Machine learning under physical constraints
Invariant representations in physics

Sixin Zhang
(sixin.zhang@toulouse-inp.fr)

Outline

Outline

Construct Invariant Representations

Linear Discriminant Analysis

Outline

- Construct invariant representations (e.g. translational, rotational invariance) using transformations of Fourier, wavelet, and wavelet scattering.
- TP: classification of cosmology dust using invariant representations in 2d (eval.)
- Final project: Molecular Energy Prediction in 3d (eval.).

References

- Notes des Cours 18'-21' par Campagne et Mallat, "Sciences des Données" du Collège de France, https://www.di.ens.fr/~mallat/CoursCollege.html
- Lecture notes of Edouard Oyallon, Advanced topics in Deep Learning, https://edouardoyallon.github.io/MAP670R-2022/index.html

Example in vision: group symmetry

- Invariance (group symmetry): A group action g on a signal x does not change an outcome
- For example, an image begin translated, rotated, dilated (zoom) or deformed does not change its object-level information.

Definition of group symmetry

- $(G, \cdot)$ is a group if $\cdot$ is a mapping from $G \times G \to G$ such that
  - $\exists \mathrm{id} \in G$, such that $g \cdot \mathrm{id} = \mathrm{id} \cdot g = g$, for any $g \in G$.
  - $(g_1 \cdot g_2) \cdot g_3 = g_1 \cdot (g_2 \cdot g_3)$, for any $g_1, g_2, g_3 \in G$.
  - $\forall g \in G$, there exists $g^{-1} \in G$ such that $g \cdot g^{-1} = g^{-1} \cdot g = \mathrm{id}$.
- Examples: invertible triangular matrices under matrix multiplication,

$$(A \cdot B) \cdot C = A \cdot (B \cdot C), \quad A^{-1}A = \mathrm{id}$$

Action of group: basic examples in physics

- Action of group on x: $g \cdot x$ for $g \in G$.
- Translation group: Let $\tau \in \mathbb{R}^d$,

$$g_\tau \cdot x(u) = x(u - \tau), u \in \mathbb{R}^d$$

- Rotation group in 2d ($d = 2$): Let r_θ be a rotation of angle $\theta \in [0, 2\pi]$,

$$g_\theta \cdot x(u) = x(r_\theta \cdot u), u \in \mathbb{R}^2$$

- Dilation group: Let $s > 0$,

$$g_s \cdot x(u) = x(u/s), u \in \mathbb{R}^d$$

Examples of ODE/PDE: translation invariance

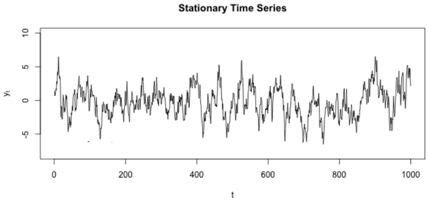
- Ordinary and partial differential equations (ODE/PDE) are often used in mathematical models of physics.
- Example of ODE: Lorenz system ($u \in \mathbb{Z}$) is **spatially translation invariant**

$$\frac{dx_t(u)}{dt} = (x_t(u+1) - x_t(u-2))x_t(u-1) - x_t(u) + F$$

  - If x_t is a solution, then $g_\tau \cdot x_t$ is also a solution ($\tau \in \mathbb{Z}$)
- Example of PDE: Homogeneous Navier-Stokes equation is also spatially translation invariant (in 2d or 3d).

Example of stationary processes

- Stationary processes exist in physics, e.g. white noise



- $\{X(u)\}_{u\in\mathbb{R}^d}$ is a stationary process if for any $u \in \mathbb{R}^d$,
  - $\mathbb{E}(X(u))$ is a constant.
  - $\mathbb{E}(X(u)X(u-\tau))$ is a function of $\tau \in \mathbb{R}^d$.
- More generally, the joint distribution in translation invariant:

$$(X(u_1), X(u_2), \cdots, X(u_n)) \underset{\text{law}}{=} (X(u_1-\tau), X(u_2-\tau), \cdots, X(u_n-\tau))$$

for any $n \geq 1$, $u_1 \in \mathbb{R}^d, \cdots, u_n \in \mathbb{R}^d, \tau \in \mathbb{R}^d$.

Example of Isotropic and self-similar processes

- The translation group of τ in **stationary processes** can be extended to the rotation group to define isotropic processes, and to the dilation group to define self-similar processes.
- For example, $\{X(u)\}_{u \in \mathbb{R}^d}$ is an isotropic process for $d = 2$, if for $u \in \mathbb{R}^2$ and $u' \in \mathbb{R}^2$,
  - $\mathbb{E}(X(g_\theta u))$ is a constant along $\theta \in [0, 2\pi]$.
  - $\mathbb{E}(X(g_\theta u)X(g_\theta u'))$ does not change with $\theta \in [0, 2\pi]$.
- In practice, we consider a finite number of angles (a discrete group), e.g. $\theta = \ell\pi/L, 0 \leq \ell < 2L$.

Construct invariant representations

- Idea: construct invariance from equi-variance.
- An invariant representation is a transformation $\Phi(x)$ of x:

$$\Phi(x) = \Phi(g \cdot x), \quad \forall g \in G$$

- An equi-variant representation is a transformation $\tilde{\Phi}(x)$ of x:

$$g \cdot \tilde{\Phi}(x) = \tilde{\Phi}(g \cdot x), \quad \forall g \in G$$

- Build Φ from $\tilde{\Phi}$ by **group averaging** (on a finite group):

$$\Phi(x) = \frac{1}{|G|} \sum_{g \in G} \tilde{\Phi}(g \cdot x)$$

Translation group: convolution and equi-variance

- Convolution $\star$ with a filter h:

$$x \star h(u) = \sum_v x(u-v)h(v)$$

- For a signal $x \in \mathbb{R}^{Nd}$ of length Nd, the convolution is assumed circular: $u, v, u-v \in \{0, \cdots, N-1\}^d$.
- In this case, we have equi-variance to discrete translation group: $\tau \in G = \{0, \cdots, N-1\}^d$,

$$(g_\tau \cdot x) \star h = g_\tau \cdot (x \star h)$$

Translation equi-variance of convolution on images

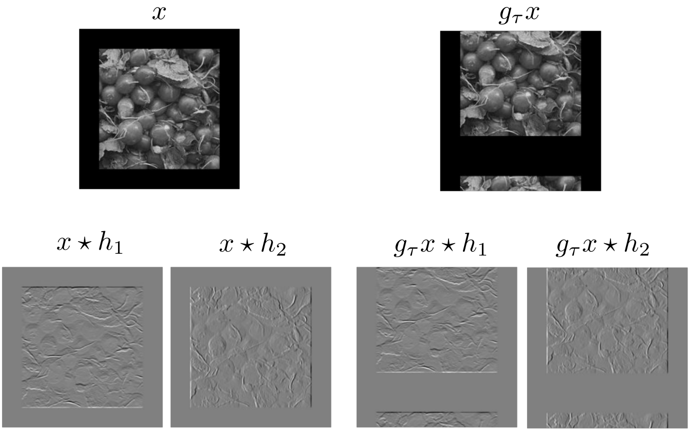


Figure: Left: **Circular convolution** of an image x with two filters h_1 and h_2. Right: the image is translated $g_\tau x$ first before the convolution.

Translation invariant representation

- Relation with CNN: $\tilde{\Phi}$ convolutional layer, Φ pooling layer

$$\tilde{\Phi}(x) = x \star h, \quad \Phi(x) = \frac{1}{N^d} \sum_{\tau} x \star h(\tau)$$

- Add non-linearity ρ: point-wise transformation such as ReLU

$$\tilde{\Phi}(x) = \{\rho(x \star h_k)\}_k$$

- If x has multiple-channels (e.g. color images), write $x = \{x_c\}$,

$$\tilde{\Phi}(x) = \left\{ \rho(\sum_c x_c \star h_{c,k}) \right\}_{k \leq K}$$

  - The summation along c combines all the channels of x.
  - This is a typical layer in CNN with K output channels.

Invariant covariance representation

- For a zero-mean stationary process observed on $u \in \{0, \cdots, N-1\}^d$, $X_N(u)$, its (translation) **invariant covariance representation** is defined by

$$\Phi(X_N) = \frac{1}{N^d} \sum_u X_N(u) X_N(u-\tau), \quad \tau \in \{0, \cdots, N-1\}^d.$$

- Due to the stationarity, it is a statistical estimator of the true covariance between $X(u)$ and $X(u-\tau)$.
- For example, $\tau = 0$ estimates the variance of $X(u)$.
- $\Phi(X_N)$ can be equivalently written using the circular convolution,

$$\frac{1}{N^d} X_N \star \tilde{X}_N(\tau), \quad \tilde{X}_N(u) = X_N(-u)$$

Outline

Linear discriminant Analysis (LDA)

- We derive a linear classifier from Bayes rule, by assuming a Gaussian model on each sample (x, y).
- Gaussian model: assume $y \in \{c_1, \cdots, c_K\}$ (K categories),

$$p(x|y = c_k) = \mathcal{N}(\mu_k, \Sigma), \quad p(y = c_k) = \pi_k.$$

- The classifier decides that $x \in \mathbb{R}^N$ is in class c_k if

$$k = \arg\max_{k'} p(y = c_{k'}|x)$$

- Two key question:
  - What is the classification rule? Is it linear in x?
  - Given M i.i.d. samples $\{(x_i, y_i)\}_{i \le M}$, how to estimate $\{\mu_k\}_k$ and Σ?

LDA as a linear classifier

- Rewrite $p(x|y = c_k) = \mathcal{N}(\mu_k, \Sigma)$,

$$p(x|y = c_k) = \frac{1}{\sqrt{|2\pi\Sigma|}} e^{-\frac{1}{2}(x-\mu_k)^\intercal \Sigma^{-1}(x-\mu_k)}$$

- The posterior $p(y = c_k|x)$ is

$$\frac{p(x|y = c_k)\pi_k}{\sum_{k'} p(x|y = c_{k'})\pi_{k'}}$$

- To maximize $p(y = c_k|x)$ with respect to k, is equivalent to

$$\min_k \frac{1}{2}(x - \mu_k)^\intercal \Sigma^{-1}(x - \mu_k) - \log(\pi_k)$$

  - The first term is the Mahalanobis distance (when Σ is p.d).

LDA as a linear classifier

- To maximize $p(y = c_k|x)$ is also equivalent to

$$\max_k g_k(x) = \langle \Sigma^{-1}\mu_k, x\rangle - \frac{1}{2}\mu_k^\intercal \Sigma^{-1}\mu_k + \log(\pi_k)$$

- The g_k is a linear discriminant function, thus LDA is a linear classifier.
- Parameter estimation: Assume $\{\mu_k\}$ are given a-prior, how to estimate Σ and μ_k from training data $\{(x_i, y_i)\}_{i\leq M}$?
- Reference: Bishop (section 4.1.4 for $K = 2$ and 4.1.6 for $K > 2$).

Parameter estimation in LDA ($K = 2$)

- Let M_1 the number of training samples in class c_1, $M_2 = M - M_1$,

$$\mu_1 = \frac{1}{M_1} \sum_{i:y_i=c_1} x_i, \quad \mu_2 = \frac{1}{M_2} \sum_{i:y_i=c_2} x_i$$

- The estimation of Σ can be obtained from Σ_1 and Σ_2, e.g.

$$\Sigma = \frac{M_1}{M}\Sigma_1 + \frac{M_2}{M}\Sigma_2$$

where

$$\Sigma_k = \frac{1}{M_k} \sum_{i:y_i=c_k} (x_i - \mu_k)(x_i - \mu_k)^T, \quad k = 1, 2$$