Gaussian Process Probes (GPP) for Uncertainty-Aware Probing (NeurIPS 2023)

Shin Yun Seop

January 8, 2024

Seoul national university, statistics, IDEA LAB

Contents

Basic concept

1. Probe
    - Investigates what task the given representation model is suitable for.
    - Specifically, the goal is to understand on what task the pre-trained representation model has been trained.
    - Through this, we enhance understanding of the specific tasks the model can perform and grasp the characteristics of the model.

Basic concept(Continued)

2. Uncertainty
   - Aleatoric uncertainty: Irreducible uncertainty induced by noisy data.
   - Epistemic uncertainty: Reducible uncertainty induced by lack of knowledge.
   - Note: High confident does not mean low uncertainty.

Gaussain Process Probes(GPP)

- GPP expand existing linear probing method by using gaussian process.
- It does not require access to training data, gradients, or the architecture of pre-trained representation model.
  (Note: This method is applied to pre-trained model.)
- It probe a model's representations of concepts and measure both epistemic uncertainty, aleatory uncertainty of probing.
- There is no need for learning this; it only requires tuning the hyperparameters based on prior knowledge or experiment.

Contents

Background: Notations for GPP

- $\mathcal{X}$: Input space(ex: Image space)
- $\phi : \mathcal{X} \to \mathbb{R}^d$: Given pre-trained model.
- $x \in \mathcal{X}$: Input of model.
- $a = \phi(x) \in \mathbb{R}^d$: Vector representation of given input.
- $D = \{(\phi(x_i), y_i)\}_{i=1}^N, x_i \in \mathcal{X}, y_i \in \{0, 1\}$: Given observations.
- $Q = \{(\phi(x_1'), y_1'), \cdots, (\phi(x_M'), y_M')\}$: Query set.
- $g \sim \mathcal{G}(\theta)$: Classifier following Beta gaussian process.
- $\theta = (\mu, k)$: Parameter for the Beta gaussian process.

Background: Beta Gaussian Process

Definition

Random element $g : \mathbb{R}^d \to [0,1]$ follow Beta Gaussian Process if

$$g = \frac{1}{1+e^{-f}}, \text{ where } f = f_\alpha - f_\beta, \text{ and}$$
$$f_\alpha \sim \mathcal{GP}(\mu, k), f_\beta \sim \mathcal{GP}(\mu, k), f_\alpha \perp\!\!\!\perp f_\beta.$$

Simply we denote g follow Beta GP as $g \sim \mathcal{G}(\theta)$, where $\theta = (\mu, k)$.
Note that μ, k are mean and kernel functions used to define gaussian process.

Adapting Beta GPs for GPP

- Let $g \sim \mathcal{G}(\theta)$ where $\mu(a) = log(\epsilon) - \frac{v}{2}$ and,

$$k\left(a, a'\right) = v \frac{a^\top a' + 1}{(\|a\|^2 + 1)^{\frac{1}{2}} \left(\|a'\|^2 + 1\right)^{\frac{1}{2}}}, \text{ where } v = \log\left(\frac{1}{\epsilon} + 1\right)$$

  for all $a, a' \in \mathbb{R}^d$ be the prior distribution of classifier.
- Note: $\epsilon > 0$ is the hyperparameter.
- Now, assume that $y|g, a \sim bernoulli(g(a))$. From this we can obtain posterior distribution of classifier when $D = \{(\phi(x_i), y_i)\}_{i=1}^N$ is given.
- Posterior distribution $g|D \sim \mathcal{G}(\theta_D)$ can be obtained as closed form.(See Appendix)

Probing and Uncertainty measures

- We define probing and uncertainty measures using the posterior prediction $g(a)$ where $g|D \sim \mathcal{G}(\theta_D),^{\forall} a \in \mathbb{R}^d$.
- Practically, we obtain posterior using observation D and probe the query set Q.

Probing and Uncertainty measures(Continued)

- **Judged probability** $:= \mathbb{E}[g(a)]$ which is the expected probability that the label of input is positive.
- **Aleatoric** $:= \mathbb{H}[y|g(a)]$, the expected entropy of the conditional distribution $p(y|g(a))$. Higher aleatoric corresponds to more noisy in the label of a.
- **Epistemic** $:= -\mathbb{H}[g(a)]$, the negative entropy of the distribution of g(a). High epistemic means that we are "highly confident" about the underlying probability.

Probing and Uncertainty measures(Continued)

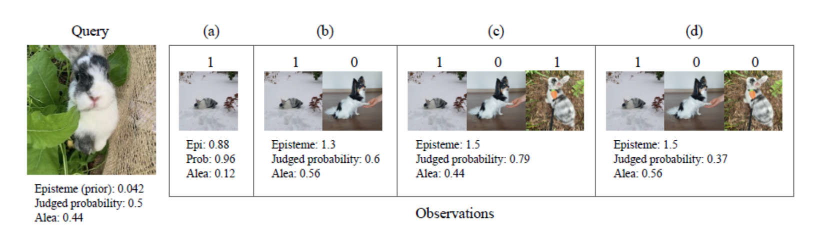

Figure 1: How epistemic, judged probability and aleatoric of GPP changes as more observations are given.

Experiment

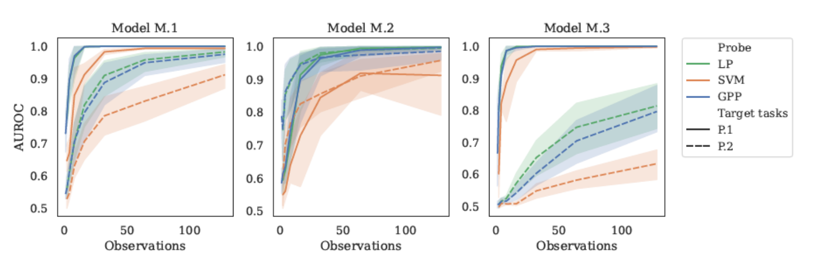

Figure 2: Number over observations versus AUROC curve. $\mathcal{M}_1$ and $\mathcal{M}_3$ are trained on color-related tasks, $\mathcal{M}_2$ is trained on geometry-related task. And, $\mathcal{P}_1$ is color-related task and $\mathcal{P}_2$ is geometry-related task.

Suggestions for Future Research

- Use hyperparameter ϵ means for given $a \in \mathbb{R}^d$, $g(a) \sim beta(\epsilon, \epsilon)$ under given mean and kernel function.
- In original beta gaussian distribution, parameters of $g(a)$ depend on a and they don't necessarily have to be the same value.
- And the tuning process of hyperparameter relies on the researcher.
  $\Rightarrow$ Can't we learn the parameters through the observation?

Contents

Reference

1. Wang, Zi, et al. "Gaussian Process Probes (GPP) for Uncertainty-Aware Probing." arXiv preprint arXiv:2305.18213 (2023).
2. Milios, Dimitrios, et al. "Dirichlet-based gaussian processes for large-scale calibrated classification." Advances in Neural Information Processing Systems 31 (2018).

Appendix: Posterior inference

- Without loss of generality, we write observations as a union of a dataset (of size n) with the positive labels only and a dataset (of size $N-n$) with negative labels only, i.e., $D=\{(a_i,y_i)\}_{i=1}^{N}=D^{+}\cup D^{-}$where $D^{+}=\{(a_i,y_i)\}_{i=1}^{n}$ and $D^{-}=\{(a_i,y_i)\}_{i=n+1}^{N}$.
- For convenience, we use the following short-hand notation:

$$v'=\log\left(\frac{1}{\epsilon+s}+1\right),\quad v''=\log\left(\frac{1}{\epsilon}+1\right),$$
$$y'=\log(\epsilon+s)-\frac{v'}{2},\quad y''=\log(\epsilon)-\frac{v''}{2}$$

Appendix: Posterior inference(Continued)

- Let

$$k(a, \boldsymbol{a}) = [k(a_i, a)]_{i=1}^{N} \in \mathbb{R}^1,$$

$$k\left(\boldsymbol{a}, a'\right) = \left[k\left(a_i, a'\right)\right]_{i=1}^{N} \in \mathbb{R}^{N\times 1},$$

$$\mu(\boldsymbol{a}) = [\mu(a_i)]_{i=1}^{|D|} \in \mathbb{R}^{N\times 1}, \quad K = [k(a_i, a_j)]_{i=1,j=1}^{N} \in \mathbb{R}^{N}$$

for any given $a, a' \in \mathbb{R}^d$ and

$$\boldsymbol{y}_\alpha = \begin{bmatrix} y'1_n \\ y''1_{N-n} \end{bmatrix} \in \mathbb{R}^{N\times 1},$$

$$K_\alpha = K + \begin{bmatrix} v'I_n & 0 \\ 0 & v''I_{N-n} \end{bmatrix} \in \mathbb{R}^{N}.$$

Appendix: Posterior inference(Continued)

- $\boldsymbol{y}_\beta, K_\beta$ are obtained by exchange the location of prime and double prime in $\boldsymbol{y}_\alpha, K_\alpha$.
- $f|D \sim \mathcal{GP}(\mu_D, k_D)$ is obtained from above notation and functions. Its mean and kernel functions are following,
$$\mu_D(a) = k(a, \boldsymbol{a})\left(K_\alpha^{-1}\left(\boldsymbol{y}_\alpha - \mu(\boldsymbol{a})\right) - K_\beta^{-1}\left(\boldsymbol{y}_\beta - \mu(\boldsymbol{a})\right)\right),$$
$$k_D\left(a, a'\right) = 2k\left(a, a'\right) - k(a, \boldsymbol{a})\left(K_\alpha^{-1} + K_\beta^{-1}\right)k\left(\boldsymbol{a}, a'\right)$$
for any given $a, a' \in \mathbb{R}^d$.
- The posterior distribution $g|D \sim \mathcal{G}(\theta_D)$ is obtained by $g = \frac{1}{1+e^{-f}}$ where $f|D \sim \mathcal{GP}(\mu_D, k_D)$.