Progress Report 1: Generative Models Using Probabilistic Principal Component Analysis

Group 9
Aashima Yuthika(1401071), Deep Patel(1401010), Deval Shah(1401060), Kirtan Modi(1401122), Yash Kotadia(1401114)
School of Engineering and Applied Sciences, Ahmedabad University
Subject: Machine Learning & Algorithms and Optimisation for Big Data
March 12, 2017

***Abstract*—The main aim of this project is to be able to build Generative models using the concept of Probabilistic Principal Component Analysis or PPCA. PPCA essentially tries to fill in the data voids using various methods and assumptions that are discussed more in detail later on in the report. These data voids are usually known as the latent (unobservable) variables. Some of the main reasons for using PPCA instead of PCA are: i) The variance-covariance matrix needs to be calculated, this can be very computation-intensive for large datasets with a high dimensions, ii) Outlying data observations can unduly affect the analysis. The Expectation Maximisation Algorithm is often used for estimating these latent variables, but it is not necessary that we do so. Hence, we have also compared results for matrix reconstruction using PPCA with and without the EM Algorithm.**

I. Introduction

Generative models are being largely studied now by the Machine Learning Community and there are many researchers working on this problem as well. For this project we will be making use of the concept of PPCA to try and achieve this. PPCA is formulated within a maximum-likelihood framework, based on a specific form of Gaussian latent variable model. This leads to multiple mixture models than can combined as a probabilistic mixture, whose parameters can be determined using an EM algorithm. We have discussed both PPCA and EM briefly in the following sections.

II. Probabilistic Principal Component Analysis

PPCA can be used as a general Gaussian density model in addition to reducing dimensions. Maximum-likelihood estimates can also be computed for elements associated with principal components. Apart from these, PPCA Captures dominant correlations with few parameters. Hence, PPCA can be used to estimate what are known as latent variables.

A latent variable model relates a *d*-dimensional observation vector *t* to a corresponding *q*-dimensional vector of latent(unobserved) variables *x*. Latent variable model with linear relationship is given by:

$$t = Wx + \mu + \varepsilon \tag{1}$$

Here, $w = d \times q$ matrix that relates 2 sets of variables.
μ = non-zero mean.
ε = error or noise
Latent variables: x $\sim$ N(0, I).
Error (or noise): $\varepsilon \sim$ N(0, ψ).
Location term (mean): μ
The *d* x *q* matrix *W* relates the set of variables, while the parameter vector μ permits the model to have non-zero mean.

The marginal distribution for the observed data **t** is readily obtained by integrating out the latent variables and is likewise Gaussian: $\mathbf{t} \sim \mathcal{N}(\mu, C)$ where the observation covariance model is specified by $C = WW^T + \sigma^2 I$.

The maximum likelihood estimator for μ is given by the mean of the data, where is S is the simple covariance matrix of the observations t_n

$$S = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mu)(t_n - \mu)^T \tag{2}$$

III. Expectation Maximisation Algorithm

The EM Algorithm basically consists of two steps - the 'E-Step' or the 'Expectation Step' that guesses a probability distribution over completions of missing data in the current model; and the 'M-Step' or the 'Maximisation Step' that re-estimates the model parameter using these completions. It is used in PPCA in the following manner:

- Initialise *W*, *z*
- Impute missing *X* from *W* and *z*
- Estimate *z* on all data based on the current *W*
- Repeat until convergence
- Estimate *W* on all data based on the current *z*

Note, that here we are assuming that we are not given a dataset with completely missing observations.

IV. EM vs PPCA

Although we have used EM algorithm inside the PPCA to estimate the latent variables it is not necessary that this is used. It can be estimated via eigen-decomposition, and incorporated in the probability model. Howvere, there may be an advantage in the EM approach for a large *d* since the presented algorithm doesn't require the computation of the *d x d* covariance matrix ($O(nd^2)$), nor its explicit eigen-decomposition ($O(d^3)$)

V. Results

We have compared the simply PPCA algorithm with the one using EM for latent variable estimation. The image used is 'London.jpg' (attached in the submitted folder). The reconstruction with simple PPCA gave an RMS error of 0.0187, and the one with EM algorithm gave an error of 0.0190. We also removed some entries from the image and performed EM on it to recover it with various parameters for the size of q and number of iterations - itr. The result images for the same are also attached in the submitted folder.

References

[1] Michael E. Tipping; Christopher M. Bishop, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Vol. 61, No.3 (1999), 611-622.
[2] https://people.cs.pitt.edu/m̃ilos/courses/cs3750-Fall2007/lectures/class17.pdf
[3] https://www.youtube.com/watch?v=ekEEG7t3sGI