Pattern Recognition and Machine Learning 6.2.

1G06Q117-5 園田 翔*

2009/6/8

1 Constructing Kernels

Kernel function

$$\begin{array}{ccccccccc} k & : & \mathcal{X}\times\mathcal{X} & \to & \mathbb{R}^N\times\mathbb{R}^N & \to & \mathbb{R} & & \\ & & \Psi & & \Psi & & \Psi & & \\ & & (\mathbf{x},\mathbf{x}') & \mapsto & (\phi(\mathbf{x}),\phi(\mathbf{x}')) & \mapsto & \sum_{i=1}^N \phi_i(\mathbf{x})\phi_i(\mathbf{x}') & =: & \langle\phi(\mathbf{x}),\phi(\mathbf{x}')\rangle_f \end{array}$$

1.1 how to construct valid kernel functions

to construct valid kernel functions

1. to choose a feature space mapping $\phi(\mathbf{x})$.
2. to construct $k(\mathbf{x},\mathbf{y})$ and find certain $\phi(\mathbf{x})$
3. to see if the Gram Matrices $\mathbf{K}_{ij} := k(\mathbf{x}_i,\mathbf{x}_j)$ for all possible $\{\mathbf{x}_n\}$ are positive semidefinite. (necessary and sufficient condition, Shawe-Taylor and Cristianini, 2004)
4. to build a kernel out of simpler ones.

Rem. We require that a kernel $k(\mathbf{x},\mathbf{x}')$

- be symmetric and positive semidefinite
- expresses the appropriate form of similarity between $\mathbf{x}$ and $\mathbf{x}'$

* 早稲田大学理工学部電気・情報生命工学科　村田研究室４年

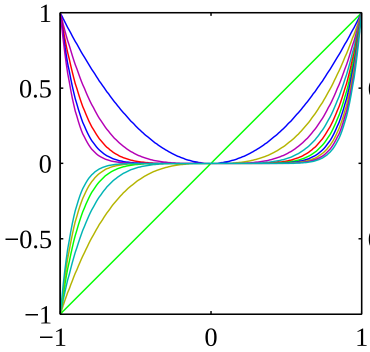

図 1.1 Figure6.1a

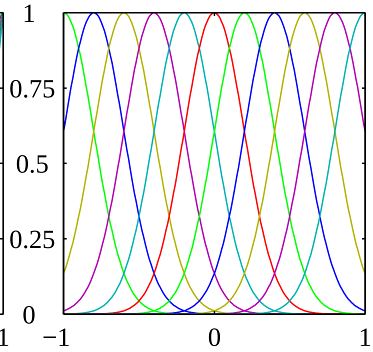

図 1.2 Figure6.1b

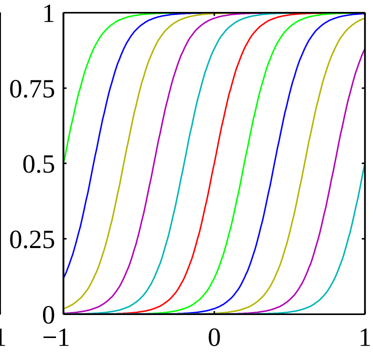

図 1.3 Figure6.1c

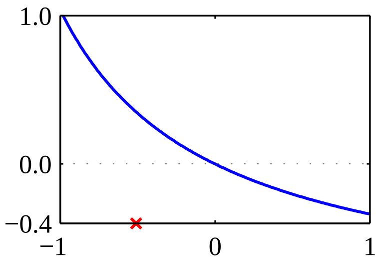

図 1.4 Figure6.1d

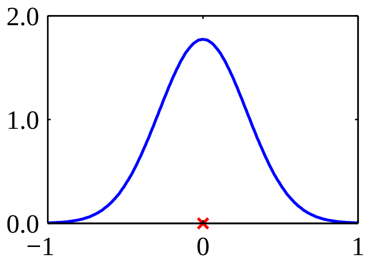

図 1.5 Figure6.1e

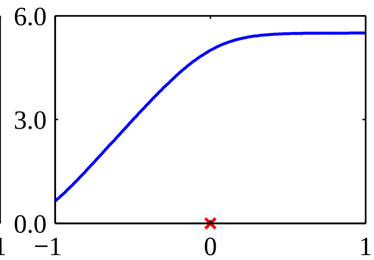

図 1.6 Figure6.1f

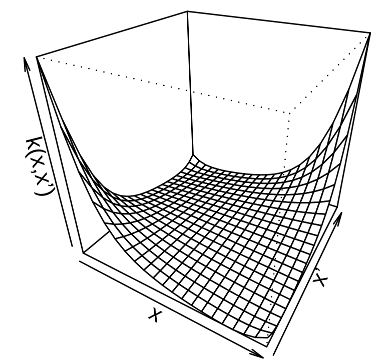

図 1.7 monomial kernel

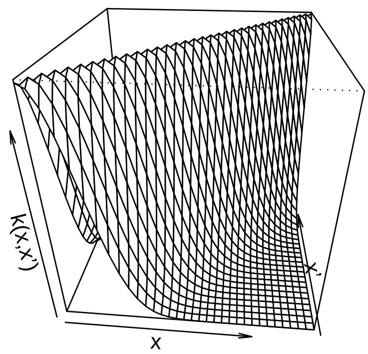

図 1.8 gaussiann kernel

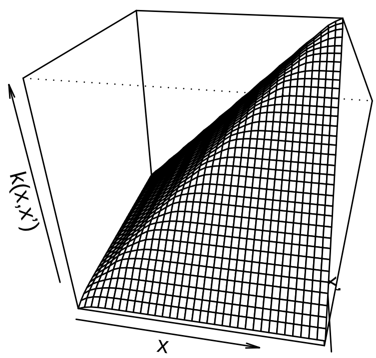

図 1.9 logistic sigmoid kernel

the followings are valid kernels.

Given $k_1(\mathbf{x},\mathbf{x}'), k_2(\mathbf{x},\mathbf{x}')$ to be valid,

$$k(\mathbf{x},\mathbf{x}') := f(\mathbf{x})k_1(\mathbf{x},\mathbf{x}')f(\mathbf{x}') \quad (\ f\ :\ \text{function}\ ) \tag{1.1}$$

$$k(\mathbf{x},\mathbf{x}') := ck_1(\mathbf{x},\mathbf{x}') \quad (\ \text{c}\ :\ \text{positive constant}\ ) \tag{1.2}$$

$$k(\mathbf{x},\mathbf{x}') := k_1(\mathbf{x},\mathbf{x}') + k_2(\mathbf{x},\mathbf{x}') \tag{1.3}$$

$$k(\mathbf{x},\mathbf{x}') := k_1(\mathbf{x},\mathbf{x}')k_2(\mathbf{x},\mathbf{x}') \tag{1.4}$$

$$k(\mathbf{x},\mathbf{x}') := q(k_1(\mathbf{x},\mathbf{x}')) \quad (\ \text{q}\ :\ \text{polynomial with nonnegatibe coefficients}\ ) \tag{1.5}$$

$$k(\mathbf{x},\mathbf{x}') := \exp(k_1(\mathbf{x},\mathbf{x}')) \tag{1.6}$$

$$k(\mathbf{x},\mathbf{x}') := k_3(\phi(\mathbf{x}),\phi(\mathbf{x}')) \quad (\ \phi(\mathbf{x}) \in \mathbb{R}^N,\ k_3(\mathbf{x},\mathbf{x}')\ \text{is a valid kernel in}\ \mathbb{R}^N) \tag{1.7}$$

$$k(\mathbf{x},\mathbf{x}') := \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}' \quad (\ \mathbf{x} \in \mathbb{R}^M, \mathbf{A}\ :\ \text{sym. pos. semidef.}\ ) \tag{1.8}$$

$$k(\mathbf{x},\mathbf{x}') := k_a(\mathbf{x}_a,\mathbf{x}'_a) + k_b(\mathbf{x}_b,\mathbf{x}'_b) \quad (\ \mathbf{x} = (\mathbf{x}_a,\mathbf{x}_b)\ ) \tag{1.9}$$

$$k(\mathbf{x},\mathbf{x}') := k_a(\mathbf{x}_a,\mathbf{x}'_a)k_b(\mathbf{x}_b,\mathbf{x}'_b) \tag{1.10}$$

Proofs

1.6

Ex1. Polynomial kernel

Polynomial kernel

$\mathbf{x}, \mathbf{x}' \in \mathbb{R}^N$, $c > 0$

$$k(\mathbf{x}, \mathbf{x}') := (\mathbf{x}^{\mathrm{T}}\mathbf{x}')^M \tag{1.11}$$

$$k(\mathbf{x}, \mathbf{x}') := (\mathbf{x}^{\mathrm{T}}\mathbf{x}' + c)^M \tag{1.12}$$

- (1.11) contains all monomials order M.
- Whereas (1.12) contains all terms up to degree M.
- If $\mathbf{x}$ and $\mathbf{x}'$ are two images, it represents a particular weighted sum of products of M pixels in the $\mathbf{x}$ with M pixels in the $\mathbf{x}'$.

Ex2. Gaussian kernel

Gaussian kernel

$\mathbf{x}, \mathbf{x}' \in \mathbb{R}^N$

$$k(\mathbf{x}, \mathbf{x}') := \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right) \tag{1.13}$$

$$= \exp\left(-\frac{\mathbf{x}^{\mathrm{T}}\mathbf{x} + (\mathbf{x}')^{\mathrm{T}}\mathbf{x}' - 2\mathbf{x}^{\mathrm{T}}\mathbf{x}'}{2\sigma^2}\right) \tag{1.14}$$

$\kappa(\mathbf{x}, \mathbf{x}')$: nonlinear kernel

$$k(\mathbf{x}, \mathbf{x}') := \exp\left(-\frac{\kappa(\mathbf{x}, \mathbf{x}) + \kappa(\mathbf{x}', \mathbf{x}') - 2\kappa(\mathbf{x}, \mathbf{x}')}{2\sigma^2}\right) \tag{1.15}$$

Ex3. Kernels over graphs, sets, strings and text documents.

The kernel defined over sets

D : fixed set

$A_1, A_2 \subset D$

$$k(A_1, A_2) := 2^{|A_1 \cap A_2|} \tag{1.16}$$

where $|A|$ denotes the number of elements in A

- Kernels can be defined over graphs, sets, strings and text documents.

Ex4. Kernels from probabilistic generative models

- Generative models can deal naturally with missing data,
  and in the case of HMMs it can handle sequences of varying length.
- Whereas Discriminative models generally give BETTER performance.
- In order to combine two approaches, we define a kernel using a generatibe model, and apply the kernel in a discriminative approach.

The kernel defined over sets

$p(\mathbf{x})$: generative model

$$k(\mathbf{x}, \mathbf{x}') := p(\mathbf{x})p(\mathbf{x}') \tag{1.17}$$

$p(i)$: positive weighting coefficients, or 'latent' variable (§9.2)
$p(\mathbf{z})$: weighting coefficients for continuous latent variable

$$k(\mathbf{x}, \mathbf{x}') := \sum_i p(\mathbf{x}|i)p(\mathbf{x}'|i)p(i) \tag{1.18}$$

$$\xrightarrow[i\to\infty]{} \int p(\mathbf{x}|\mathbf{z})p(\mathbf{x}'|\mathbf{z})p(\mathbf{z})d\mathbf{z} \tag{1.19}$$

HMM (§13.2)
$\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_L\}$: input data consists of ordered sequences.
$\mathbf{Z} = \{\mathbf{z}_1, \cdots, \mathbf{z}_L\}$: corresponding sequence of hidden states.

$$k(\mathbf{X}, \mathbf{X}') := \sum_{\mathbf{Z}} p(\mathbf{X}|\mathbf{Z})p(\mathbf{X}'|\mathbf{Z})p(\mathbf{Z}) \tag{1.20}$$

- (1.17) represents that $\mathbf{x}$ and $\mathbf{x}'$ are similar if they have high probabilities.
- (1.18) is equivalent, if normalized, to a mixture distribution.
- A popular generative model for sequences is the HMM, which expresses the distribution $p(\mathbf{X})$ as a marginalization over $\mathbf{Z}$.
- (1.20) measures the similarity of two sequences.

Ex5. Fisher kernel

Fisher kernel

$p(\mathbf{x}|\theta)$: θ-parametrized generative model

Fisher score :

$$\mathbf{g}(\theta,\mathbf{x}) := \nabla_\theta \ln p(\mathbf{x}|\theta) \tag{1.21}$$

Fisher information matrix :

$$\mathbf{F} := \mathbb{E}_\mathbf{x}\left[\mathbf{g}(\theta,\mathbf{x})\mathbf{g}(\theta,\mathbf{x})^\mathrm{T}\right] \tag{1.22}$$

$$= \int \begin{pmatrix} \frac{\partial \ln p(\mathbf{x}|\theta)}{\partial\theta_1}\frac{\partial \ln p(\mathbf{x}|\theta)}{\partial\theta_1} & \cdots & \frac{\partial \ln p(\mathbf{x}|\theta)}{\partial\theta_1}\frac{\partial \ln p(\mathbf{x}|\theta)}{\partial\theta_P} \\ \vdots & \ddots & \vdots \\ \frac{\partial \ln p(\mathbf{x}|\theta)}{\partial\theta_P}\frac{\partial \ln p(\mathbf{x}|\theta)}{\partial\theta_1} & \cdots & \frac{\partial \ln p(\mathbf{x}|\theta)}{\partial\theta_P}\frac{\partial \ln p(\mathbf{x}|\theta)}{\partial\theta_P} \end{pmatrix} p(\mathbf{x}|\theta)d\mathbf{x} \tag{1.23}$$

Fisher kernel :

$$k(\mathbf{x},\mathbf{x}') := \mathbf{g}(\theta,\mathbf{x})^\mathrm{T}\mathbf{F}^{-1}\mathbf{g}(\theta,\mathbf{x}') \tag{1.24}$$

- It measures the similarity between $\mathbf{x}$ and $\mathbf{x}'$ induced by the generative model $p(\mathbf{x}|\theta)$.
- It can be motivated from the perspective of information geometry.(Amari, 1998)
- form-invariant under a nonlinear re-parametrization : $\theta \to \psi(\theta)$

  [Proof.]

  Let $f(\theta) := \ln p(\mathbf{x}|\theta)$, $\widetilde{f}(\psi(\theta)) := f(\theta)$

$$\mathbf{g}(\theta,\mathbf{x}) = \frac{\partial f(\theta)}{\partial\theta} = \frac{\partial \widetilde{f}(\psi(\theta))}{\partial\theta} = \mathcal{J}\mathbf{h}(\psi,\mathbf{x}) \quad \left(\mathcal{J} := \left(\frac{\partial\psi}{\partial\theta}\right)^\mathrm{T}, \mathbf{h}(\psi,\mathbf{x}) := \frac{\partial\widetilde{f}(\psi)}{\partial\psi}\right) \tag{1.25}$$

  Therefore,

$$\begin{aligned} \mathbf{F} &= \mathbb{E}_\mathbf{x}[\mathbf{g}(\theta,\mathbf{x})\mathbf{g}(\theta,\mathbf{x}')^\mathrm{T}] && (1.26)\\ &= \mathbb{E}_\mathbf{x}[\mathcal{J}\mathbf{h}(\psi,\mathbf{x})\mathbf{h}(\psi,\mathbf{x}')^\mathrm{T}\mathcal{J}^\mathrm{T}] && (1.27)\\ &= \mathcal{J}\mathbb{E}_\mathbf{x}[\mathbf{h}(\psi,\mathbf{x})\mathbf{h}(\psi,\mathbf{x}')^\mathrm{T}]\mathcal{J}^\mathrm{T} && (1.28) \end{aligned}$$

  Then,

$$\begin{aligned} \mathbf{g}(\theta,\mathbf{x})^\mathrm{T}\mathbf{F}^{-1}\mathbf{g}(\theta,\mathbf{x}') &= \mathbf{h}(\psi,\mathbf{x})^\mathrm{T}\mathcal{J}^\mathrm{T}\left(\mathcal{J}^\mathrm{T}\right)^{-1}\left(\mathbb{E}_\mathbf{x}[\mathbf{h}(\psi,\mathbf{x})\mathbf{h}(\psi,\mathbf{x}')^\mathrm{T}]\right)^{-1}\mathcal{J}^{-1}\mathcal{J}\mathbf{h}(\psi,\mathbf{x}') && (1.29)\\ &= \mathbf{h}(\psi,\mathbf{x})^\mathrm{T}\left(\mathbb{E}_\mathbf{x}[\mathbf{h}(\psi,\mathbf{x})\mathbf{h}(\psi,\mathbf{x}')^\mathrm{T}]\right)^{-1}\mathbf{h}(\psi,\mathbf{x}') && (1.30) \end{aligned}$$

  Q.E.D.
- In practice, we substitute the sample average for the proper $\mathbf{F}$.

$$\mathbf{F} \simeq \frac{1}{N}\sum_{n=1}^{N}\mathbf{g}(\theta,\mathbf{x_n})\mathbf{g}(\theta,\mathbf{x_n})^\mathrm{T} \tag{1.31}$$

  This is the covariance matrix of the Fisher scores. Thus the kernel corresponds to a whitening of these scores.

- or, more simply replace $\mathbf{F} \to \mathbf{I}$. This is NO MORE form-invariant.
- Fisher kernels applied to document retrieval.(Hofmann, 2000)

Ex6. Sigmoidal kernel

Sigmoidal kernel

$$k(\mathbf{x}, \mathbf{x}') := \tanh(a\mathbf{x}^{\mathrm{T}}\mathbf{x}' + b) \tag{1.32}$$

- This is NOT positive semidefinite in general.
- superficial resemblances between SVMs and NNs.
- some Baysian NNs have deeper links to kernel methods. (§6.4.7)