T-122.102 Special Course in Information Technology

Information diffusion kernels

Based on the technical report by John Lafferty and Guy Lebanon, (2004)
Diffusion Kernels on Statistical Manifolds (CMU-CS-04-101)

Sven Laur
Helsinki University of Technology
swen@math.ut.ee,slaur@tcs.hut.fi

Outline

- The problem and motivation
- From data to distribution
- What is a reasonable geometry over the distributions?
  - ⋆ Coordinates, tangent vectors, distances etc.
- Why heat diffusion?
  - ⋆ Geodesic distance *vs*. Mercer kernel, Gaussian kernels.
- Building a model
- Extracting an approximate kernel

How to build kernels for discrete data structures?

- Simple embedding of discrete vectors to $\mathbb{R}^n$
  - ⋆ Works with vectors of fixed length
  - ⋆ It is *ad hoc* technique
- Embedding via generative models
  - ⋆ Theoretically sound
  - ⋆ What should be the right proximity measure?
  - ⋆ Proximity measure should be independent of parameterization!

Parameterization invariant kernel methods

- Fisher kernels

$$K(\boldsymbol{x},\boldsymbol{y}) = \langle \nabla \ell(\boldsymbol{x}|\theta), \nabla \ell(\boldsymbol{y}|\theta) \rangle$$

- Information diffusion kernels

$$K(\boldsymbol{x},\boldsymbol{y}) = ???$$

- Mutual information kernels (Bayesian prediction probability)

$$K(\boldsymbol{x},\boldsymbol{y}) = \Pr\left[\boldsymbol{y}|\boldsymbol{x}\right] \propto \int p(\boldsymbol{y}|\theta)p(\boldsymbol{x}|\theta)p(\theta)d\theta$$

integrated over model class $\mathcal{P}$ with prior probability $p(\theta)$.

Text classification

- Bag of word approach produces a count vector $(x_1, \ldots, x_n)$

- Let the model class be a multinomial distribution.

- MLE estimate is

$$\widehat{\theta}_{\mathsf{tf}}(\boldsymbol{x}) = \frac{1}{x_1 + \cdots + x_n}(x_1, \ldots, x_n).$$

- Second embedding is inverse document frequency weighting

$$\widehat{\theta}_{\mathsf{tfidf}}(\boldsymbol{x}) = \frac{1}{x_1 w_i + \cdots + x_n w_n}(x_1 w_i, \ldots, x_n w_n)$$
$$w_i = \log(1/f_i)$$

What is a statistical manifold?

- Statistical manifold is a family of probability distributions

$$\mathcal{P} = \{p(\cdot|\theta) : \mathcal{X} \to \mathbb{R} : \theta \in \Theta\} ,$$

where Θ is open subset of $\mathbb{R}^n$.

- The parameterization must be unique

$$p(\cdot|\theta_1) \equiv p(\cdot|\theta_2) \qquad \Longrightarrow \qquad \theta_1 = \theta_2$$

- Parameters θ can be treated as the coordinate vector of $p(\cdot|\theta)$

Set of admissible coordinates and distributions

- The parameterization ψ is admissible iff ψ as a function of primary parameters θ is C^∞ smooth.

- The set of admissible parameterization is an invariant.

- We consider only such manifolds where log-likelihood function $\ell(\boldsymbol{x}|\theta) = \log p(\boldsymbol{x}|\theta)$ is C^∞ differentiable w.r.t θ.

- The multinomial family satisfies the C^∞ requirement

$$\ell(\boldsymbol{x}|\theta) = \log \prod_{j=1}^{m} \theta_{x_j} = \sum_{j=1}^{m} \log \theta_{x_j}.$$

Geometry ≈ distance measure

- Distance measure determines geometry. This can be reversed.

- Recall that the length of a path $\gamma : [0, 1] \to \mathcal{P}$

$$d(p, q) = \int_0^1 \|\dot{\gamma}(t)\| dt = \int_0^1 \sqrt{\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle} dt,$$

  where $\dot{\gamma}(t)$ is a tangent vector.

- But the set $\mathcal{P}$ does not have any geometrical structure!!!

- We redefine (tangent) vectors—vectors will be operators.

What is a vector?

- Vector will be operator that maps C^∞ functions $f : \mathcal{P} \to \mathbb{R}$ to reals. For fixed coordinates θ and point p natural maps $(\frac{\partial}{\partial \theta_i})_p$ emerge
$$\left(\frac{\partial}{\partial \theta_i}\right)_p (f) = \frac{\partial f}{\partial \theta_i}\bigg|_p .$$
They will be basis of tangent space.

- For arbitrary differentiable γ we can express
$$f(\gamma(t))' = \left[\theta_1(t)'\left(\frac{\partial}{\partial \theta_1}\right)_{\gamma(t)} + \cdots \theta_n(t)'\left(\frac{\partial}{\partial \theta_n}\right)_{\gamma(t)}\right](f).$$
The operator in the square brackets does not depend on f and has right type—it will be a speed/tangent vector.

Is this a reasonable definition?

- The speed vector $\dot\gamma(t)$ uniquely characterizes the rate of change of arbitrary admissible function f

$$\dot\gamma(t)(f) = f(\gamma(t))'_t$$

- There is a one-to-one correspondence

$$\dot\gamma(t) \xmapsto{\theta} (\dot\theta_1(t), \ldots, \dot\theta_n(t)) \in \mathbb{R}^n.$$

- The are coordinate transformation formulas between different bases

$$\left(\frac{\partial}{\partial\theta_i}\right)_{i=1}^n \quad \text{and} \quad \left(\frac{\partial}{\partial\psi_i}\right)_{i=1}^n$$

- We really cannot expect more, if there is no geometrical structure!!!

Kullback-Leibler divergence

- The most reasonable distance measure between adjacent distributions p and q is the weighted Kullback-Leibler divergence

$$J(p, q) = D_{p||q} + D_{q||p}$$
$$= \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x} + \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x},$$

- It quantifies additional utility if we use wrong distribution.

- In discrete case it means that we need $J(p, q)$ times more bits for encoding.

What is a reasonable distance metrics?

Consider an infinitesimal movement along the curve $\gamma(t)$.

- The corresponding change of coordinates is from θ to $\theta + \dot{\theta}\Delta t$ and the distance formula gives

$$d(p,q)^2 \approx \Delta t^2 \|\dot{\gamma}(t)\|^2 = \Delta t^2 \sum_{i,j=1}^{n} \dot{\theta}_i \dot{\theta}_j \left\langle \frac{\partial}{\partial \theta_i}, \frac{\partial}{\partial \theta_j} \right\rangle$$

- Under mild regularity conditions

$$J(p,q) \approx \Delta t^2 \sum_{i,j=1}^{n} \dot{\theta}_i \dot{\theta}_j g_{ij}, \quad g_{ij} = \int p(\boldsymbol{x}) \cdot \frac{\partial \ell(\boldsymbol{x}|\theta)}{\partial \theta_i} \cdot \frac{\partial \ell(\boldsymbol{x}|\theta)}{\partial \theta_j} d\boldsymbol{x}.$$

- Hence, the local requirement $d^2(p,q) \approx J(p,q)$ fixes geometry

$$\left\langle \frac{\partial}{\partial \theta_i}, \frac{\partial}{\partial \theta_j} \right\rangle = g_{ij}.$$

Limitations of geodesic distance

- Geodesic distance $d(p, q)$ is the shortest path between p and q.
- Geodesic distance cannot be always used for SVM kernels
  - ⋆ SVM kernel (Mercer kernel) is a computational shortcut of

    $$K(\boldsymbol{x}, \boldsymbol{y}) = \Psi(\boldsymbol{x}) \cdot \Psi(\boldsymbol{y}),$$

    where $\Psi : \mathbb{R}^n \to \mathbb{R}^d$ is a smooth enough function.
  - ⋆ If geodesic distance corresponds to a Mercer kernel then there must be only one shortest path between two points.

Classification via temperature

- Consider two classes ”hot” and ”cold”, i.e. each data point has a an initial amount of heat λ_i concentrated around a small neighborhood.
- All other points have zero temperature.
- Fix a time moment t. All points below zero belong to the class ”cold” and others to the class ”hot”.
- Heat gradually diffuses over the manifold. If $t \to \infty$ all points have constant temperature. Varying t gives different levels of smoothing.
- Large t gives flatter decision border that is classification is more robust, but also a less sensitive.

How to model heat diffusion?

- Classical heat diffusion is given by partial differential equations

$$\frac{\partial f}{\partial t} - \Delta f = 0$$
$$f(x, 0) = f(x)$$

and by Dirichlet' or von Neumann boundary conditions.

- In non-Euclidean geometry Laplace operator has a nasty form

$$\Delta f = \det G^{-1/2} \sum_{i,j=1}^{n} \frac{\partial}{\partial \theta_j} \left[g^{ij} \det G^{1/2} \frac{\partial f}{\partial \theta_i} \right]$$

where g^{ij} are elements of inverse Fisher matrix G.

Extracting the kernel

- In the Euclidean space $\mathbb{R}^n$

$$\Delta f = \frac{\partial^2 f}{\partial x_1^2} + \cdots + \frac{\partial^2 f}{\partial x_n^2}.$$

- The solution corresponding to initial condition $f(\boldsymbol{x})$

$$f(\boldsymbol{x}, t) = (4\pi)^{-n/2} \int \exp\Big(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{4t}\Big) f(\boldsymbol{y}) d\boldsymbol{y}$$

- Alternatively

$$f(\boldsymbol{x}, t) = \int K_t(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{y}) d\boldsymbol{y} \qquad K_t(\boldsymbol{x}, \boldsymbol{y}) = \exp\Big(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{4t}\Big)$$

- In SVM-s $f = \lambda_1 \delta_{\boldsymbol{x}_1} + \cdots + \lambda_k \delta_{\boldsymbol{x}_k}$ and integral collapses to a sum.

Central theoretical result

Theorem

Let M be a complete Riemannian manifold. Then there exists a kernel function K (heat kernel), which satisfies the following properties:

(1) $K(\boldsymbol{x}, \boldsymbol{y}, t) = K(\boldsymbol{y}, \boldsymbol{x}, t)$;

(2) $\lim_{t \to 0} K(\boldsymbol{x}, \boldsymbol{y}, t) = \delta(\boldsymbol{x}, \boldsymbol{y})$;

(3) $(\Delta - \frac{\partial}{\partial t}) K(\boldsymbol{x}, \boldsymbol{y}, t) = 0$;

(4) $K(\boldsymbol{x}, \boldsymbol{y}, t) = \int K(\boldsymbol{x}, \boldsymbol{z}, t - s) K(\boldsymbol{z}, \boldsymbol{y}, s) d\boldsymbol{z}$.

The assertion means:

(1) if q converges parameter-wise p then $J(p, q) \to 0$;

A "slight" drawback!

- There are few know closed form solutions of heat diffusion kernel.
- The approximation makes things complicated

$$K_t(\boldsymbol{x},\boldsymbol{y}) \approx K_t^{(m)} = (4\pi t)^{-n/2} \exp\Big(-\frac{d^2(\boldsymbol{x},\boldsymbol{y})}{4t}\Big) \Big[\psi_0(\boldsymbol{x},\boldsymbol{y}) + \psi_1(\boldsymbol{x},\boldsymbol{y})t + \cdots + \psi_m(\boldsymbol{x},\boldsymbol{y})t^m\Big],$$

  where $d(\boldsymbol{x},\boldsymbol{y})$ corresponds to geodesic distance.
- Nasty but closed form formula for approximation terms exist.
- The approximation error is $O(t^m)$.
- The approximation does not have to be a Mercer kernel.

Example: Geometry of multinomials

It is straightforward to compute Fisher information matrix of multinomial family

$$g_{ij} = \begin{cases} 0, & \text{if } i \neq j, \\ 1/\theta_i, & \text{if } i = j. \end{cases}$$

- There is no known closed form solutions.
- We need an easy way to compute geodesic distances.

Isometry—a way to simplify things

- Isometry is C^∞ differentiable map $F : \mathcal{P} \to \mathcal{S}$ that preserves lengths of paths.

- The model will be $n+1$ dimensional positive orthant in $\mathbb{R}^{n+1}$

$$\mathcal{S}^+ = \left\{(x_1, \dots, x_{n+1}) : x_1^2 + \cdots + x_{n+1}^2 = 4\right\}.$$

- It is easy to verify that

$$F(\theta_1, \dots, \theta_n) = (2\sqrt{\theta_1}, \dots, 2\sqrt{\theta_{n+1}})$$

preserves lengths, ie. the length of vectors along curves are always same.

Example: Distances of trinomials

Explicit form of multinomial kernel

- Since the shortest paths on the spheres are big circles

$$
\begin{aligned}
d(\theta, \theta') &= 2 \arccos(\langle F(\theta), F(\theta') \rangle) \\
&= 2 \arccos\Big(\sqrt{\theta_1 \theta_1'} + \cdots + \sqrt{\theta_{n+1} \theta_{n+1}'}\Big),
\end{aligned}
$$

where $\theta_{n+1} = 1 - \theta_1 - \ldots - \theta_m'$ and $\theta_{n+1} = 1 - \theta_1 - \ldots - \theta_m'$.

- For the first order approximation $O(t)$ it is sufficient to use

$$
K_t(\theta, \theta') = (4\pi t)^{-n/2} \exp\Big(-\frac{\arccos^2(\sqrt{\theta}, \sqrt{\theta'})}{t}\Big).
$$

- Compared with Gaussian kernel works better if the data is close to edges.

Gaussian *vs.* heat kernel

Conclusion

- Information geometry provides parameterization independent kernels.

- Devising a kernel for more complex models requires *enormous intellectual effort*.

- However, nothing stops us from using already derived kernels.

- SLT bounds are available — the asymptotic generalization performance is essentially the same as Gaussian kernels with the same dimension.