Faculté des Sciences

1ière année Master Informatique (ASD)

Matière : Apprentissage et complexité

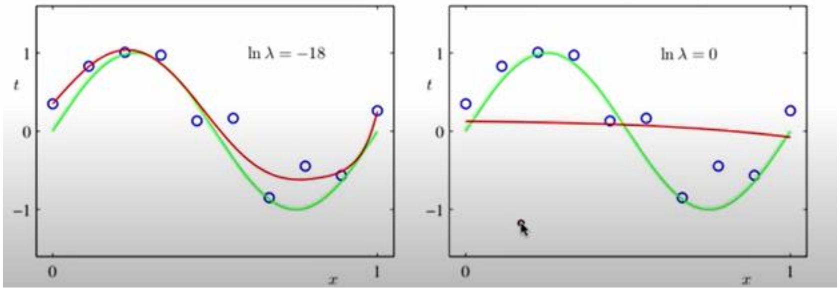
Année : 2023/2024

TD 3

Exercice 1

En faisant apprendre une modèle de régression polynomiale de degrés M sur un ensemble d'apprentissage avec deux valeurs de régularisation $\lambda=10^{-8}$ et $\lambda=1$.

Les figures (a) et (b) illustrent le modèle de la régression (courbe en rouge) après la régularisation et le résultat de prédiction sur des données d'apprentissage (courbe en vert) avant la régularisation.



(a) (b)

Interpréter les résultats obtenus.

Exercice 2 :

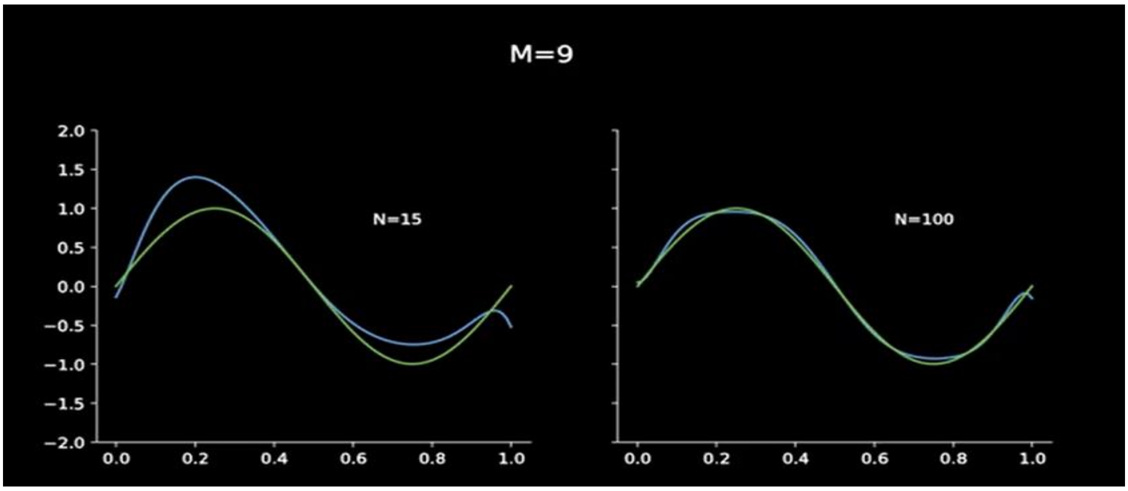
1. Soit le tableau suivant qui représente la variation des coefficients de quatre modèles polynomiaux de régression ayant de différents degrés M (sans régularisation).

	M=0	M=1	M=3	M=9
w_0	-0.035921	0.783019	0.011334	0.117859
w_1	0.000000	-1.637881	9.292162	-32.240315
w_2	0.000000	0.000000	-26.789442	552.780526
w_3	0.000000	0.000000	17.037287	-2730.404851
w_4	0.000000	0.000000	0.000000	4771.346668
w_5	0.000000	0.000000	0.000000	2008.069636
w_6	0.000000	0.000000	0.000000	-19321.678538
w_7	0.000000	0.000000	0.000000	28345.892698
w_8	0.000000	0.000000	0.000000	-17836.898852
w_9	0.000000	0.000000	0.000000	4242.454497

Que remarquez-vous ?

2. Que proposez-vous pour surpasser le problème du sur-apprentissage pour le polynôme de degrés 9 ?

3. Nous avons augmenté la taille des données d’apprentissage pour le polynôme de degrés 9, la figure suivante illustre le résultat de la génération du modèle à partir de données d’apprentissage de taille 15 et 100 exemples.



Que remarquez-vous ?

4. En utilisant un polynôme de degrés 9 et en variant les différentes valeurs de régularisation, le tableau suivant démontre les coefficients de régression obtenus.

	$\lambda = 0$	$\lambda = 0.00000000152$	$\lambda = 1$
	ln λ=-1000	ln λ=-18	ln λ=0
w_0	0.117859	0.084455	0.357091
w_1	-32.240315	-8.198305	-0.332004
w_2	552.780526	185.185112	-0.403486
w_3	-2730.404851	-860.896418	-0.304300
w_4	4771.346668	1438.245443	-0.191590
w_5	2008.069636	-422.121842	-0.097300
w_6	-19321.678538	-1042.178572	-0.023986
w_7	28345.892698	231.682021	0.031876
w_8	-17836.898852	1113.582842	0.074296
w_9	4242.454497	-635.940161	0.106591

Que remarquez-vous ?

Exercice 3 :

En considérant un modèle d'apprentissage non linéaire, nous cherchons à faire des prédictions sur un ensemble de test en utilisant un polynôme des degrés (k) égaux à 1 et 2. En utilisant la mesure AIC, quel est le modèle optimal le mieux adapté à notre cas en prenant en compte les valeurs SSE=12 pour k=1 et SSE=8 pour k=2, avec un nombre d'exemples d'apprentissage égal à 4.

Exercice 4 :

Appliquer la sélection en avant et en arrière des variables indépendantes (Blood Pressure ; BP, age, score math, sex) sur le jeu de données suivant en utilisant le critère AIC et modèle de régression linéaire multiple.

BP	Age	Score mat	Sex
124	22	65	Man
117	33	34	Women
127	35	35	Man
132	44	43	Man
117	42	54	Women
122	52	60	Women
119	62	60	Women
134	65	42	Man

Exercice 5 :

Nous souhaitons appliquer le modèle de la régression entre le nombre d'heures travaillées par jour (Y) dans une compagnie et un ensemble de variables (X1, X2, X3, X4, X5, X6, X7). En utilisant la méthode de la sélection stepewise de variables , identifier parmi les sept variables indépendantes utilisées celles sont les plus pertinentes pour la prédiction.

Obs.	Day	Y	X1	X2	X3	X4	X5	X6	X7
1	M	128.5	7,781	100	886	235	644	56	737
2	T	113.6	7,004	110	962	388	589	57	1,029
3	W	146.6	7,267	61	1,342	398	1,081	59	830
4	Th	124.3	2,129	102	1,153	457	891	57	1,468
5	F	100.4	4,878	45	803	577	537	49	335
6	S	119.2	3,999	144	1,127	345	563	64	918
7	M	109.5	11,777	123	627	326	402	60	335
8	T	128.5	5,764	78	748	161	495	57	962
9	W	131.2	7,392	172	876	219	823	62	665
10	Th	112.2	8,100	126	685	287	555	86	577
11	F	95.4	4,736	115	436	235	456	38	214
12	S	124.6	4,337	110	899	127	573	73	484
13	M	103.7	3,079	96	570	180	428	59	456
14	T	103.6	7,273	51	826	118	463	53	907
15	W	133.2	4,091	116	1,060	206	961	67	951
16	Th	111.4	3,390	70	957	284	745	77	1,446
17	F	97.7	6,319	58	559	220	539	41	440
18	S	132.1	7,447	83	1,050	174	553	63	1,133
19	M	135.9	7,100	80	568	124	428	55	456
20	T	131.3	8,035	115	709	174	498	78	968

Sachant que :

X1= nombre des transactions réalisées

X2= nombre des cadeaux vendus

X3=nombre de clients

X4=nombre des items retourné par les clients

X5=nombre de chèques encaissés

X6=nombre de mails traités

X7=nombre des tickets vendus

Exercice 6 :

Soit le tableau suivant qui représente des données recueillies après de six individus, incluant trois variables : la pression artérielle systolique (SBP), le niveau de cholestérol (Chol) et l'âge.

SBP	Chol	Age
120	126	38
125	128	40
130	128	42
121	130	42
135	130	44
140	132	46

1. Calculer le coefficient de corrélation de Pearson entre le niveau de cholestérol (Chol) et l'âge. Quelle conclusion en tirez-vous ?
2. Chercher les paramètres du modèle de régression linéaire multiple.
3. Calculer les valeurs propres et les vecteurs propres de la matrice de covariance correspondante.
4. Calculer les deux composantes principales.
5. En utilisant le modèle de régression en composantes principales, faites la prédiction de la pression artérielle systolique d'une personne avec un niveau de cholestérol de 125 et âgée de 40 ans.
6. Représenter graphiquement la pression artérielle systolique avec la première composante principale.