



ESI-SBA- ÉCOLE SUPÉRIEURE EN INFORMATIQUE 08-MAI-1945

Machine Learning | n.dif@esi-sba.dz

TD2

Exercise 1

Based on the following dataset and the algorithms: Naïve Bayes (with Laplace smoothing), 3-NN (Euclidian), ID3, et C4.5 classifies the sample : Faible moyen ville non ?

Explain why the Laplace smoothing is necessary in this situation.

Convert the resulting ID3 and C4.5 trees into decision rules

Montant	Age	Lieu	Etude	Classe
Moyen	Moyen	Village	Oui	Oui
Elevé	Moyen	Bourg	Non	Non
Faible	Agé	Bourg	Non	Non
Faible	Moyen	Bourg	Oui	Oui
Moyen	Jeune	Ville	Oui	Oui
Elevé	Agé	Ville	Oui	Non
Moyen	Agé	Ville	Oui	Non
Faible	Moyen	Village	Non	Non

Exercise 2

Based on the following dataset and the algorithms: Mixed NaiveBayes, 1-NN (Manhattan) classifies the sample: Burn 170 85?.

Choose the best threshold for taille using the bi-partition strategy in C4.5.

Couleur	Taille	Poids	Classe
Brun	145	45	F
Blond	160	50	F
Blond	150	60	H
Blond	185	85	H
Brun	180	90	H
Noir	165	55	F
Noir	175	80	F

Brun	155	75	F
Noir	160	65	H

Exercice 3 : Logistic regression

Based on the logistic regression algorithm and the following dataset

1. Plot the samples and verify if they are linearly separable.
2. predict the value of the parameters $\theta_0, \theta_1, \theta_2$ in the first iteration, all parameters are initialized with 0 and le learning rate $\alpha = 0.01$.
3. Compute the cross-entropy loss in the first iteration.

age	cholestrol	stroke
2	1	0
4	1.5	1
3	1	0
3.5	0.5	1
2	0.5	0
5.5	1	1
1	1	0

Exercice 4 : Evaluation metrics

- 1- Compute the accuracy, precision, recall, specificity, micro f1-score, and macro f1-score, for the following confusion matrix.
- 2- Witch metric is better for unbalanced datasets: micro f1-score or macro f1-score? Explain.

		<i>gold labels</i>		
		urgent	normal	spam
<i>system output</i>	urgent	8	10	1
	normal	5	60	50
	spam	3	30	200