Landmarking Manifolds with Gaussian Processes

Dawen Liang, John Paisley
Dept. of Electrical Engineering, Columbia University
Overview

The problem we try to address

• Data points are from low-dimensional (nonlinear) manifold embedded in the high-dimensional space

• Find a small set of locations (landmarks) spaced out along this manifold to summarize the data
Motivation

Active learning with Gaussian processes

Active learning problem:

- Given a dataset \((x_1, y_1), \ldots, (x_n, y_n)\), pick a new location \(x \in D\) to query the corresponding \(y\), such that a large amount of information is gained according to some measure.
Motivation

Active learning with Gaussian processes

• Gaussian Processes: prior over functions
  • Mean function \( m(\cdot) \)
  • Covariance function (kernel) \( k(\cdot, \cdot) \) (we choose Gaussian kernel)
Motivation

Active learning with Gaussian processes

• Gaussian Processes: prior over functions
  • Mean function $m(\cdot)$
  • Covariance function (kernel) $k(\cdot, \cdot)$ (we choose Gaussian kernel)

• Given the dataset $y = [y_1, \ldots, y_n]^T$
  $y \sim \mathcal{N}(m, K)$
  
  where $m = [m(x_1), \ldots, m(x_n)]^T$
  $K_{i,j} = k(x_i, x_j)$
Motivation

Active learning with Gaussian processes

• Find $x \in \mathcal{D}$ with the highest posterior uncertainty
Motivation

Active learning with Gaussian processes

• Find $x \in \mathcal{D}$ with the highest posterior uncertainty

$$y(x) | y \sim \mathcal{N}(\xi(x), \Sigma(x)),$$

$$\xi(x) = k(x, \mathcal{D}_n)K_n^{-1}y,$$

$$\Sigma(x) = k(x, x) - k(x, \mathcal{D}_n)K_n^{-1}k(x, \mathcal{D}_n)^T$$
Motivation

Active learning with Gaussian processes

• Find $x \in \mathcal{D}$ with the highest posterior uncertainty

\[
y(x) | y \sim \mathcal{N}(\xi(x), \Sigma(x)), \\
\xi(x) = k(x, \mathcal{D}_n)K_n^{-1}y, \\
\Sigma(x) = k(x, x) - k(x, \mathcal{D}_n)K_n^{-1}k(x, \mathcal{D}_n)^T
\]

Objective
Motivation

Active learning with Gaussian processes

- The newly selected $\mathbf{x} \in \mathcal{D}$ will be pushed away by the ones that are already selected, so the entire space where the data resides is efficiently explored.
Motivation

Active learning with Gaussian processes

- The newly selected $x \in \mathcal{D}$ will be pushed away by the ones that are already selected, so the entire space where the data resides is efficiently explored.
Proposed Method
Manifold landmarking with Gaussian processes

There are possible scenarios that we don’t want to use actual data points to landmark the manifold:
Proposed Method

Manifold landmarking with Gaussian processes

There are possible scenarios that we don’t want to use actual data points to landmark the manifold:

• The data points are not densely sampled — it would be too restrictive (e.g. in high-dimensional space).
Proposed Method
Manifold landmarking with Gaussian processes

There are possible scenarios that we don’t want to use actual data points to landmark the manifold:

• The data points are not densely sampled — it would be too restrictive (e.g. in high-dimensional space).

• *A priori* we believe it makes more sense to allow landmarks to not correspond exactly to single data point (e.g. faces, documents).
Proposed Method

Manifold landmarking with Gaussian processes

Setup:

\( \mathcal{M} \): a manifold in ambient space \( \mathbb{S} \) (not necessarily \( \mathbb{R}^d \))

\( \mu \): a probability distribution on the ambient space with support on the manifold

\( \mathcal{N} \): a zero-mean noise process
Proposed Method
Manifold landmarking with Gaussian processes

Setup:

\( \mathcal{M} \): a manifold in ambient space \( \mathbb{S} \) (not necessarily \( \mathbb{R}^d \))

\( \mu \): a probability distribution on the ambient space with support on the manifold

\( \mathcal{N} \): a zero-mean noise process

\[
x = \hat{x} + \epsilon \in \mathbb{S} \quad \hat{x} \sim \text{i.i.d. } \mu \\
\epsilon \sim \text{i.i.d. } \mathcal{N}
\]

• The data points are sampled from the manifold, corrupted by zero-mean noise process.
Proposed Method
Manifold landmarking with Gaussian processes

• Define a manifold-supported kernel function for $\forall t, t' \in S$

\[
k(t, t') = \int_{\hat{x} \in S} \phi_{\hat{x}}(t) \phi_{\hat{x}}(t') d\mu(\hat{x})
\]

where $\phi_{\hat{x}}(t) = \exp\{-\|t - \hat{x}\|^2/\eta\}$

Intuition: Two data points are considered “close” according to the path between them along the manifold.
Proposed Method

Manifold landmarking with Gaussian processes

- Define a manifold-supported kernel function for \( \forall t, t' \in \mathcal{S} \)

\[
k(t, t') = \int_{\hat{x} \in \mathcal{S}} \phi_{\hat{x}}(t)\phi_{\hat{x}}(t')d\mu(\hat{x})
\]

where \( \phi_{\hat{x}}(t) = \exp\{-||t - \hat{x}||^2/\eta\} \)

Intuition: Two data points are considered “close” according to the path between them along the manifold.

- Construct a plug-in estimator:

\[
k(t, t') \approx \frac{1}{N} \sum_{i=1}^{N} \phi_{x_i}(t)\phi_{x_i}(t') := \frac{1}{N} \phi(t)^T \phi(t'),
\]
Proposed Method
Manifold landmarking with Gaussian processes

Proposed method

\[ k(t, t') = \int_{\hat{x} \in S} \phi_{\hat{x}}(t) \phi_{\hat{x}}(t') d\mu(\hat{x}) \]

- The kernel function is evaluated at \( \forall t, t' \in S \)
- Data is used to approximate the integral

Active learning with GPs

\[ k(t, t') = \int_{\hat{x} \in \mathbb{R}^d} \phi_{\hat{x}}(t) \phi_{\hat{x}}(t') d\hat{x} \]

- The kernel function is evaluated at two data points
Proposed Method

Manifold landmarking with Gaussian processes

- We can optimize the similar objective with the newly defined kernel over the continuous ambient space

\[ t_{n+1} = \arg \max_{t \in \mathcal{S}} k(t, t) - k(t, \mathcal{T}_n) K_n^{-1} k(t, \mathcal{T}_n)^T \]

- A stochastic projected gradient method is developed for large datasets
Experiments

Qualitative evaluation

- Face datasets:
  - Yale (2,475 images)
  - PIE face (11,554 images)
- Documents:
  - NYT (1.8 million articles)
Experiments
PIE face t-SNE 2D embedding
Experiments

“Topics” on NYT

- Ambient space: the intersection of unit sphere with positive orthant
- The squared root of normalized word histogram
Experiments

MNIST digits classification

What is the benefit of allowing landmarks to move along the continuous ambient space?

- On MNIST, given landmarks \( T_n = \{t_1, \ldots, t_n\} \), derive landmark-based feature for image \( x_d \) and classify with logistic regression.

\[
\vec{w}(x_d) = [\phi_{t_1}(x_d), \ldots, \phi_{t_n}(x_d)]^T
\]

- Compare with random selection and active learning.

![Test Accuracy Graph](image-url)
Experiments

More constrained ambient space

- Automatic music tagging on the Million Song Dataset (370K songs for selecting landmarks and training classifier) with normalized vector quantization histograms.

- Compare with active learning, nonnegative matrix factorization and K-means on both annotation (F-score) and retrieval (AROC & MAP).
Experiments

More constrained ambient space
Summary

We present an algorithm for finding landmarks along a manifold that capture the low-dimensional nonlinear structure of the data.

- The landmarks are allowed to move along the continuous ambient space by optimizing an objective.
- The landmarks are learned sequentially and the new one will be “repelled” by the existing ones.
- We derive a stochastic algorithm for learning landmarks with large datasets.
Thanks!

Python code available: www.github.com/dawenl/manifold_landmarks