MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/∼ pam

Aula 16

16 de maio de 2024

Sumário

1. ML e Séries Temporais
2. ML e Classificação de ST

Modelos lineares e não lineares em ML

- Métodos Supervisionados em ML incluem dois grupos:
  - i) Modelos lineares, estimação por regularização (ridge, lasso, elastic nets etc).
  - ii) Modelos não lineares: RN feedforward (shallow and deep), florestas aleatórias e boosting, CNN e GAN.
- Outros métodos:
  - iii) Ensemble-based methods: complete subset regression, bagging.
  - iv) Hybrid methods: combinam modelos lineares e não lineares.

Modelos para Séries Temporais

Tarefas da Análise de Séries Temporais (ST):

- Previsão de ST.
- Agrupamentos (Clustering).
- Tarefas de reconhecimento em medicina, indústria, verificação de assinaturas, detecção de doenças do cérebro.
- Classificação de ST

Modelos para Séries Temporais

- Considere a uma ST multivariada $\mathbf{x}_t, t = 1, \ldots, T$ em que $\mathbf{x}_t$ contém valores de d variáveis e uma ST univariada Y_t, $t = 1, \ldots, T$.
- O objetivo é prever Y_t, para horizontes $h = 1, \ldots, H$, com base em valores de $\mathbf{x}_t$(e possivelmente em valores passados de Y_t).
- Uma suposição básica é que o processo $\{Y_t, \mathbf{x}_t\}$, $t \geq 1$, é fracamente estacionário com valores em $\mathbb{R}^{d+1}$.
- O modelo a ser considerado é

$$Y_t = f(\mathbf{x}_t) + e_t, \quad t = 1, \ldots, T. \tag{1}$$

Aqui, f é uma função desconhecida, $E(e_t) = 0$ e $\text{Var}(e_t) < \infty$.

Modelos para Séries Temporais

- Queremos estimar f e usar o modelo para fazer previsões h passos à frente,

$$Y_{t+h} = f(\mathbf{x}_t) + e_{t+h}, \quad h = 1, \ldots, H, \ t = 1, \ldots, T.$$

- Para avaliar o método de previsão, as medidas mais usadas são o EQMP (erro quadrático médio de previsão) e o EAMP(erro absoluto médio de previsão)
- Escolhendo-se uma função perda $L(f, \hat{f})$, o objetivo é selecionar f de um conjunto de modelos que minimiza o risco $E[L(f, \hat{f})]$.

Classificação de ST

- O problema da classificação de ST pode ser definido como segue: Dado um conjunto de classes $\mathcal{C}$, um conjunto de treinamento $\mathcal{T}$ de ST X_t, associadas com sua classe de rótulos $y(X_t) \in \mathcal{C}$, isto é, , $\mathcal{T} = \{(X_1, y(X_1), \ldots, (X_m, y(X_m))\}$, o objetivo é encontrar uma função f (o **classificador** ou **modelo**) tal que $f(X) = y(X)$, para alguma ST $X \notin \mathcal{T}$.
- A semântica de classes varia de aplicação para aplicação:
  i) no caso de diagnóstico de uma doença,uma classe positiva e uma negativa podem ser consideradas; uma ST registrada para pacientes afetados pela doença pertencem à classe positiva, enquanto uma ST de uma pessoa não diagnosticada com a doença pertence à classe negativa;
  ii) no caso de verificação de uma assinatura ou identificação de uma pessoa, as classes podem corresponder a pessoas diferentes.
- A função f pode ser estimada de várias maneiras, eg, via RN, SVM ou árvores. Em qualquer caso, o processo de encontrar a f apropriada é chamado **treinamento**.
- Para avaliar o erro de classificação, este é medido num novo conjunto de ST, chamado de **conjunto de teste**.

Classificação de ST

- Medidas de qualidade:
  i) Acurácia (a taxa de casos corretamente classificados);
  ii) F-measure (a média harmônica de precisão e recall)
  iii) Área sob a curva ROC;
  iv) Área sob a curva precision-recall.
- Atenção em dividir os dados, VC pode ser necessária.

J. C. Prandini, P. A. Morettin and C. Chiann (2024). The area under Normal ROC curves. To appear, SPJMS.

Técnicas de Classificação de ST

- Classificação baseada em características (feature-based):
  - i) extração de características (feature extraction) + técnicas usuais (SVM, Bayes, decision trees)
  - ii) possíveis características: min, max, média, dp, número de mudanças de sinais, etc; distâncias de outras ST.
- Deep convolutional neural networks. A. Krizhevsky et al. (2012) (The Alex Net)
- Similarity-based classification (nearest neighbors and extensions, Dynamic Time Warping (DTW)). Sakoe and Chiba (1978), Buza (2018).

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, Volume ASSP-26, 43–49.

Buza, K. (2018). Time series classification and its applications. In 8th International Conference on Web Intelligence, Mining and Semantics.

Preprocessamento

- Na maioria das aplicações, a ST original é preprocessada; o objetivo é transformar os dados a uma forma que permita resolver o problema de classificação de forma acurada. Por exemplo, no caso de amostragem irregular de ST, usar técnicas de interpolação.
- As técnicas mais usadas de preprocessamento:
  i) **Transformar a ST para o domínio da frequência ou escala** (Transformada de Fourier ou de ondaleta). Isso pode destacar propriedades relevantes. Por exemplo, em reconhecimento de falas (discursos), o sinal correspondente a sim ou não pode ser distinguido baseado em suas transformadas de Fourier devido à presença ou ausência decomponentes de baixa e alta frequências correspondentes às vozes de s e ã.
  ii) **Médias de observações consecutivas** e associar um valor discreto (simbólico) a elas reduz o ruído e o comprimento da ST, reduzindo recursos computacionais. Essa técnica é chamada **symbolic aggregate approximation** (SAA).
  iii) **Mudança de valor** pode ser mais descritivo que o próprio valor, eg, dado $X = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, considere $X' = \{(x_2 - x_1, y_2 - y_1), \ldots, (x_n - x_{n-1}, y_n - y_{n-1}))\}$. Isso corresponde a derivar a ST ao longo do tempo.
  iv) **Alinhamento**

Classificadores para ST

Classificação de ST é desafiadora devida a vários fatores:

i) Observações subsequentes são correlacionadas, o que não é modelada pelos classificadores convencionais, tais como regressão logística, Bayes ou SVM.

ii) O comprimento de uma ST é, usualmente, não uniforme, eg, a escrita ou comprimento de uma série de batimentos cardíacos em mlseg varia. Em contraste, os classificadores acima mencionados supõem que o número de característica de entrada (preditores) é fixo.

iii) Uma maneira de contornar essas dificuldades é extrair um número fixo de características (tais como mínimo, máximo, média, dp) para cada ST e usar essas características num classificador convencional.

iv) Características úteis podem ser usadas na presença ou ausência de padrões locais ou semelhanças a alguma(s) ST selecionada(s).

v) Trabalhos recentes em classificação de ST são baseados em Deep Learning (DL). Enquanto CNN têm um bom desempenho em termos de acurácia de classificação, conjuntos de treinamento muito grandes (da ordem de milhões de imagens) podem ser necessários e pode ser difícil interpretar o modelo resultante (e.g. ImageNet contest).

DL - CNN

- 2012-2015 classification results in ImageNet contest:
  - 2012: Supervision (Toronto, AlexNet): % error: 15.3 (CNN, 7 layers)
  - 2013: Clarifai (NYU): % error: 11.7 (CNN)
  - 2014: GoogLeNet : % error: 6.6 (CNN)
  - 2015: ResNet: % error: under 3.5 (CNN, 150 layers)
- AlexNet: there were about 1.2 million training images, 50 thousand validation images, and 150 thousand testing images. Classification into 1000 classes. Training took 6 days.

Dynamic Time Warping- DTW

- Distâncias não euclidianas: edit, Hamming etc
- DTW é usada em classificadores recentes de ST. Meszlényi et al, 2017.
- DTW é uma medida de similaridade entre ST e foi introduzida na literatura por Vintsyuk (1968) e Sakoe and Chiba (1978), em ambos os casos em aplicações com falas.
- Enquanto kNN é intuitivo em espaços vetoriais, em princípio pode ser aplicado a qualquer tipo de dados. 1-nearest neighbor (1-NN) com DTW é um classificador competitivo. Xi et al. (2006).
- O único requisito é que uma **medida de distância** apropriada seja definida e que possa ser usada para detrminar os casos de treinamento mais similares.
- No caso de ST os casos são ST e uma das distâncias mais usadas é DTW.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, Volume ASSP-26, 43–49.

Vintsyuk, T. K. (1968). Speech discrimination by dynamic programming. *Cybernetics*, **4**, 52–57.

Dynamic Time Warping

- Usualmente, ao calcular a distância entre duas ST X_1 e X_2, comparamos o $i$ésimo elemento de X_1 com o i-ésimo elemento de X_2 e agregamos os resultados de tal comparação.
- Na realidade, quando observamos o mesmo fenômenos várias vezes, não esperamos que ele ocorra sempre na mesma posição do tempo, e a duração do evento pode variar um pouco.
- DTW captura a similaridade das formas de duas ST de tal maneira que permita **alongamentos**: a i-ésima posição de X_1 é comparada com a $i^{'}$-ésima posição de X_2 e $i^{'}$ pode ou não ser igual a i.
- DTW usa a distância edit : isso significa que, conceitualmente, podemos considerar calcular a distância DTW de duas ST X_1 e X_2, de comprimentos T_1 e T_2, transformando X_1 em X_2. Sakoe and Chiba, 1978.

Dynamic Time Warping

- DTW usa o enfoque de **programação dinâmica** (*dynamic programming*). O cálculo do custo da transormação mínima é feito preenchendo-se as entradas de uma matriz $T_1 \times T_2$.
- Cada entrada corresponde à distância entre uma subsequência de X_1 e uma subsequência de X_2
- Dois passos são possíveis, ambos associados a um custo: **substituição** (*replacement*) e **alongamento** (*elongation*).
- O custo de transformar X_1 em X_2 é a soma dos custos de todos os passos possíveis.
- Há muitas possibilidades de transformar X_1 em X_2. DTW calcula aquela com custo mínimo.
- Para detalhes, veja Buza (2018).
- Pacote dwt do R.

Formulação do problema

- O problema de otimização é (Tavenard, 2021):

$$DTW_q(\mathbf{x}, \mathbf{y}) = \min_{\pi \in \mathcal{A}(\mathbf{x},\mathbf{y})} \left(\sum_{(i,j)\in\pi} d(\mathbf{x}_i, \mathbf{y}_i)^q \right)^{1/q}. \quad (2)$$

  Aqui um **caminho de alinhamento** (*alignment path*) π de comprimento K é uma sequência de K pares de índices $((i_0, j_0), \ldots, (i_{K-1}, j_{K-1}))$ e $\mathcal{A}(\mathbf{x}, \mathbf{y})$ é o conjunto de todos os caminhos admissíveis.

- Para ser considerado admissível, um par deve satisfazer as condições:
  - i) começo (resp. fim) da ST são coincidentes:
    $\pi_0 = (0, 0)$;
    $\pi_{K-1} = (n-1, m-1)$.
  - ii) A sequência é monotônica crescente em i e j e todos os índices no tempo devem aparecer pelo menos uma vez, isto é,
    $i_{k-1} \leq i_k \leq i_{k-1} + 1$,
    $j_{k-1} \leq j_k \leq j_{k-1} + 1$

Tavenard, R. (2021). An introduction do Dynamic Time Warping. In https://rtavenar.github.io/blog/dtw.html.

Formulação do problema

- Outra maneira de representar um caminho DWT é usar uma matriz binária, cujas entradas não nulas são aquelas correspondentes a uma coincidência entre elementos da ST.
- Esta representação é relacionada à representação da sequência de índices acima por meio de:

$$(A_\pi)_{i,j} = \begin{cases} 1, \text{ if } (i,j) \in \pi, \\ 0, \text{ caso contrário.} \end{cases} \tag{3}$$

  Isso é ilustrado na Figura 1, na qual entradas não nulas na matriz binária são representadas por pontos e a sequência equivalente de coincidências é produzida à direita.
- Usando notação matricial, DWT pode ser escrita como a minimização de um produto interno entre matrizes:

$$DTW_q(\mathbf{x}, \mathbf{y}) = \min_{\pi \in \mathcal{A}(\mathbf{x},\mathbf{y})} \langle A_\pi, D_q(\mathbf{x}, \mathbf{y}) \rangle^{1/q}, \tag{4}$$

  na qual $D_q(\mathbf{x}, \mathbf{y})$ armazena as distâncias $d(\mathbf{x}_i, \mathbf{y}_j)^q$.

Dynamic Time Warping

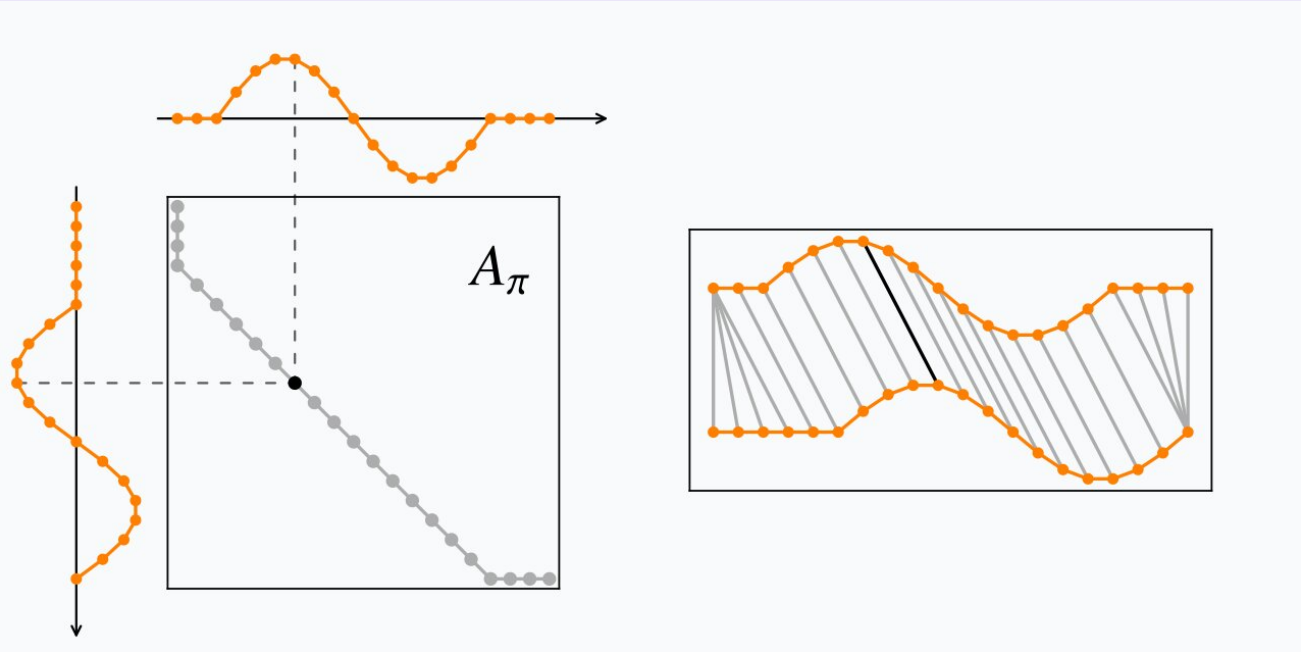

Figura 1: Caminho DTW representado como matriz binária (esquerda). Cada ponto no caminho indica uma entrada não nula em A_π, logo a coincidência entre um elemento de **x** com um elemento de **y**. Note a correspondência entre esta representação e aquela à direita.

DTW: Solução algoritmica

- Se bem que a otimização em (4) seja uma minimização sobre um conjunto finito, o número de caminhos possíveis (Delannoy number) torna-se muito grande mesmo para comprimentos de ST moderados. Supondo que n e m sejam da mesma ordem, existem $O\left(\frac{(3+2\sqrt{2})^n}{\sqrt{n}}\right)$ diferentes caminhos em $\mathcal{A}(\mathbf{x}, \mathbf{y})$, o que torna o problema intratável ao listar todos os caminhos sequencialmente para calcular o mínimo.
- Felizmente, existe uma solução exata do problema de otimização, e este pode ser obtido usando adynamic programming. Esta consiste em ligar a solução de um problema a soluções de sub-problemas mais fáceis e guardando as soluções para uso posterior
- No caso de DTW, nos baseamos em:

$$R_{i,j} = DTW_q(\mathbf{x}_{\rightarrow i}, \mathbf{y}_{\rightarrow j})^q, \tag{5}$$

  na qual a notação $\mathbf{x}_{\rightarrow i}$ denota a ST $\mathbf{x}$ observada até o instante i (inclusive).
- Então, podemos observar que

$$\begin{aligned} R_{i,j} \quad = \quad & \min_{\pi \in \mathcal{A}(\mathbf{x}_{\rightarrow i}, \mathbf{y}_{\rightarrow j})} \sum_{(k,\ell)\in\pi} d(\mathbf{x}_k, \mathbf{y}_\ell)^q = \\ & d(\mathbf{x}_i, \mathbf{y}_j)^q + \min_{\pi \in \mathcal{A}(\mathbf{x}_{\rightarrow i}, \mathbf{y}_{\rightarrow j})} \sum_{(k,\ell)\in\pi[:-1]} d(\mathbf{x}_k, \mathbf{y}_\ell)^q = \\ & d(\mathbf{x}_i, \mathbf{y}_j)^q + \min(R_{i-1,j} R_{i,j-1}, R_{i-1,j-1}). \end{aligned} \tag{6}$$

Dynamic Time Warping

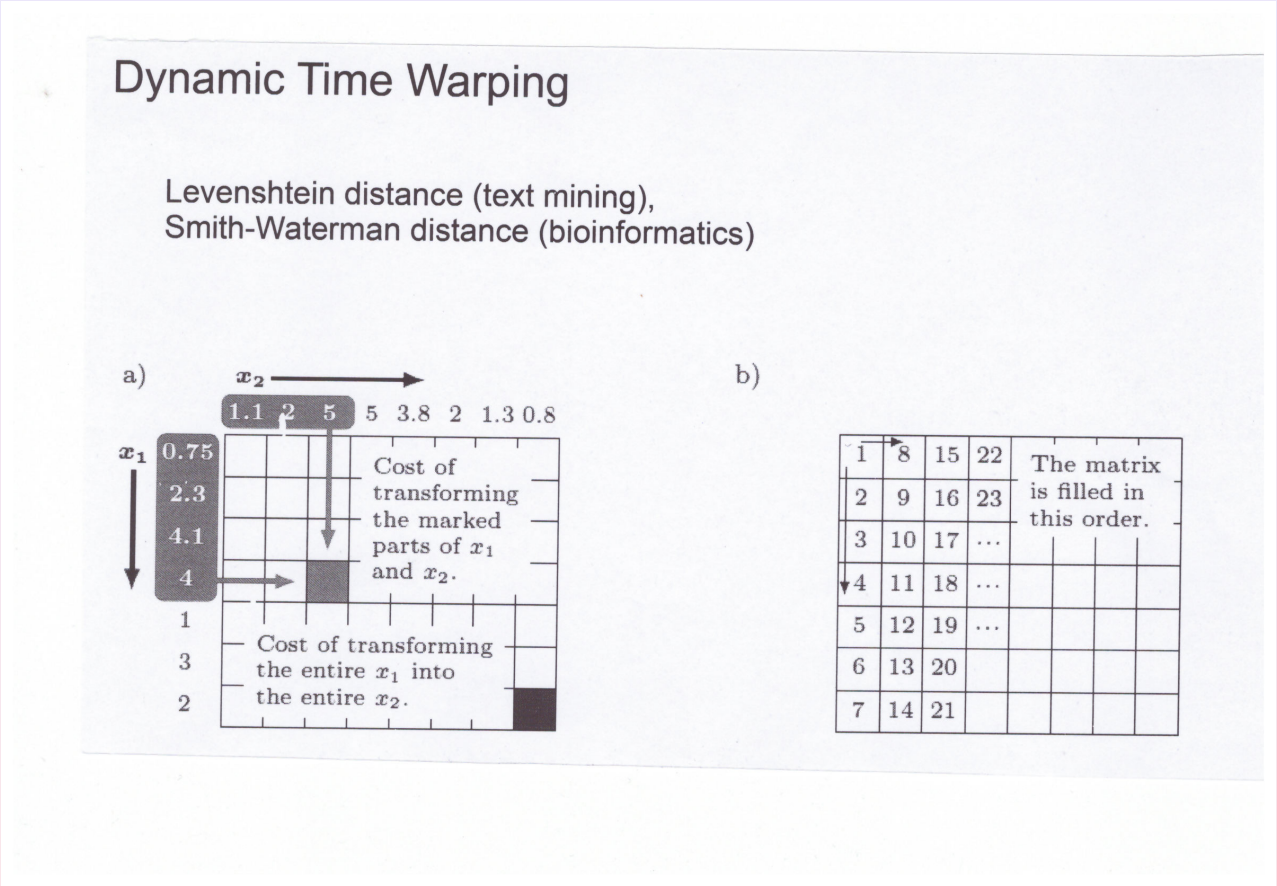

Figure 15: DWT.

Dynamic Time Warping

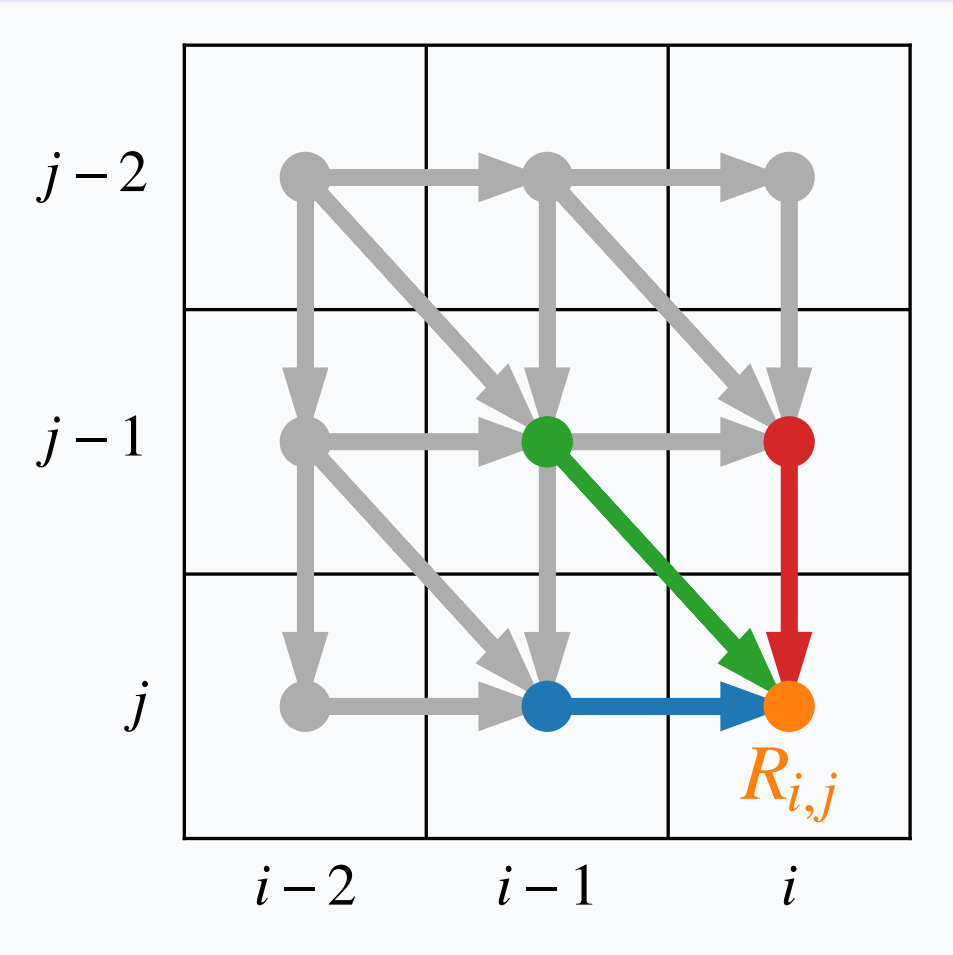

Figure 16: Transições em DWT.

Dynamic Time Warping

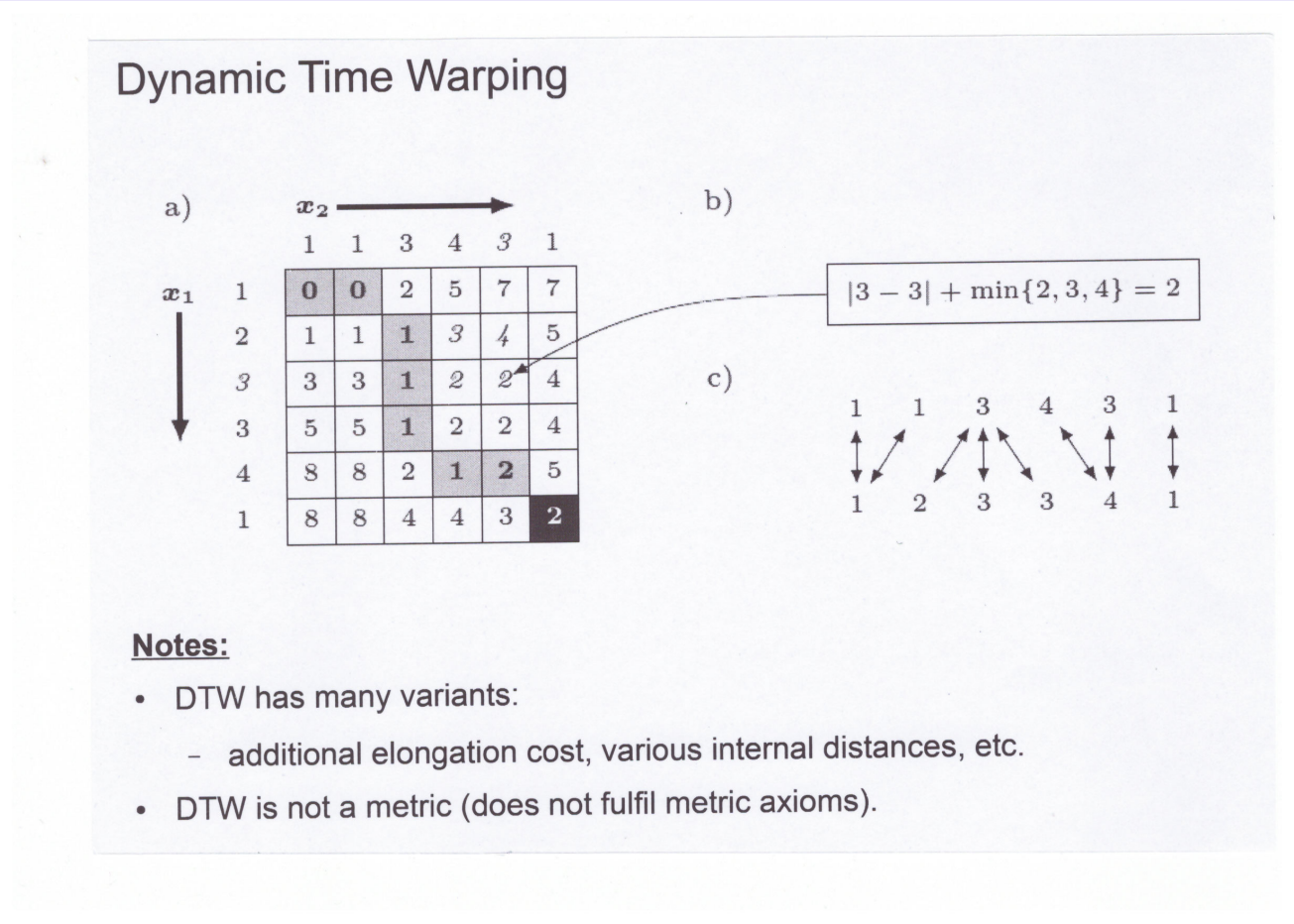

Figure 17: Cálculo da DWT.

Dynamic Time Warping

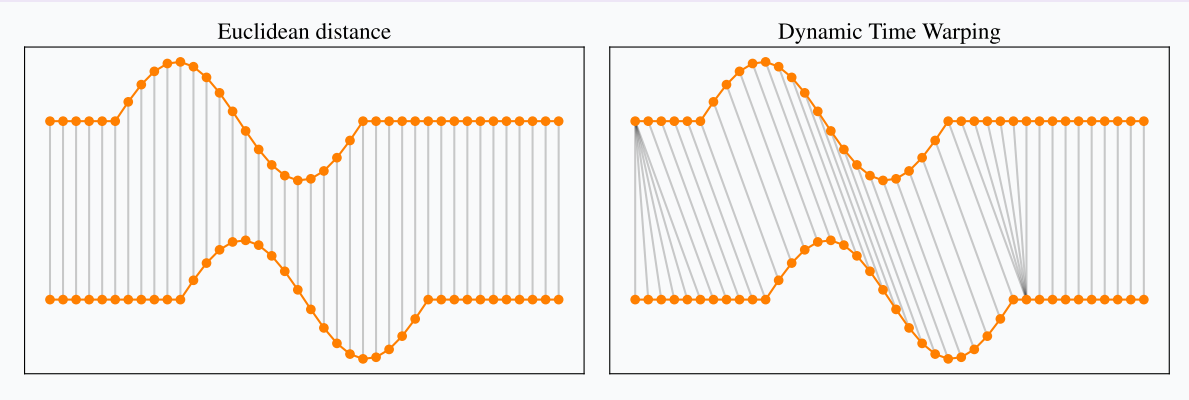

Figure 18: DTW versus Euclidean.

Clustering: Edit vs Euclidean

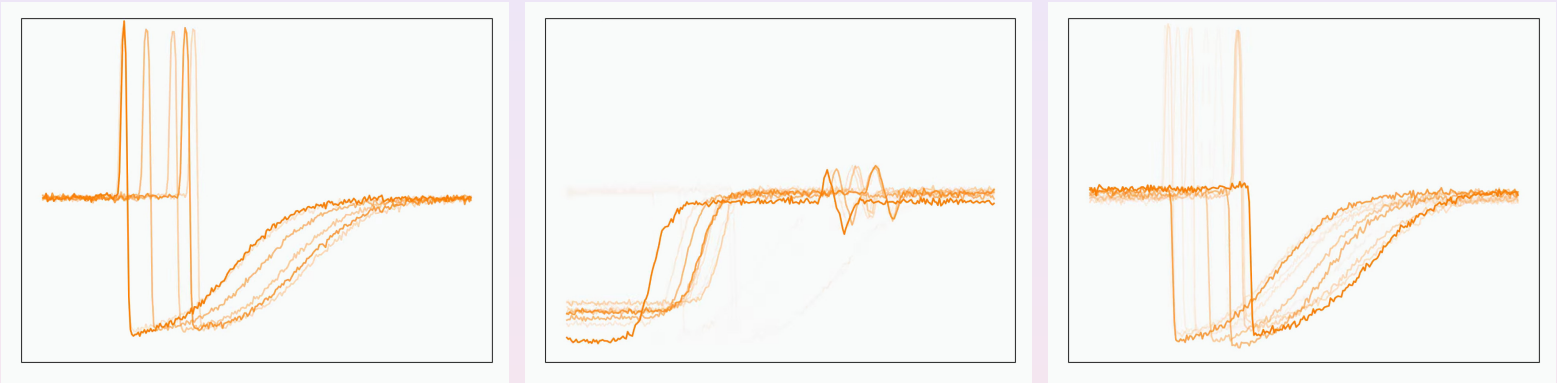

Figure 19: Three types of generated time series.

Clustering: Edit vs Euclidean

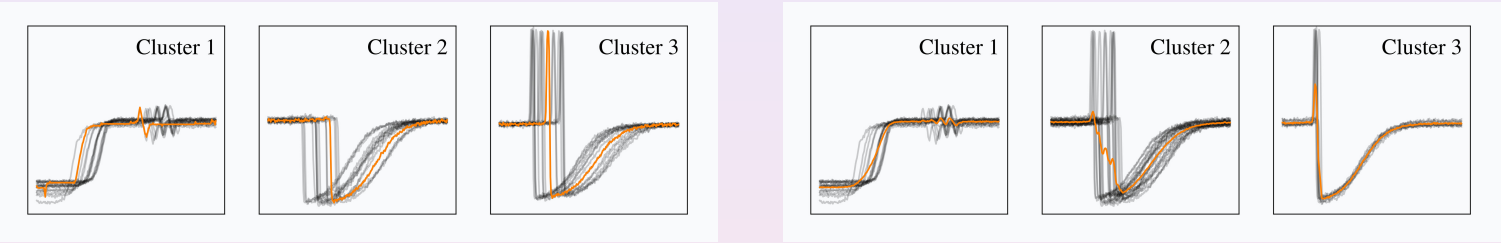

Figure 20: Clustering: Edit versus Euclidean.

Alternative approaches

Apesar das evidências a favor de classificadores 1-NN com distências Euclidiana ou DTW , há pesquisas sobre enfoques alternativos:

- Shapelets. Introduzidas por Ye and Keogh (2009, 2011); Shapelet: uma subsequência (ST) que é identificada como representativa da filiação a uma classe.
- DTW ponderada.
- Bagging, boosting ou florestas construídas sobre sumários estatísticos.
- Fusão de medidas de distâncias alternativas.

Ye, L. and Keogh, E. (2010). Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. *Data Mining and Knowledge Discovery*, **22**, 149–182.

Referências

Hastie, T., Tibshirani, R. and Friedman, J. (2017). *The Elements of Statistical Learning*, 2nd Edition, Springer.

Giorgino, T. (2009). Computing and visualizing DTW alignments in R: The dtw Package. *Journal of Statistical Software*, **31**, 1–24.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.

Masini, R.P., Medeiros, M.C. and Mendes, E.F. (2021). Machine learning advances for time series forecasting. *Journal of Economic Surveys*, 1-36, Wiley. DOI: 10.1111/joes.12429.

Morettin, P. A. e Singer, J. M. (2022). *Estatística e Ciência de Dados*. LTC: Rio de Janeiro.