# 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

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

    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 $Y$.

- Given a particular feature $X_i$, the uncertainty of $Y$ changes
  - Suppose we quantify using conditional entropy $H(Y|X_i)$, defined as

$$H(Y|X_i) = -\sum_{y \in \mathcal{Y}, x_i \in \mathcal{X}} p(x_i, y) \log_2 \frac{p(x_i, y)}{p(x_i)} \tag{3}$$

    where $\mathcal{X}$ and $\mathcal{Y}$ denote the support sets of $X_i$ and $Y$.

- The reduction in uncertainty is the informativeness of feature $X_i$

$$I(X_i, Y) = H(Y) - H(Y|X_i) \tag{4}$$

  where $I(X_i, Y)$ is the mutual information which quantifies the reduction in uncertainty in $Y$ after seeing feature $X_i$.

**Questions:**

1. Given the definition of mutual information as in (4), show its calculation as in (1).
2. Given the definition of mutual information as in (4), derive a similar formula as equation (1) for $I(Y, X_i)$. Conclude that the mutual information is symmetric, i.e.,

$$I(X_i, Y) = I(Y, X_i). \tag{5}$$

3. Now, let's look at an example. Given a dataset as below.

| day | outlook | temperature | humidity | wind | play |
|---|---|---|---|---|---|
| 1 | sunny | hot | high | weak | no |
| 2 | sunny | hot | high | strong | no |
| 3 | overcast | hot | high | weak | yes |
| 4 | rain | mild | high | weak | yes |
| 5 | rain | cool | normal | weak | yes |
| 6 | rain | cool | normal | strong | no |
| 7 | overcast | cool | normal | strong | yes |
| 8 | sunny | mild | high | weak | no |
| 9 | sunny | cool | normal | weak | yes |
| 10 | rain | mild | normal | weak | yes |
| 11 | sunny | mild | normal | strong | yes |
| 12 | overcast | mild | high | strong | yes |
| 13 | overcast | hot | normal | weak | yes |
| 14 | rain | mild | high | strong | no |

   We want to decide whether to play or not to play Tennis on a Saturday. This is a binary classification problem (play vs no-play). Each input (a Saturday) has four features: Outlook, Temp, Humidity, Wind.
   Now compute $I(\mathit{outlook}, Y)$, $I(\mathit{temp}, Y)$, $I(\mathit{humidity}, Y)$, and $I(\mathit{wind}, Y)$, respectively. Given your results, which feature is the most informative feature?

## 2 Decision Trees [40 pts]

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy). Leaf node (e.g., Play) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

**Algorithm:**

The core algorithm for building decision trees called ID3 by J.R.Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. ID3 uses Entropy and Information Gain to construct a decision tree.

**Question:**

Construct the decision tree using the training dataset as above.