Landmarking Manifolds with Gaussian Processes

Dawen Liang, John Paisley

Wittawat Jitkrittum

Gatsby Machine Learning Journal Club

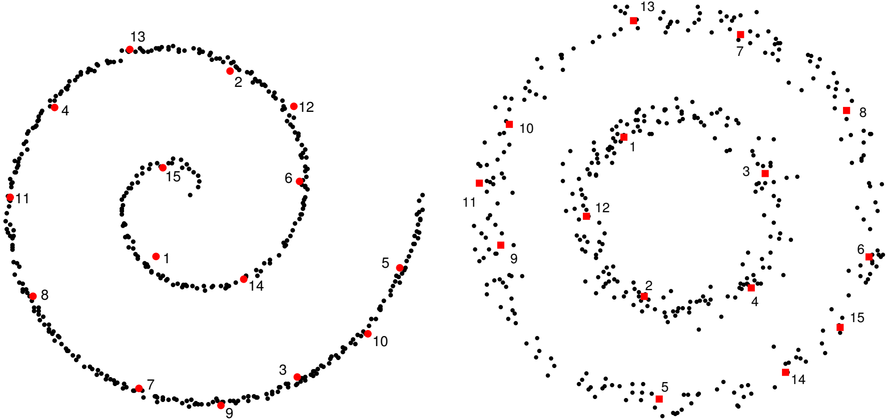
4 Aug 2015

Overview

```
Landmarking Manifolds with Gaussian Processes
Dawen Liang, John Paisley
ICML 2015.
```

- **Goal**: Find a few points characterizing the structure of the manifold
  - Documents: bag-of-word landmarks = topics
  - Faces: landmarks = distinct facial features
- **Idea**:
  1. Gaussian process
  2. $x_{n+1} \leftarrow \arg\max_x$ predictive variance$(x \mid x_1, \ldots, x_n)$. x not from a finite set.
  3. Repeat
- Based on active learning idea
- A new landmark is "repelled" by those already selected

Example: Manifold Landmarking



(b) Manifold landmarking

Gaussian Process (GP)

- Paired data: $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$.
- Let K_n be the kernel matrix on $\{x_i\}_{i=1}^n$ i.e., $(K_n)_{ij} = k(x_i, x_j)$.
- Let $Y := (y_1 | \cdots | y_n)^\top$.
- Given $\mathcal{D}_n$, $y(x)$ at a new test point:

$$
\begin{aligned}
y(x) \mid Y &\sim \mathcal{N}(\xi(x), \Sigma(x)), \\
\text{predictive mean:} \quad \xi(x) &= k(x, \mathcal{D}_n) K_n^{-1} Y, \\
\text{predictive variance:} \quad \Sigma(x) &= k(x, x) - k(x, \mathcal{D}_n) K_n^{-1} k(x, \mathcal{D}_n)^\top,
\end{aligned}
$$

where $k(x, \mathcal{D}_n) := (k(x, x_1), \ldots, k(x, x_n))$.

Active Learning with GP

- Find the next $x_{n+1} \in \mathcal{D}$ (finite set) to query y_{n+1} by

$$x_{n+1} = \arg\max_{x \in \mathcal{D}} \Sigma(x),$$

$$\text{predictive variance:} \quad \Sigma(x) = k(x,x) - k(x,\mathcal{D}_n) K_n^{-1} k(x,\mathcal{D}_n)^\top,$$

- Note $\Sigma(x)$ does not depend on Y.
- Choose from a finite set $\mathcal{D}$. Drawbacks:
  - High-dimensional data are not usually densely sampled.
  - Perhaps a landmark should not correspond to an observation e.g., landmark = local average of faces.
- **Proposal**: Find $x_{n+1} = \arg\max_x \Sigma(x)$ by gradient ascent.

Kernel k

- Low-dimensional manifold $\mathcal{M}$ in an ambient space $\mathbb{S}$
- $\mu, \mathcal{N} :=$ distributions on $\mathbb{S}$. Support of μ is $\mathcal{M}$.
- $\mathcal{N}$: a zero mean noise process.
- Assume the observed data point $x = \hat{x} + \epsilon \in \mathbb{S}$ where $\hat{x} \overset{i.i.d.}{\sim} \mu$ and $\epsilon \overset{i.i.d.}{\sim} \mathcal{N}$.
- Kernel for $t, t' \in \mathbb{S}$:

$$k(t,t') = \int_{\hat{x}\in\mathbb{S}} \phi_{\hat{x}}(t)\phi_{\hat{x}}(t')\, d\mu(\hat{x}),$$
$$\phi_{\hat{x}}(t) = \exp\left(-\|t-\hat{x}\|^2/\eta\right).$$

- We do not have μ or $\hat{x} \sim \mu$. Approximate with observations $\{x_i\}_{i=1}^N$:

$$k(t,t') \approx \frac{1}{N}\sum_{i=1}^{N} \phi_{x_i}(t)\phi_{x_i}(t') := \frac{1}{N}\vec{\phi}(t)^\top \vec{\phi}(t'),$$

where $\vec{\phi}(t) := (\phi_{x_i}(t), \ldots, \phi_{x_N}(t))^\top$.

Kernel k

- Low-dimensional manifold $\mathcal{M}$ in an ambient space $\mathbb{S}$
- $\mu, \mathcal{N} :=$ distributions on $\mathbb{S}$. Support of μ is $\mathcal{M}$.
- $\mathcal{N}$: a zero mean noise process.
- Assume the observed data point $x = \hat{x} + \epsilon \in \mathbb{S}$ where $\hat{x} \overset{i.i.d.}{\sim} \mu$ and $\epsilon \overset{i.i.d.}{\sim} \mathcal{N}$.
- Kernel for $t, t' \in \mathbb{S}$:

$$k(t,t') = \int_{\hat{x} \in \mathbb{S}} \phi_{\hat{x}}(t)\phi_{\hat{x}}(t')\, d\mu(\hat{x}),$$
$$\phi_{\hat{x}}(t) = \exp\left(-\|t - \hat{x}\|^2/\eta\right).$$

- We do not have μ or $\hat{x} \sim \mu$. Approximate with observations $\{x_i\}_{i=1}^N$:

$$k(t,t') \approx \frac{1}{N}\sum_{i=1}^{N} \phi_{x_i}(t)\phi_{x_i}(t') := \frac{1}{N}\vec{\phi}(t)^\top \vec{\phi}(t'),$$

where $\vec{\phi}(t) := (\phi_{x_i}(t), \ldots, \phi_{x_N}(t))^\top$.

Finding Landmarks with Stochastic Gradient

- Given n selected landmarks $\mathcal{T}_n = \{t_1, \ldots, t_n\}$,

$$t_{n+1} = \arg\max_{t \in \mathbb{S}} k(t,t) - k(t, \mathcal{T}_n) K_n^{-1} k(t, \mathcal{T}_n)^\top$$
$$\approx \arg\max_{t \in \mathbb{S}} \vec{\phi}(t)^\top \vec{\phi}(t) - \vec{\phi}(t)^\top \Phi(\Phi^\top \Phi)^{-1} \Phi^\top \vec{\phi}(t) := \arg\max_{t \in \mathbb{S}} f_n(t),$$

  where $\Phi = \left[\vec{\phi}(t_1) | \cdots | \vec{\phi}(t_n)\right] \in \mathbb{R}^{N \times n}$.

- Rewrite $f_n(t)$:

$$f_n(t) = \sum_{i=1}^{N} \sum_{j=1}^{N} M_{ij} \phi_{x_i}(t) \phi_{x_j}(t),$$
$$M_{ij} = \delta_{ij} - \left(\Phi(\Phi^\top \Phi)^{-1} \Phi^\top\right)_{ij},$$
$$\nabla_t f_n(t) = -\sum_{i=1}^{N} \sum_{j=1}^{N} \frac{4M_{ij}}{\eta} \left[t - \frac{x_i + x_j}{2}\right] \phi_{x_i}(t) \phi_{x_j}(t).$$

- To handle large N, use stochastic gradient.

Finding Landmarks with Stochastic Gradient

- Given n selected landmarks $\mathcal{T}_n = \{t_1, \ldots, t_n\}$,

$$t_{n+1} = \arg\max_{t \in \mathbb{S}} k(t,t) - k(t, \mathcal{T}_n) K_n^{-1} k(t, \mathcal{T}_n)^\top$$
$$\approx \arg\max_{t \in \mathbb{S}} \vec{\phi}(t)^\top \vec{\phi}(t) - \vec{\phi}(t)^\top \Phi (\Phi^\top \Phi)^{-1} \Phi^\top \vec{\phi}(t) := \arg\max_{t \in \mathbb{S}} f_n(t),$$

  where $\Phi = \left[\vec{\phi}(t_1) | \cdots | \vec{\phi}(t_n) \right] \in \mathbb{R}^{N \times n}$.

- Rewrite $f_n(t)$:

$$f_n(t) = \sum_{i=1}^{N} \sum_{j=1}^{N} M_{ij} \phi_{x_i}(t) \phi_{x_j}(t),$$
$$M_{ij} = \delta_{ij} - \left(\Phi (\Phi^\top \Phi)^{-1} \Phi^\top \right)_{ij},$$
$$\nabla_t f_n(t) = - \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{4 M_{ij}}{\eta} \left[t - \frac{x_i + x_j}{2} \right] \phi_{x_i}(t) \phi_{x_j}(t).$$

- To handle large N, use stochastic gradient.

Algorithm (projected gradient)

Algorithm 1 Manifold landmarking with GPs

1: To find landmark t_{n+1} given $t_1, \dots, t_n$, initialize $t_{n+1}^{(1)}$ and do the following:
2: **for** $s = 1, \dots S$ **do**
3: Randomly subsample a set B_s of observations $x \in \mathcal{D}$.
4: For each t_k, construct $\vec{\phi}_s(t_k)$ using $x \in B_s$ and the function $\phi_x(t_k) = \exp(-\|x - t_k\|^2/\eta)$.
5: Define the matrix $\Phi = [\vec{\phi}_s(t_1), \dots, \vec{\phi}_s(t_n)]$ and set $M = I - \Phi(\Phi^T\Phi)^{-1}\Phi^T$.
6: Let $f_n(t, B_s) = \sum_{x_i, x_j \in B_s} M_{ij}\phi_{x_i}(t)\phi_{x_j}(t)$.
7: Calculate $\gamma = t_{n+1}^{(s)} + \rho_s \nabla_t f_n(t, B_s)|_{t_{n+1}^{(s)}}$ using Equation (10) and step size ρ_s.
8: Project γ onto $\mathbb{S} \subseteq \mathbb{R}^d$ to obtain $t_{n+1}^{(s+1)}$.
9: **end for**

- Subsample $B_s \subset \{x_1, \dots, x_N\} := \mathcal{D}$.
- Step size ρ_s such that $\sum_s |\rho_s| = \infty$ and $\sum_s \rho_s^2 < \infty$.
- Eq. 10 = $\nabla_t f_n(t)$.

Experiments

- Images, text and music data
- Step size: $\rho_s = (10 + s)^{-0.51}$
- 1000 gradient steps for each landmark
- Batch size: $|B_s| = 1000$
- Kernel width: $\eta = \sum_i \hat{\sigma}_i^2$
    - $\hat{\sigma}_i^2$ is the variance of the i^{th} dimension.

Yale Face Dataset

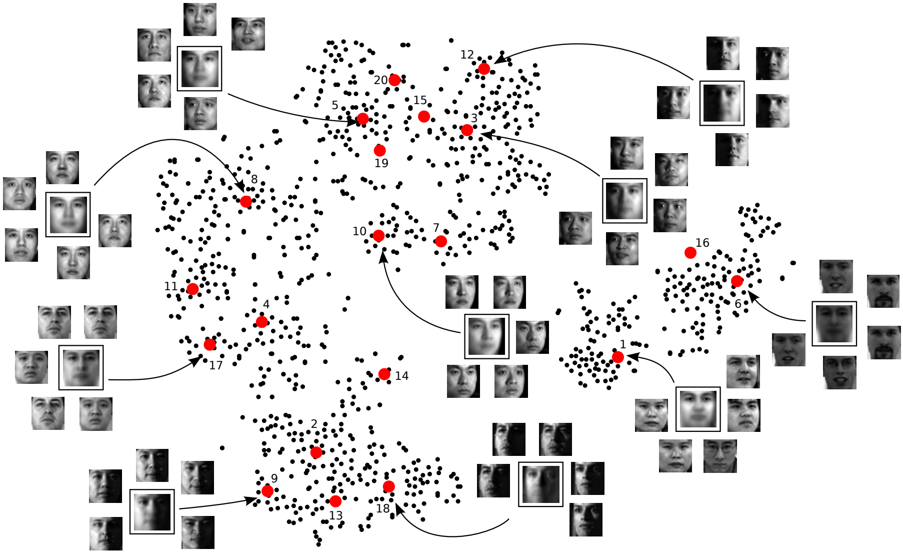
- Yale faces database. 2475 images of size 42×48.
- 165 images of various illuminations for 15 people.

Figure 2. The first eight landmarks from the Yale faces dataset.

- Does not correspond to any single person in the dataset.

PIE Faces Dataset

- 11,554 images of size 64×64. 68 people.
- 2D embedding of 1000 random images by t-SNE.



Documents: New York Times, 20 Newsgroup

- d^{th} document:

$$x_d(j) = \sqrt{(\text{\#occurrences of word j})/n_d},$$

  $n_d :=$ #words in document d.
- Without $\sqrt{\ \ }$, x_d is a discrete distribution over words.
- Landmark t is in the same space.
  - Can be interpreted as a topic (distribution over words) as in LDA.

Data

- New York Times: 1.8 million documents. Vocab. size: 8000.
- 20 Newsgroup vocab. size: 1545.

Documents: New York Times, 20 Newsgroup

- d^{th} document:

$$x_d(j) = \sqrt{(\text{\#occurrences of word j})/n_d},$$

  $n_d :=$ #words in document d.
- Without $\sqrt{}$, x_d is a discrete distribution over words.
- Landmark t is in the same space.
  - Can be interpreted as a topic (distribution over words) as in LDA.

Data

- New York Times: 1.8 million documents. Vocab. size: 8000.
- 20 Newsgroup vocab. size: 1545.

New York Times, 20 Newsgroup Results

Table 1. (top) The "most probable" words for the first 11 landmarks learned on the 1.8 million document New York Times dataset. (bottom) The first 12 landmarks from the 20 Newsgroup dataset.

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}
percent	inc	beloved	street	treasury	republican	minutes	mrs	game	percent	film
going	net	notice	sunday	bills	house	add	daughter	season	market	life
national	share	paid	music	rate	bush	oil	graduated	team	stock	man
public	reports	deaths	avenue	bonds	senate	salt	married	games	billion	story
life	earns	wife	theater	bond	political	cup	son	play	yesterday	book
ago	qtr	loving	art	notes	government	pepper	father	second	prices	movie
house	earnings	mother	museum	municipal	democrats	tblspoon	yesterday	left	quarter	love

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}
good	windows	team	turkish	encryption	god	ftp	car	israel	nasa	scsi	gun
make	dos	game	turks	key	jesus	file	good	israeli	gov	drive	guns
ve	card	year	armenia	technology	bible	pub	cars	jews	space	ide	weapons
work	mb	games	soviet	government	christ	mail	price	arab	long	mb	crime
back	system	season	today	chip	christians	program	buy	state	orbit	hard	control

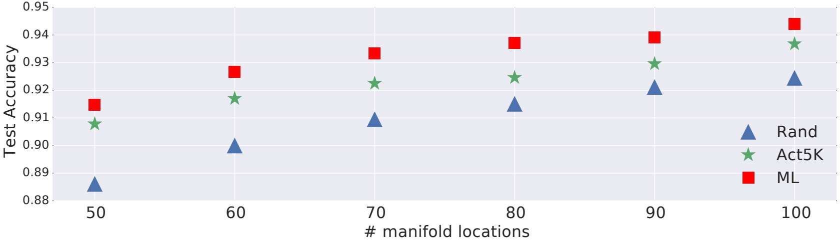
- Showing top 5,7 highest coordinates of t_i.
- Landmarks correspond to thematically meaningful concepts.

MNIST Classification with Landmarks

- Quantitatively evaluate the landmarks on handwritten digit classification problem (MNIST).
- Given landmarks $\mathcal{T}_n = \{t_i\}_{i=1}^n$, compute feature for image x_d:

$$\vec{w}(x_d) = \left[\phi_{t_1}(x_d), \ldots, \phi_{t_n}(x_d)\right]^\top .$$

- ℓ_2-regularized logistic regression.
- Train/validate/test sizes = 50,000/10,000/10,000.



- **Rand**: random n data points as landmarks.
- **Act5K**: GP active learning with the same kernel. Subsample data to 5000 images.

Automatic Music Tagging

- audio content $\mapsto$ semantic tags (e.g., classic, slow)
- Million Song Dataset. 561 tags. Train/test: 371,209/2,757.

Feature construction:

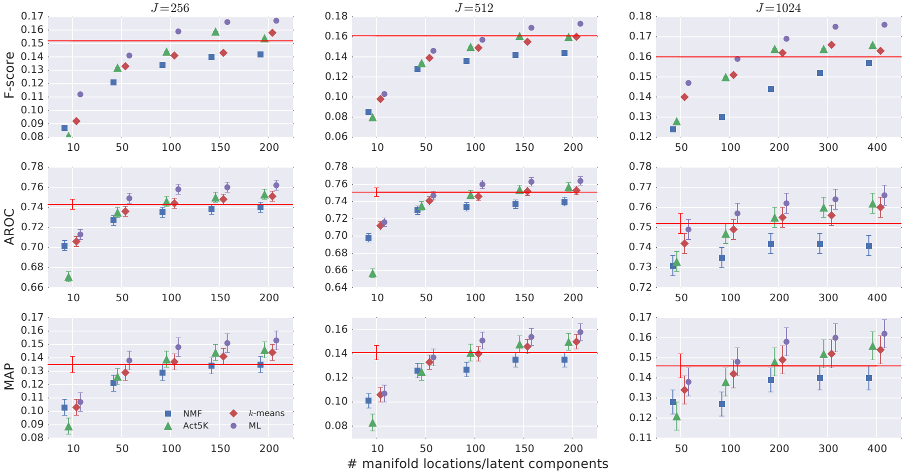
1. Echo Nest's timbre features (similar to MFCC). Multiple vectors per song.
2. k-means on all the vectors with J clusters (codewords).
3. For each song, assign each feature vector to the closest cluster. Song $x_d =$ histogram of cluster identities (J bins).

- Each tag is treated as a binary classification task. ℓ_2-regularized logistic regression.
- Use $\vec{w}(x_d)$ as before.

Annotation and Retrieval

- Use F-score to measure ability to annotate song. F-score computed from average per-tag precision, recall.
- Retrieval: given a query tag, provide a list of related songs.
  - Rank each song by the predicted probability.
  - Compute AROC and Mean Average Precision.

Music Tagging Results



- Red line = logistic regression on raw VQ features.
- NMF = Non-negative matrix factorization.
- k-means = treat centroids as landmarks.
- High codebook size (J) does not improve the performance.

References I

a

- a