

Diffusion Kernels on Statistical Manifolds

Juan Carlos Arango Parra

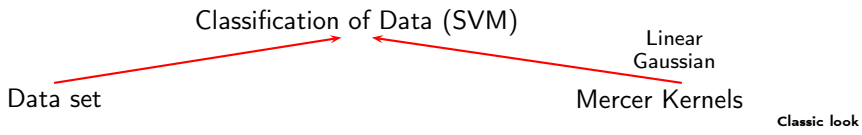
September 29, 2017

Advisers: PhD Gabriel Ignacio Loaiza Ossa
PhD Carlos Alberto Cadavid Moreno

Doctoral Seminar 1
Universidad EAFIT
Department of Mathematical Sciences
PhD in Mathematical Engineering
2017

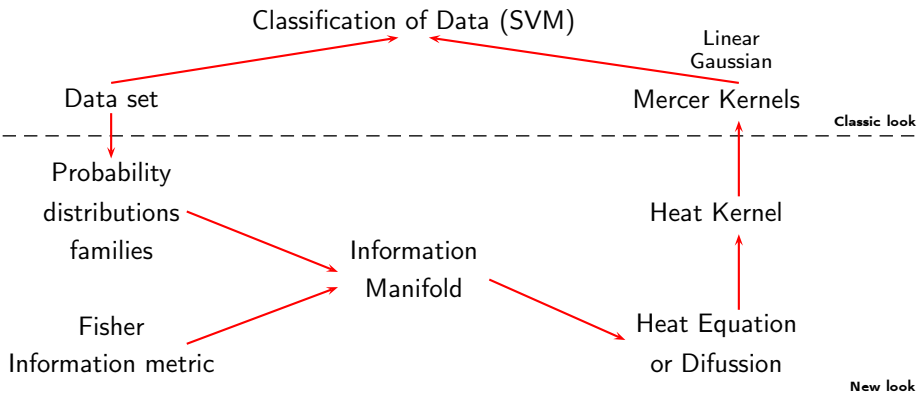
Objective of the article

Lafferty and Lebanon propose in their article *Diffusion Kernels on Statistical Manifold* (Laferty and Lebanon, 2005) the following way of work



Objective of the article

Lafferty and Lebanon propose in their article *Diffusion Kernels on Statistical Manifold* (Laferty and Lebanon, 2005) the following way of work



Support Vector Machine - SVM

Mathematically, a support vector machine builds a hyperplane that can be used to do classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the greatest distance to the closest training data points of any class (called functional margin), since in general, the larger the margin, the less generalization error of the classifier. (Cárdenas, 2015; Scikit Learn, 2017).

What is Kernel in mathematics?

Let X be non-empty set. A symmetric function $K : X \times X \rightarrow \mathbb{R}$ is called positive definite Kernel in X if

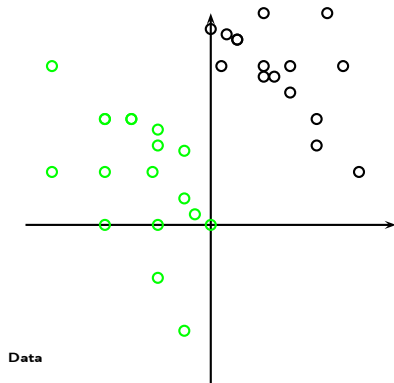
$$\sum_{i=1}^n \sum_{j=1}^m c_i c_j K(x_i, x_j) \geq 0$$

is verified for any $n, m \in \mathbb{N}$, each $x_i \in X$ and $c_i \in \mathbb{R}^+$. Examples of Kernels are

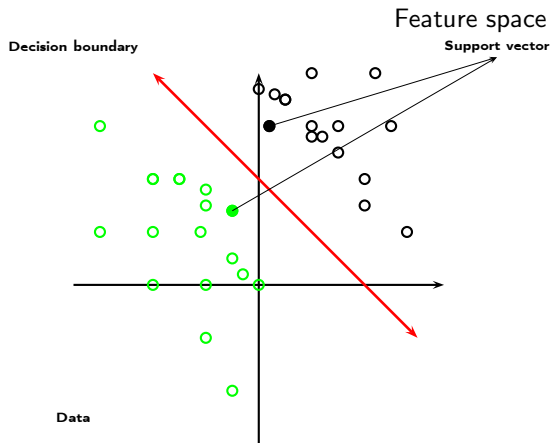
- 1 Linear: $K(x, y) = x^T y$, where $x, y \in \mathbb{R}^d$.
- 2 Polynomial: $K(x, y) = (x^T y + r)^n$, where $x, y \in \mathbb{R}^d$ and $r > 0$.
- 3 Gaussian (RBF): $K(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$.
- 4 Sigmoid: $K(x, y) = \tanh(\alpha x^T y + r)$ where α and r are constant.

Example SVM

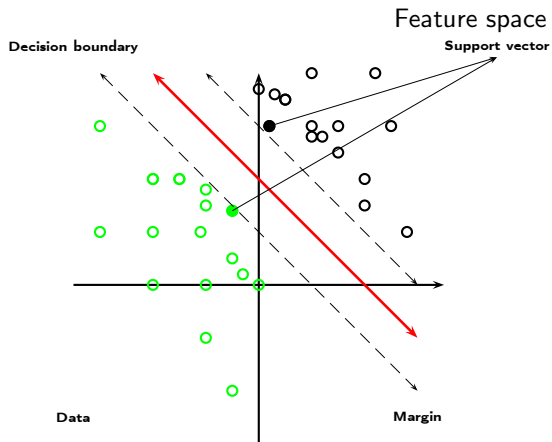
Feature space



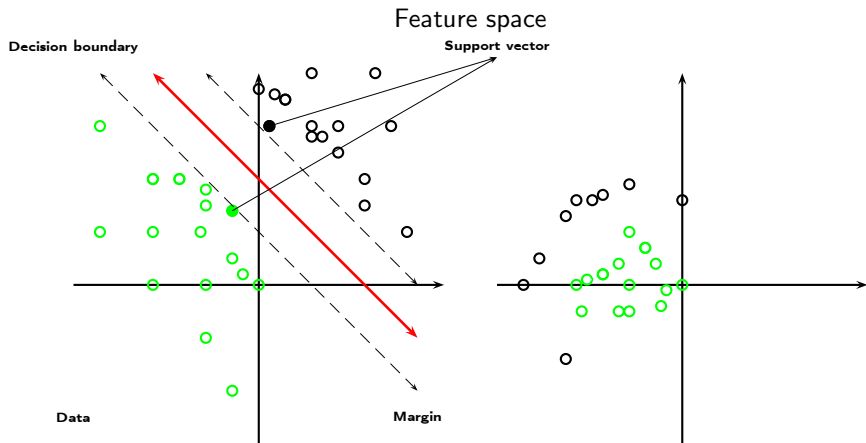
Example SVM



Example SVM



Example SVM



Example SVM

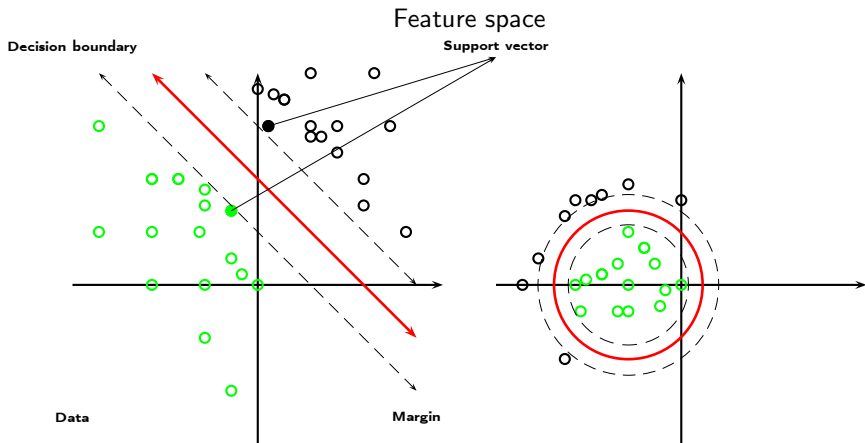


Figure 1: Ideas about the operation of SVM

Riemannian manifold

Let M be a topological space. A chart in M consists of a pair (U, ϕ) , where U is an open set in M and ϕ is a bijection of U to some open set A of \mathbb{R}^n . Given a set of indexes I , a collection of charts on M

$$\mathcal{A} = \{(U_i, \phi_i) : i \in I\}$$

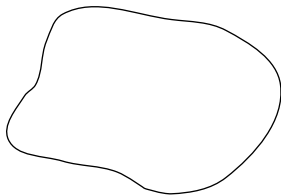
where $\phi_i : U_i \rightarrow A_i$ is called a C^p -atlas with $p \geq 0$, when the following conditions are verified

- 1 the family $\mathcal{F} = \{U_i : i \in I\}$ is a covering of M ,
- 2 for all $i, j \in I$, $\phi_i(U_i \cap U_j)$ is an open set,
- 3 for all $i, j \in I$, with $U_i \cap U_j \neq \emptyset$, the mapping (called transition function)

$$\phi_j \circ \phi_i^{-1} : \phi_i(U_i \cap U_j) \rightarrow \phi_j(U_i \cap U_j) ,$$

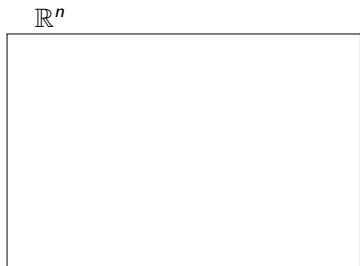
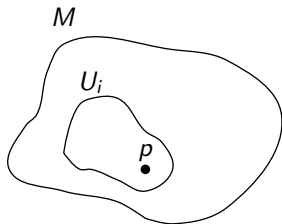
is an isomorphism of class C^p . If the transition function are differentiable then the manifold is differentiable (Loaiza and Quiceno, 2013)

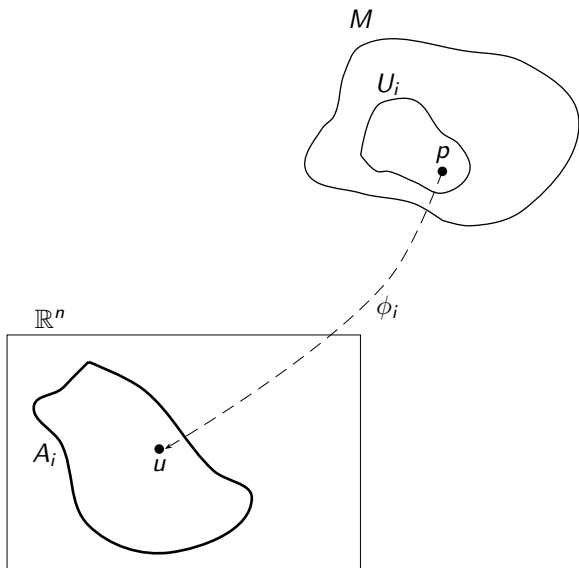
M

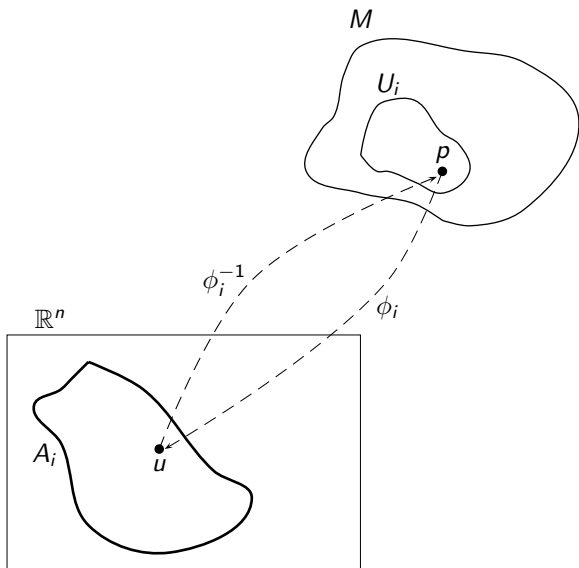


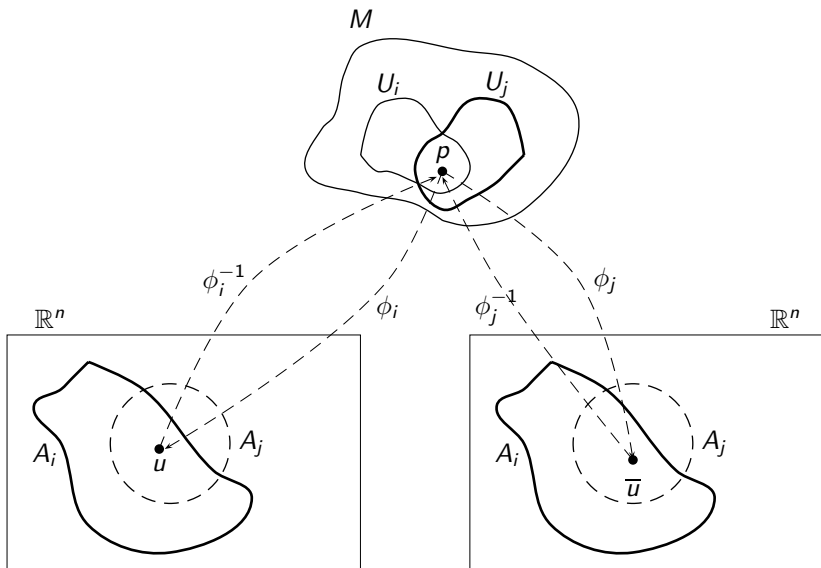
\mathbb{R}^n











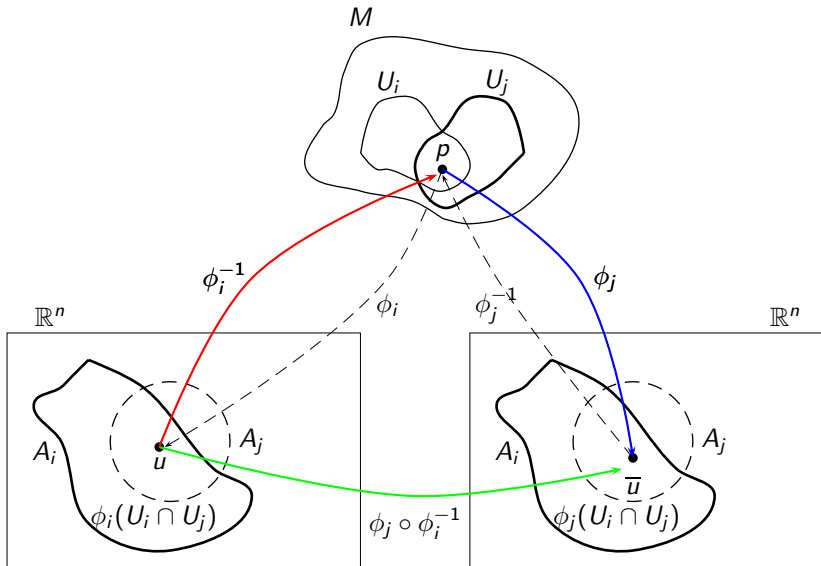


Figure 2: Topological Manifold

Riemannian Metric

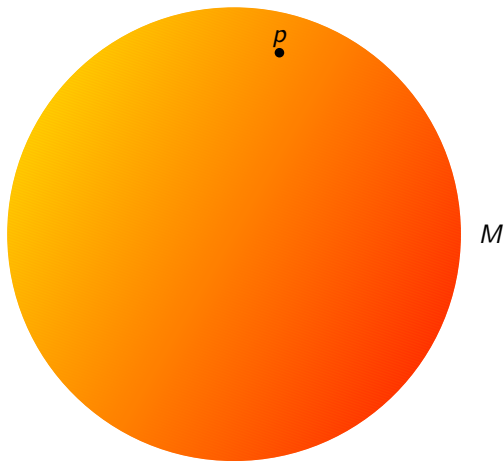
A Riemannian metric on a differentiable manifold M is a mapping that assigns each point p of the variety M an inner product g_p on the tangent space $T_p M$; thus, for every two vector fields X_p, Y_p en M , the function $g_p(X_p, Y_p) = \langle X_p, Y_p \rangle_p$ is smooth. A M variety provided with a Riemannian metric is called *Riemannian Manifold*. The metric does not depend on the choice of the local coordinates, in addition it can be expressed

$$g_p(X_p, Y_p) = \sum_{i,j}^n v_i w_j \underbrace{\langle \partial_{i|p}, \partial_{j|p} \rangle}_{g_{ij}} \quad \text{with } \partial_i = \frac{\partial}{\partial x_i}$$

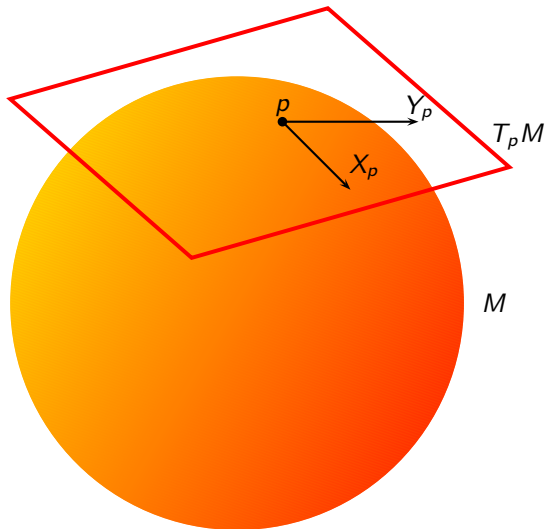
where v_i y w_j are the coefficients in the representation of X_p y Y_p in the canonical basis of the tangent space $T_p(M)$ given by $\{\partial_{1|p}, \dots, \partial_{n|p}\}$. The terms $g_{ij}(p)$ represent the entries of the matrix $g(p)$ which is symmetric and definite positive.

The existence (always exists at least one) of a metric allows defining the length of vectors (of curves) . The curve for which the shortest distance is presented is called *geodesic* (Burns and Gidea, 2005).

Example of a Riemannian manifold: spherical surface



Example of a Riemannian manifold: spherical surface



Example of a Riemannian manifold: spherical surface

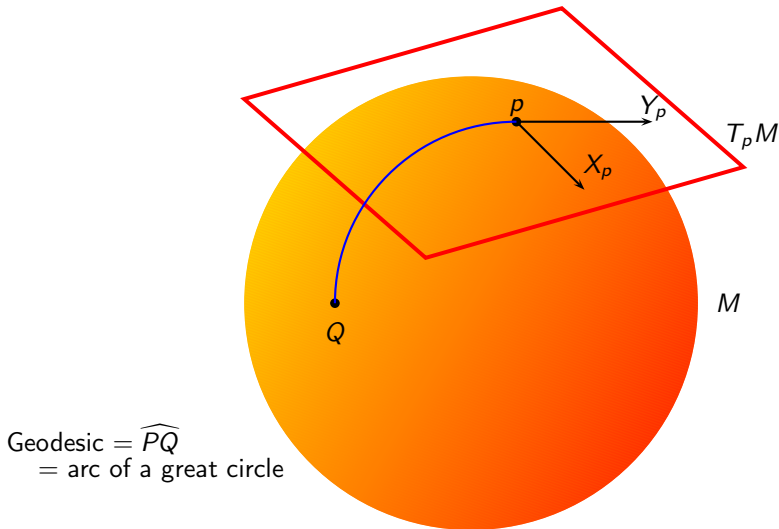


Figure 3: Space tangent and geodesic in a sphere

Laplacian Operator

Let (M, g) a Riemannian manifold. For any smooth function f over M , the gradient is defined as the vector field $grad(f)$ in $T(M)$ that satisfies $\langle grad(f), X \rangle_g = X(f)$ for all $X \in T(M)$, in local coordinates, the gradient is written as $(grad(f))_i = \sum_j g^{ij} \frac{\partial f}{\partial x_j}$ where g^{ij} are the components of the inverse of the matrix $g = [g_{ij}]$, then

$$grad(f) = \sum_{i,j=1}^n \underbrace{\left(g^{ij} \frac{\partial f}{\partial x_j} \right)}_{(grad(f))_i} \partial_i .$$

In these same local coordinates, the divergent of X is written

$$div(X) = \frac{1}{\sqrt{\det g}} \sum_i \frac{\partial}{\partial x_i} \left(\sqrt{\det g} X_i \right) .$$

The Laplacian or Laplace-Beltrami operator on (M, g) of a smooth function $f : M \rightarrow \mathbb{R}$ is defined as

$$\Delta_g f = div(grad(f)) = \frac{1}{\sqrt{\det g}} \sum_j \frac{\partial}{\partial x_j} \left(\sum_i g^{ij} \sqrt{\det g} \frac{\partial f}{\partial x_i} \right) .$$

Heat Equation

The heat equation on (M, g) is the partial differential equation $\frac{\partial f}{\partial t} = \Delta_g f$. A solution to the problem with initial condition

$$\begin{cases} \frac{\partial f}{\partial t} = \Delta_g f \\ f(\cdot, 0) = f_0 \in L^2(M) \end{cases} \quad (1)$$

is a continuous function $f : M \times [0, \infty) \rightarrow \mathbb{R}$ denoted $f(x, t)$ which describes the temperature at a point x at time t beginning with an initial heat distribution described by the initial condition $f(x, 0)$. This solution is such that for each fixed $t > 0$, $f(\cdot, t)$ is a function C^2 , and for each $x \in M$, $f(x, \cdot)$ is C^1 (Cadavid and Vélez, 2014).

Heat Kernel

The Heat Kernel $K_t(x, y)$ is the solution to the initial condition heat equation given by the Dirac delta function δ_x . This Heat Kernel allows to describe the other solutions of the heat equation by means of convolution as

$$f(x, t) = \int_M K_t(x, y) f_0(y) dy .$$

It also satisfies the following properties

- 1 $K_t(x, y) = K_t(y, x)$ (Symmetric)
- 2 $(\Delta - \frac{\partial}{\partial t}) K_t(x, y) = 0$ (Solution of the Heat Equation)
- 3 $K_t(x, y) = \int_M K_{t-s}(x, z) K_s(z, y) dz$ for any $s > 0$ (Definite positive)

According to these properties, the Heat Kernel is also a Mercer Kernel.

In the case of a n -dimensional flat Euclidean space, the Heat Kernel has the form

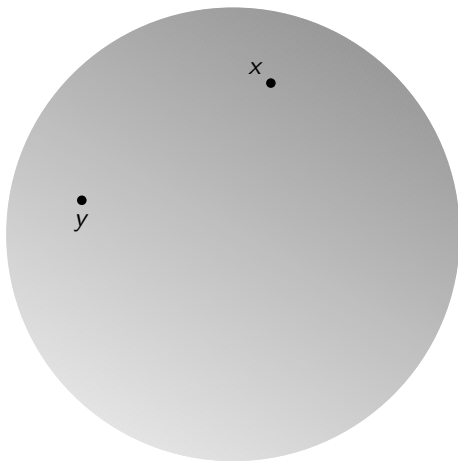
$$K_t(x, y) = \underbrace{\frac{1}{(4\pi t)^{n/2}} \exp\left(-\frac{\|x - y\|^2}{4t}\right)}_{\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)} = \frac{1}{(4\pi t)^{n/2}} \exp\left(-\frac{d^2(x, y)}{4t}\right)$$

where $\|x - y\|^2$ is the square of the Euclidean distance between points x and y . The parametrix expansion approximates the heat kernel locally as a correction to this Euclidean heat Kernel, is written

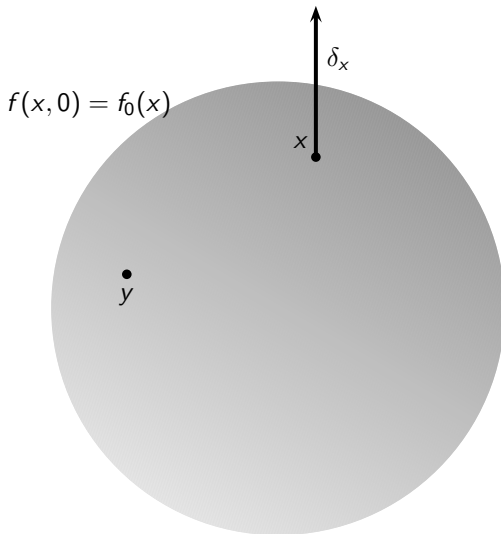
$$P_t^m(x, y) = \frac{1}{(4\pi t)^{n/2}} \exp\left(-\frac{d^2(x, y)}{4t}\right) (\Psi_0(x, y) + \dots + \Psi_m(x, y)t^m)$$

for currently unspecified functions $\Psi_k(x, y)$, where $d^2(x, y)$ denote the square of the geodesic distance on the manifold.

Interpretation of the Heat Equation using the Dirac Delta



Interpretation of the Heat Equation using the Dirac Delta



Interpretation of the Heat Equation using the Dirac Delta

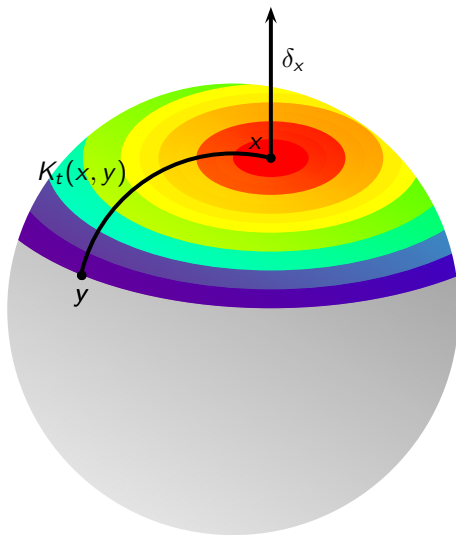


Figure 4: Heat Kernel

Particular cases

- 1 Multinomial Simplex.
- 2 Normal Distribution.

Fisher Information Metric

Let $\mathcal{F} = \{p(\cdot|\theta) : \theta \in \Theta\}$ a statistical family n -dimensional on a certain set X , where Θ is an open set in \mathbb{R}^n , $\theta = (\theta_1, \dots, \theta_n)$ and there is σ -measure μ in X such that for each $\theta \in \Theta$, $p(\cdot|\theta)$ is a probability density with respect to μ . Is denoted $\partial_i = \frac{\partial}{\partial \theta_i}$ and $\ell_\theta(x) = \log p(x|\theta)$. The Fisher information metric is defined in terms of the matrix $g(\theta)$ where its components are given by

$$g_{ij}(\theta) = E_\theta[\partial_i \ell_\theta \partial_j \ell_\theta] = \int_X p(x|\theta) \partial_i \ell_\theta(x) \partial_j \ell_\theta(x) d\mu(x).$$

Other equivalences of this Fisher information matrix are

$$g_{ij}(\theta) = 4 \int_X \partial_i \sqrt{p(x|\theta)} \partial_j \sqrt{p(x|\theta)} d\mu(x) = - \int_X p(x|\theta) \partial_j \partial_i \ell_\theta(x) d\mu(x).$$

Multinomial Simplex

The multinomial simplex, or simplex, is the set of all nonnegative vectors $\theta = (\theta_1, \dots, \theta_n)$ such that $0 \leq \theta_i \leq 1$ and their sum is one, we write

$$M = P_n = \left\{ \theta \in \Theta : \sum_{i=1}^{n+1} \theta_i = 1 \right\}.$$

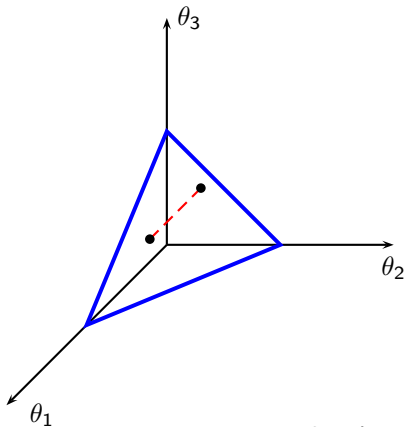
Through the transformation, $z_i = F(\theta_i) = 2\sqrt{\theta_i}$ (isometry), each point in the n-multinomial is applied at a point of the n-sphere of radius 2, hence the multinomial information geometry is the geometry of the sphere in the first octant of the Euclidean space, where the geodesic distance between two points θ y θ' is given by

$$d(\theta, \theta') = 2 \arccos \left(\sum_{i=1}^{n+1} \sqrt{\theta_i \theta'_i} \right).$$

The solution to the heat equation through the parametrix is expressed as

$$K_t(\theta, \theta') \approx (4\pi t)^{-n/2} \exp \left(-\frac{1}{t} \arccos^2 \left(\sum_{i=1}^{n+1} \sqrt{\theta_i \theta'_i} \right) \right).$$

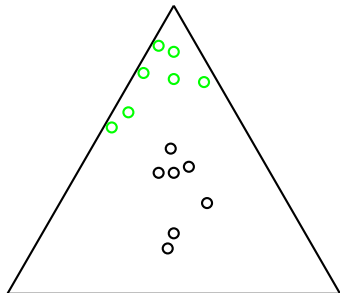
If $n = 2$, the 2-simplex is of the form $\theta_1 + \theta_2 + \theta_3 = 1$ (trinomial distribution), is a 2-dimensional manifold (a triangle) as shown in the graph. The mapping $z_i = 2\sqrt{\theta_i}$ is the sphere of radius 2 which is also 2-dimensional.



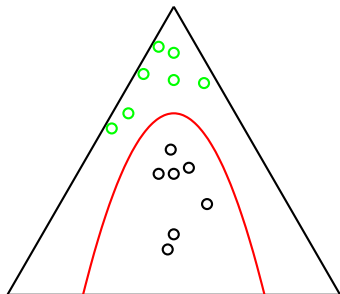
$$z_1^2 + z_2^2 + z_3^2 = 4(\theta_1 + \theta_2 + \theta_3)$$
$$z_1^2 + z_2^2 + z_3^2 = 4$$

Figure 5: Simplex 2-dimensional

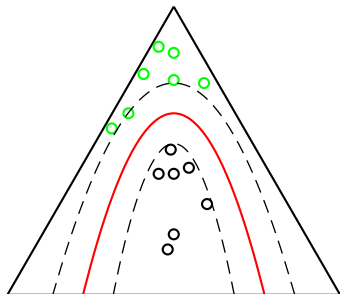
Trinomial distribution



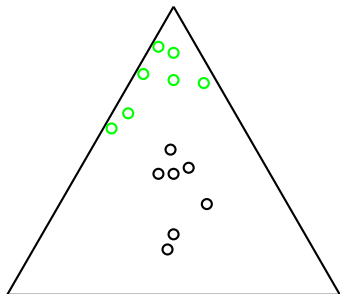
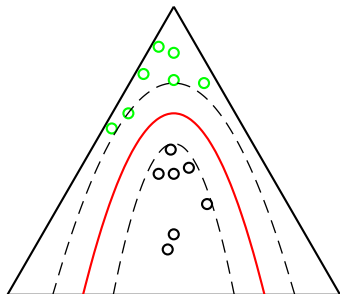
Trinomial distribution



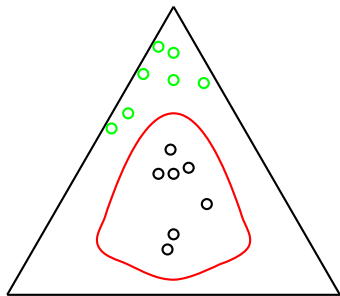
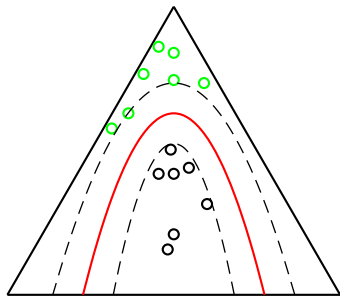
Trinomial distribution



Trinomial distribution



Trinomial distribution



Trinomial distribution

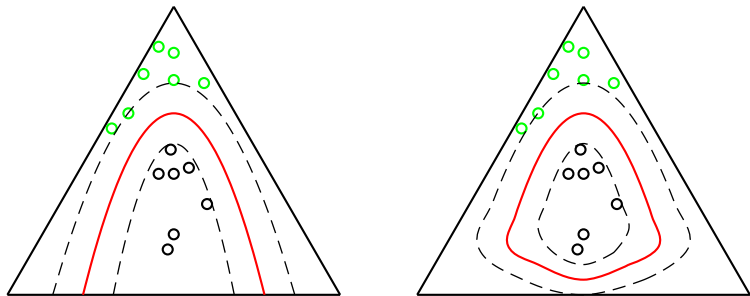


Figure 6: Kernel Gaussian vs Diffusion Kernel

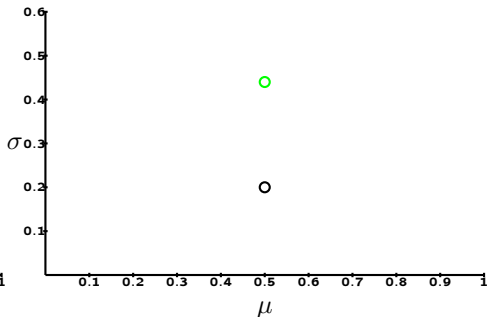
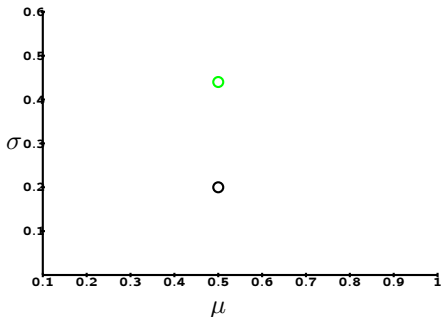
Normal Distribution

In this case the family of normal distributions $\mathcal{N}(\mu, \sigma I_{n-1}) = \{p(\cdot | \theta) : \theta = (\mu, \sigma)\}$ where $\mu \in \mathbb{R}^{n-1}$ is the mean and σ is the variance. In this space it is possible to show that the coefficients of the Fisher information matrix are written in the form $g_{ij} = \frac{\sqrt{2}}{\sigma^2} \delta_{ij}$; this metric confers to this manifold the structure of the superior plane in a hyperbolic space \mathbb{H}^n , where the heat kernel takes the form

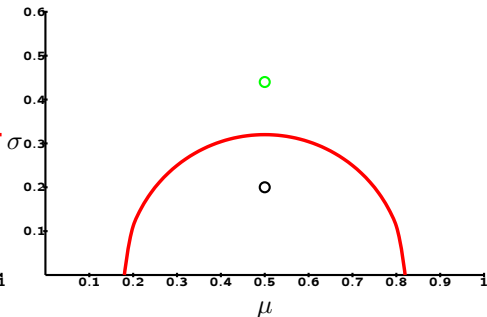
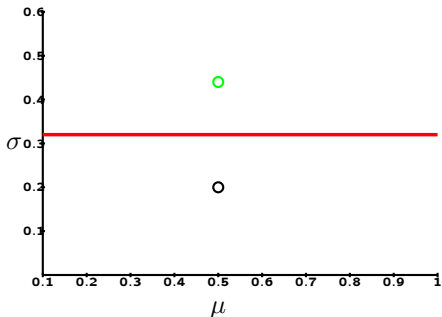
$$K_t(x, x') = \begin{cases} \frac{(-1)^m}{(2\pi)^m} \frac{1}{\sqrt{4\pi t}} \left(\frac{1}{\sinh \rho} \frac{\partial}{\partial \rho} \right)^m \exp\left(-m^2 t - \frac{\rho^2}{4t}\right) & \text{Si } n = 2m + 1 \\ \frac{(-1)^m}{(2\pi)^m} \frac{\sqrt{2}}{\sqrt{4\pi t^3}} \left(\frac{1}{\sinh \rho} \frac{\partial}{\partial \rho} \right)^m \int_{\rho}^{\infty} \frac{s \exp\left(-\frac{(2m+1)^2 t}{4} - \frac{s^2}{4t}\right)}{\sqrt{\cosh s - \cosh \rho}} ds & \text{Si } n = 2m + 2 \end{cases}$$

where $\rho = d(x, x')$ is the geodesic distance between the two points in the plane \mathbb{H}^n .

In the case where $n = 1$, that is, $m = 0$ then $K_t(x, x') = \frac{1}{\sqrt{4\pi t}} \exp\left(-\frac{d^2(x, x')}{4t}\right)$ which is the Gaussian Kernel in \mathbb{R} . The following graph shows how the decision boundaries would be in the hyperbolic space for these normal distributions.



In the case where $n = 1$, that is, $m = 0$ then $K_t(x, x') = \frac{1}{\sqrt{4\pi t}} \exp\left(-\frac{d^2(x, x')}{4t}\right)$ which is the Gaussian Kernel in \mathbb{R} . The following graph shows how the decision boundaries would be in the hyperbolic space for these normal distributions.



In the case where $n = 1$, that is, $m = 0$ then $K_t(x, x') = \frac{1}{\sqrt{4\pi t}} \exp\left(-\frac{d^2(x, x')}{4t}\right)$ which is the Gaussian Kernel in \mathbb{R} . The following graph shows how the decision boundaries would be in the hyperbolic space for these normal distributions.

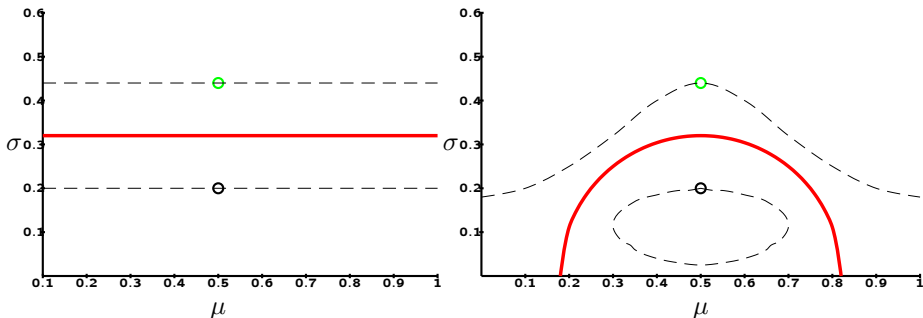








Figure 7: Hyperbolic Space \mathbb{H}^2

What follows after this conceptual review of the article?

- 1 To delve into the mathematical problem of Text Classification with SVM.
- 2 Make an implementation in Python or MatLab of SVM that shows the advantages of diffusion kernels.
- 3 Find other diffusion kernels for different distribution families of probability (Poisson, Beta, Hypergeometric, etc.)

-  Burns; Keith and Gidea; Marian. *Differential Geometry and Topology. With a View to Dynamic System*. Studies in advance mathematics. Chapman & Hall / CRC, 2005.
-  Cadavid; Carlos and Vélez; Juan Diego. *A Remark on the Heat Equation and Minimal Morse Functions on Tori and Spheres*. Ingeniería y Ciencia, Vol. 09, No. 17, Enero-Junio 2013. EAFIT.
-  Cardenas Montes; Miguel. *Support Vector Machine. Graphs, Statistics and Data Mining with Python*. Presentation, November of 2015.
http://wwae.ciemat.es/~cardenas/docs/curso_MD/svm.pdf
-  Loaiza; Gabriel and Quiceno; Héctor. *A q -exponential statistical Banach manifold*. Journal of Mathematical Analysis and Applications, 398, 2013.
-  Laferty; John and Lebanon; Guy. *Diffusion Kernels on Statistical Manifolds*. Journal of Machine Learning Research, 6 (2005), pp. 129-163.
-  *1.4. Support Vector Machines*. Scikit Learn.
<http://scikit-learn.org/stable/modules/svm.html>. September 23/2017.