Exercise 11 : Bayes rule, Decision-tree, Probability, Outlier Detection

Exercise 11-1 : Bayes rule

Consider a discrete random variable θ with a sample space $\{a, b\}$ and a probability mass function with $Pr(\theta = a) > 0.5$. Also consider a likelihood function

$$Pr(X|\theta) = \theta^X (1 - \theta)^{(1-X)}$$

which is applied to the data set $\mathcal{D} = \{X_1 = 0, X_2 = 1, X_3 = 1\}$. Compute the posterior probability $Pr(\theta = a|\mathcal{D})$ using the Bayes rule.

Exercise 11-2 : Decision-tree

ID	X_1	X_2	X_3	Y
1	A	C	E	-1
2	A	C	F	$+1$
3	A	D	E	-1
4	A	D	F	$+1$
5	A	C	F	-1
6	B	C	E	-1
7	B	D	E	$+1$
8	B	C	E	$+1$
9	B	D	F	$+1$
10	B	C	F	-1

A decision tree is being trained on the above data set. As root of the tree, the attribute X_1 was already selected. Which attributes are selected as test nodes at the next level based on the Gini index and Information Gain ?

(a) For the branch of $X_1 = A$.

(b) For the branch of $X_1 = B$.

Exercise 11-3 : Probability

Consider two independent random variables

$$X \in \{1, 2, 3, 4, 5, 6\}$$
$$\text{and}$$
$$Y \in \{1, 2, 3, 4, 5, 6, 7, 8\}$$

with probability distribution $Pr[X = k] = 1/6$ for all $k \in \{1, 2, 3, 4, 5, 6\}$ and $Pr[Y = k] = 1/8$ for all $k \in \{1, 2, 3, 4, 5, 6, 7, 8\}$. Calculate the following probability statements about the random variable $Z = X + Y$?

(a) $Pr[Z = 4]$

(b) $Pr[Z > 5]$

(c) $Pr[Z < 7]$

(d) $Pr[Z = 7]$

Exercise 11-4 : Tools : Outlier Detection (LOF,OPTICS)

(a) Load python packages : `make-classification, metrics, pyplot`. Then load `quantile, where` from `numpy`, finally load `LocalOutlierFactor` from `sklearn.neighbors` and `OPTICS` from `sklearn.cluster`.

(b) Make a random dataset with 200 samples and two number of classes, plot the dataset.

(c) Fit the `Local Outlier Factor(LOF)` algorithm on the dataset and check behaviour with different neighbourhood sizes. Extract negative outputs of model and put them as outlier.

(d) Visualize both abnormal and normal data with different colors.

(e) Run `OPTICS` algorithm on the same dataset. Find the distace of each sample from core using `core-distances-`, put 2 percent of data with highest distance from core as outlier.

(f) Visualize both abnormal and normal data with different colors.