CS5350: Machine Learning *Due end of semester*

Additional Reading

Dirichlet Process dp.pdf

This is acutally a challenge as I know nothing about stochastic process. But I'm interested in it because the DP aims to solve the model complexity problem, which is often hard to solve in most of the class projects. For example in project 5, the number of classes (number of latent vairables z) have to be decided by some criteria like BIC or AIC. DP is able to get a good z under Bayesian framework. Actually with DP model, we assume a infinitely number of mixed model. But it is tractable because the number of the components used as a priori is small. It is logarithmic of the number of observed data points.

Moreover, with DP we also do not have to assume a Gaussian Mixed model. DP is actually a distributio over distribution. That is, each 'event' in DP is also a distribution.

So, in project 5 for example, if we do not assume the data are drawn from a mix-Gaussian distribution, we can assume the distribution that the data are drawn from as G. Then we assume G itself is drawn from a Dirichlet Process, i.e. $G\tilde{}DP(\alpha, H)$. The $H$is the base distribution, and the mean of DP. As the paper mentioned, the DP is a discrete distribution, and this property is impoartant in applying DP for infereence of the G from the observed data set $\boldsymbol{\theta}$.

If we assume the base distribution is Gaussian, and we want to use Bayesian method to compare if this assumption is better than other possible distribution. Then DP can help in that to choose as large class models as possible to compare the Gaussian with. The DP can be seen as a 'relaxation' to the based distribution. And if the base distribution performs better than the DP relaxed model, we have confidence than this base model is a correct choice.

The future work include the efficient of computation for inference by DP model. This paper did not talk much about them. And I may need more reading for this.

Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data NIPS2003_AA42.pdf

This short paper is more difficult than I originally expected. Basically it talked about an alternative method to get PCA. As this is closely related to various explanation of PCA on Bishop's textbook I also read associated chapter of the book (chapter 12). First let me talk about the book. In the book, the author computes the projection matrix from two criteria: One is maximize the variance of the points projected on the new base vectors. That is, get a direction so the projection on the direction has maximal variance. The second criteria is to minimize the covariance between the original data and projected data on subspace. Also the author the maximum likelihood sollution and EM solution. The ML solution assume a latent varialbe $\boldsymbol{z}$ is a $\mathcal{N}(0, I)$ distribution, and $\boldsymbol{z}$ is actually the new data point on the subspace. so $\boldsymbol{x}$ is the data point in original space, we have $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z} + \mu$. hence the probability of $\boldsymbol{x}$ given $\boldsymbol{z}$ is also a Gaussian distribution. If we marginize $\boldsymbol{z}$ and get $p(\boldsymbol{x})$, it also a Gaussian model because it's acutally a sum of a Gaussian $p(\boldsymbol{x}|\boldsymbol{z})$ over all $\boldsymbol{z}$, and its covariance are governed by $\boldsymbol{z}$'s μ, σ and also the projection matrix $\boldsymbol{W}$. So the ML solution just need to maximize the $\ln p(\boldsymbol{X}|\mu, \boldsymbol{W}, \sigma^2)$ with regard to $\boldsymbol{W}$, μ and σ^2.

Now back to this paper. The paper did not marginize the latent variable z and maximize the $p(\boldsymbol{x})$ with regard to the papameters $\boldsymbol{W}$. Instead, for $p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{W})$, it tries to marginize $\boldsymbol{W}$, and get a Gaussian distribution $p(\boldsymbol{x}|\boldsymbol{z})$ (Here I used different notation with the paper: $\boldsymbol{x}$ is observed data on original space, and $\boldsymbol{z}$ is the latenet variable — the projected data points in subspace. The paper use $\boldsymbol{X}$ as the latenet variable and $\boldsymbol{Y}$ as the observed data.) To do this, the paper assume the papameter, i.e. the projection matrix $\boldsymbol{W}$'s each row $\boldsymbol{w}_i$ is Gaussian distribution. (I'm not sure if this is the correct assumption.) Then the paper tries to maximize the $p(\boldsymbol{x}|\boldsymbol{z})$ with regard to latent variable $\boldsymbol{z}$, and finally find $\boldsymbol{z}$ is governed by the eigenvectors of $\boldsymbol{z}\boldsymbol{z}^\top$.

There are two advantage of doing this. First It's easy to replace $\boldsymbol{z}\boldsymbol{z}^\top$ (the $\boldsymbol{Y}\boldsymbol{Y}^\top$ in the paper) with some other kernel functions and get kernel PCA. Another good thing is with this method it's natural to extend to non-linear mapping from $\boldsymbol{z}$ to $\boldsymbol{x}$.

If we related this with what we learned on class: in EM method for mix-Gaussian model, can we marginize the parameters — $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$, and π_k first, and try to maximize the $p(\boldsymbol{x}|\boldsymbol{z})$ with repect to $\boldsymbol{z}$? I doubt it. This is different with the models in the paper, where the author assume the papameter $\boldsymbol{W}$ is normal distribution. But how can we know the distribution of $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$ in our mix-Gaussian model? I may need to read more about the 'conjugate prior' for answering this question.