Introduction

What?
- Transfer Learning: A model learned from one set can be a prior for the second.
- Statistics on Manifolds: The data on manifold is mapped into tangent space by the Log map, from this we consider the covariance, PCA, Linear Regression.

Why?
- Is model learned in one region are utilized in another?

How?
- Let $M$ be manifold and $T_p(M)$ and $T_q(M)$ be tangent spaces.
- Suppose we have a data in $T_p(M)$.
- Parallel Transport data to $T_q(M)$ then learn and use this model in $T_q(M)$. This scales poorly.
- Different approach is Model Transport: the model is learned in $T_q(M)$ then transport the model to $T_p(M)$.
- Computational complexity is free from the size of data set.

Geometry on the sphere

- A geodesic on the sphere is a great circle, parallel transport given by rotation.

Discussion and conclusion

- We have studied the geometry of finite probability measures and its Transfer learning.
- The tangent space $T_pP_+(I)$ is trivial.
- Geodesics and Parallel transport are solved explicitly.
- In general, Parallel transport has no closed form, we have to approximate it.
- In the Box-plots, we see that the transferred models are comparable to the usual one.

The geometry of finite probability measures

Let $I = \{1, \ldots, n, n+1\}$ be a finite set.

- Measure space: $P_+(I) = \{\mu \mid \mu_i \geq 0, \forall i \in I, \text{ and } \sum_{i \in I} \mu_i = 1\}$.
- Tangent space: $T_I P_+(I) = \{\mu \times S_\mu(I), \text{ where } S_\mu(I) = \{\mu_i \mid \mu_i \geq 0, \sum_{i \in I} \mu_i = 0\}\}$.
- Fisher-Rao metric: $g_\mu(x, Y) = \sum_{i \in I} \frac{\mu_i}{\mu} X_i d\gamma, \forall \mu = \sum_{i \in I} \mu_i Y_i d\gamma \in T_I P_+(I)$.
- Levi-Civita connection $\nabla$: Let $X, Y$ be constant vector fields, $\frac{\partial \mu}{\partial \mu}$ is Radon-Nykodym derivative. Then
  \[
  \nabla_X Y_{\mu} = -\frac{1}{2} \left( dX_{\mu} - dY_{\mu} \right) q, \quad \text{for } i \in I.
  \]
- Geodesics: Given $\mu \in P_+(I)$ and a unit tangent vector $X \in T_I P_+(I)$. Then there exist a unique geodesic $\alpha(t) = \sum_{i \in I} \alpha_i(t) d\gamma$ in $P_+(I)$, starting from a point $\mu$ with direction $X$, where
  \[
  \alpha_i(t) = \left( \cos \frac{t}{2} + \frac{X_i}{\mu_i} \sin \frac{t}{2} \right)^2, \quad \text{for } i \in I.
  \]
- Distance: For every $\mu, \nu \in P_+(I)$ we have $d(\mu, \nu) = 2 \arccos \left( \sum_{i \in I} \sqrt{\mu_i \nu_i} \right)$.
- Exponential map: $e_\mu(X) = \sum_{i \in I} \left( \cos \frac{X_i}{\mu_i} + \frac{X_i}{\mu_i} \sin \frac{X_i}{\mu_i} \right) \delta_i$, where $\delta_i \in P_+(I)$.
- Log map: $\log_\mu(\nu) = \frac{1}{\sqrt{\mu}} \sum_{i \in I} \left( \sqrt{\frac{\mu_i}{\nu_i}} - \sum_{j \in I} \sqrt{\mu_{ij}} \right) \delta_i$, where $\delta_i \in P_+(I)$.
- Parallel transport: Let $\mu, \nu \in P_+(I)$ and $\alpha : [0, E] \to P_+(I)$ be a geodesic curve such that $\alpha(0) = \mu$ and $\alpha(E) = \nu$. The parallel transport, $\Gamma_{\mu \to \nu}(X) = T_{\mu} P_+(I) \to T_{\nu} P_+(I)$, given by
  \[
  \Gamma_{\mu \to \nu}(X) = \sum_{i \in I} \sqrt{\nu} \left( -C \sqrt{\nu_i} \left( 2 \sin \frac{X_i}{\mu} - \frac{X_i}{\mu} \cos \frac{E}{2} \right) + \frac{X_i}{\mu_i} \right) \right) \delta_i,
  \]
  where $\tau$ is the unit tangent vector $\tau = \log_\mu(\nu)/E$ and $C = \frac{1}{9} b_i(x, \tau)$.

- Isometry: By the map $\Phi(\mu) = 2 \sum_{i \in I} \sqrt{\nu_i} \delta_i$, $P_+(I)$ is isometric to the sphere
  \[
  S(0,2,1) = \left\{ f \in \mathbb{R}^{n+1} \mid f_i > 0, \forall i \in I, \text{ and } \sum_{i \in I} (f_i)^2 = 4 \right\}.
  \]

Pull-back from the sphere, we find again the geodesics and parallel transport.

Transfer Learning

- We have the data $(x_i, y_i)^{n+1} \subset T_p P_+(I)$ and its labels $(y_i)^{n+1} \in I$.
- A linear regression model in $T_p P_+(I)$ has the following form
  \[
  X \mapsto Y(X) = X^T a + a_0 = \left( X, G a \right) + a_0,
  \]
  where $a_0 \in \mathbb{R}$, while $a$ and $G a = c$ are considered as tangents vectors on $T_p P_+(I)$, and $G$ is the matrix induced by Riemannian metric.
- Let $\text{loss}$ be the loss function associated with $y_i$, e.g., $\text{loss} : y_i \mapsto (y_i - y_i)^2$.
- Transport the Model:
  \[
  (\beta, \alpha) = \arg\min_{\beta, \alpha} \sum_{i \in I} \text{loss}(X^T a + a_0) \Rightarrow G \Gamma_{\alpha \to \beta} = \arg\min_{\beta, \alpha} \sum_{i \in I} \text{loss}(\Gamma(X)) T b + \beta_0,
  \]
  where $\Gamma$ is the parallel transport from $T_p P_+(I)$ to $T_p P_+(I)$.

Box-plots of the score for the Logistic Regression Model

The Models learned and test in the same spaces

The transferred models learned from S0

S0 is the uniform measure, while C1, ..., C10 are chosen randomly.

Acknowledgments

- This work was funded by ANR-IA project.

References