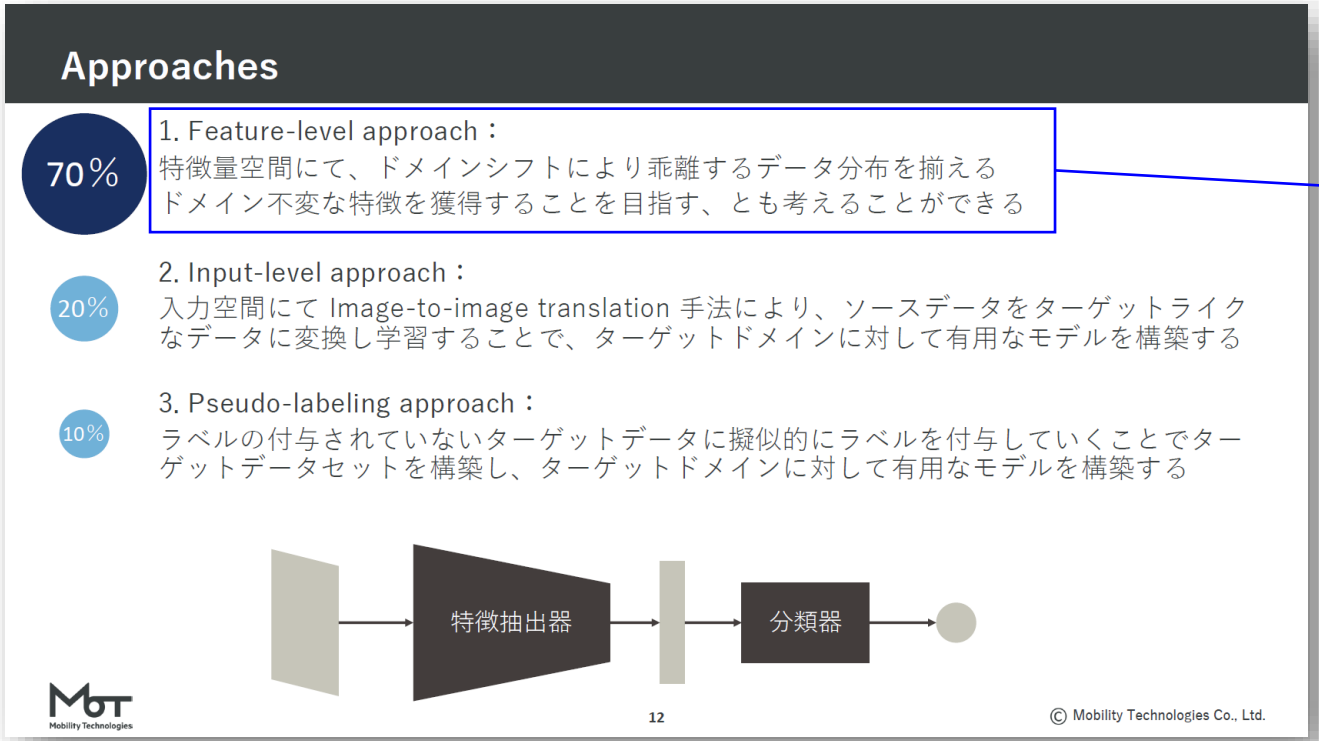
Deep Domain Confusion[Tzeng2014] 説明資料

Deep Domain Confusionの位置づけ



https://speakerdeck.com/takarasawa_/domain-adaptation

1. Feature-level approach

特徴量空間にて、ドメインシフトにより乖離するデータ分布を揃える
ドメイン不変な特徴を獲得することを目指す、とも考えることができる

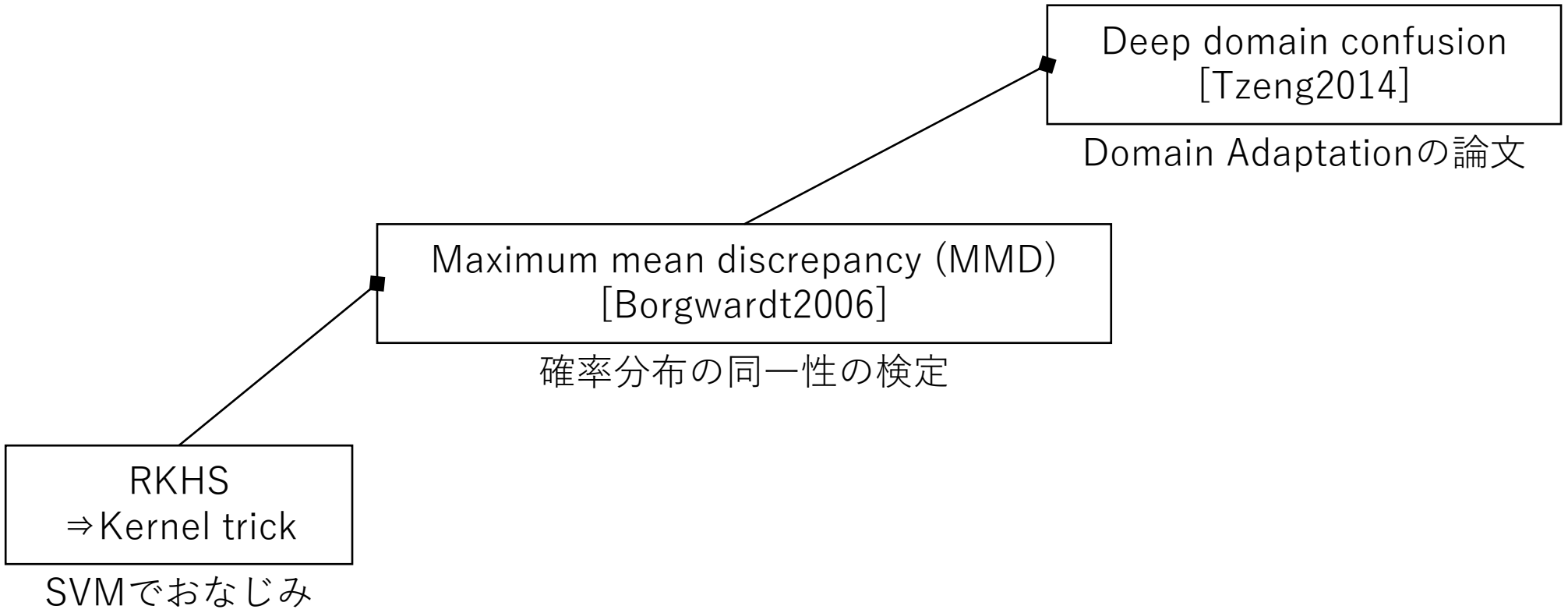
1. Statistic Criterion（MMD、CORAL）：分布間距離を測る指標を用いる
2. Adversarial-based：敵対的学習を用いる
3. Reconstruction-based：再構成誤差を用いる
4. Normalization-based：正規化層を用いる

1-1. Statistic Criterion

特徴量空間におけるデータ分布の乖離への直接的なアプローチ
分布間距離を測る指標を制約として加え、両ドメインの分布間距離が小さくなるように学習

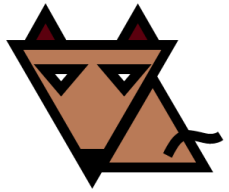
DAでよく用いられる：
- MMD-based
- CORAL

今日の説明Map

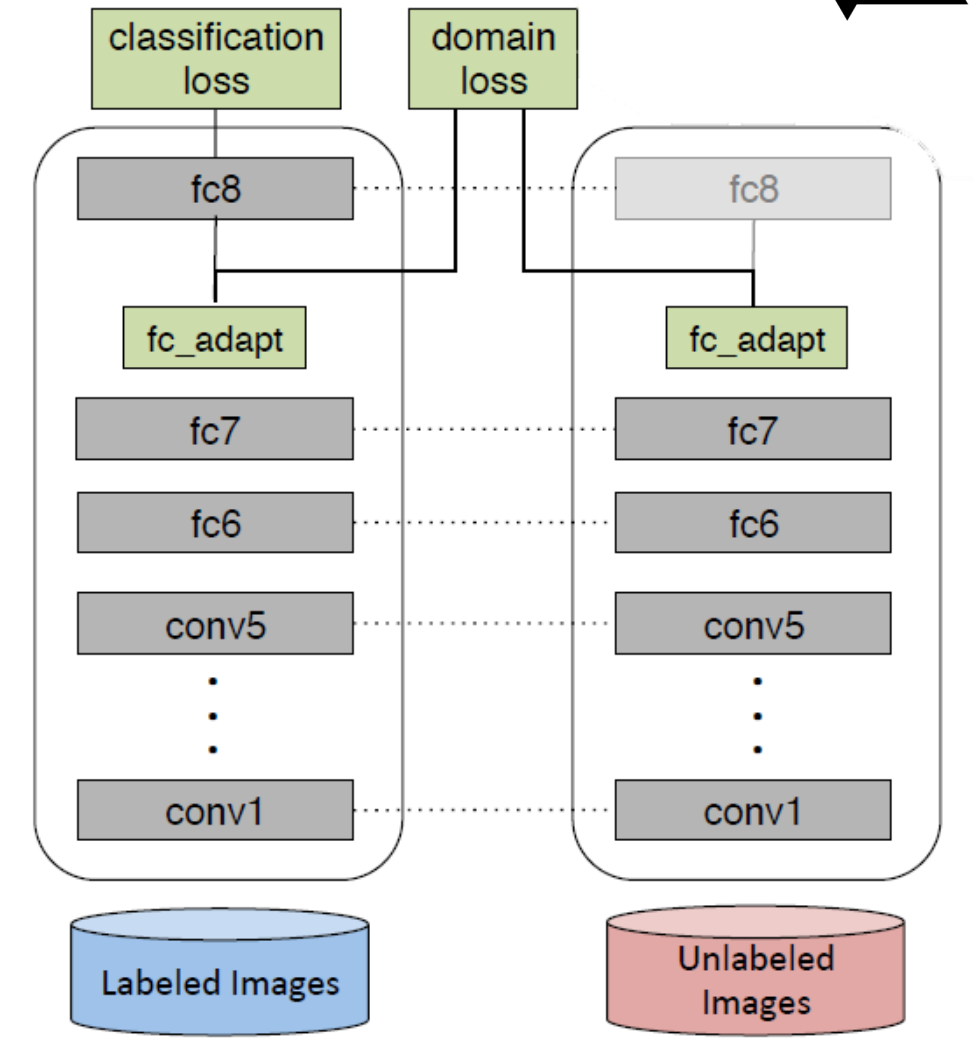

Deep domain confusion
[Tzeng2014]
Domain Adaptationの論文
Maximum mean discrepancy (MMD)
[Borgwardt2006]
確率分布の同一性の検定
RKHS
⇒Kernel trick
SVMでおなじみ

Deep Domain Confusionの説明

Deep Domain Confusion



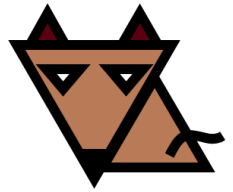
- FC間に**Adaptation Layer**を挿入
- Source-Target間の**Domain Loss**を使用
  - 分布の近さの指標である**MMD**で最適化
- Source-Targetは同じArchitectureで、WeightもShare
- Target側はラベル無しでもよいし、少量のラベルがついていてもよい



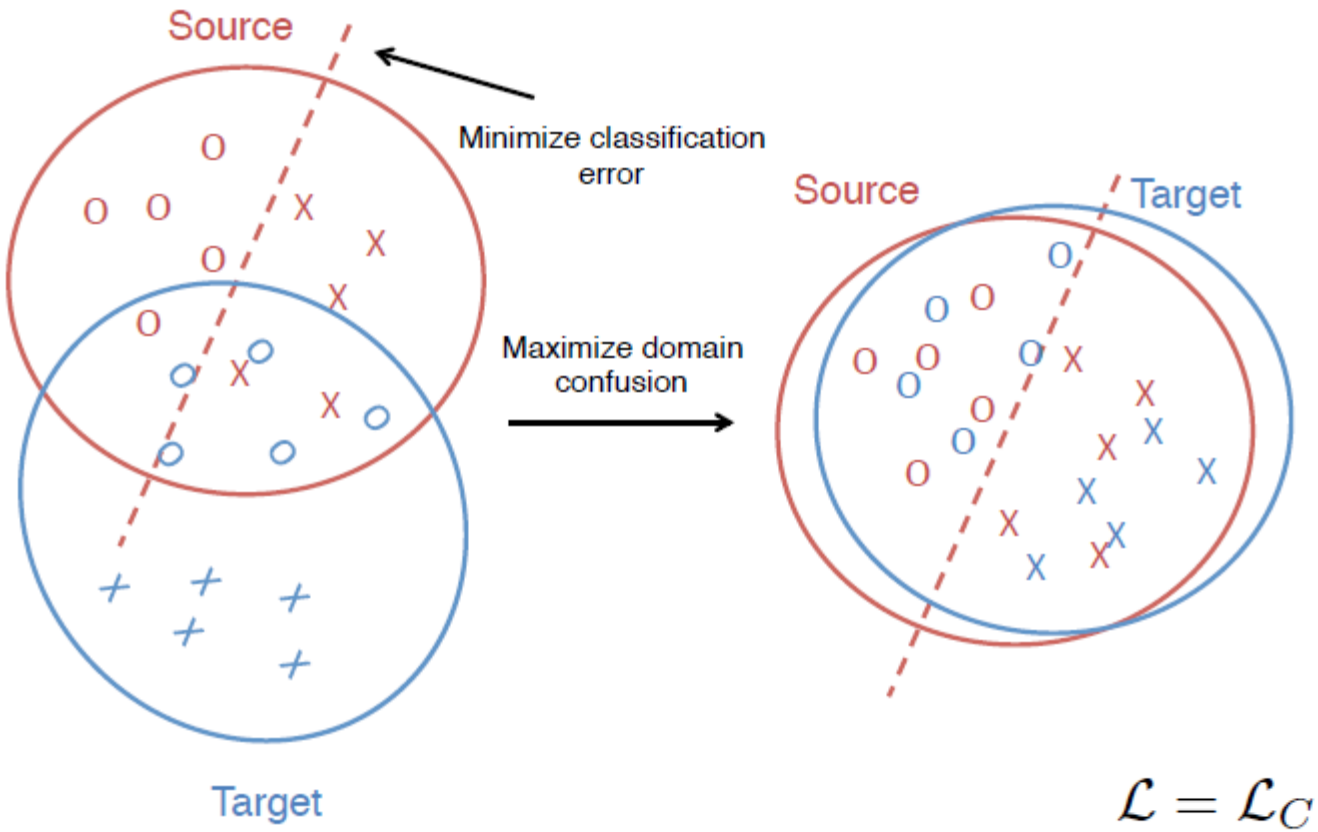
Source domain
Taraget domain w/ labels

Taraget domain w/o labels

基本的な考え方

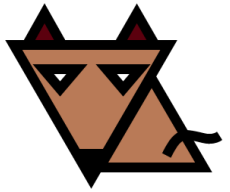


- 分類エラーを最小化しつつ、Domain間の分布を近づけることで、Target domainでも分類がうまくいくはずだ

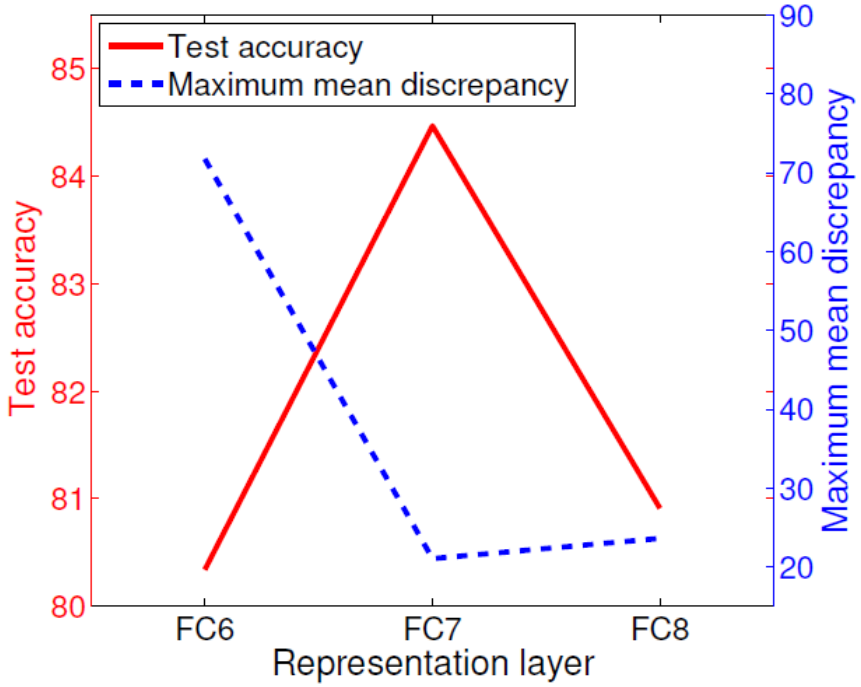


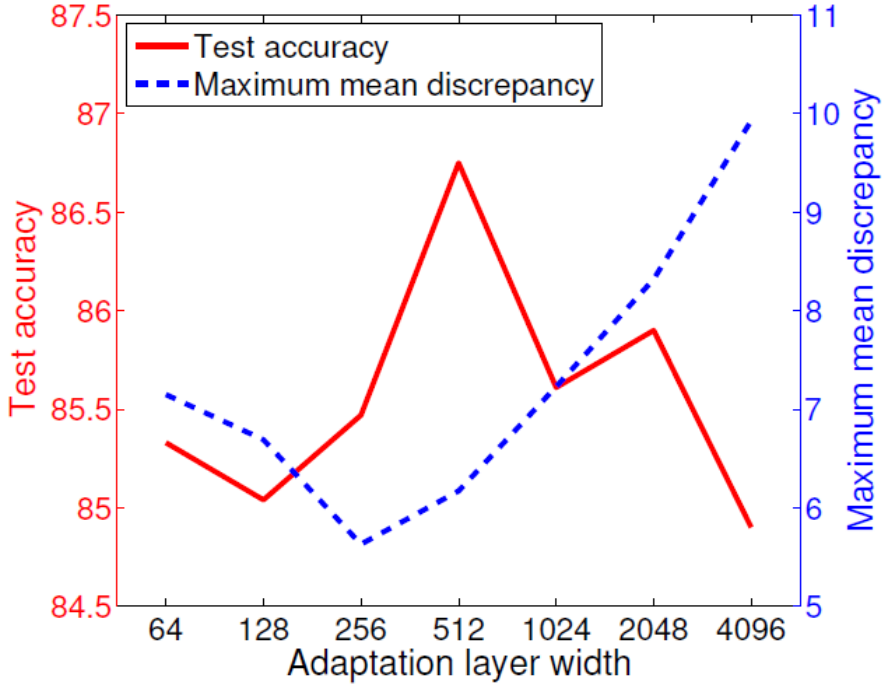
$$\mathcal{L} = \mathcal{L}_C(X_L, y) + \lambda \mathrm{MMD}^2(X_S, X_T)$$

Network architecture

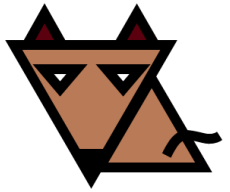


- 5層CNN＋3層FC(4096, 4096, |C|)
- MMDを用いてAdaptation Layerの挿入位置と次元を選択





結果



- Domain adaptation用のOffice dataset[Saenko]で比較
  - Office内にある物体を含む画像のDataset

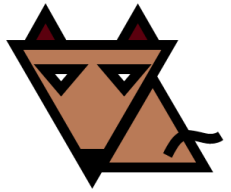
Supervised

	$A \to W$	$D \to W$	$W \to D$	Average
GFK(PLS,PCA) [16]	46.4 ± 0.5	61.3 ± 0.4	66.3 ± 0.4	53.0
SA [13]	45.0	64.8	69.9	59.9
DA-NBNN [31]	52.8 ± 3.7	76.6 ± 1.7	76.2 ± 2.5	68.5
DLID [8]	51.9	78.2	89.9	73.3
$DeCAF_6$ S+T [11]	80.7 ± 2.3	94.8 ± 1.2	–	–
DaNN [14]	53.6 ± 0.2	71.2 ± 0.0	83.5 ± 0.0	69.4
Ours	$\mathbf{84.1 \pm 0.6}$	$\mathbf{95.4 \pm 0.4}$	$\mathbf{96.3 \pm 0.3}$	**91.9**

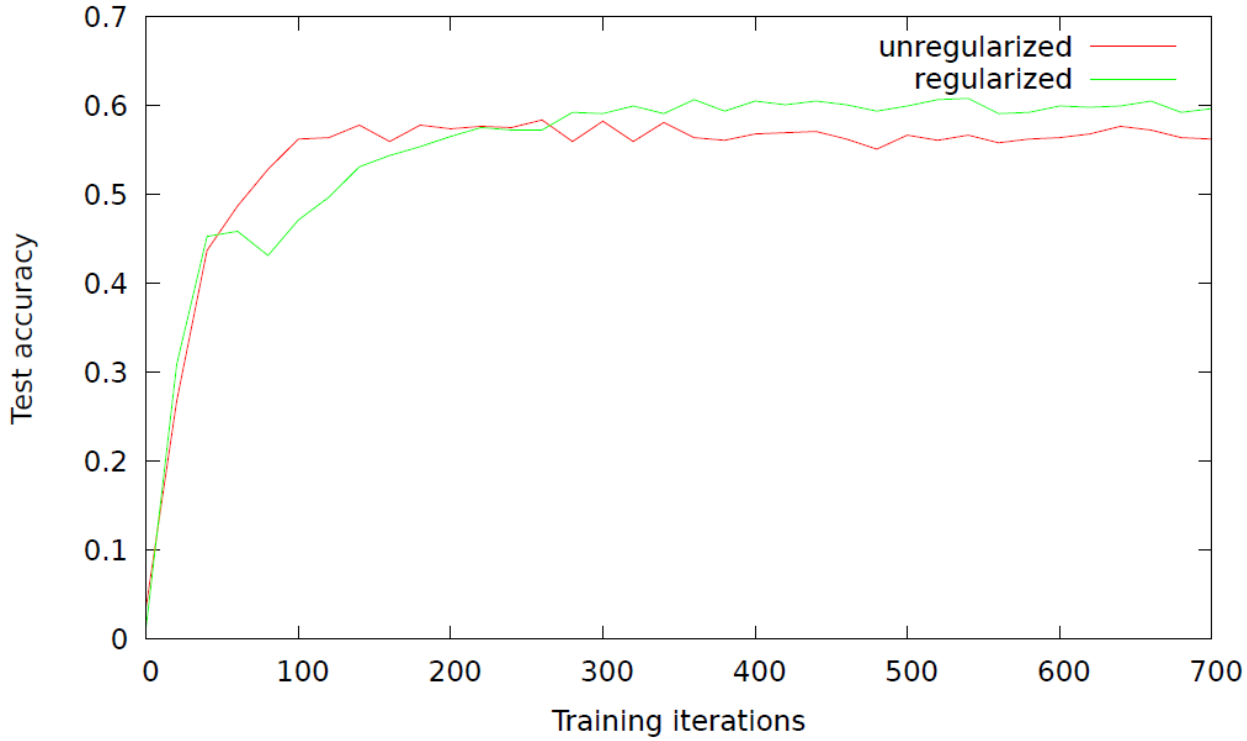
Unsupervised

	$A \to W$	$D \to W$	$W \to D$	Average
GFK(PLS,PCA) [16]	15.0 ± 0.4	44.6 ± 0.3	49.7 ± 0.5	36.4
SA [13]	15.3	50.1	56.9	40.8
DA-NBNN [31]	23.3 ± 2.7	67.2 ± 1.9	67.4 ± 3.0	52.6
DLID [8]	26.1	68.9	84.9	60.0
$DeCAF_6$ S [11]	52.2 ± 1.7	91.5 ± 1.5	–	–
DaNN [14]	35.0 ± 0.2	70.5 ± 0.0	74.3 ± 0.0	59.9
Ours	$\mathbf{59.4 \pm 0.8}$	$\mathbf{92.5 \pm 0.3}$	$\mathbf{91.7 \pm 0.8}$	**81.2**

MMDによる正則化の効果



- 立ち上がりは遅いが、最終的には高いAccuracy
- Amazon⇒WebcamのAdaptation

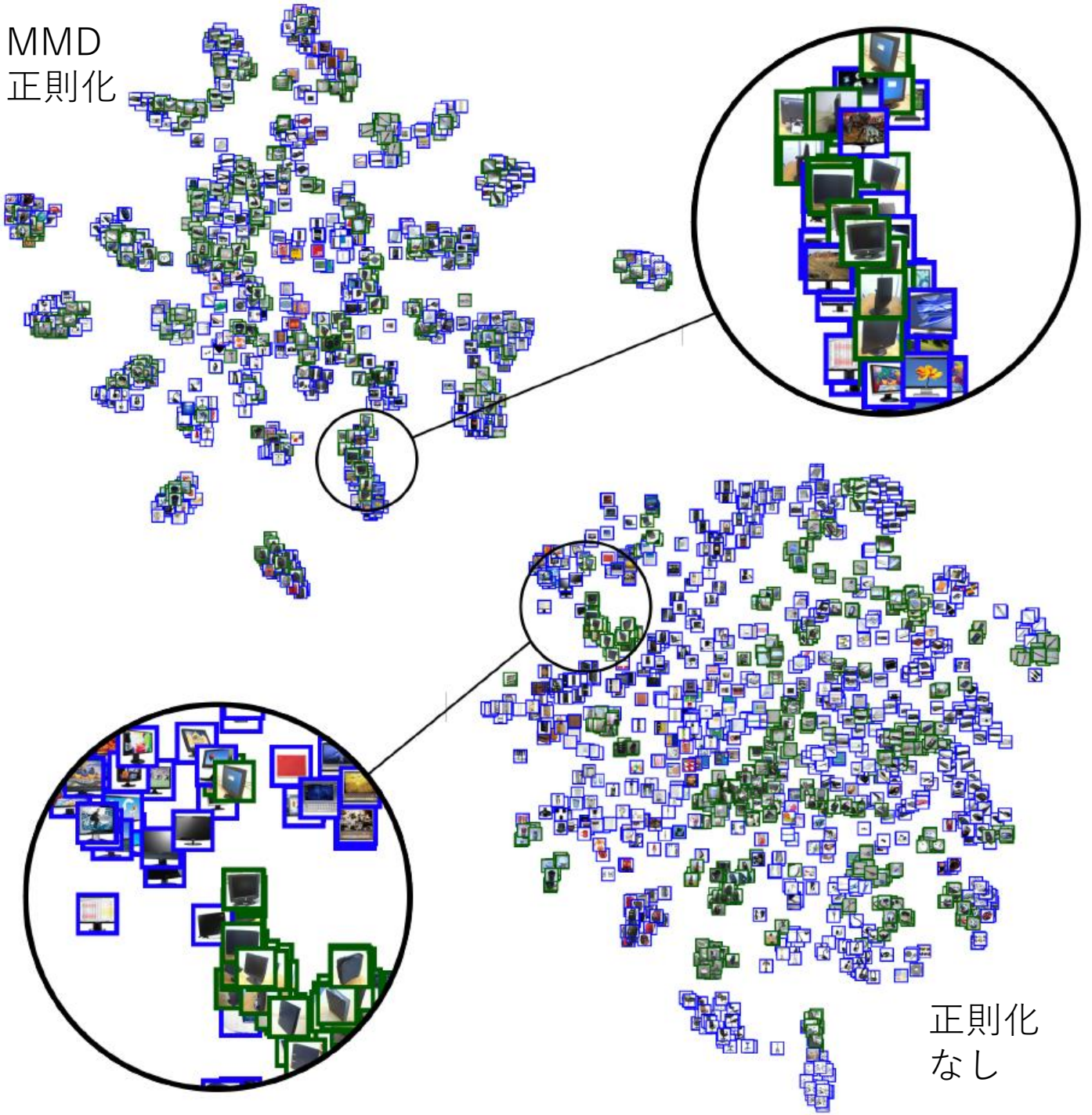


正則化後の分布

- MMD正則化後の分布（上）と正則化なしの分布（下）を比較
- t-SNE*により可視化
  - 近くのデータが近くなるように可視化
- Domain
  - 青: Amazon
  - 緑: Webcam

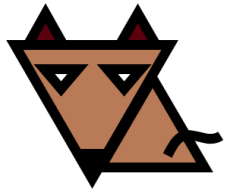
- 青と緑が近くに分布していることが分かる
- モニター同士も近いように見える

*t-sne[Van der Maaten]

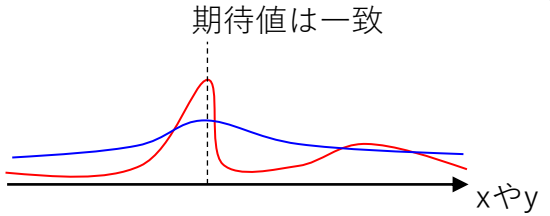


MMDの説明

MMD - Maximum Mean Discrepancy



- 2つの確率分布の同一性の判定のための指標



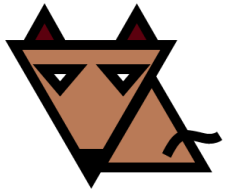
- ある確率分布の期待値と、別の確率分布の期待値が同じだったとしても、それらが同一な分布であるとは言えない

- **さまざまな関数で元データを別空間に写像**(特徴マッピング)したものの期待値同士の差異の中で、最も大きくなるような差異の量

$$\mathrm{MMD}[\mathcal{F}, p, q] := \sup_{f \in \mathrm{F}} \left(\mathbf{E}_p[f(x)] - \mathbf{E}_q[f(y)] \right)$$

- MMDがゼロなら同一とみなす
  - 直観的には、**平均が同じでも同じ分布とは言えないが、例えば平均も分散も歪度も尖度も・・・全部一緒なら一緒とみなせる**、というような考え方

MMD（続き）



- そもそも確率分布が分からないので、代わりに得られているサンプルの平均で代用

$$\mathrm{MMD}[\mathcal{F}, X, Y] := \sup_{f \in \mathrm{F}} \left(\frac{1}{m} \sum_{i=1}^{m} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} f(y_i) \right)$$

- ただし、$\mathcal{F}$が任意の関数集合では、いくらでも大きな差分になるような関数を選べる
- そこで、 $\mathcal{F}$は**ある関数空間中の単位超球**という制約をつける
- それでもまだ無限の関数の選択肢があるが、このある関数空間の特性を用いてMMDを求める

再生核ヒルベルト空間（RKHS）

再生核ヒルベルト空間

Def. 集合 Ω 上の再生核ヒルベルト空間（Reproducing kernel Hilbert space, RKHS) H とは，Ω 上の関数からなるヒルベルト空間であって，任意の $x \in \Omega$ に対し $\phi_x \in H$ があって，

$$\langle f, \phi_x \rangle = f(x) \qquad (\forall f \in H) \qquad \text{（再生性）}$$

を満たすものをいう.

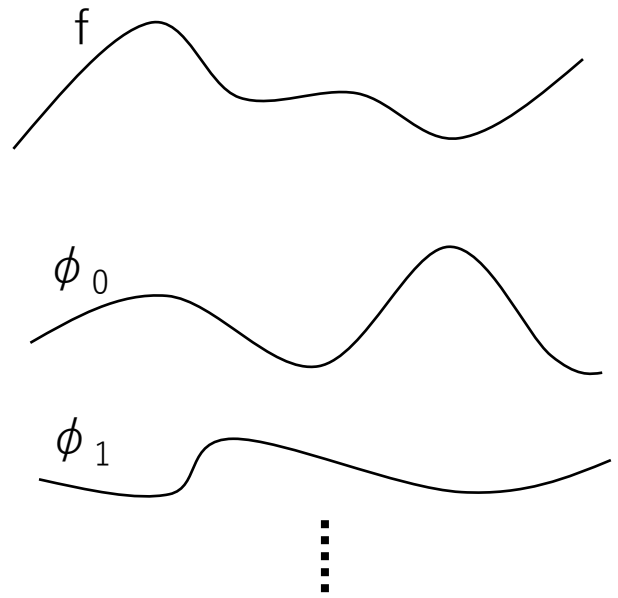
- 再生核:
$$k(y,x) := \phi_x(y)$$
このとき
$$\langle f, k(\cdot, x) \rangle = f(x)$$
- 再生核は存在すれば一意的. (証明略 [Exercise])
- 再生核ヒルベルト空間は「関数空間」. しかしL^2とは大きく異なる.

注）$k(\cdot, x)$： x を固定した第1変数を変数とする1変数関数

10

[Borgwardt2006]上の表記

$$f(x) = \langle f \phi(x) \rangle_{\mathrm{H}}$$



https://www.ism.ac.jp/~fukumizu/ISM_lecture_2010/Kernel_2_basics.pdf

カーネルトリック

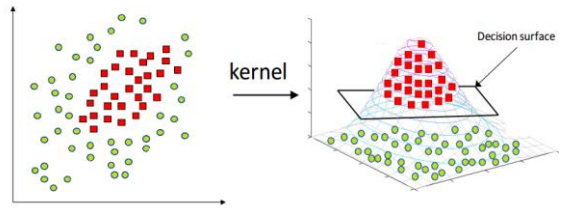
$$\langle f, \phi_x \rangle = f(x)$$

- fがϕ_yの時、

$$\langle \phi_y, \phi_x \rangle = \phi_y(x) = k(y, x)$$

参考：SVMでの考え方

元の空間では線形分離できないが、非線形写像した高次元空間では線形分離できる。計算は内積だけ必要なので、カーネルのみでOK。

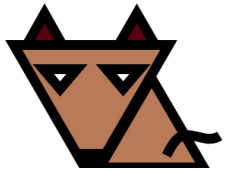


- kをカーネル関数という
- xをHに写像して内積をとるかわりに、カーネル関数の計算のみをすればよい（⇒計算が楽）
- MMDではガウスカーネル(RBFカーネル)を用いる

[Borgwardt2006]上の表記変数が違うので注意

$$k(x, x') = \exp(-\|x - x'\|^2/(2\sigma^2))$$

Gauss kernel



ガウスカーネルとは

ベクトルの場合: $e^{-a\|x-x'\|^2}$

・$K(x,x') = e^{-a(x-x')^2}$

という式で定義される二変数関数のことをガウスカーネルと言います。a は正の定数です。関数の入力は x と x' で、出力はスカラーです。このページでは一次元のガウスカーネルについて説明します（x と x' はスカラーとします）。

・ガウス分布（正規分布）の確率密度関数に似ています。

・ガウスカーネル $K(x,x')$ は x と x' の「近さ」を表します。

・$x = x'$ のとき $K(x,x') = 1$ で、$x \neq x'$ のときは $K(x,x') < 1$ です。

https://mathwords.net/gausskernel

ガウスカーネルの特徴ベクトルとは

データ x に対する特徴ベクトルが $\vec{\phi}(x)$ であるとき、それに対応するカーネル関数は、

$K(x,x') = \vec{\phi}(x)^{\top}\vec{\phi}(x')$

となります（カーネルは特徴ベクトルの内積です）。

では、逆に、カーネル関数がガウスカーネル $K(x,x') = e^{-a(x-x')^2}$ であるとき、それに対応する特徴ベクトルはどうなるでしょうか？

つまり、

$e^{-a(x-x')^2} = \vec{\phi}(x)^{\top}\vec{\phi}(x')$

となる特徴ベクトル $\vec{\phi}(x)$ はどうなるでしょうか。

実は、第 r 成分が

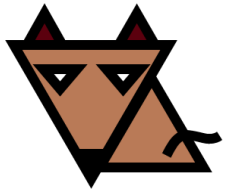
$\phi_r(x) = Ce^{-2a(x-r)^2}$

であるようなベクトル $\vec{\phi}(x)$ となります。

ただし、$C = \left(\frac{4a}{\pi}\right)^{\frac{1}{4}}$ です。

また、r は実数全体を動きます。つまり、$\vec{\phi}(x)$ は（連続無限個成分があるような）無限次元ベクトルになります。

Gauss kernel



証明

証明したいことは、

$$e^{-a(x-x')^2} = \vec{\phi}(x)^\top \vec{\phi}(x')$$

です。

$\vec{\phi}(x)$ は「連続無限個」成分があるベクトルです。そのため、上式の右辺、つまり特徴ベクトルの内積は、積分になります：

$$\int_{-\infty}^{\infty} \phi_r(x)\phi_r(x')dr$$

（各成分の積の和というイメージです）

$\phi_r(x)$ を代入して変形（指数の中身を平方完成）すると、

$$\int_{-\infty}^{\infty} Ce^{-2a(x-r)^2} Ce^{-2a(x'-r)^2} dr$$

$$= C^2 \int_{-\infty}^{\infty} e^{-4a(r-\frac{x+x'}{2})^2 - a(x-x')^2} dr$$

$$= C^2 e^{-a(x-x')^2} \int_{-\infty}^{\infty} e^{-4a(r-\frac{x+x'}{2})^2} dr$$

$$= e^{-a(x-x')^2} C^2 \sqrt{\frac{\pi}{4a}}$$

$$= e^{-a(x-x')^2}$$

となりました。

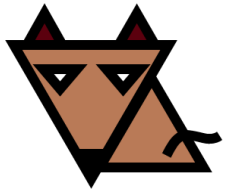
ただし、積分には、公式

$$\int_{-\infty}^{\infty} e^{-A(r-k)^2} dr = \sqrt{\frac{\pi}{A}}$$

を使いました。

https://mathwords.net/gausskernel

MMDの話に戻ると



- 関数集合$\mathcal{F}$が、Universal RKHS上の単位超球であれば、MMDがゼロの時2つの確率分布は一致する
- MMDはマッピング関数ϕの期待値を用いて簡略化される

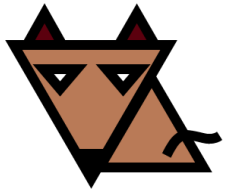
THEOREM 2.2. *Let p, q be Borel probability measures on $\mathcal{X}$ a compact subset of a metric space, and let $\mathcal{H}$ be a universal reproducing kernel Hilbert space with unit ball $\mathcal{F}$. Then $\mathrm{MMD}[\mathcal{F}, p, q] = 0$ if and only if $p = q$.*

Moreover, denote by $\mu_p := \mathbf{E}_p[\phi(x)]$ the expectation of $\phi(x)$ in feature space (assuming that it exists).[2] *Then one may rewrite MMD as*

$$\mathrm{MMD}[\mathcal{F}, p, q] = \|\mu_p - \mu_q\|_{\mathrm{H}}.$$

$$\begin{aligned}\mathrm{MMD}[\mathcal{F}, p, q] &= \sup_{\|f\|_{\mathcal{H}} \le 1} \mathbf{E}_p[f(x)] - \mathbf{E}_q[f(y)] \\ &= \sup_{\|f\|_{\mathcal{H}} \le 1} \mathbf{E}_p[\langle \phi(x), f \rangle_{\mathcal{H}}] - \mathbf{E}_q[\langle \phi(y), f \rangle_{\mathcal{H}}] \\ &= \sup_{\|f\|_{\mathcal{H}} \le 1} \langle \mu_p - \mu_q, f \rangle_{\mathcal{H}} = \|\mu_p - \mu_q\|_{\mathcal{H}}.\end{aligned}$$

Kernel関数を用いたMMD計算



- Kernel関数を用いることで、ヒルベルト空間の計算をすることなくMMDの2乗を計算することができる

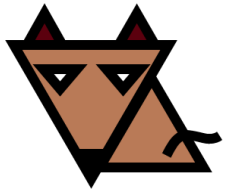
COROLLARY 2.3. *Under the assumptions of theorem 2.2 the following is an unbiased estimator of* $\mathrm{MMD}^2[\mathcal{F}, p, q]$:

$$\mathrm{MMD}^2[\mathcal{F}, X, Y] = \frac{1}{m(m-1)} \sum_{i \neq j}^{m} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j}^{n} k(y_i, y_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(x_i, y_j).$$

$$\begin{aligned} \mathrm{MMD}^2[\mathcal{F}, p, q] &:= \langle \mu_p - \mu_q, \mu_p - \mu_q \rangle_{\mathcal{H}} \\ &= \langle \mu_p, \mu_p \rangle_{\mathcal{H}} + \langle \mu_q, \mu_q \rangle_{\mathcal{H}} - 2\langle \mu_p, \mu_q \rangle_{\mathcal{H}} \\ &= \mathbf{E}_p \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} + \mathbf{E}_q \langle \phi(y), \phi(y') \rangle_{\mathcal{H}} \\ &\quad - 2\mathbf{E}_{p,q} \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}, \end{aligned}$$

まとめ

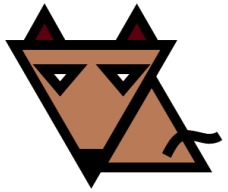
議論と今後



- Negative面
  - 性能は微妙
  - 分布を合わせるだけで、分布内の判別面も同じになる保証はない
  - ResultsのAccuracyがあんまりよくない
  - DAのためにLayerを追加しないといけないのはいまひとつ
  - 画像処理に適応する場合、明示的な特徴量層がないので、どうすればよいか？それ以外でも、全結合層を持たないNetworkではどうすれば？
- Positive面
  - 特徴量の分布を近づけるという考え方自体は使えそう（＋αは必要だろう）

- 今後
  - MMDベース手法のより新しい論文や関連論文を読みたい
  - RKHSまわりをしっかり理解したい

Reference



- Domain adaptation
  - https://speakerdeck.com/takarasawa_/domain-adaptation
- Deep domain confusion
  - Tzeng, Eric, et al. "Deep domain confusion: Maximizing for domain invariance." arXiv preprint arXiv:1412.3474 (2014).
- Office dataset
  - K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In Proc. ECCV, 2010.
- t-SNE
  - Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.11 (2008).
- MMD
  - Borgwardt, Karsten M., et al. "Integrating structured biological data by kernel maximum mean discrepancy." Bioinformatics 22.14 (2006): e49-e57.
  - https://slidesplayer.net/slide/17714999/
    - RKHSとMMDについて
  - https://www.ism.ac.jp/~fukumizu/ISM_lecture_2010/Kernel_2_basics.pdf
    - RKHSについて
  - https://mathwords.net/gausskernel
    - ガウスカーネルについて