13

Transfer and Self-learning in Probabilistic Models

Fetze Pijlman[1,2]

[1]Signify, The Netherlands
[2]Eindhoven University of Technology, The Netherlands

Abstract

Transfer learning and self-learning are well known techniques that can separately improve probabilistic models. In this article it is investigated how to combine both methods in a single approach enabling transfer learning over different environments in which self-learning models are deployed. It is found that such learnings can be enabled through prior optimisation.

Keywords: probabilistic models, Bayesian, online learning, self-learning, prior optimisation, transfer learning.

13.1 Motivation

Probabilistic models are of interest for constructing classifiers or estimating parameters (see e.g. Murphy [1]). Examples of such applications are image classification in cameras and noise level estimation in radar sensors. These sensors are often deployed in diverse environments. This means that the underlying model could benefit from self-learning. Baye's rule offers such a method

$$P\left(\theta \mid X\right) = \frac{P(X|\theta)P(\theta)}{P(X)}, \tag{13.1}$$

where θ are the model parameters, X are the observations, and P(θ) is the prior distribution. Note that the observations X may also contain feedback

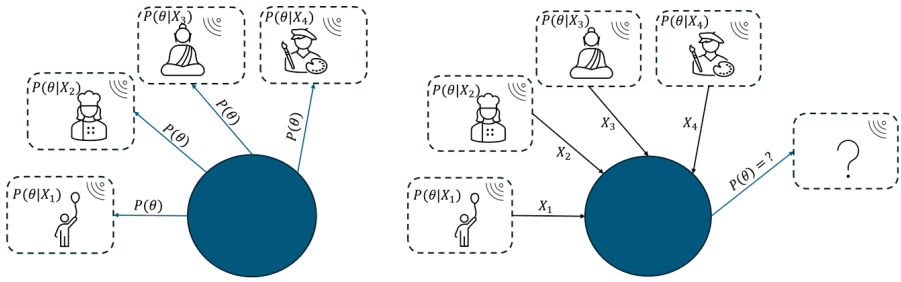

Figure 13.1 On the left-hand side a prior was sent to various diverse environments where in each environment a posterior was estimated. On the right-hand side the question is how to choose the prior for a new unknown environment when sensor events from other environments are known.

and interactions from users. A self-learning example is a motion sensor that self-learns its false positive rate.

The prior distribution is our initial expectation on the parameters prior to the observation of new data. When no historical data is available, an expert choice should be made for the prior. However, in some situations data may be available from other diverse environments. The question here is how to choose a prior for a new unseen environment when data is available from other diverse environments. This problem is illustrated in Figure 13.1.

In our motion sensor example, it may be believed that sensors in some applications have a false positive rate of once per week while in other applications the false positive rate is once per month. Note that for a new sensor deployment one cannot simply take an average of the posteriors from different environments as each posterior represents a different deployment: some posteriors may be very concentrated around a particular value which may slow down adaptation to a new environment and posteriors for different environments may have seen different amount of data (how to average that?). Moreover, averaging different posteriors is fundamentally wrong; similarly, one cannot take the average of an apple and an orange.

To solve this problem Xuan [2] discusses a graphical model containing common and custom nodes. A challenge with such an approach is that the graph itself will also change as more data becomes available. Similarly, Suder [3] splits up the parameter space into common and custom parameters. A challenge with this approach is how to decide which parameters are common or which ones are not. Pautrat [4] discusses the grouping of similar environments, each having their own prior. A challenge with such an approach is

how to decide on the number of groups and when to decide on forming a new group when a new environment is substantially deviating.

Although the posteriors are fundamentally different for different environments, knowledge from different environments is still useful for determining an optimal prior for a new environment. Although the environments may be substantially different there is one aspect that they share in common which is the prior from which the probabilistic model could have started. This commonality offers an opportunity for transfer learning and it will be studied in this article.

13.2 Prior Optimisation

A choice for prior is similar to a choice for model. When having a set of models $\mathcal{M}_i$ then the probability for a model to hold given the data is given by (see e.g. Murphy [1])

$$P\left(\mathcal{M}_i|X\right)=\frac{P(X|\mathcal{M}_i)\ P(\mathcal{M}_i)}{P(X)}. \tag{13.2}$$

If a prior is parametrized through a hyperparameter μ then the best hyperparameter can be selected through

$$\text{argmax}_{\mu} P\left(X|\mu\right)\ P\left(\mu\right). \tag{13.3}$$

In some cases, there may be no direct expectations for $P\left(\mu\right)$ as given in Equation 1.3 but there may be prior expectation for parameter θ. In those cases, an effective prior on μ can be derived. Prior to the knowledge on the prior $P\left(\theta\right)$ we take $P\left(\mu\right)$ to be uniform (first line equation below right-hand-side). The probabilities for observing a distribution $P\left(\theta\right)$ from μ is simply the product of $P\left(\theta_i \mid \mu\right)$ weighted by the amount $P\left(\theta_i\right)\Delta\theta$ (second line involves a product integral)

$$\begin{aligned} P(\mu \mid P[\theta]) &\propto P(P[\theta] \mid \mu) \\ &= \lim_{\Delta\theta\to 0}\prod_i P\left(\theta_i \mid \mu\right)^{P(\theta_i)\Delta\theta} \\ &= \exp\int d\theta P(\theta)\log P(\theta \mid \mu), \end{aligned} \tag{13.4}$$

where in the last step the Geometric integral relation from Volterra was used. Note that the same result can be obtained by using variational Bayesian

approximation. A variational approximation of a posterior is given by

$$P(\mu \mid X) \approx \operatorname{argmax}_{q(\mu)} \int d\mu\, q(\mu) \left(\log P(X \mid \mu) - \log \frac{q(\mu)}{P(\mu)} \right). \tag{13.5}$$

Using the prior $P(\theta)$ for updating our knowledge on $P(\mu)$ gives

$$\begin{aligned} P(\mu \mid P[\theta]) &\approx \operatorname*{argmax}_{q(\mu)} \int d\mu q(\mu) \left(\int d\theta P(\theta) \log P(\theta \mid \mu) - \log \frac{q(\mu)}{P(\mu)} \right) \\ &\propto \exp \int d\theta P(\theta) \log P(\theta \mid \mu), \end{aligned} \tag{13.6}$$

where in the last line we assumed the initial prior $P(\mu)$ to be flat. Equation 1.4 gives for the effective prior $P(\mu)$

$$P(\mu \mid P[\theta]) = \frac{\exp \int d\theta P(\theta) \log P(\theta \mid \mu)}{\int d\mu' \exp \int d\theta P(\theta) \log P(\theta \mid \mu')} \tag{13.7}$$

In Equation 1.3, X contains all of the available data. It will be assumed that captured diversity in environments is representative for the underlying population of environments. The known environments are allowed to have different amounts of observations. As the known environments are assumed to be representative, an optimal prior for known environments will be an optimal prior for a new unknown environment.

Taking the logarithm of Equation 1.3 and labelling data from environment i by X_i one finds

$$\mu = \operatorname{argmax}_{\mu} \left[\textstyle\sum_i \log P(X_i|\mu) + \log P(\mu) \right]. \tag{13.8}$$

Using

$$\log P(X) = \int d\theta\, q(\theta) \left[\log P(X \mid \theta) - \log \frac{P(\theta|X)}{P(\theta)} \right], \tag{13.9}$$

with

$$\int d\theta\, q(\theta) = 1, \tag{13.10}$$

one finds

$$\begin{aligned} \mu \approx \operatorname*{argmax}_{\mu} \Bigg[\sum_i \int d\theta_i P(\theta_i \mid X_i, \mu) \Bigg\{ \log P(X_i \mid \theta_i) \\ - \log \frac{P(\theta_i \mid X_i, \mu)}{P(\theta_i \mid \mu)} \Bigg\} + \log P(\mu) \Bigg]. \end{aligned} \tag{13.11}$$

In case the posterior for an environment cannot be exactly determined one can approximate by (see Attias [5], Morey [6], and Tran [7])

$$\mu \approx \underset{\mu}{\operatorname{argmax}} \max_{q(\theta_i)} \left[\sum_i \int d\theta_i q(\theta_i) \left\{ \log P(X_i \mid \theta_i) - \log \frac{q(\theta_i)}{P(\theta_i \mid \mu)} \right\} + \log P(\mu) \right] \tag{13.12}$$

Note that $q(\theta_i)$ approximates the posterior for each environment i in terms of the reversed KL-divergence. The reversed KL-divergence may lead to modes in $q(\theta_i)$ introducing errors. For this reason the posterior $q(\theta_i)$ should have sufficient freedom to describe all typical groups of customers.

It is of interest to study the optimal prior in case of a single deployment (or in the limit that all deployments have the same environment) and for $P(\mu)$ being constant. If one has conjugate priors then one can iteratively solve Equation 1.11 by choosing at each iteration the prior $P(\theta_i|\mu)$ to be equal to the posterior from the previous iteration $P(\theta_i|X_i, \mu)$. In the limit of convergence, the logarithm containing ratio of posterior over prior will vanish and the posterior $P(\theta_i|X, \mu)$ will strongly peak around $\mu = \operatorname{argmax}_{\theta_i} P(X|\theta_i)$. In other words, in absence of any prior belief then the best prior for a new environment that is the same as the existing environment is a strongly peaked function around the most likely parameters (as expected).

Note that strongly concentrated priors delay learnings from new data. Peaking of priors is reduced when a belief for $P(\mu)$ is included and when known environments are diverse.

13.3 Example Categorical Distribution

For categorical distributions the conjugate prior is a Dirichlet distribution of which its hyperparameter is often denoted by $\boldsymbol{\alpha}$ instead of μ that was used in this paper upto now. Given applications i some observed categories j and assuming a flat prior such that the last term $\log P(\mu)$ in Equation 1.11 disappears one obtains

$$\boldsymbol{\alpha} = \operatorname{argmax}_{\boldsymbol{\alpha}} \sum_i \log \frac{\Gamma\left(\sum_j \alpha_j\right) \prod_j \Gamma(\alpha_j + n_{ij})}{\Gamma\left(\sum_j \alpha_j + n_{ij}\right) \prod_j \Gamma(\alpha_j)}. \tag{13.13}$$

Consider the following example. In a first application one encounters 1000 items of category 0 and 2 items of category 1 while in a second

application one encounters 10 items of category 0 and 2 items of category 1. Using the equation above one finds the optimal prior to be a Dirichlet distribution with $\boldsymbol{\alpha} = (5.8, 0.41)$. If the previous two applications are random draws from an underlying population of applications, then the obtained Dirichlet distribution would be optimal for a new application. The ratio of the hyperparameters indicates the expected ratio of classes, the sum of the hyperparameters is an indicator for the knowledge strength.

Consider now another example in which one encounters in the first application 50 items of category 0 and 3 items of category 1, and in the second application 5 items of category 0 and 20 items of category 1. In this example the optimal hyperparameters for the Dirichlet distribution are $\boldsymbol{\alpha} = (0.87, 0.60)$. The low numbers indicate a lack of knowledge which is expected given the large differences in observations between the two applications.

13.4 Conclusions and Discussion

Self-learning probabilistic models are useful for learning context and thereby self-learning increases performance in diverse environments. Similarly, transfer learning can help to use learned knowledge from one environment for another environment. In this article these two aspects are integrated into a single approach by optimising the prior.

It was found that an optimal prior can be obtained by maximising a sum of evidences over environments. The amount of data per environment is allowed to vary. In order to optimise the prior, observations from different environments need to be shared. For cameras this means sharing of images, for radar sensors it means sharing of radar signals. Sharing of sensor events may involve privacy aspects.

Although the approach leads to the best (or most optimal) prior one does need to realize that this approach carries the risk of overfitting. The situation is similar to variational Bayesian methods that yield the "best approximated" posterior which in reality may be far away from the true posterior.

Acknowledgements

EdgeAI “Edge AI Technologies for Optimised Performance Embedded Processing” project has received funding from Key Digital Technologies Joint Undertaking (KDT JU) under grant agreement No 101097300. The KDT

JU receives support from the European Union's Horizon Europe research and innovation program and Austria, Belgium, France, Greece, Italy, Latvia, Luxembourg, Netherlands, Norway.

References

[1] K.P. Murphy, "Machine learning - a probabilistic perspective", The MIT press (2012)
[2] J. Xuan, J. Lu, G. Zhang, "Bayesian Transfer Learning: An Overview of Probabilistic Graphical Models for Transfer Learning". arXiv:2109.13233 (2021)
[3] P.M. Suder, J. Xu, and D.B. Dunson, "Bayesian Transfer Learning". arXiv:2312.13484 (2023)
[4] R. Pautrat, K. Chatzilygeroudis and J.B. Mouret, "Bayesian Optimization with Automatic Prior Selection for Data-Efficient Direct Policy Search". arXiv 1709.06919 (2017)
[5] H. Attias. "A variational bayesian framework for graphical models". *neurIPS* (1999).
[6] R. D. Morey, J.W. Romeijn, J.N. Rouder, The philosophy of Bayes factors and the quantification of statistical evidence, Journal of Mathematical Psychology (2016)
[7] M.N. Tran, Trong-Nghia Nguyen, Viet-Hung Dao, A practical tutorial on Variational Bayes, arXiv: 2103.01327 (2021)