NEML CW

Ben James, Johns Noble

March 2025

Contents

1 Introduction

It is often very natural to attempt to model data using gaussian distribution. Gaussian distributions are well defined for data points which lie in euclidean space for any number of dimensions. We however know that most data should be modelled as to be taken from a manifold. There are methods to do manifold learning but once we have a manifold the notion of probability distributions need to be altered such that we are defining a probability measure across a manifold. It does not make sense for a region outside the manifold to possess non-zero measure. The Locally Adaptive Normal Distribution (LAND) aims to attempt to do this. Once we have a notion of normal distribution on a manifold given some mean and covariance, We can try and fit distribution to the data using MLE approaches. Finally, we will be able to use the fact that we can fit gaussians to data to be able to perform EM algorithms over data to fit a mixture of LANDS over data. TODO: Write something about how manifold learning is handled i.e. Do we know the manifold which data will be distributed on or is this something that needs to be learnt using graphs?

2 Constructing Metrics

In order to define the idea of distances over manifolds and in order to attempt to capture local behaviour of data, we need to a Riemannian Metric which acts

on tangent vectors. The Metric $\mathbf{M(x)}$ gives an inner product $\langle \mathbf{u},\ \mathbf{M(x)v} \rangle$ This allows us to define geodesics on our manifold as the path which minimises the length:

$$\hat{\gamma} = \underset{\gamma}{argmin} \int_0^1 \sqrt{\langle \gamma'(t), \mathbf{M}(\gamma(t))\gamma'(t) \rangle} dt,\ \gamma(0) = \mathbf{x},\ \gamma(1) = \mathbf{y}$$

Once we have $\hat{\gamma}$ it is natural to define our Exponential map given a $\mathbf{v} \in T_{\mathbf{x}}M$ such that $\mathbf{y} = Exp_{\mathbf{x}}(\mathbf{v}) \in M$ where $\hat{\gamma}(t) = Exp_{\mathbf{x}}(t \cdot \mathbf{v})$. We can then begin to define the *Log* mapping as: $Log_{\mathbf{x}}(y) = v \in T_{\mathbf{x}}M$. In order to calculate each geodesic, we must express solve the equation relating christofell symbols. This works out to be a second order differential ODE that in practice can be solved numerically.

A generalised view of thinking about the normal distribution is the distribution over a given space such that given a known mean and covariance for which we have maximum entropy. This way of defining the normal distribution allows us to extend to non euclidean spaces. It can be shown that given a Log map, the distribution that satisfies this is as follows:

$$p_M(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = exp(-\frac{1}{2}\langle Log_{\boldsymbol{\mu}}(\boldsymbol{x}), \boldsymbol{\Sigma}^{-1} Log_{\boldsymbol{\mu}}(\boldsymbol{x}) \rangle)$$

We also need to make considerations on how we can learn a metric tensor so that we can define a riemannian metric. The most obvious thing we can do is to use a single global metric tensor such that $dist^2(\boldsymbol{x}_i, \boldsymbol{x}_j = (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{x}_j)$ where M is symetric and positive definite. The issue with this is that a single global metric is never enough to capture the non linearity of manifolds. We can attempt to also learn several metric tensors for different sections of data. This involves picking a few centres $(\boldsymbol{x_1}, \boldsymbol{x_2} ... \boldsymbol{x_r})$, defining a fixed metric $\boldsymbol{M}_r$ for each data point picking the closest metric. We can generalise this sort of approach by letting:

$$\boldsymbol{M}(\boldsymbol{x}) = \Sigma_{r=1}^{R} \frac{w_r(\boldsymbol{x})}{\Sigma_j w_j(\boldsymbol{x})}$$

Intuitively, we can thinkg of the ratio of w_i to be the responsibility of each centre for $\boldsymbol{x}$. For example in the case of picking the nearest tensor, we can say that:

$$w_r(x) = \begin{cases} 1, & \|\boldsymbol{x} - \boldsymbol{x}_r\|^2_{\boldsymbol{M}_r} \leq \|\boldsymbol{x} - \boldsymbol{x}_j\|^2_{\boldsymbol{M}_j}, \forall j \\ 0, & \text{otherwise} \end{cases}$$

We can experiment with using different w functions and by using various number of centres. To summarise, a metric tensor is constructed using the inverse covariance of training data in a local space. The authors of LAND decided that when constructing metric tensors, that we should only consider the diagonals of the covariance to avoid issues with overfitting and to speed up computation. They also used for function $w_n(\boldsymbol{x}) = exp(-\frac{\|\boldsymbol{x}_n - \boldsymbol{x}\|_2^2}{2\sigma^2})$. Defining for the d value in the diagonal of $\boldsymbol{M}$ to be

$$\boldsymbol{M}(\boldsymbol{x}) = (\Sigma_{n=1}^{N} w_n(\boldsymbol{x})(x_{nd} - x_d)^2 + \rho)^{-1}$$

The ρ is here for numerical stability and σ can be tweaked depending on the densitty and dimensions of the data. In our experiments we begin by assuming that the data lies on a sphere so that we could use already existing Exponential and Log mappings and the metric tensor over a sphere. We were able to write an implementation that learnt a riemannian metric using the formula above.

3 Parametric Inference

We are tasked with finding the mean and covariance matrix of the distribution and we have implemented this using numerical methods. The objective function we are trying to maximise is as follows:

$$\underset{\boldsymbol{\mu}\in M\boldsymbol{\Sigma}\in S^D_{++}}{\operatorname{argmin}} (\phi(\boldsymbol{\mu}\boldsymbol{\Sigma}) = \frac{1}{2N}\Sigma_{n=1}^{N}\langle Log_{\boldsymbol{\mu}}(\boldsymbol{x}_n), \boldsymbol{\Sigma}^{-1}Log_{\boldsymbol{\mu}}(\boldsymbol{x}_n)\rangle + log(C(\boldsymbol{\mu}, \boldsymbol{\Sigma}))$$

Where C here is the normalisation constant of normalisation to ensure its a probability distribution. To compute this value we require to be able to compute $E_{N(0,\boldsymbol{\Sigma})}[\sqrt{|\boldsymbol{M}Exp_{\boldsymbol{\mu}}(\boldsymbol{v})|}]$, intuitively can be thought of as the the expectation of the volume of the manifold. Since this is value is intractable we use monte carlo simulation techniques to be able to compute this by sampling over $N(0, \boldsymbol{\Sigma})$. We are able to work out two grad functions for the objective, one for the mean, one for the covariance. The grad function for the mean is in th form $-\boldsymbol{\Sigma}^{-1}(...)$. Simply applying gradient descent with this grad function however results in unstable convergence due to the condition number of $\boldsymbol{\Sigma}$. We end up chosing a direction which is independent from $\boldsymbol{\Sigma}$ by taking the direction to be $-\boldsymbol{\Sigma}\nabla\phi$ which can be proved to give a direction of objective function decrease. We also must be careful with applying gradient descent over the covariance matrix. This is because these covariance matrices must neccesarily be symmetric positive definit. In order to ensure that this always hold we can decompose $\boldsymbol{\Sigma} = PDP^T$ and therefore $\boldsymbol{\Sigma}^{-1} = PD^{-1}P^T = (PD^{-\frac{1}{2}})(PD^{-\frac{1}{2}})^T = A^TA$ Representing the objectives with $\boldsymbol{A}$ allows us to optimise over $\boldsymbol{A}$ and ensure our covariance matricies are valid.

4 Extending to Mixture Models

5 Conclusion

We were able to show that we could perform gradient descent in the space of parameters to find an approximation for the MLE of a normal distribution on a given manifold. We were also able given data points to learn a manifold and infer the geodesics to produce an exponential and log mapping by solving ODEs.

Interesting cases we found were ones were cases where the variance of the distribution is high comparitive to the size of a spherical manifold. We found in these cases that generally convergence is slower and often inaccurate. This should make sense since we no longer have an injective exponential map and

the tails of the normal distribution are able to "wrap" back around. This effect causes the overall distribution to look less like a normal distribution and more like a distribution with heavy tails. We found that setting the start variance to 0.7 was a good value. We can come up with these values by establishing bounds on the probability that a point appears over π distance away in the tangent space.