{1, \ldots, K\}$.

- For each value of the label $Y = k$
  - Estimate density $p(Y = k)$
- For each feature X_i, $i = \{1, \ldots, d\}$
  - Estimate its density $p(X_i)$
  - For each value of the label $Y = k$, estimate the density $p(X_i|Y = k)$
  - Score feature X_i, $i = \{1, \ldots, d\}$, using

$$I(X_i, Y) = \sum_{x_i \in \mathcal{X}, y \in \mathcal{Y}} p(x_i, y) \log_2 (\frac{p(x_i, y)}{p(x_i)p(y)}) \tag{1}$$

    where $\mathcal{X}$ and $\mathcal{Y}$ denote the support sets of X_i and Y.
- Choose those feature X_i with high score I_i

Insight: Informativeness of a feature

- We are uncertain about label Y before seeing any input.
  - Suppose we quantify using entropy $H(Y)$, defined as

$$H(Y) = -\sum_{y \in \mathcal{Y}} p(y) \log_2 p(y) \tag{2}$$

    where $\mathcal{Y}$ denotes the support sets of Y.