Improving Fairness in Stochastic Variational Gaussian Processes
CID: 01342416

Abstract

Gaussian processes are a machine learning method that makes use of a Bayesian framework to make predictions. A scaleable approximation to Gaussian processes, namely "stochasitic variational Gaussian processes" (SVGP) exhibits state-of-the art performance at both regression and classification tasks.

In preliminary experiments, however, SVGP appears to suffer from poor peformance on fairness metrics such as "equality of opportunity difference" for datasets that contain sensitive attributes. In this paper, a method for improving fairness in SVGP through reweighting data samples is introduced. Furthermore, a parameter for the fairness-accuracy trade-off is introduced, allowing for a completely customisable model.

1. Introduction

Gaussian process (GP) inference has proven to be remarkably useful as a machine learning technique in tasks such as supervised regression and classification [8]. Unlike neural networks, which perform well on large noiseless datasets, GPs are non-parametric models that fit within the Bayesian framework, and therefore excel at regression tasks that suffer from a lack of data that are noisy, and that require uncertainty estimates on predictions [10].

As increasingly important decisions are delegated to machine learning algorithms, it becomes increasingly necessary unpack the idea of *fairness* in the context of AI. A model should not base these important decisions on attributes such as race or sex.

With the beneficial properties of GPs, it seems strange that such few works have delved into the challenge of making GPs fairer. In 2020, Tan et al. published a paper in which they introduce "the first fair Gaussian process" (FGP), which takes a mathematically rigorous approach to adjusting GPs such that certain fairness criteria like equality of opportunity (EOP) are satisfied [9].

In this paper, a more experimental and creative approach to improving fairness in GPs is taken. In particular, fairness in stochastic variational Gaussian processes (SVGP) [4], a sparse GP method that allows for non-Gaussian likelihoods and better scalablility, is explored.

2. Background

2.1. Gaussian Processes

While Gaussian distributions are defined by a mean value and covariance matrix, Gaussian processes can be entirely described by mean and convariance *functions*: $\mathcal{GP}(\mu(x), \kappa(x, x'))$. GPs represent distributions over possible functions that can by updated using Bayes rule after considering the training data. The kernel function κ tells us what kinds of functions we expect to sample from our Gaussian process, while the mean function μ tells the function about which these samples are centred [1].

In GP regression, the aim is to infer the noise-free *latent* function $f(\cdot)$ that underpins the N datapoints $\{\mathbf{X}, \mathbf{y}\}$, and hence to determine latent function values $\mathbf{f}_* \triangleq f(\mathbf{X}_*)$ (and new target values $\mathbf{y}_*$) at new test locations $\mathbf{X}_*$. Without loss of generality, our GP prior is usually chosen with $\mu = 0$, as we generally have no preconceived notions on how we expect the model to behave. Furthermore, any prior knowledge can generally be factored into our prior by picking the appropriate kernel function.

Exact inference is possible in the homoskedastic case i.e. the case where the noise across the samples is uniform and Gaussian such that the likelihood is given by: $p(\mathbf{y} \mid \mathbf{f}) = \mathcal{N}(y; \mathbf{0}, \mathbb{I}_N \sigma_y^2)$, where $\mathbf{f} \triangleq f(\mathbf{X})$. With a prior on f of the form:

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K}) \tag{1}$$

we obtain the following *predictive* posterior distribution:

$$p\left(\mathbf{y}_* \mid \mathbf{y}\right) = \mathcal{N}\left(\mathbf{y}_*; \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*\right) \tag{2a}$$
$$\boldsymbol{\mu}_* \triangleq \mathbf{K}_*^T \mathbf{K}_y^{-1} \mathbf{y} \tag{2b}$$
$$\boldsymbol{\Sigma}_* \triangleq \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}_y^{-1} \mathbf{K}_* + \mathbb{I}_N \sigma_y^2 \tag{2c}$$

where $\mathbf{K} = \kappa(\mathbf{X}, \mathbf{X})$, $\mathbf{K}_* = \kappa(\mathbf{X}_*, \mathbf{X})$, $\mathbf{K}_{**} = \kappa(\mathbf{X}_*, \mathbf{X}_*)$, and $\mathbf{K}_y = \mathbf{K} + \mathbb{I}_N \sigma_y^2$.

The log marginal likelihood, given by:

$$\log p(\mathbf{y}) = -\frac{1}{2} \mathbf{y} \mathbf{K}_y^{-1} \mathbf{y} - \frac{1}{2} \log |\mathbf{K}_y| - \frac{N}{2} \log(2\pi) \tag{3}$$

is the objective function used during optimisation to find the optimal kernel hyperparameters $\boldsymbol{\theta}$. This optimisation process automatically regularises our probabilistic model by maximising the likelihood of $\boldsymbol{\theta}$ given the data.

[1] See Rasmussen and Williams [8] for a more comprehensive description of GPs.

2.2. Stochastic Variational Gaussian Processes

In the case that data is abundant, computation of new predictions using exact Gaussian processes becomes expensive. With a dataset of N training points, algorithmic complexity scales as $\mathcal{O}(N^3)$, rendering the intermediate matrix calculations infeasible for even moderately sized training datasets of a few thousand points [7, 4]. Furthermore, in the case of classification, analytical computation of the marginal likelihood and posterior is intractable because of the use of a non-Gaussian likelihood.

Stochastic variational Gaussian processes (SVGP) [4, 5] is a sparse Gaussian process method that solves both these problems by summarising the data with M *inducing inputs* $\mathbf{X}_m$ and *inducing variables* $\mathbf{f}_m$, and employing variational inference to minimise the $\mathcal{KL}$-divergence between the true posterior, $p(\mathbf{f} \mid \mathbf{y})$ and the approximate posterior $q(\mathbf{f})$. This minimisation is equivalent to the maximisation of a lower bound on the marginal likelihood or *evidence lower bound* (ELBO); trainable parameters of the model are learned through gradient-based optimisation using the ELBO as the loss function.

3. Methods

3.1. Datasets

Two datasets were chosen for carrying out SVGP binary classification.

***Adult* income dataset** This dataset contains samples with input features such as sex, race, age and years in education and output targets for income salary. The desireable outcome is when the sample is classed as having a salary of larger than 50k USD per year based on these features. For this dataset, sex was chosen as the sensitive feature of interest.

***COMPAS* recidivism dataset** This dataset contains data samples for criminals, to determine the likelihood that criminal will reoffend within a two-year period. The desireable outcome for a sample is to be classed as to not reoffend. Race was chosen as the sensitive feature of interest.

3.2. Scoring Metrics

The following scoring metrics were used to evaluate the performance of the model:

- Accuracy score - ratio of correct predictions to total test samples.
- Equality of opportunity (EOP) difference - the true positive rate difference between the two classes of the sensitive attribute [3].
- Ratio of positives predictions (RPP) - ratio of number of positively classed predictions to positively labelled test samples.

[2]See GPflow source code for more information and implementation.

3.3. Fairness-based Reweighting

Reweighting is a data preprocessing technique that adjusts the individual samples contribution to the learning of the model such that the sensitive feature (e.g. sex) remains independent from the target label (e.g. income) [6]. For each data sample $\{\mathbf{x}_i, y_i\}$, an associated weight w_i is calculated.

Reweighting in the SVGP model was done by multiplying the individual samples' contributions to the ELBO loss function by its associated weight w_i.

4. Preliminary Experiments

4.1. Effect of Regularisation on Scoring Metrics

The value for ϵ in the likelihood[2], which represents the fraction of incorrectly labelled datapoints in the dataset [2], can be interpreted as a regularisation parameter in the SVGP model.

On both datasets, maximum accuracy was achieved using a small regularisation parameter of $\epsilon < 0.1$, with an accuracy score of 81% on the *Adult* dataset and 65% on the *COMPAS dataset*. However, this value for ϵ yielded EOP difference scores of -0.45 and -0.21 respectively, the significant negative value indicating a bias towards the privileged groups (males and Caucasian respectively).

At first glance, for the *Adult* dataset, increasing ϵ appeared to have a beneficial impact to the EOP difference score. However, it was also noticed that the number of positive (desirable) predictions made by the model was also reducing, and as a result, this improvement to the fairness was rendered trivial; the model converged to a predictor that would predict a negative output regardless of the input.

A similar effect was observed for the *COMPAS* dataset, but with a *worsening* of the fairness score with increased regularisation. The changes made to ϵ appeared to bring no benefits to the model, neither in terms of fairness nor accuracy (see Appendix, figure 1 for plots).

The likely reason for this discrepancy in the model behaviour between the two datasets is that the majority of the *Adult* dataset has negative labels, while the majority of the *COMPAS* dataset have positive labels.

5. Fairer SVGP

Reweighting the data resulted in improved fairness for both datasets with a negligible drop of accuracy score to a minimum of 98% of the full plain SVGP model. In the case of the *Adult* dataset, we see that this reweighting results in more samples being classed positively than without reweighting. As these new weights are multiplied by the samples contribution to the ELBO loss function, an increased weight for a female sample might allow it to cross the decision boundary during optimisation of the model. Therefore the increase in positive predictions is likely to be entirely of female samples.

5.1. Accuracy-Fairness Trade-off

While the original weights for the *COMPAS* dataset did slightly improve the fairness of the model, the EOP score was still far from being considered ideal.

A parameter for varying the "strength" of the reweighting is introduced. For sample weight w_i, we can transform the weight to $\tilde{w}_i$ with parameter α using:

$$\tilde{w}_i = \alpha(w_i - 1) + 1 \tag{4}$$

This simple adjustment to the weights allows us to parametrise the trade-off between accuracy and fairness, while ensuring that $\sum \tilde{w}_i = \sum w_i = N$. The weights can be made weaker using $\alpha < 1$ or stronger using $\alpha > 1$. With α the weights become 1 for all samples.

A strengthening of the weights for the *COMPAS* samples with factor $\alpha = 1.25$ yielded the fairest model, with still a negligible impairment to the accuracy. Furthermore, the number of positive predictions was nearer the number of test labels, as a result of the strengthened reweighting.

ϵ	α	Accuracy	EOP difference	RPP
0.1	0	0.807	-0.447	0.589
0.1	1	0.791	0.0057	0.726

Table 1: Scores of SVGP classifier on the *Adult* dataset. $\alpha = 0$ corresponds to weights of 1 for all samples. Reweighting makes a significant improvement to the fairness of the model.

ϵ	α	Accuracy	EOP difference	RPP
0.1	0	0.652	-0.207	1.157
0.1	1	0.650	-0.151	1.150
0.1	1.25	0.645	0.012	1.037

Table 2: Scores of SVGP classifier on the *COMPAS* dataset. $\alpha > 1$ corresponds to weights that are "stronger" than the originally calculated weights. Reweighting makes a significant improvement to the fairness of the model.

A graph showing how varying α affects the fairness and accuracy of the model for both datasets is plotted in figure 2.

The disadvantage of this data preprocessing method is that for the fairest value for α to be obtained, cross validation is necessary as a new model needs to be constructed in order to test a new value for α. Furthemore, with introduction of new data, adjustments to the model might need to be made again.

In the case that we desire an automatic adjustment to the model to satisfy fairness criteria (without an accuracy-fairness tradeoff), the construction of the loss function itself needs to be changed.

6. Conclusion

Reweighting the data can be applied to SVGP classification to make an improvement to the fairness of the model with a very small compromise to the quality of predictions. This method is very easy to implement as only a small adjustment to the model is needed compared to plain SVGPs.

7. Specifications

GPflow [2], a library that can employ sparse GP models, was used for these experiments.

Weights for the data samples were computed using the AIF-360 toolkit [1].

A descriptive Jupyter notebook, in which there are Python implementations of the experiments above, can be found here.

References

[1] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, Oct. 2018. 3

[2] A. G. de G. Matthews, M. van der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León-Villagrá, Z. Ghahramani, and J. Hensman. Gpflow: A gaussian process library using tensorflow. *Journal of Machine Learning Research*, 18(40):1–6, 2017. 2, 3

[3] M. Hardt, E. Price, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 2

[4] J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data, 2013. 1, 2

[5] J. Hensman, A. Matthews, and Z. Ghahramani. Scalable Variational Gaussian Process Classification. In G. Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 351–360, San Diego, California, USA, 09–12 May 2015. PMLR. 2

[6] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2011. 2

[7] J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(65):1939–1959, 2005. 2

[8] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. MIT Press, London, England, 2005. 1

[9] Z. Tan, S. Yeom, M. Fredrikson, and A. Talwalkar. Learning fair representations for kernel models, 2020. 1

[10] M. Van der Wilk. *Sparse Gaussian process approximations and applications*. PhD thesis, University of Cambridge, 2019. 1

A. Fairness and Accuracy Plots

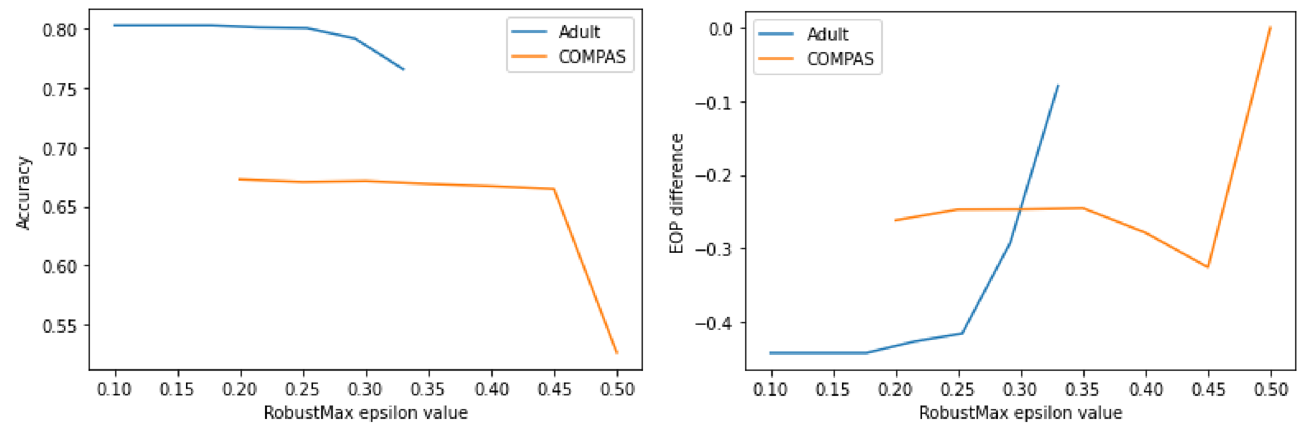

Figure 1: The effect of varying the regularisation parameter ϵ on fairness and accuracy for an SVGP model for *Adult* and *COMPAS* datasets

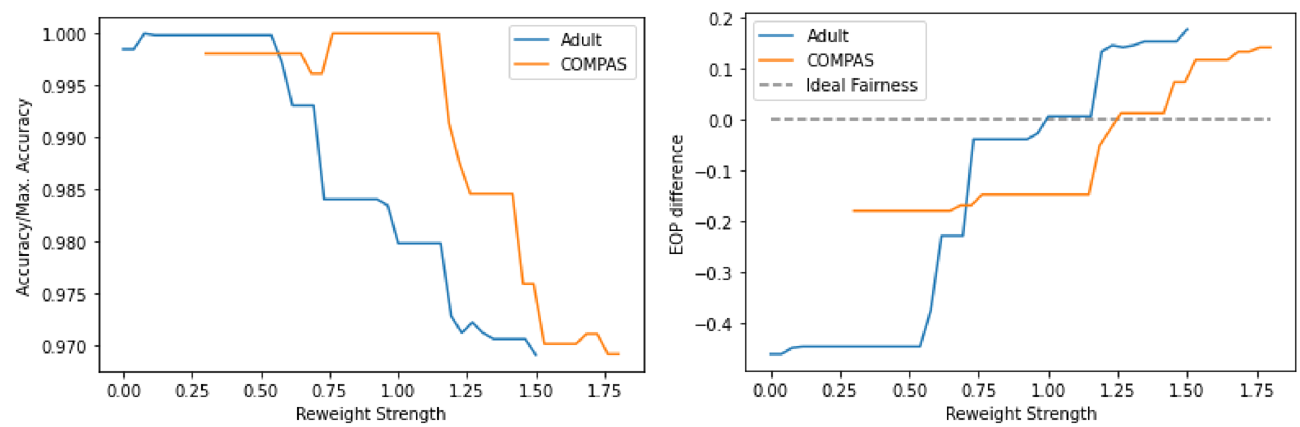

Figure 2: The effect of varying the reweighting strength parameter α (from 4) on fairness and accuracy for an SVGP model for *Adult* and *COMPAS* datasets