Soft-DTW: a Differentiable Loss Function for Time-Series

Jingchang Liu

July 20, 2017

University of Science and Technology of China

Table of Contents

A brief introduction to DTW(Dynamic time warping)

Background

- Application: discrepancy of two time series.
- Motivation: Align different indexes of two time series. The length of the sequence can different.

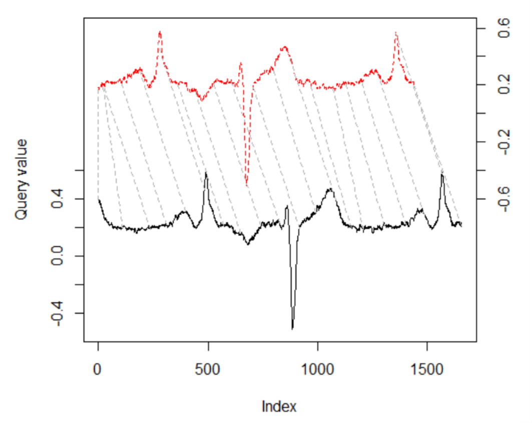

Figure 1: Diagram of DTW

Warping path

- $x = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^{p \times n}$, $y = (y_1, y_2, \cdots, y_m) \in \mathbb{R}^{p \times m}$
- Warping path: $W = w_1 w_2 \cdots w_N$, $w_k = (i, j)$
- $w_k = (i, j) \rightarrow w_{k+1} = (i+1, j)$ or $w_{k+1} = (i, j+1)$ or $w_{k+1} = (i+1, j+1)$

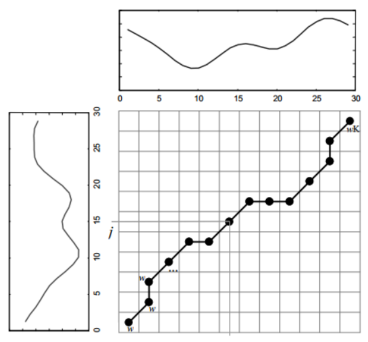

Figure 2: Diagram of warping path

Formulations

- Cost matrix: $\triangle(x, y) := [\delta(x_i, y_j)]_{ij} \in \mathbb{R}^{n\times m}$
- Set of binary alignment matrices: $\mathcal{A}_{n,m} \subset \{0,1\}^{n\times m}$
- Formulation: $DTW(x, y) := \min_{A\in\mathcal{A}_{n,m}} < A, \triangle(x, y) >$
- $r_{ij} = DTW(x_{1:i}, y_{1:j})$

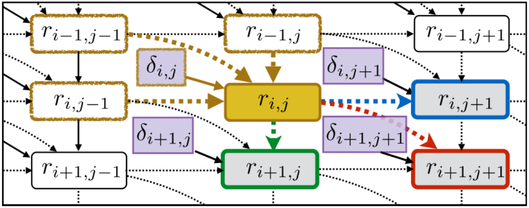

Figure 3: Diagram of iteration

Iteration

$r_{i,j} := \delta_{i,j} + \min\{r_{i-1,j-1}, r_{i-1,j}, r_{i.j-1}\}$

Algorithm 1 Calculate DTW

Require: $x = (x_1, x_2, \cdots, x_n)$, $y = (y_1, y_2, \cdots, y_m)$

Ensure: $r \in \mathbb{R}^{n \times m}$

```
r := [0..n, 0..m]
for i := 1 to n do
    for j := 1 to m do
        r_{i,j} := δ_{i,j} + min(r_{i,j}, r_{i,j−1}, r_{i−1,j−1})
    end for
end for
```

Warp window

- alignment path is unlikely to very far from the diagonal.

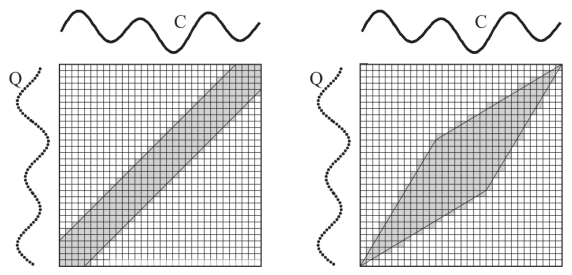

Figure 4: Diagram of warp window

Soft-DTW

Background

Averaging

- $y_1, \cdots, y_n$: A family of N times series , m_i: length of y_i, x: a single barycenter time seies.
- $\min_{x \in \mathbb{R}^{p \times n}} \sum_{i=1}^{N} \frac{\lambda_i}{m_i} DTW\left(x, y_i\right)$

Time series prediction

- $\min_{\theta \in \Theta} \sum_{i=1}^{N} DTW\left(f_\theta\left(x_i^{1,t}\right), x_i^{t+1,n}\right)$
- f_θ: Prediction function, $x_i^{t+1,n}$: prediction sequence,

Generalized min operator

- Generalized min operator:

$$\min{}^{\gamma}\left\{a_1, a_2, \cdots, a_n\right\} := \begin{cases} \min_{i \leq n} a_i, \gamma = 0, \\ -\gamma \log \sum_{i=1}^{n} e^{-a_i/\gamma}, \gamma > 0 \end{cases}$$

- $$DTW_\lambda\left(x, y\right) = \min{}^{\gamma}\left\{< A, \triangle\left(x, y\right) >, A \in \mathcal{A}_{n,m}\right\}$$

- $$\nabla_x DTW_\gamma\left(x, y\right) = \left(\frac{\partial \triangle\left(x, y\right)}{\partial x}\right)^T \frac{\partial DTW_\lambda\left(x, y\right)}{\partial \triangle\left(x, y\right)}$$

- $$e_{i,j} := \frac{\partial r_{n,m}}{\partial r_{i,j}}$$

- $$\frac{\partial r_{n,m}}{\partial \delta_{i,j}} = \frac{\partial r_{n,m}}{\partial r_{i,j}} \frac{\partial r_{i,j}}{\partial \delta_{i,j}} = e_{i,j}$$

Dynamic programming to derivative

Chain rule

$$\underbrace{\frac{\partial r_{n,m}}{\partial r_{i,j}}}_{e_{i,j}} = \underbrace{\frac{\partial r_{n,m}}{\partial r_{i+1,j}}}_{e_{i+1,j}} \frac{\partial r_{i+1,j}}{\partial r_{i,j}} + \underbrace{\frac{\partial r_{n,m}}{\partial r_{i,j+1}}}_{e_{i,j+1}} \frac{\partial r_{i,j+1}}{r_{i,j}} + \underbrace{\frac{\partial r_{n,m}}{\partial r_{i+1,j+1}}}_{e_{i+1,j+1}} \frac{\partial r_{i+1,j+1}}{\partial r_{i,j}}$$

Calculate $\frac{\partial r_{i+1,j}}{\partial r_{i,j}}$

1.

$$\begin{aligned} r_{i+1,j} &= \delta_{i+1,j} + \min{}^{\lambda} \{r_{i,j-1}, r_{i,j}, r_{i+1,j-1}\} \\ &= \delta_{i+1,j} - \gamma \log \left(e^{-r_{i,j-1}/\gamma} + e^{-r_{i,j}/\gamma} + e^{-r_{i+1,j-1}/\gamma}\right) \end{aligned}$$

2.

$$\frac{\partial r_{i+1,j}}{\partial r_{i,j}} = e^{-r_{i,j}/\gamma} / \left(e^{-r_{i,j-1}/\gamma} + e^{-r_{i,j}/\gamma} + e^{-r_{i+1,j-1}/\gamma}\right)$$

Derivation

- $$\begin{aligned}\gamma \log \frac{\partial r_{i+1,j}}{\partial r_{i,j}} &= \min^{\gamma}\{r_{i,j-1}, r_{i,j}, r_{i+1,j-1}\} - r_{i,j} \\ &= r_{i+1,j} - r_{i,j} - \delta_{i+1,j}\end{aligned}$$
- $$\gamma \log \frac{\partial r_{i,j+1}}{\partial r_{i,j}} = r_{i,j+1} - r_{i,j} - \delta_{i,j+1}$$
- $$\gamma \log \frac{\partial r_{i+1,j+1}}{\partial r_{i,j}} = r_{i+1,j+1} - r_{i,j} - \delta_{i+1,j+1}$$

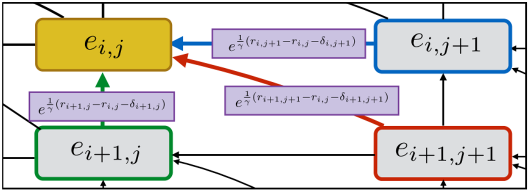

Figure 5: Iteration of derivative

Inputs: $\mathbf{x}, \mathbf{y}$, smoothing $\gamma \geq 0$, distance function δ.
$\Delta = [\delta(x_i, y_j)]_{i,j}$.
$r_{0,0} = 0; r_{i,0} = r_{0,j} = \infty; i \in [\![n]\!], j \in [\![m]\!]$.
for $j = 1, \ldots, m$ **do** ▷ Forward recursion
  for $i = 1, \ldots, n$ **do**
    $r_{i,j} = \delta_{i,j} + \min^{\gamma}\{r_{i-1,j-1}, r_{i-1,j}, r_{i,j-1}\}$
  end for
end for
$\delta_{i,m+1} = \delta_{n+1,j} = 0, i \in [\![n]\!], j \in [\![m]\!]$
$e_{i,m+1} = e_{n+1,j} = 0, i \in [\![n]\!], j \in [\![m]\!]$
$r_{i,m+1} = r_{n+1,j} = -\infty, i \in [\![n]\!], j \in [\![m]\!]$
$\delta_{n+1,m+1} = 0, e_{n+1,m+1} = 1, r_{n+1,m+1} = r_{n,m}$
for $j = m, \ldots, 1$ **do** ▷ Backward recursion
  for $i = n, \ldots, 1$ **do**
    $a = \exp \frac{1}{\gamma}(r_{i+1,j} - r_{i,j} - \delta_{i+1,j})$
    $b = \exp \frac{1}{\gamma}(r_{i,j+1} - r_{i,j} - \delta_{i,j+1})$
    $c = \exp \frac{1}{\gamma}(r_{i+1,j+1} - r_{i,j} - \delta_{i+1,j+1})$
    $e_{i,j} = e_{i+1,j} \cdot a + e_{i,j+1} \cdot b + e_{i+1,j+1} \cdot c$
  end for
end for
Output: $\mathbf{dtw}_{\gamma}(\mathbf{x}, \mathbf{y}) = r_{n,m}$
$\nabla_{\mathbf{x}}\, \mathbf{dtw}_{\gamma}(\mathbf{x}, \mathbf{y}) = \left(\frac{\partial \Delta(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}}\right)^T E$

Figure 6: Computes $DTW_{\gamma}(x, y)$ and $\nabla_x DTW_{\gamma}(x, y)$

Experiments

Averaging experiment

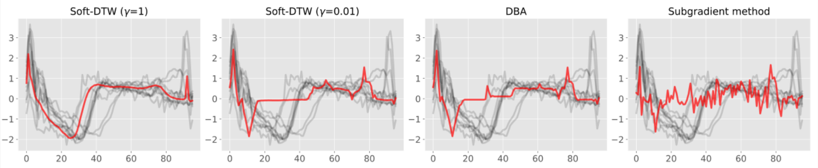

Figure 7: Comparison between our proposed soft barycenter and the barycenter obtained by DBA and the subgradient method, on the ECG200 dataset

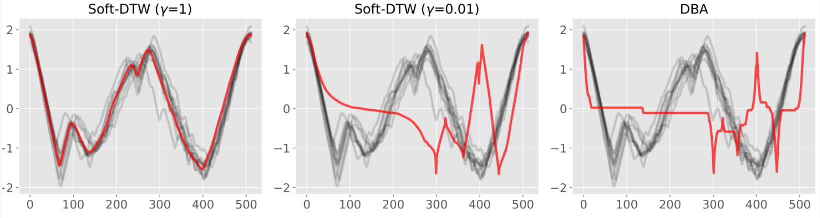

Figure 8: Herring

Averaging experiment II

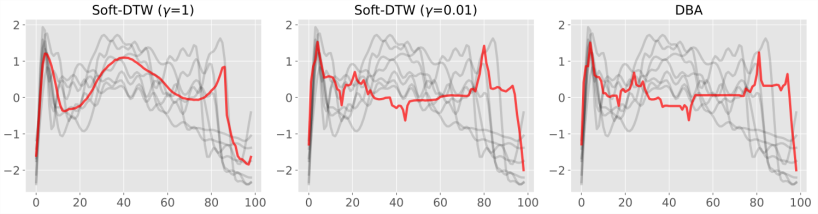

Figure 9: Medical Images

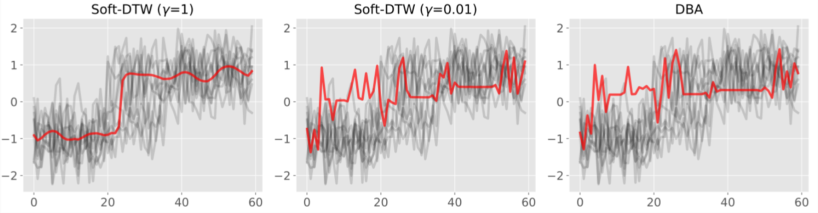

Figure 10: Synthetic Control

Prediction experiments

Dataset	Soft-DTW loss $\gamma = 1$	$\gamma = 0.1$	$\gamma = 0.01$	$\gamma = 0.001$	Euclidean loss
50words	6.473	**4.921**	4.999	6.489	18.734
Adiac	0.094	**0.074**	0.078	0.109	0.103
ArrowHead	1.851	**1.708**	1.933	1.909	2.073
Beef	12.229	**8.688**	10.244	9.126	22.228
BeetleFly	35.037	25.439	27.588	**23.494**	50.610
BirdChicken	31.878	19.914	25.100	**14.981**	30.693
CBF	10.802	**9.263**	9.595	10.151	12.868
Car	1.724	2.307	2.202	**1.318**	1.588
ChlorineConcentration	7.876	2.108	2.331	1.735	**0.769**
CinC_ECG_torso	45.675	26.337	**23.567**	24.550	48.171
Coffee	0.914	0.727	1.662	1.883	**0.660**
Computers	92.584	84.723	78.953	**75.435**	235.208
Cricket_X	9.394	8.042	**7.123**	7.226	12.080
Cricket_Y	11.989	9.643	**9.534**	9.545	15.002
Cricket_Z	9.161	6.889	**6.585**	7.200	11.003
DiatomSizeReduction	1.182	0.922	**0.820**	0.897	1.203
DistalPhalanxOutlineAgeGroup	0.426	0.291	0.541	0.496	**0.231**
DistalPhalanxOutlineCorrect	0.494	0.476	0.564	0.591	**0.351**
DistalPhalanxTW	0.441	0.330	0.305	1.214	**0.231**
ECG200	1.874	**1.716**	1.884	1.734	1.905
ECG5000	4.895	4.705	4.543	**4.441**	5.463
ECGFiveDays	1.834	1.944	1.699	**1.642**	2.220
Earthquakes	74.738	59.973	60.877	**57.827**	147.980
ElectricDevices	20.186	**15.125**	15.218	15.287	37.121
FISH	0.464	0.429	**0.354**	0.459	0.462
FaceAll	9.317	7.451	7.902	**7.276**	10.716
FaceFour	**19.564**	20.881	28.150	28.839	46.841
FacesUCR	15.359	**14.643**	16.143	17.428	28.576
Gun_Point	0.896	**0.805**	0.923	0.834	0.858
Ham	20.154	17.931	17.786	**17.413**	24.340

Figure 11: Time-series prediction: DTW loss achieved when using random init

Prediction experiments

Dataset	Soft-DTW loss $\gamma = 1$	$\gamma = 0.1$	$\gamma = 0.01$	$\gamma = 0.001$	Euclidean loss
50words	6.330	5.628	4.885	**4.553**	18.734
Adiac	0.082	0.076	**0.064**	0.079	0.103
ArrowHead	1.823	2.016	**1.762**	2.106	2.073
Beef	7.250	6.940	7.146	**3.757**	22.228
BeetleFly	32.430	**26.600**	27.199	29.003	50.610
BirdChicken	24.952	22.600	**19.914**	20.540	30.693
CBF	10.744	8.978	9.215	**8.398**	12.868
Car	0.906	0.812	**0.709**	0.740	1.588
ChlorineConcentration	6.018	0.979	**0.695**	0.698	0.769
CinC_ECG_torso	29.892	**18.638**	19.635	19.191	48.171
Coffee	0.870	0.582	0.511	**0.496**	0.660
Computers	86.619	**79.250**	82.215	81.417	235.208
Cricket_X	10.954	8.200	**7.932**	8.296	12.080
Cricket_Y	11.901	10.150	10.265	**9.574**	15.002
Cricket_Z	9.714	7.760	**7.544**	8.041	11.003
DiatomSizeReduction	0.964	**0.852**	0.874	0.869	1.203
DistalPhalanxOutlineAgeGroup	0.403	0.206	**0.175**	0.177	0.231
DistalPhalanxOutlineCorrect	0.515	0.310	0.300	**0.262**	0.351
DistalPhalanxTW	0.468	0.228	0.186	**0.178**	0.231
ECG200	1.907	1.541	1.565	**1.536**	1.905
ECG5000	4.737	4.190	4.398	**4.148**	5.463
ECGFiveDays	1.584	1.396	**1.322**	1.335	2.220
Earthquakes	71.461	**55.819**	56.504	57.153	147.980
ElectricDevices	19.499	15.045	**14.999**	15.228	37.121
FISH	0.439	0.353	0.319	**0.318**	0.462
FaceAll	9.309	8.687	**7.803**	7.853	10.716
FaceFour	20.483	**20.411**	21.259	21.444	46.841
FacesUCR	14.984	14.530	**14.403**	14.729	28.576
Gun_Point	0.447	0.368	0.300	**0.297**	0.858
Ham	16.152	14.717	**12.252**	13.424	24.340
Haptics	15.177	14.275	12.394	**11.931**	23.130
Herring	0.310	0.305	0.292	**0.249**	0.865
InsectWingbeatSound	3.104	2.346	2.186	**2.036**	5.437
ItalyPowerDemand	0.802	**0.595**	0.623	0.654	0.881
LargeKitchenAppliances	61.531	63.834	59.116	**57.219**	266.853
Lighting2	65.602	62.240	61.561	**60.826**	147.668

Figure 12: Time-series prediction: DTW loss achieved when using Euclidean init

Q & A