Faculté des Sciences

1ière année Master Informatique (ASD)

Matière : Apprentissage et complexité

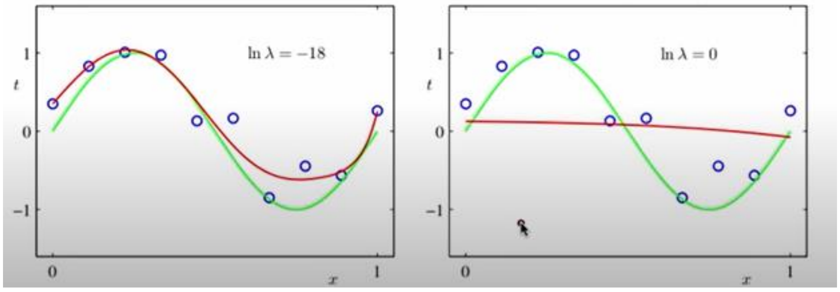
Année : 2022/2023

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 λ=10^{-8} et λ=1.

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



(a) (b)

Interpréter les résultats obtenus.

Avec λ=10^{-8} ($\ln \lambda = -18$), on a un modèle de régression qui s'adapte au mieux aux données d'apprentissage. Le modèle a plus de capacité.

Avec une régularisation élevé λ=1 ($\ln \lambda = 0$), on remarque que lorsque on augmente la régularisation, plus on simplifie le modèle de la régression. Ce dernier a moins de capacité (flexibilité) à prédire.

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 ?

On voit lorsqu'on augmente la capacité du modèle pour quantité de données d'apprentissage, on observe que les valeurs des paramètres augmentent. Si M=9, le modèle souffre du sur-apprentissage.

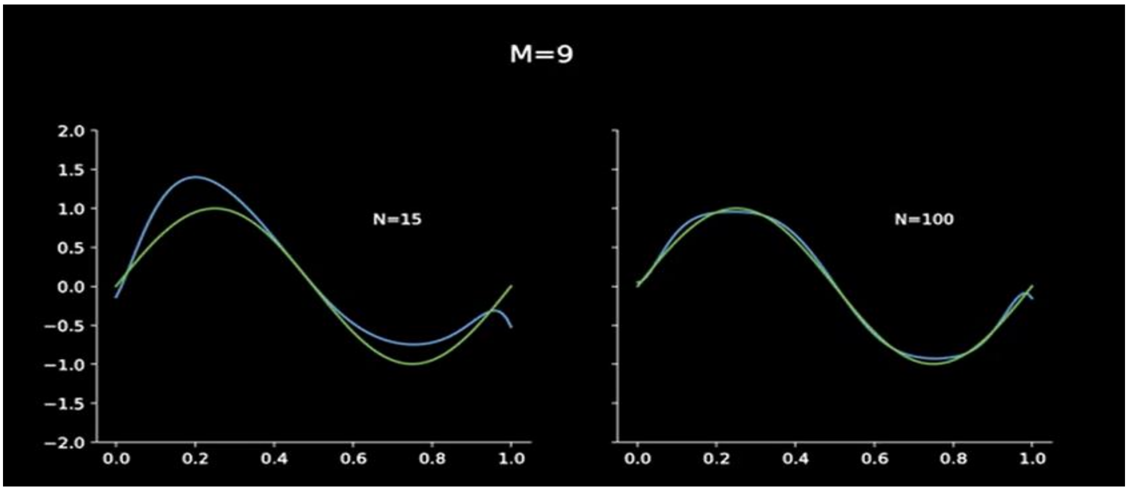
2. Que proposez-vous pour surpasser le problème du sur-apprentissage pour le polynôme de degrés 9 ?

   On propose deux solutions :

   (1) Régularisation du modèle

   (2) Augmenter la taille des données d'apprentissage

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 ?

   En augmentant la taille des données d'apprentissage, on peut résoudre le problème du sur-apprentissage du modèle polynomiale.

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 ?

Si lamda =0 (pas de régularisation), on obtient des valeurs de coefficients élevé (positives ou négatives). → Problème du sur apprentissage du modèle (biais faible et variance forte).

Si lamda =0.0000000152, nous remarquons les valeurs des poids sont significativement réduit à des valeurs proches de 0. → Le modèle n'apprend rien (biais fort et variance faible).

Si lamda=1, nous obtenons un modèle optimal (avec des coefficients plus réduits) avec une bonne généralisation (biais et variance faibles).

Exercice 3 :

$$AIC_{k=1} = 4 * 2\log\left(\frac{12}{4}\right) + 2 * 2 = 12.7944$$

$$AIC_{k=2} = 4 * 2\log\left(\frac{8}{4}\right) + 2 * 3 = 11.5456$$

Le modèle le plus optimale est le polynôme de degré 2 (avec la valeur AIC la plus basse).

Exercice 4 :

1. Backward selection

$$y = \beta_0 + \beta_1\ age + \beta_2 score + \beta_3\ sex$$

Variables	**AIC**
$age + score + sex (y = \beta_0 + \beta_1\ age + \beta_2 score + \beta_3\ sex)$	69.96
-score ($y = \beta_0 + \beta_1\ age + +\beta_3\ sex$)	**52.15**
-age($y = \beta_0 + \beta_1\ age + \beta_3\ sex$)	62.18

-sex($y = \beta_0 + \beta_1\ age + \beta_3\ sex$)	71.63

$$y = \beta_0 + \beta_1\ age + \beta_3\ sex$$

Variables	AIC
$age + sex$	**52.15**
-age	53.07
-sex	63.19

Le meilleur modèle de régression est : $y = \beta_0 + \beta_1\ age + \beta_3\ sex$

1. Forward selection

$$y = \beta_0$$

Variables	AIC
y	58.10
+score	62.94
+age	63.19
+sex	**53.07**

$$y = \beta_0 + \beta_1 sex$$

Variables	AIC
y	53.07
+score	62.18
+age	**52.15**

$$y = \beta_0 + \beta_1 sex + \beta_2\ age$$

Variables	AIC
y	**53.07**
+score	69.96

Le meilleur modèle de régression est : $y = \beta_0 + \beta_1 sex + \beta_2\ age$

Exercice 5 :

Etape 1 : faire la régression de Y avec chaque variable indépendante (X1, X2, X3, X4, X5, X6, X7)**.**

Résultat de la régression de Y avec X1 :

	Coefficients	T-value
Intercepte	117 ,7281	17.1927
Coefficient X1	**-0.0001**	**-0.0541**

Résultat de la régression de Y avec X2 :

	Coefficients	T-value
Intercepte	105.7621	18.3834
Coefficient X2	**0.1277**	**2.1655**

Résultat de la régression de Y avec X3 :

	Coefficients	T-value
Intercepte	93.7389	13.9711
Coefficient X3	**0.0302**	**3.6786**

Résultat de la régression de Y avec X4 :

	Coefficients	T-value
Intercepte	114.7259	23.3657
Coefficient X4	**0.0125**	**0.6018**

Résultat de la régression de Y avec X5:

	Coefficients	T-value
Intercepte	88.1930	14.8124
Coefficient X5	**0.0491**	**5.1312**

Résultat de la régression de Y avec X6:

	Coefficients	T-value
Intercepte	84.9667	10.3869
Coefficient X6	**0.5562**	**4.0717**

Résultat de la régression de Y avec X7:

	Coefficients	T-value
Intercepte	103.3229	23.4758
Coefficient X7	**0.0186**	**3.5591**

Etape 2 : générer le modèle à partir de deux variables : $y = \beta_0 + \beta_1 X$?

Y, regressed only on:	X1	X2	X3	X4	X5	X6	X7
b_1	-0.001	0.128	0.030	0.013	0.049	0.556	0.019
t-statistic	-0.054	2.165	3.679	0.602	5.131	4.072	3.559
P value	0.957	0.035	0.001	0.550	0.000	0.000	0.001

X5 a la plus grande valeur absolue de t-value (et p-value <0.05), donc on définit le modèle comme suit :

$$y = \beta_0 + \beta_1 X_5$$

Etape 3 : générer le modèle à partir de trois variables : $y = \beta_0 + \beta_1 X_5 + \beta_2 X?$

Y, regressed on:	X5+X1	X5+X2	X5+X3	X5+X4	X5+X6	X5+X7
b_2	0.001	0.132	0.014	-0.032	0.358	0.007
t-statistic	1.422	2.817	1.680	-1.742	2.762	1.248
P value	0.162	0.007	0.099	0.087	0.008	0.2179

X2 a la plus grande valeur absolue de t-value (et p-value <0.05), donc on définit le modèle comme suit :

$$= \beta_0 + \beta_1 X_5 + \beta_2 X_2$$

Etape 4 : générer le modèle à partir de quatre variables : $y = \beta_0 + \beta_1 X_5 + \beta_2 X_2 + \ \beta_3 X?$

Y, regressed on:	X5+X2+X1	X5+X2+X3	X5+X2+X4	X5+X2+X6	X5+X2+X7
b_3	0.0014	0.0083	-0.0347	0.2582	0.0048
t-statistic	1.5004	0.9744	-2.0238	1.9001	0.8854
P value	0.1401	0.3347	0.0486	0.0634	0.3804

$$y = \beta_0 + \beta_1 X_5 + \beta_2 X_2 + \ \beta_3 X_4$$

Etape 5 : générer le modèle à partir de quatre variables : $y = \beta_0 + \beta_1 X_5 + \beta_2 X_2 + \ \beta_3 X_4 + \beta_4 X?$

Y, regressed on:	X5+X2+X4+X1	X5+X2+X4+X3	X5+X2+X4+X6	X5+X2+X4+X7
b_4	0.0015	0.0155	0.2523	0.0051
t-statistic	1.7579	1.8066	1.9156	0.9664
P value	0.0853	0.0772	0.0615	0.3388

Au niveau de significativité de 5%, on compare la valeur p avec 0,05. Si la valeur p est inférieure à 0,05, on rejette l'hypothèse nulle et considère la variable comme statistiquement significative.

Pour la combinaison 1, 0,085>0,050, donc elle n'est pas considérée comme statistiquement significative au niveau de 5%.

Pour la combinaison 2, 0,0772>0,050, ce qui signifie également qu'elle n'est pas statistiquement significative au niveau de 5% (la même remarque pour les autres combinaisons de variables).

Donc, aucun des quatre variables ajoutées au modèle n'a une valeur t-statistique significative. La régression en avant est terminée dans ce niveau.

Le modèle final de régression multiple:

$$y = \beta_0 + \beta_1 X_5 + \beta_2 X_2 + \beta_3 X_4 + \varepsilon$$

Exercice 6:

$$SBP = intercept + chol + age$$

1. Cor (chol, age)=0.94→ Ce modèle souffre de la colinéarité.
2. $SBP = intercept + chol + age$

$$SBP = intercept + PC1$$

3. Calculer les valeurs propres et les vecteurs propres de la matrice de covariance correspondante:

$\alpha_1 = 0.589,\ \alpha_2 = 0.808 \rightarrow\ 0.589^2 + 0.808^2 = 1$

4. $PC1 = 0.589.chol + 0.808.age$

SBP	Chol	Age	PC1	PC2
120	126	38	105	-79
125	128	40	108	-80
130	128	42	109	-79
121	130	42	111	-80
135	130	44	112	-79
140	132	46	115	-80

Variance			12.08	0.32

$$PC1 = 0.589.chol + 0.808.age$$

Si chol=126, age=38 → $PC1 = 105$

On doit garder la composante principale avec le maximum de variance.

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.

$$SBP = \beta_0 + \beta_1 PC1$$

$PC1 = 0.589.chol + 0.808.age$ → β_0 =-83.9, β_1 =1.932

→ SBP =-83.9+1.932 PC1

SBP =-83.9+1.932 ($0.589.chol + 0.808.age$)

SBP =-83.9+1.14.chol+1.56 age

Si chol= 125 et age= 40 → SBP =-83.9+1.14.chol+1.56 age= 121

6. Représenter graphiquement la pression artérielle systolique avec la première composante principale.