Séries Temporelles - cours 2

RCP217

Clément Rambour

Plan

Résumé du dernier cours

Statistical Forecasting

- **Motivation** : utiliser le contexte global
  - Utilisation de toute la série : débruitage
  - Utilisation du passé : prédiction
- Extraire les statistiques du processus sous réserve de stationnarité

Résumé du dernier cours

Prédiction progressive

$$\hat{x}_t = \boldsymbol{\varphi}^T \boldsymbol{x} \quad \text{avec} \quad \boldsymbol{\varphi} = \underset{\boldsymbol{\varphi}}{\operatorname{argmin}} \, \mathbb{E}\left[(x_t - \boldsymbol{\varphi}^T \boldsymbol{x})^2\right]$$

où $\boldsymbol{x} = [x_{t-1}, \cdots, x_{t-p}]$

Résumé du dernier cours

Prédiction progressive

$$\hat{x}_t = \boldsymbol{\varphi}^T \boldsymbol{x} \quad \text{avec} \quad \boldsymbol{\varphi} = \underset{\boldsymbol{\varphi}}{\operatorname{argmin}}\ \mathbb{E}\left[(x_t - \boldsymbol{\varphi}^T \boldsymbol{x})^2\right]$$

où $\boldsymbol{x} = [x_{t-1}, \cdots, x_{t-p}]$

Algorithmes

- Résolution des équations de Yule-Walker
- Gram-Schmidt ou des innovations
- Levinson-Durbin

Résumé du dernier cours

Processus Auto Regréssif Moving Average (ARMA)

Un processus est ARMA d'ordre $(p, q) \in \mathbb{N}^2$ s'il est du stationnaire au sens large et solution de l'équation de récurrence

$$X_t = Z_t + \sum_{k=1}^{p} \varphi_k X_{t-k} + \sum_{k=1}^{q} \theta_k Z_{t-k}$$

où $\forall t, Z_t \sim \mathcal{N}(0, \sigma^2)$ et $\forall i, j, \varphi_i$ and $\theta_j \in \mathbb{R}$.

Résumé du dernier cours

Processus Auto Regréssif Moving Average (ARMA)

Un processus est ARMA d'ordre $(p, q) \in \mathbb{N}^2$ s'il est du stationnaire au sens large et solution de l'équation de récurrence

$$X_t = Z_t + \sum_{k=1}^{p} \varphi_k X_{t-k} + \sum_{k=1}^{q} \theta_k Z_{t-k}$$

où $\forall t, Z_t \sim \mathcal{N}(0, \sigma^2)$ et $\forall i, j, \varphi_i$ and $\theta_j \in \mathbb{R}$.

Existence, causalité et inversion

- Le polynôme $\Phi(z) = \sum_{k=1}^{p} \varphi_k z^k$ n'a pas de racine de module 1
- Si $\Phi(z)$ n'a pas de racine $|z| < 1$ alors la solution est causale
- Si $\Theta(z)$ n'a pas de racine $|z| < 1$ alors la solution est inversible
- Estimation des paramètres par Least Square or Maximum Likelihood

Résumé du dernier cours

Kalman Filter

$$\begin{cases} x_{t+1} & = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{u}_t + \boldsymbol{w}_t \\ y_{t+1} & = \boldsymbol{C}\boldsymbol{x}_t + \boldsymbol{n}_t \end{cases}$$

- Estimer l'état d'un système à partir de mesures incomplètes
- Estimation de l'erreur de prédiction
- Correction de la prédiction
- Online

$$\hat{\boldsymbol{x}}_t = \hat{\boldsymbol{x}}_{t-1} + \boldsymbol{K}_t(\boldsymbol{y} - \boldsymbol{C}\hat{\boldsymbol{x}}_{n|n-1})$$

Maintenant

- Comment comparer des séries temporelles entre elles ?
  - Distance
  - Similarité
  - Extraction de *features*

Maintenant

- Comment comparer des séries temporelles entre elles ?
  - Distance
  - Similarité
  - Extraction de *features*
- Si on sait comparer on peut regrouper
  - Clustering

Maintenant

- Comment comparer des séries temporelles entre elles ?
  - Distance
  - Similarité
  - Extraction de *features*
- Si on sait comparer on peut regrouper
  - Clustering
- Apprentissage profond pour remplacer les approches développées précédemment ?
  - Réseaux récurrents
  - RNN Encodeurs-Décodeurs
  - Introduction à l'attention

Plan

Comparaison de ST

Comment comparer deux séries temporelles ?

Comparaison de ST

Comment comparer deux séries temporelles ? Problèmes courants

- Translations
- Différence d'échelle (*scaling*)
- Décalages (*shift*)
- Longueur
- Occlusions
- Complexité

Distance Euclidienne

Définition

Définition naturelle de la distance entre deux vecteurs, la distance entre deux séries temporelles $\{x_t\}_{0\leq t\leq N}$ et $\{y_t\}_{0\leq t\leq N}$ s'écrit :

$$\ell_2 = \sqrt{\sum_{t=1}^{N}(x_t - y_t)^2}$$

Distance Euclidienne

Définition

Définition naturelle de la distance entre deux vecteurs, la distance entre deux séries temporelles $\{x_t\}_{0\leq t\leq N}$ et $\{y_t\}_{0\leq t\leq N}$ s'écrit :

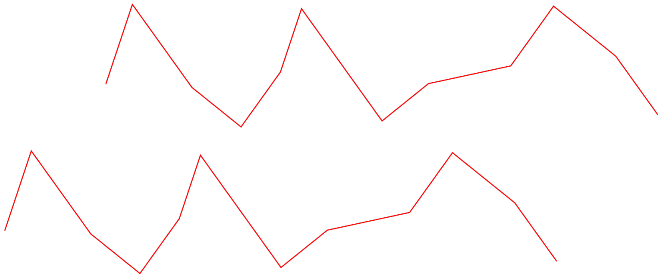
$$\ell_2 = \sqrt{\sum_{t=1}^{N}(x_t - y_t)^2}$$

- Facile à interpréter
- Rapide à implémenter et calculer

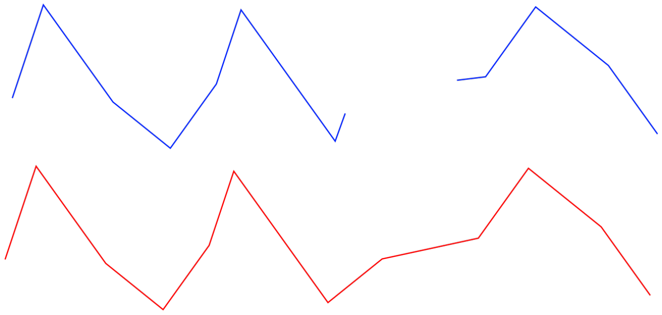
- Nécessite des prétraitement pour l'invariance à l'échelle, la translation, aux décalages ou à la longueur
- Information moyenne : pas de notion de forme

Exemples

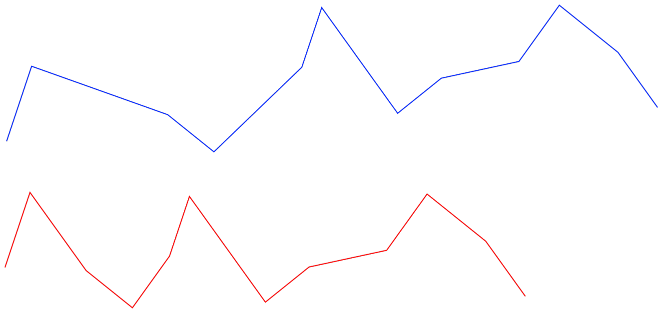
Exemples



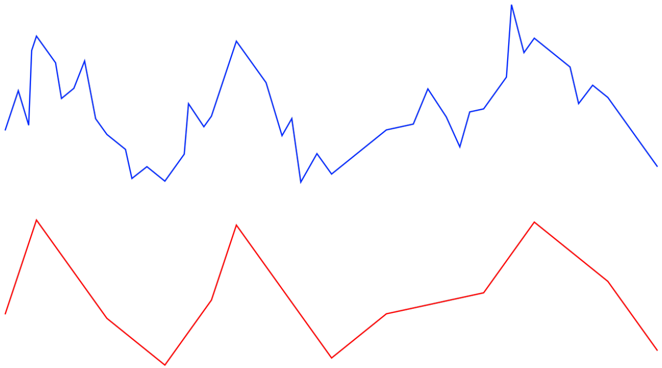
Exemples



Exemples

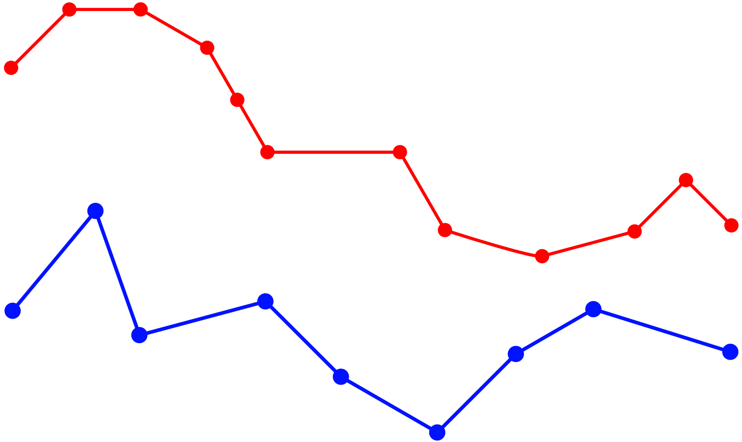


Exemples



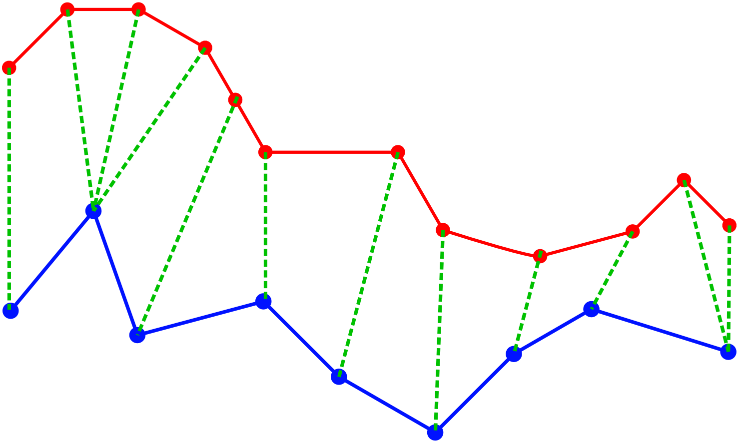
Dynamic Time Warping

Motivation : relier chaque point avec son plus proche voisin dans la série comparée



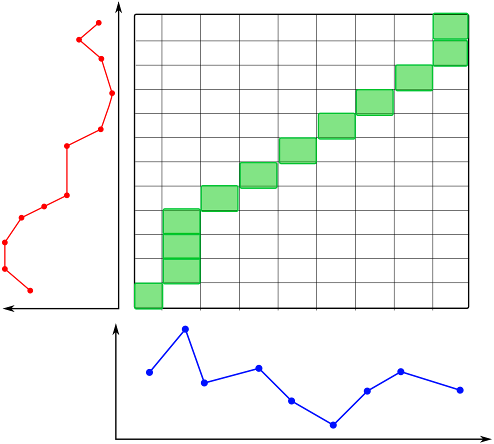
Dynamic Time Warping

Motivation : relier chaque point avec son plus proche voisin dans la série comparée



Dynamic Time Warping

On doit calculer la distance entre chaque paire → matrice d'alignements



Dynamic Time Warping

On doit calculer la distance entre chaque paire → matrice d'alignements

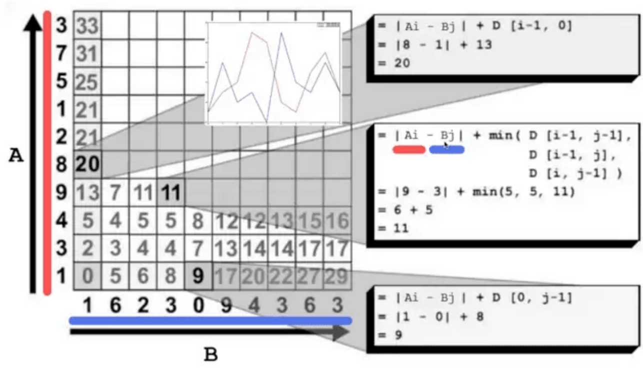


FIGURE – Thales Sehn Körting - DTW vidéo

Dynamic Time Warping

Inputs: $\mathbf{x}, \mathbf{y}$, smoothing $\gamma \geq 0$, distance function δ
$r_{0,0} = 0; r_{i,0} = r_{0,j} = \infty; i \in [\![n]\!], j \in [\![m]\!]$
for $j = 1, \ldots, m$ **do**
  for $i = 1, \ldots, n$ **do**
    $r_{i,j} = \delta(x_i, y_j) + \min^{\gamma}\{r_{i-1,j-1}, r_{i-1,j}, r_{i,j-1}\}$
  end for
end for
Output: $(r_{n,m}, R)$

Figure – REF

Other shape measurement

Cross-Correlation based measurement

Beyond Raw Data

Features et modèle

- Auto-corrélation
- Min-max values
- ARMA modèle coefficients

Plan

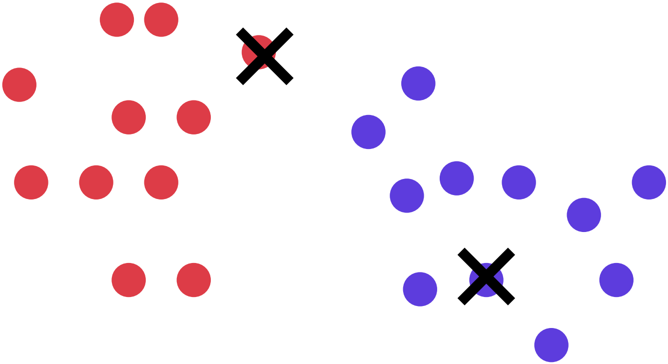
Partitional Clustering

Motivations : regrouper les séries temporelles en se basant sur une distance ou un critère de similarité

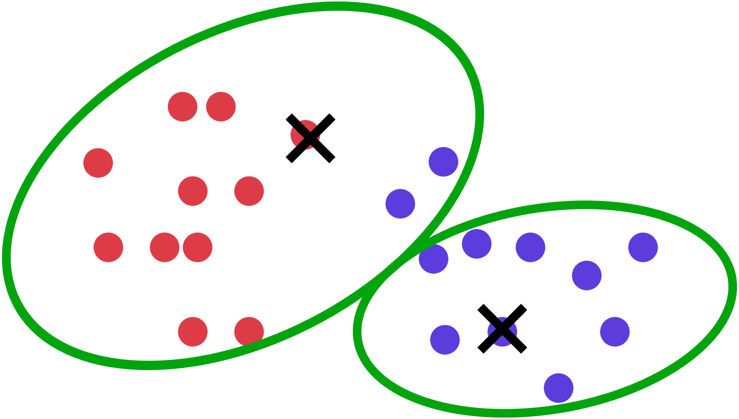
- Algorithmes simples à mettre en oeuvre
- Populaires
- Convergence lente

- K-Means
- K-Medoids

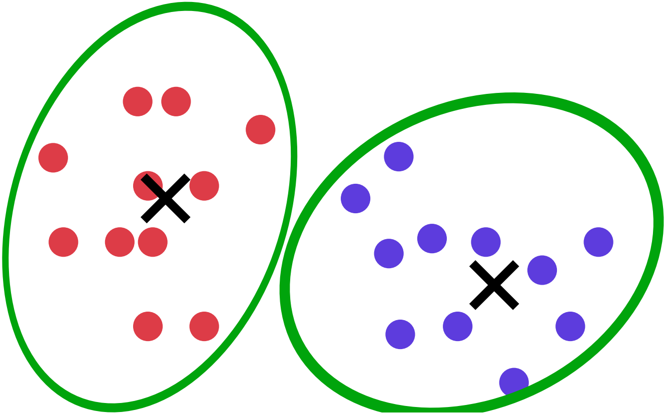
K-Means



K-Means



K-Means



Partitional Clustering

- K-Means nécessite une distance
- Les centres de classe ne sont pas forcément représentatifs
- K-Medoids : similaire à K-Means mais centre de classes $\in$ points
- K-Medoids : plus représentatif que K-Means et peut être calculé avec une similarité

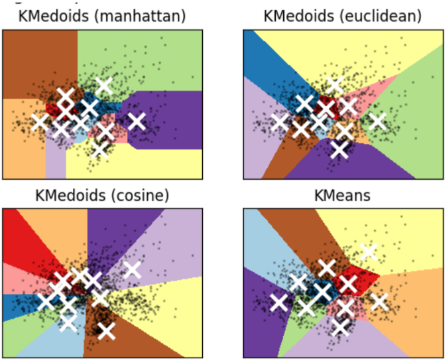


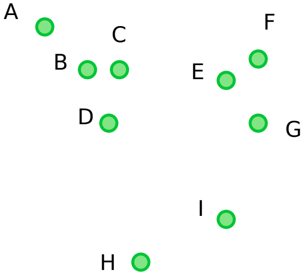
Figure – ref : scikit learn

Agglomerative Hierarchical

Motivations : Construire une hiérarchie de clusters

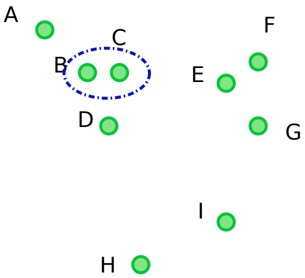
- **Agglomération** : Chaque élément est considéré comme un cluster. Les clusters sont fusionnés lorsque l'on monte dans la hiérarchie.
- **Division** : Tous les éléments appartiennent au même cluster et sont séparés récursivement en descendant la hiérarchie.

Agglomerative Hierarchical : Single Linkage Clustering

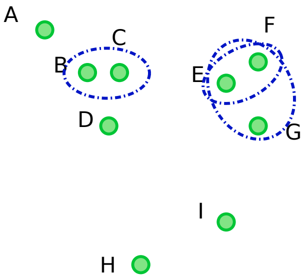


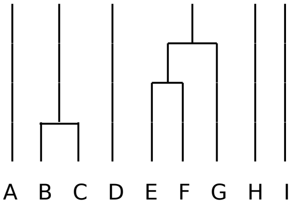
A B C D E F G H I

Agglomerative Hierarchical : Single Linkage Clustering

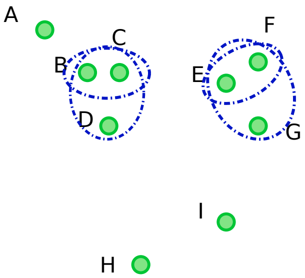


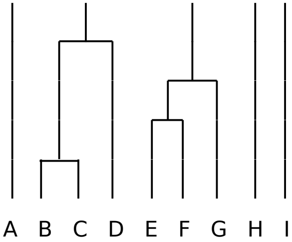
Agglomerative Hierarchical : Single Linkage Clustering



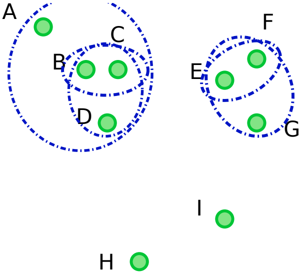


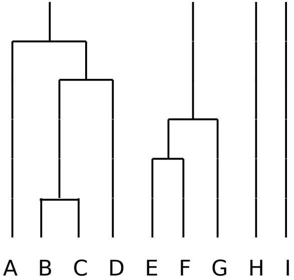
Agglomerative Hierarchical : Single Linkage Clustering



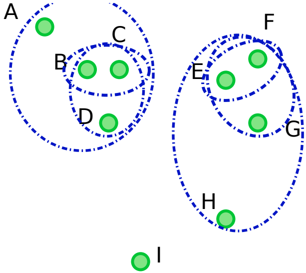


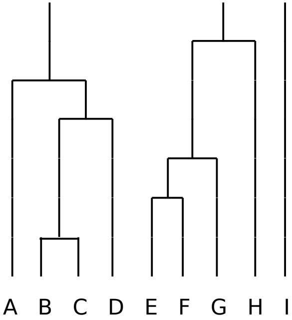
Agglomerative Hierarchical : Single Linkage Clustering



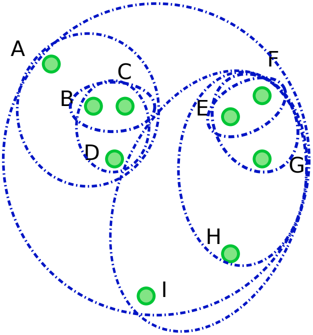


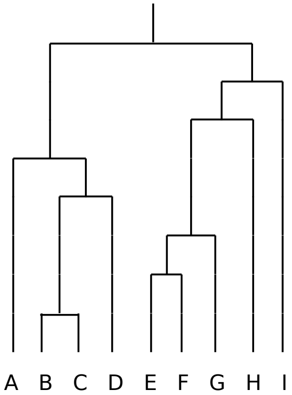
Agglomerative Hierarchical : Single Linkage Clustering



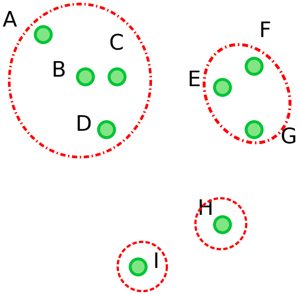


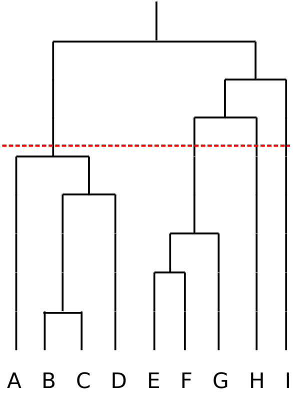
Agglomerative Hierarchical : Single Linkage Clustering





Agglomerative Hierarchical : Single Linkage Clustering



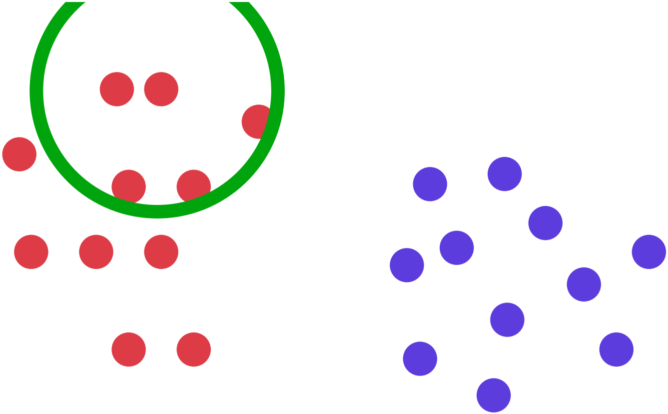


Density based clustring

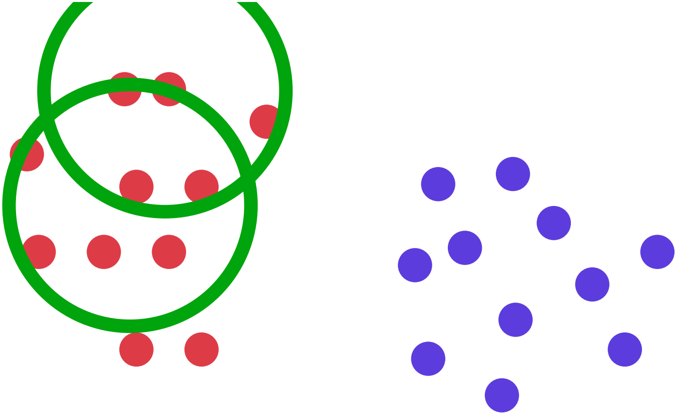
Motivation : séparer les clusters en fonction de leur densité

- DBSCAN
- OPTIC
- Density Peaks clustering

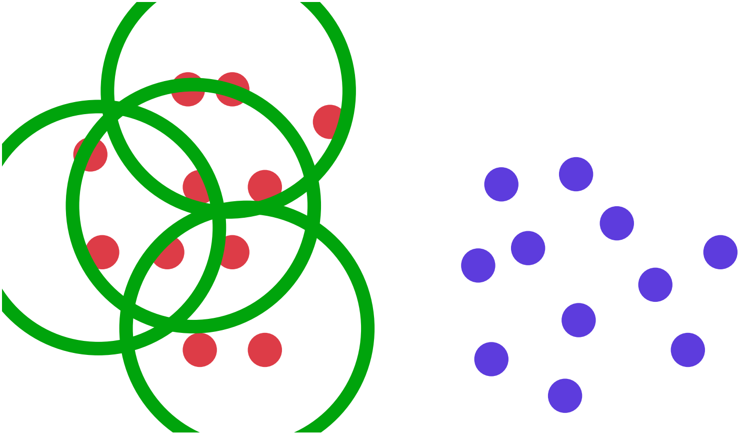
Density based clustring : DBSCAN



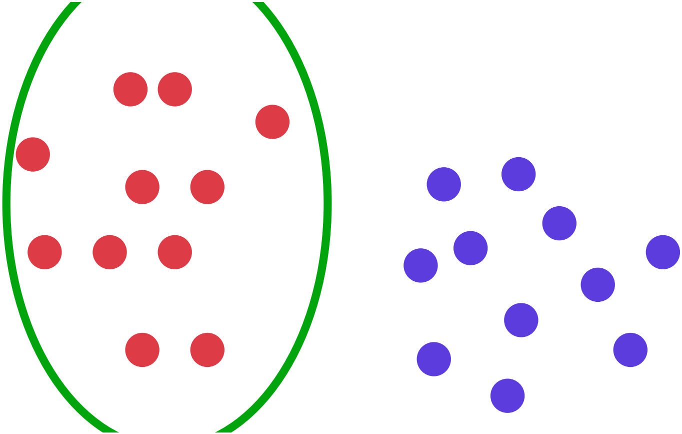
Density based clustring : DBSCAN



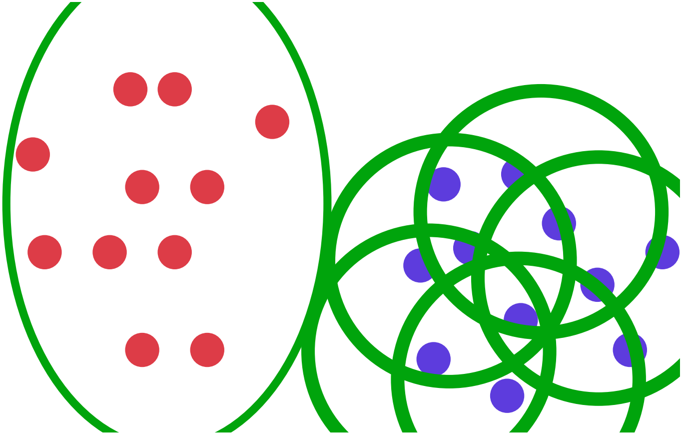
Density based clustring : DBSCAN



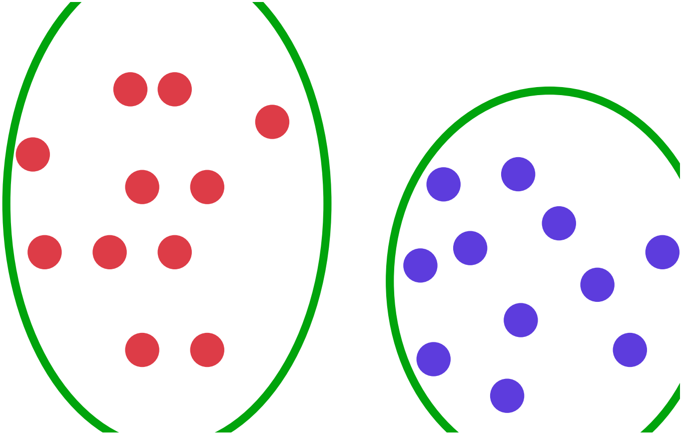
Density based clustring : DBSCAN



Density based clustring : DBSCAN



Density based clustring : DBSCAN



Plan

Deep Learning strategies

- Convolutional Neural Networks
- Recurrent Neural Networks
- Estimation $n + 1$
- Multi-horizon
- Probabilistic outputs

Convolutional Neural Networks

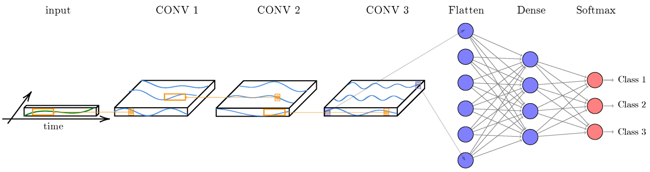


Figure – C. Pelletier et Al. 2019

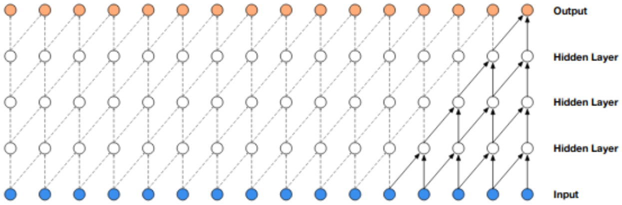


Figure – Van Den Oord et Al. 2016

Convolutional Neural Networks

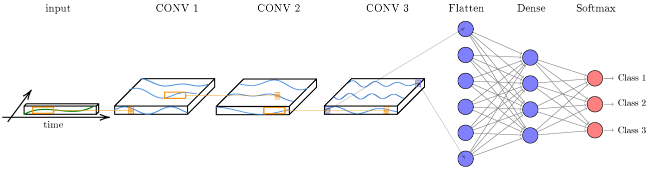


FIGURE – C. Pelletier et Al. 2019

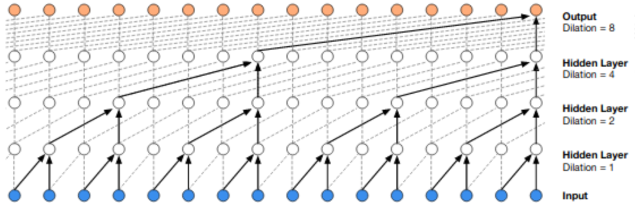


FIGURE – Van Den Oord et Al. 2016

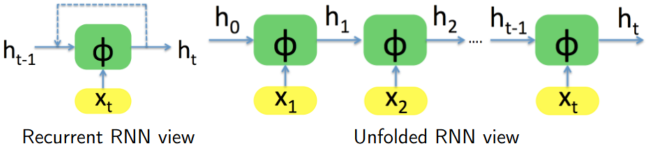
Convolutional Neural Networks

- Plus stable, meilleur estimation locale et plus compact que FCN

- Taille des entrées limités
- Décisions indépendantes entre différents pas
- Ne peut pas prendre des séquences de taille variable

Recurrent NN

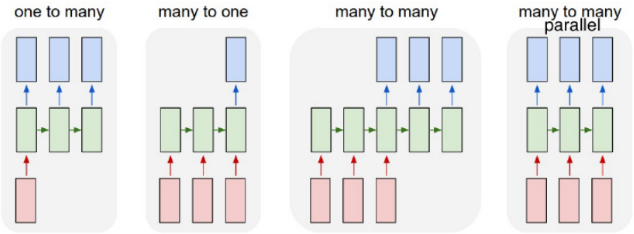
- Entrée : $\{x_t\}_{0\leq t\leq N}$
- Hidden state : $\{h_t\}_{0\leq t\leq N}$
- RN cell : $h_t = \Phi_t(x_t, h_{t-1})$ (récursif)
- $\phi_t = \phi$ les paramètres sont partagés dans le temps



RNN

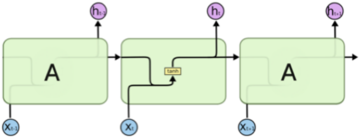
Différentes correspondances (*mapping*) pour différentes tâches

- many-to-one : prédiction
- one-to-Many : image captioning
- many-to-many parallel : traduction, sous titrage
- many-to-many : classification



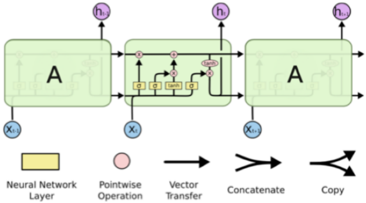
LSTM

- Recap: Vanilla RNN cell



$$h_t = \tanh\left(W\begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}\right)$$

- Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997]



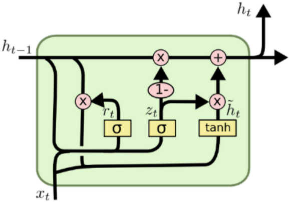
$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$

$$c_t = f \odot c_{t-1} + i \odot g$$

$$h_t = o \odot \tanh(c_t)$$

Credit: C. Olah

GRU



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

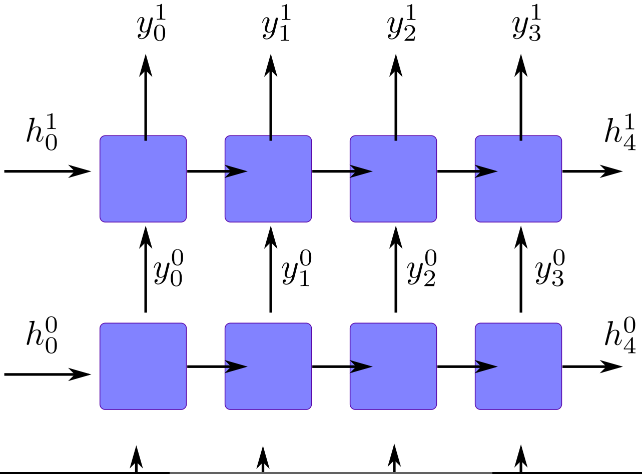
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

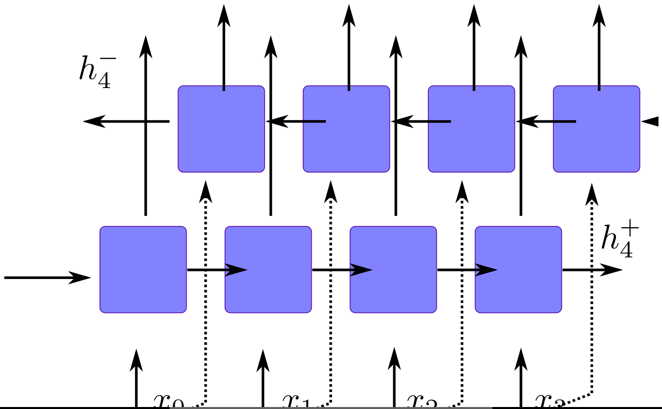
Multi-layers RNN

- Comme pour les CNN, il est possible d'empiler des block récurrents
- La sortie d'un bloc à un niveau est transmise en entrée du bloc correspondant au niveau suivant



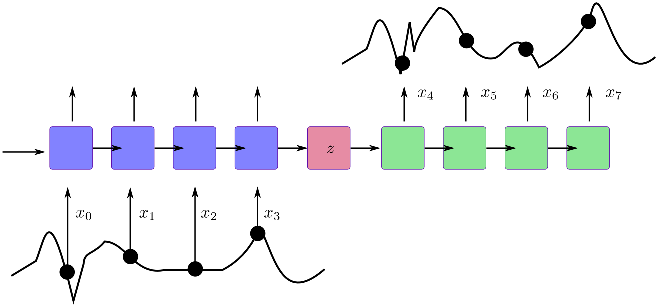
Forward and Backward RNN

- Plus de contexte peut être obtenu si le réseau parcourt la séquence dans les deux sens
- On a alors deux vecteur contextuels h^+ et h^-



seq2seq

- L'architecture en *encoder-decoder* est généralisable aux RN
- L'encodage se fait à travers le vecteur caché transmis à chaque étape de lecture de la séquence.
- Le décodage est effectué à partir du contexte global
- Dans une optique de prédiction, la dernière entrée peut aussi servir d'entrée au décodeur



seq2seq

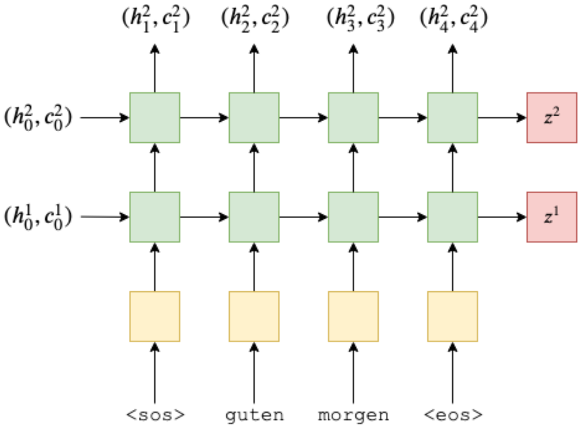


FIGURE – Ref : https ://github.com/bentrevett/pytorch-seq2seq

seq2seq

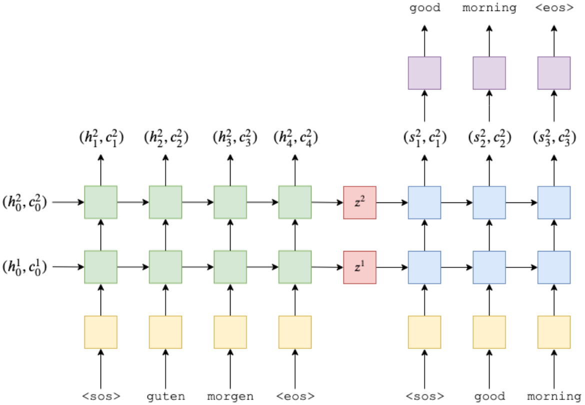


Figure – Ref : https ://github.com/bentrevett/pytorch-seq2seq

Plan

Attention seq2seq

Problème des seq2seq :

- Compression d'information très contraignante
- Difficile de trop augmenter la taille du vecteur caché

Attention

- Utiliser touts les états cachés
- Pondérer leur utilisation en fonction de leur utilité

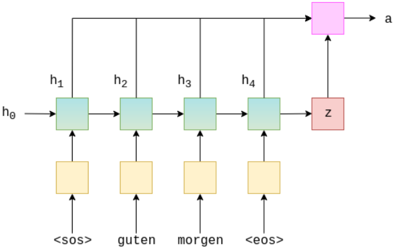


Figure – Ref : https://github.com/bentrevett/pytorch-seq2seq

Attention seq2seq

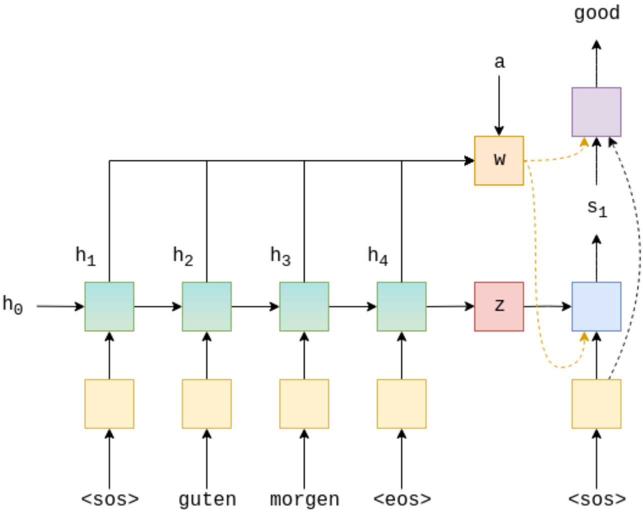


FIGURE – Ref : https ://github.com/bentrevett/pytorch-seq2seq

Transformers

- Architecture complètement connectée
- Toute l'information est mise en correspondance avec elle même
- Permet de repondérer les données en fonction de leur redondance et intérêt pour une tâche donnée