Telecom Paris
2019/2020

IA710 : Apprentissage Statistique
Pavlo Mozharovskyi, François Roueff

TP noté n° 6 : Séries Temporelles

Vous devez envoyer votre fichier sous format `ipynb` avant le mercredi 23/10/2019 23h59 à l'email suivant : pavlo.mozharovskyi@telecom-paris.fr. Chaque question est évaluée sur 1 point, à l'exception des questions 7, 9 et 13 qui sont évaluées sur 2 points. 3 point sont donnés pour :

- aspect global de présentation : qualité de rédaction, orthographe, présentation, graphes, titres, etc...,
- aspect global du code : indentation, style, lisibilité du code, commentaires adaptés,
- absence de bugs.

- Découverte de Python -

Consulter les pages suivantes pour démarrer ou bien trouver quelques rappels de Python :

— http://www.python.org
— http://scipy.org
— http://www.numpy.org
— http://scikit-learn.org/stable/index.html
— http://www.loria.fr/~rougier/teaching/matplotlib/matplotlib.html
— http://jrjohansson.github.io/

- Données "Syndromes grippaux" -

1) Importez les données du fichier `Openhealth_S-Grippal.csv`. Consultez http://www.openhealth.fr/ias pour plus d'information sur les données. Affichez les 5 premières lignes du jeu de données. On va travailler seulement avec la variable `IAS_brut`, sous la forme d'une série temporelle, dont on précisera la fréquence d'échantillonage et les dates de début et de fin. Pour cette variable, dans ce jeu de données, les cas de nullité correspondent à des données manquantes. Combiende données manquantes comporte-t-elle ? Pour les traitements des questions suivantes, éliminez les données manquantes par imputation, c'est-à-dire en déduisant une valeur plus "raisonnable" aux dates non-observées à partir des valeurs aux dates observées : par exemple par la moyenne de deux dates les plus proches.
2) Tracez la série temporelle considérée en fonction du temps. Commentez brièvement ce que vous observez.
3) Tracez l'histogramme de la loi marginale. Quel impact la distribution observée par l'histogramme a sur la trajectoire représentée à la question précédente ?
4) Reprenez les questions précédentes après tranformations des observations par le logarithme naturel. Commentez.
5) En utilisant `signal.periodogram()` (du package signal importé depuis scipy), tracez le périodogramme des données. Expliquez les pics les plus significatifs que vous observez. Attention au fait que dans la communauté signal les spectres sont normalisées pour des fréquences absolues dans $[-1/2, 1/2]$ plutôt que $[-\pi, -\pi]$.
6) Pour supprimer une tendance périodique de période T dans un signal $(x_t)_{t\in\mathbb{Z}}$ il suffit d'appliquer le filtre Δ_T défini par
$$[\Delta_T x]_t = x_t - x_{t-T}$$
Utilisez cette méthode pour supprimer la tendance périodique de période 1 an. Quel pic du périodogramme cette méthode a fait disparaître ?

7) On vous propose un exercise de prédiction uniquement à partir de la tendance périodique. On va prédire les valeurs de `IAS_brut` pour les dates du 1er avril 2013 au 17 avril 2014 en se basant sur les observations précédentes (du 1er juillet 2009 au 31 mars 2013) pour estimer la tendance périodique des données. Pour cela, pour un nombre d'harmoniques fixé (disons `n_harm`), construisez les variables explicatives pour la période d'apprentissage : pour chaque $k = 1, 2, ...$ `n_harm`, ajoutez deux variables explicatives, $x_{i,2k-1} = \cos(t_i \cdot k \cdot \frac{2\pi}{T})$ et $x_{i,2k} = \sin(t_i \cdot k \cdot \frac{2\pi}{T})$, où t_i est le moment de temps (on peut utiliser les nombres entiers au lieu de dates) et $T = 365$. Estimez la tendance périodique en régressant les données d'apprentissage sur ces variables explicatives.

   Avec ce modèle linéaire, prédisez les valeurs du variable `IAS_brut` pour la période du 1er avril 2013 au 17 avril 2014.

   Tracez les valeur de la série temporelle et votre prédiction pour l'ensemble de données (vous pouvez utilisez les couleurs différentes pour pour les partie apprentissage et prédiction). Puis, donnez le risque quadratique de la prédiction et tracez les résidus pour la période prédite.

   Essayez les valeur de `n_harm` pour voir comment varie le risque quadratique de la prédiction.

- Données "Trafic internet" -

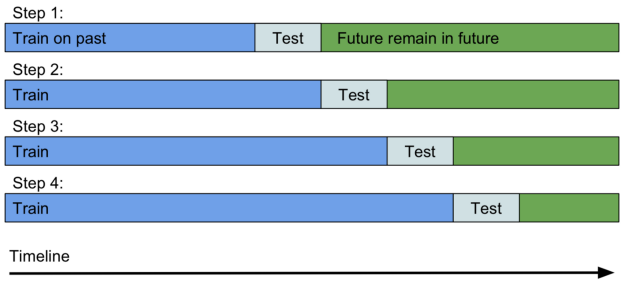
8) Importez les données du fichier `lbl-tcp-3.tcp`. Chaque ligne du tableau correspond à un paquet et la première variable et le moment de sa transmission ; consultez ftp://ita.ee.lbl.gov/html/contrib/LBL-TCP-3.html pour plus d'information sur les données. Affichez 5 premières lignes du jeu de données.

   Tout d'abord, créez la série temporelle, où chaque observation correspond au nombre de paquets transmis dans un intervalle de 10 secondes. Il y a 720 intervalles, donc vous devez obtenir 720 observations.

   Tracez la série temporelle obtenue.

9) Dans cette question on vous propose de construire un modèle auto-régressif AR(p). L'apprentissage du modèle ne demande pas beaucoup de temps et s'effectue avec $1 - 2$ ligne de code à l'aide de la fonction `ARIMA()` importée de `statsmodels.tsa.arima_model`. Le choix du paramètre p (d'une gamme de valeur prédéfinies, disons de 1 à p_{max}) est moins simple. On va aborder trois possibilité : critère d'information d'Akaike (AIC), critère d'information bayésien (BIC) et validation croisée/backtesting.

   Le critères d'information AIC et BIC sont normalement implémentés dans le logiciel et sont donnés directement après l'apprentissage du modèle. Pour faire la validation croisée pour une série, on la coupe en n_{chunks} chunks/folds et chaque foi utilise k premiers chunks pour entrainer le modèle et le chunk numéro $k + 1$ pour le tester (par exemple regarder le risque quadratique) ; voir le dessin ci-dessous.



   Pour une gamme de valeur choisie, effectue les trois méthodes de la sélection du modèle.

10) En ce basant sur les résultats de la question précédente, sélectionnez l'ordre p du modèle AR à estimer et affichez les paramètres correspondants.

    Tracez les résidus. Tracez l'estimation de la densité de résidus et la densité de la distribution normale (avec moyenne et écart-type estimés) sur le même graphique et comparez les visuellement. Commentez.

- Données “Southern oscillation index” -

11) Importez les données du fichier `soi.tsv`. Consultez http://www.bom.gov.au/climate/glossary/soi.shtml pour plus d'information sur les données. Supprimez les données manquantes.

12) A l'aide des fonctions `plot_acf()` et `plot_pacf()` importées de `statsmodels.graphics.tsaplots`, tracez la fonction d'autocorrélation et la fonction d'autocorrélation partielle.

13) En ce basant sur la question précédente, choisissez l'ordre du processus auto-régressive AR(p). Entrainer le modèle AR choisi. Tracez les résidus. Tracez l'estimation de la densité de résidus et la densité de la distribution normale (avec moyenne et écart-type estimés) sur le même graphique et comparez les visuellement. Commentez.

14) Tracez le périodogramme ; superposez le à la densité spectrale du modèle estimé à la question précédente. On utilisera qu'un modèle AR(p) de coefficients auto-régressifs $\phi_1, \ldots, \phi_p$ satisfaisant l'équation AR

$$X_t = \sum_{k=1}^{p} \phi_k X_{t-k} + \epsilon_t \ ,$$

avec (ϵ_t) bruit blanc de variance σ^2, a pour densité spectrale

$$f(\lambda) = \frac{\sigma^2}{2\pi} \left| 1 - \sum_{k=1}^{p} \phi_k \mathrm{e}^{-\mathrm{i}\, k\, \lambda} \right|^{-2} .$$

(pour des fréquences absolues normalisées entre $-\pi$ et π) ou

$$f(\omega) = \sigma^2 \left| 1 - \sum_{k=1}^{p} \phi_k \mathrm{e}^{-2\mathrm{i}\pi\, k\, \omega} \right|^{-2} .$$

(pour des fréquences absolues normalisées entre $-1/2$ et $1/2$, convention usuelle en traitement du signal.)