UNIVERSITÉ DE GENÈVE

ANALYSE ET TRAITEMENT DE L'INFORMATION
14X026

TP 4: LDA, k-means and autoregression

Author: Julien Python [120 Crédits]

E-mail: julien.python@etu.unige.ch

November 5, 2020

UNIVERSITÉ
DE GENÈVE
FACULTÉ DES SCIENCES
Département d'informatique

Rappel sur la matrice de confusion :

On a donc

True negatives False negatives

True positives False positives

n = 165	Predicted: No	Predicted: Yes
Actual: No	50	10
Actual: Yes	5	100

Comme on peut le voir sur l'exemple à droite.

Dans notre cas, au lieu d'être no et yes, ce sera 3 et 7 par exemple.

Exercice 1

Dans ce premier exercice, nous appliquons d'abord la méthode de PCA (Principal component analysis) à nos données avec les 50 premiers vecteurs propres, pour 1000 instances de 3 et 1000 instances de 7, à l'aide des données MNIST.

Dans la partie 1 de cet exercice, nous appliquons la méthode LDA (Linear discriminant analysis). On obtient ainsi la matrice de confusion :

[[950 60]

[61 967]]

Dans la seconde partie, on utilise la méthode du k-NN (k nearest neighboorood). On obtient ainsi la matrice de confusion :

[[872 138]

[82 946]]

Mais afin d'être plus général et de ne pas nous baser sur un seul cas, on a aussi calculé des données statistiques pour 50 itérations :

Pour matrice de confusion:

LDA
1-NN
---
mean:
[[[949.24 60.76]
[35.74 992.26]]

[[958.28 51.72]
[34.36 993.64]]]
---
median:
[[[949.   61.]
[35.5 992.5]]

[[972.5   37.5]
[25.5 1002.5]]]
---
min:
[[[914 39]
[13 974]]

[[882 29]
[10 932]]]
---
max:
[[[971   96]
[54 1015]]

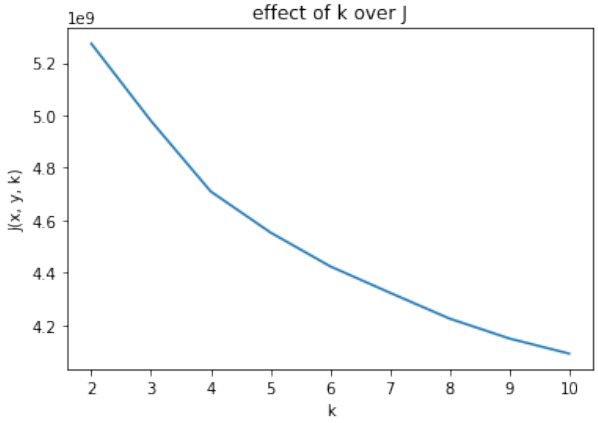
[[981 128]
[96 1018]]]

Comme on peut le voir ci-dessus, les résultats sont très proches. On observe tout de même une espérance avec plus de faux négatifs avec LDA (60 par rapport à 51 pour 1-NN). Ce résultat est confirmé par la médianne. Mais c'est intéressant car le maximum de faux négatif est plus grand pour le 1-NN, tout comme le minimum de faux négatif est plus petit pour le 1-NN. Ceci se généralise pour les trois autres termes de la matrice de confusion.

Si nous faisons l'experience une seule fois avec 1-NN et n'avons pas de chance, on pourra avoir un résultat moins bon qu'avec LDA. Mais globalement pour notre exemple précis de l'exercice 1, on obtient de meilleurs résultats avec 1-NN. De plus, 1-NN est plus rapide.

Exercice 2

Premièrement, on prend un échantillon de 2'000 instances contenant des 3 et des 7. Ce qui change ici par rapport à l'exercice 1 est que nous n'aurons presque jamais le même nombre de 3 et de 7 dans notre échantillon.



Le graphe ci-dessus représente la somme des distances au carré entre chaque sample x_i et le centre de son cluster. On remarque que plus on augmente le nombre de cluster, plus la distance J diminue. Et puisque J est une fonction monotone décroissante, on a que J va converger comme on l'observe sur le graphe à droite ci-dessus. Mais dans notre problème, on souhaiterait deux clusters, un pour 3 et un pour 7.

Pour la deuxième partie avec k=2, en exécutant on a par exemple le résultat suivant :

```
[[  52  958]
 [1004   24]]
```

Ainsi, ce résultat est très mauvais par exemple, car les clusters sont inversés, donc le cluster centroid qui devrait être au centre de la majorité des 3 est au centre de la majorité des 5, et inversemment. Après avoir fait plusieurs essais (voir la suite de l'exercice 2), on remarque qu'il y a des grandes disparités entre les essais. Ceci pourrait être expliqué par le fait que l'algorithme k-means dépend d'où sont placés initiallement les clusters centroid. S'ils sont bien placés au départ, il nous rendra un bon résultat et inversemment pour ce problème par exemple. Ceci peut être expliqué par le fait que le nombre 3 est assez proche du nombre 7.

Données statistiques concernant matrice de confusion pour 50 essais avec k=2:

3 vs. 7
---
mean:
[[559.44 450.56]
[454.98 573.02]]
---
median:
[[958. 52.]
[26. 1002.]]
---
min:
[[45 46]
[22 22]]
---
max:
[[964 965]
[1006 1006]]
---
var:
[[203957.0064 203957.0064]
[236050.8996 236050.8996]]

3 vs. 5
---
mean:
[[489.82 520.18]
[456.28 435.72]]
---
median:
[[352. 658.]
[530.5 361.5]]
---
min:
[[213 169]
[322 319]]
---
max:
[[841 797]
[573 570]]
---
var:
[[52557.4276 52557.4276]
[9559.6416 9559.6416]]

Comme on l'observe ci-dessus avec l'esperance et la médianne, 3vs7 obtient une meilleure réussite par rapport à la matrice de confusion comparé à 3vs5. J'ai recherché la variance qui est une donnée statistique importante our notre cas et elle est élevée dans les deux cas, mais encore plus élevée dans 3vs7, et donc on peut avoir des résultats très espacés si on répète l'expérience.

Ceci peut donc s'expliquer par le fait que le nombre 3 est plus proche du nombre 5 que du nombre 7. Car pour le nombre 3, seule la petite barre en haut à gauche/droite se décale pour former soit un 3, soit un 5, alors que pour modifier un 3 en 7 c'est plus compliqué, il faut un plus grand changement, et donc les clusters centroid ont plus de difficultés.

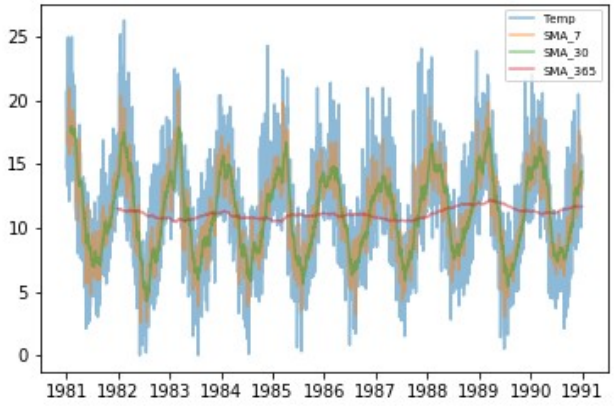
Exercice 3

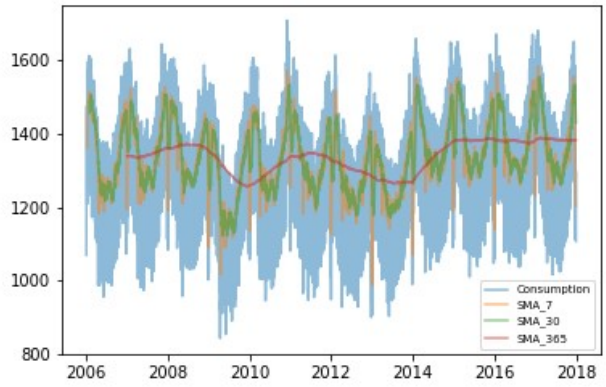
Comme on peut le voir sur ces trois figures, on observe des résultats intéressants. Sur la figure des températures, il n'y a pas de tendence particulière, on voit que le SMA annuel est plutot plat. A l'aide des SMA de 7 et 30 jours, on voit clairement que les saisons jouent un rôle important dans la température, car on obtient une courbe à la forme sinusoidale qui se répète de manière très similaire chaque année.

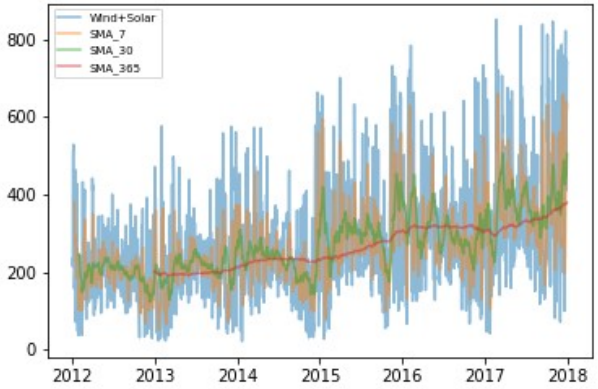
Pour la consommation, il n'y a pas de tendance particulière, le résultat est plutôt plat, on observe avec le SMA 365 qu'il y a eu deux petites baisses en 2008-9 et 2013. Pour la période de 2008-9, ceci peut probablement s'expliquer comme conséquence de la crise des subprimes. On observe chaque année un moment où la consommation est très faible grâce au SMA_7, ce qui peut être expliqué par des périodes de vacances et donc beaucoup d'entreprises ferment.

Pour le vent et le solaire, on observe une tendance à la hausse d'année en année. On voit aussi une saisonnalité annuelle avec un shéma qui semble se répéter à l'aide du SMA_30.

Evidemment, puisque les SMA utilisent des données passées, ils sont légèrement en retard par rapport au données journalières dans leur changement de direction.





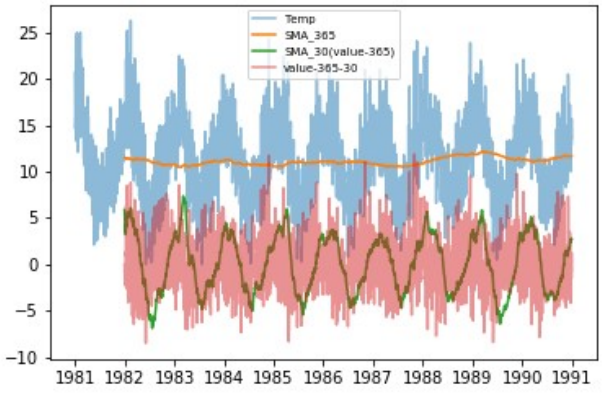


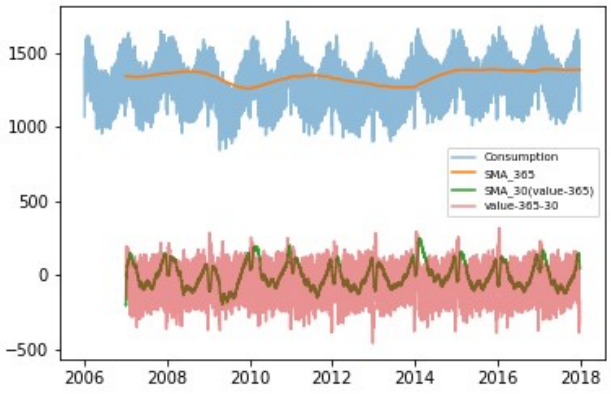
On va expliquer les valeurs pour la température, ce sera la même chose pour les deux autres.

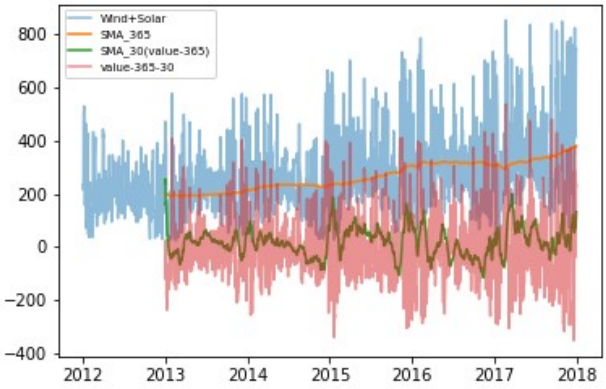
On a donc déjà vu la température et la SMA_365 (moyenne de l'année écoulée, ie trend).

Pour SMA_30(temp-365), on soustrait d'abord SMA_365 à temp "on centre les données sur l'axe y en 0", puis on effectue le SMA sur 30 jours. C'est donc la variation de température par rapport à "son esperance" (seasonality)

Pour value-365-30, c'est donc égal à x_t-SMA_365-z_t , et donc on peut parler du bruit (noise, residual) sur nos données.







En calculant l'autocorrelation entre les données et les données trimestrielles, semestrielles et annuelle, on obtient :

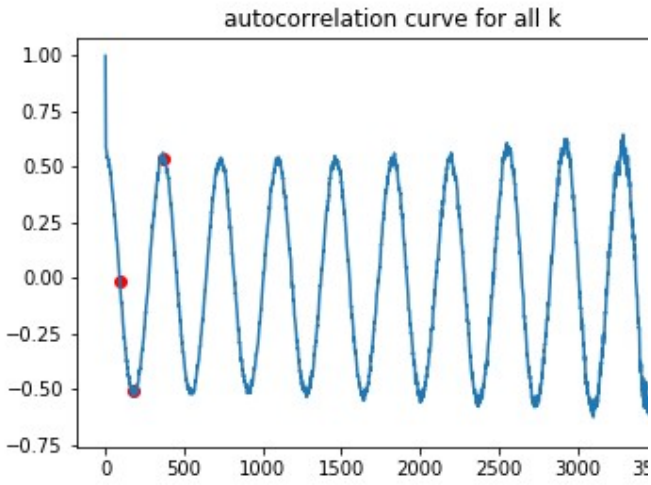
corr(data,shift_92) = -0.0139

corr(data,shift_182) = -0.5055

corr(data,shift_365) = 0.5335

Ainsi, il y a très peu de correlation entre les saisons, mais une correlation négative entre les deux périodes opposées de l'année, donc par exemple si le 1et janvier la temperature diminue, il est assez probable que 1er juin de la même année la température augmente. Et pour le dernier, cela veut dire que si par exemple la température diminue le 1er avril de l'année N, il est assez probable que la température diminue le 1er avril de l'année N+1.

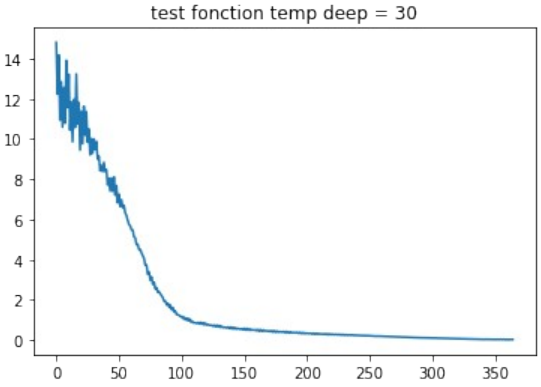
Ces résultats correspondent bien avec l'intuission qu'on pourrait avoir par rapport aux trois données présentées ci-dessus. On a donc ces trois points calculés en rouge dans le même ordre que présenté ci-dessus.

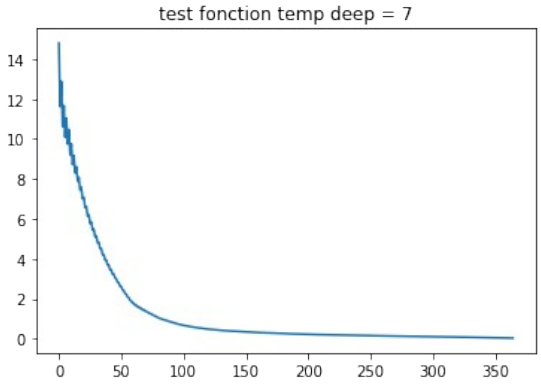


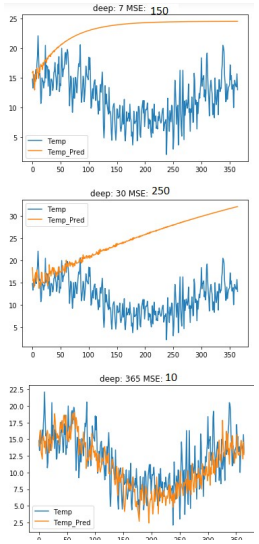
On obtient à nouveau une forme sinusoidale, et donc on a bien des données avec une périodicité d'autocorrelation de 365 jours.

Pour la dernière question, j'ai essayé (sans réussite) à avoir un plot comme celui que vous avez mis sur chamilo. Voici mes tentatives :

J'ai d'abord essayé avec les fonctions données dans l'énoncé, ci-dessous je plot l'estimation obtenue et on remarque bien que ca n'as pas du tout la forme que ca devrait avoir. Mais ca ressemble quand même au début du coup je pense que je ne suis vraiment pas loin de la solution, surtout avec le comportement au début de la fonction par rapport à la figure tout à droite du dessous disponible sur moodle. Rem: pour 365 jours je n'obtenais pas un résultat que je pouvais plotter.







J'ai donc essayé une autre méthode plus directe (différente de celle énoncée dans l'énoncé et tentée ci-dessus) et j'ai obtenu les graphes suivants , qui donnent une bonne approximation générale.