Progress report

Lukas Zorich

April 28, 2017

1 Background

Let $X = \{x_1, \ldots, x_N\}$ a collection of N points. Now, lets suppose that the features of x_j are updated, and lets call $x_j^\star$ the updated point. Lastly, lets define $X_{-j} = \{x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_N\}$.

The posterior before x_j moves, and using $P(\Theta)$ as the prior is

$$P(\Theta \mid X) \propto P(X \mid \Theta) P(\Theta)$$

,

where $P(X \mid \Theta) = \prod_i^N P(x_i \mid \Theta)$.

Now, after x_j moves to $x_j^\star$, the posterior is

$$\begin{aligned} P(\Theta \mid X_{-j}, x_j^\star) &\propto P(X_{-j}, x_j^\star \mid \Theta) P(\Theta) && (1)\\ &\propto P(X_{-j} \mid \Theta) P(x_j^\star \mid \Theta) P(\Theta) && (2)\\ &\propto \frac{P(X \mid \Theta)}{P(x_j \mid \Theta)} P(x_j^\star \mid \Theta) P(\Theta) && (3)\\ &\propto \frac{P(x_j^\star \mid \Theta)}{P(x_j \mid \Theta)} P(X \mid \Theta) P(\Theta) && (4)\\ & && (5) \end{aligned}$$

which can be written as,

$$P(\Theta \mid X_{-j}, x_j^\star) \propto \frac{P(x_j^\star \mid \Theta)}{P(x_j \mid \Theta)} P(\Theta \mid X) \tag{6}$$

Now, if we instead of considering one data point x_j, we consider a batch of S data points $X_J = \{x_j, x_{j+1}, \ldots, x_{j+S-1}, x_{j+S}\}$ that moved from x_{j+k} to $x_{j+k}^\star$, for $k = 0, \ldots, S$. And lets define $X_{-J} = \{x_1, \ldots, x_{j-1}, x_{j+S+1}, \ldots, x_N\}$. Then, Eq. 6 for the batch of S data point and because the prior $P(\Theta)$ gets canceled, can be written as,

$$P(\Theta \mid X_{-J}, X_J^\star) \propto \frac{P(\Theta \mid X_J^\star)}{P(\Theta \mid X_{-J})} P(\Theta \mid X) \tag{7}$$

.

Inspired by the work done in [1], we assume that we approximate the posterior using **variational inference**. Also, we assume that $P(\Theta)$ is an exponential family distribution for Θ with sufficient statistic $T(\Theta)$ and natural parameter ξ_0. We suppose further that if $q(\Theta)$ is the approximate posterior obtained using variational inference, then $q(\Theta)$ is also in the same exponential family with a parameter ξ such that

$$q(\Theta) \propto \exp(\xi \cdot T(\Theta)). \tag{8}$$

Similar to [1], when we make this assumptions the update in Eq. 7 becomes

$$P(\Theta \mid X_{-J}, X_J^\star) \approx \exp([\xi - \xi_J + \xi_J^\star] \cdot T(\Theta)) \tag{9}$$

,

where ξ is the natural parameter of $q(\Theta) \approx P(\Theta \mid X)$, and ξ_J and $\xi_J^\star$ corresponds to the natural parameter of $q_J(\Theta) \approx p(\Theta \mid X_J)$ and $q_J^\star(\Theta) \approx p(\Theta \mid X_J^\star)$ respectively.

Using this approach, we can update the posterior when data "moves" without the need to go through the whole dataset, instead we just need to go through the data points that moved to obtain ξ_J and $\xi_J^\star$.

2 Application to my model

For trying the proposed approach, I'm starting only with a single GMM (I prefer to start small).

For a GMM, the update would be

$$P(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda} \mid X_{-J}, X_J^\star) \approx \frac{q_{-J}^\star(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})}{q_{-J}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \tag{10}$$

$$\tag{11}$$

where

$$q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{\pi}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \tag{12}$$

$$= Dir(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \mathcal{N}(\boldsymbol{\mu} \mid \mathbf{m}, (\beta \boldsymbol{\Lambda})^{-1}) \mathcal{W}(\boldsymbol{\Lambda} \mid \mathbf{W}, \nu). \tag{13}$$

2.1 Updates

2.1.1 Dirichlet

The natural parameter for the dirichlet is:

$$\xi = \boldsymbol{\alpha} - 1 \tag{14}$$

,

hence, the update is:

$$\boldsymbol{\alpha}' = \boldsymbol{\alpha} - \boldsymbol{\alpha}_J + \boldsymbol{\alpha}_J^\star \tag{15}$$

2.1.2 Normal-Wishart

The natural parameter for the Normal-Wishart distribution is:

$$\xi = \begin{bmatrix} \beta\boldsymbol{\mu} \\ \beta \\ \boldsymbol{\Lambda}^{-1} + \beta\boldsymbol{\mu}\boldsymbol{\mu}^T \\ \nu + 2 + p \end{bmatrix} \tag{16}$$

,

hence, the updates are:

$$\boldsymbol{\mu}' = \frac{1}{\beta'}(\beta\boldsymbol{\mu} - \beta_J\boldsymbol{\mu}_J + \beta_J^\star\boldsymbol{\mu}_J^\star) \tag{17}$$

$$\beta' = \beta - \beta_J + \beta_J^\star \tag{18}$$

$$\boldsymbol{\Lambda}^{-1\prime} = (\boldsymbol{\Lambda}^{-1} + \beta\boldsymbol{\mu}\boldsymbol{\mu}^T) - (\boldsymbol{\Lambda}_J^{-1} + \beta_J\boldsymbol{\mu}_J\boldsymbol{\mu}_J^T) + (\boldsymbol{\Lambda}_J^{-1\star} + \beta_J^\star\boldsymbol{\mu}_J^\star\boldsymbol{\mu}_J^{T\star}) - \beta'\boldsymbol{\mu}'\boldsymbol{\mu}^{T\prime} \tag{19}$$

$$\nu' = \nu - \nu_J + \nu_J^\star \tag{20}$$

$$\tag{21}$$

3 Experiments

Write your conclusion here. Hola [1] blabla

4 Next steps

References

[1] Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C., and Jordan, M. I. Streaming variational bayes. In *NIPS* (2013), C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, Eds., pp. 1727–1735.