

Exercise: Entropy
Oct 1st, 2018
Deadline: Oct 3rd, 2018

Table 1 shows the values of annotated variables for some patients. Calculate the entropy for all pairs of variables, excluding the identifiers (d^*).

Training	fever	vomiting	diarrhea	shivering	Classification
d_1	no	no	no	no	healthy (H)
d_2	average	no	no	no	influenza (I)
d_3	high	no	no	yes	influenza (I)
d_4	high	yes	yes	no	salmonella poisoning (S)
d_5	average	no	yes	no	salmonella poisoning (S)
d_6	no	yes	yes	no	bowel inflammation (B)
d_7	average	yes	yes	no	bowel inflammation (B)

Table 1: Annotated variables for patients

According to the entropy values calculated, which is the best variable to predict if a person is healthy or not?

Table 2 shows other set of data where one of the variables (x_2) is numerical. How would you calculate entropy in this context (it is not allowed to consider each one of the numerical values as a unique value)?

Sample	x_1	x_2	x_3	Class
1	A	70	true	C_1
2	A	90	true	C_2
3	A	85	false	C_2
4	A	95	false	C_2
5	A	70	false	C_1
6	B	90	true	C_1
7	B	78	false	C_1
8	B	65	true	C_1
9	B	75	false	C_1
10	C	80	true	C_2
11	C	70	true	C_2
12	C	80	false	C_1
13	C	80	false	C_1
14	C	96	false	C_1

Table 2: Annotated variables for some observations