Intruduction to Machine Learning

Dr S.Amini

department: Electrical Engineering

Amirreza Velae 400102222
github repository

Project

April 21, 2023

Intruduction to Machine Learning

Project

Amirreza Velae 400102222
github repository

Theory Question 1.

In your own words, explain how the MM algorithm can deal with non-convex optimization objective functions by considering simpler convex objective functions.

solution

MM algorithm is an iterative algorithm for optimizing a nonconvex or nonconcave $l(\theta)$ that has a complex form. For example if our goal is finding maximum of a function we can consider a simpler concave function $Q(\theta, \theta^{(t)})$ that depends on a certain parameter. Function must be tight lowerbound that means in a certain point $l(\theta^{(t)} = Q(\theta^{(t)}, \theta^{(t)}))$ and also $l(\theta \leq Q(\theta, \theta^{(t)}))$. In each iteration we choose the max of $Q(\theta, \theta^{(t)}))$ as next level parameter $\theta^{(t+1)} = argmax_\theta Q(\theta, \theta^{(t)})$ and we move on to reach the maximum of the main function $l(\theta)$ and according to the following inequality we act correctly:
$l(\theta^{(t+1)}) \leq Q(\theta^{(t+1)}, \theta^{(t)}) \leq Q(\theta^{(t)}, \theta^{(t)}) = l(\theta^{(t)})$ also we can do same for finding minimum but with convex function.

Theory Question 2.

Briefly explain how the formula for mixture models:

$$p(y_n; \theta) = \sum_{k=1}^{K} p_Z(z_k; \theta) p_{Y|Z}(y|Z = z_k; \theta) \tag{1}$$

is the same as the sum over all possible values of $Z^{(i)}$ in equation

$$ln(p_Y(y_n; \theta)) = \sum_{i=1}^{N} ln(p(y_n^{(i)}; \theta)) = \sum_{i=1}^{N} ln(\sum_{k=1}^{K} p_{Y,Z}(y_n^{(i)}, z_n^{(i)} = k; \theta)) \tag{2}$$

Explain why it's easier to optimize $p_{Y,Z}(y_n^i, z_n^i = K; \theta)$ than $p(y_n; \theta)$ in the context of mixture models.

solution

The model is simplified by Bayes' rule:
$p(y, z) = p(y|z)p(z)$
$p(y_n; \theta) = \sum_{z_n} p_{Y,Z}(y_n, z_n; \theta) = \sum_{k=1}^{K} p_Z(z_k; \theta) p_{Y|Z}(y|Z = z_k; \theta)$
It is easier to optimize $p_{Y,Z}(yn, zn;)$ because in this case we know which distribution each data belongs to so optimizing become easier because we can optimize each distribution separately and find the parameters of each one.
But it is hard to optimize if we just know that a data comes from sumation of some distribution and don't know anything about each distribution because we can not fit a single distribution to model.

Theory Question 3.

Read about variational inference (or variational bayesian methods) and compare it with the procedure we used for the EM algorithm (You might want to check Wikipedia for this!)

solution

Variational Bayes (VB) is often compared with expectation maximization (EM). The actual numerical procedure is quite similar, in that both are alternating iterative procedures that successively converge on optimum parameter values. The initial steps to derive the respective procedures are also vaguely similar, both starting out with formulas for probability densities and both involving significant amounts of mathematical manipulations. However, there are a number of differences. Most important is what is being computed. EM computes point estimates of posterior distribution of those random variables that can be categorized as ”parameters”, but only estimates of the actual posterior distributions of the latent variables (at least in ”soft EM”, and often only when the latent variables are discrete). The point estimates computed are the modes of these parameters; no other information is available. VB, on the other hand, computes estimates of the actual posterior distribution of all variables, both parameters and latent variables. When point estimates need to be derived, generally the mean is used rather than the mode, as is normal in Bayesian inference. Concomitant with this, the parameters computed in VB do not have the same significance as those in EM. EM computes optimum values of the parameters of the Bayes network itself. VB computes optimum values of the parameters of the distributions used to approximate the parameters and latent variables of the Bayes network. For example, a typical Gaussian mixture model will have parameters for the mean and variance of each of the mixture components. EM would directly estimate optimum values for these parameters. VB, however, would first fit a distribution to these parameters — typically in the form of a prior distribution, e.g. a normal-scaled inverse gamma distribution — and would then compute values for the parameters of this prior distribution, i.e. essentially hyperparameters. In this case, VB would compute optimum estimates of the four parameters of the normal-scaled inverse gamma distribution that describes the joint distribution of the mean and variance of the component.

Simulation Question 1.

Each distribution has 200 data points that are concatenated in a two-dimensional array and given to you. Plot the data with three different colors in a graph.

End of Project