Uncertainty in Gradient Boosting via Ensembles

Liudmila Prokhorenkova

Yandex Research

Joint work with Andrey Malinin and Aleksei Ustimenko

ICLR 2021

Predictive uncertainty

It is important to detect when ML model is uncertain in its prediction:

- Take safer actions
- Ask for human intervention
- Use active learning

Predictive uncertainty

It is important to detect when ML model is uncertain in its prediction:

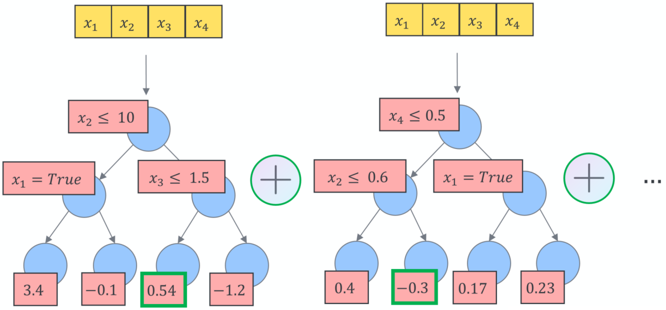
- Take safer actions
- Ask for human intervention
- Use active learning

Types of uncertainty:

- Data uncertainty: class overlap or noise in the data
- Knowledge uncertainty: lack of training data in a region

Gradient boosted decision trees

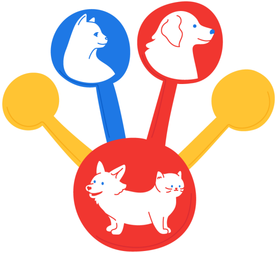
- GBDT is an additive ensemble of decision trees
- Training is iterative
- Each tree corrects errors of previously built ensemble



Data uncertainty

For classification:

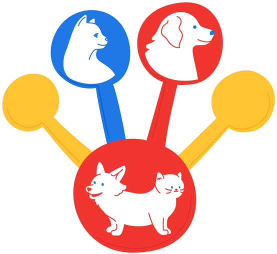
- Train a model with negative log-likelihood (cross-entropy) loss
- For each example, we get a distribution over class labels
- *Data uncertainty*: entropy of this distribution



Data uncertainty

For classification:

- Train a model with negative log-likelihood (cross-entropy) loss
- For each example, we get a distribution over class labels
- *Data uncertainty*: entropy of this distribution



For regression:

- Assume normal distribution of target given features
- Optimize negative log-likelihood
- Estimate mean and variance of the normal distribution
- *Data uncertainty*: variance of this distribution

For GBDT this was done in NGBoost [Duan et al., 2020]

Knowledge uncertainty

Can be estimated via *ensembles* [Lakshminarayanan et al., 2020]:

- Model posterior $\mathrm{p}(\boldsymbol{\theta}|\mathcal{D}) = \frac{\mathrm{p}(\mathcal{D}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})}{\mathrm{p}(\mathcal{D})}$
- Ensemble of models $\{\mathrm{P}(y|\boldsymbol{x};\boldsymbol{\theta}^{(m)})\}_{m=1}^{M}$ sampled from $\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})$
- Knowledge uncertainty is a level of "disagreement" of models

Data, knowledge, and total uncertainty

Classification:

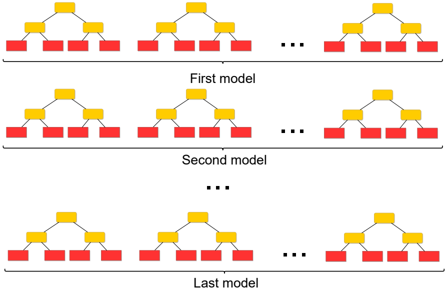
$$\underbrace{\mathcal{H}\big[\mathrm{P}(y|\boldsymbol{x},\mathcal{D})\big]}_{\text{Total Uncertainty}} = \underbrace{\mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}\big[\mathcal{H}[\mathrm{P}(y|\boldsymbol{x};\boldsymbol{\theta})]\big]}_{\text{Expected Data Uncertainty}} + \underbrace{\mathcal{I}\big[y,\boldsymbol{\theta}|\boldsymbol{x},\mathcal{D}\big]}_{\text{Knowledge Uncertainty}}$$

Regression:

$$\underbrace{\mathbb{V}_{\mathrm{p}(y|\boldsymbol{x},\mathcal{D})}[y]}_{\text{Total Uncertainty}} = \underbrace{\mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}\big[\mathbb{V}_{\mathrm{p}(y|\boldsymbol{x},\boldsymbol{\theta})}[y]\big]}_{\text{Expected Data Uncertainty}} + \underbrace{\mathbb{V}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}\big[\mathbb{E}_{\mathrm{p}(y|\boldsymbol{x},\boldsymbol{\theta})}[y]\big]}_{\text{Knowledge Uncertainty}}$$

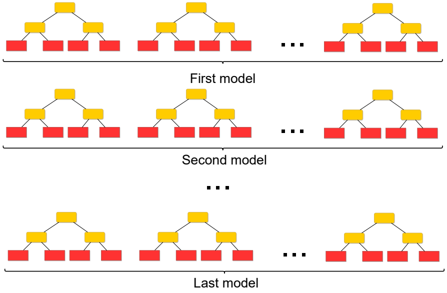
Ensembles of SGB models

- Generate independent Stochastic Gradient Boosting (SGB) models
- Use not too large sample rate, e.g., 0.5
- Obtained models are sampled from some distribution



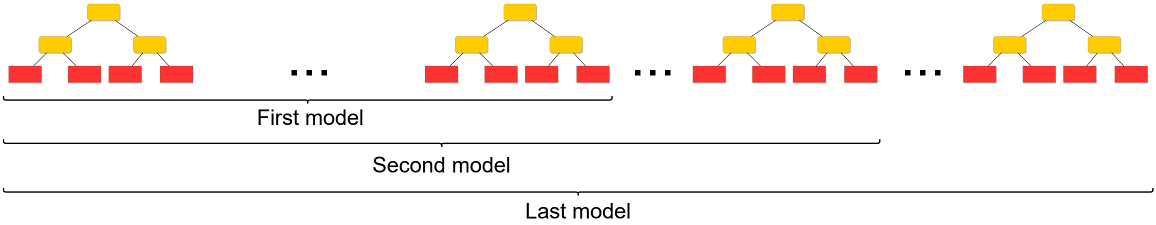
Ensembles of SGLB models

- Use Stochastic Gradient Langevin Boosting [Ustimenko et al., 2020]
- Properly set parameters
- Generate independent SGLB models
- Obtained models are asymptotically sampled from the true posterior



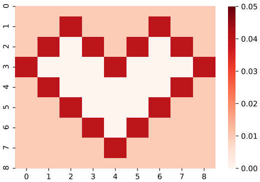
Virtual ensembles

Virtual ensembles



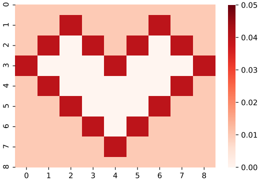
- Generate one SGLB model
- Consider ensemble $\Theta_{T,K} = \{\boldsymbol{\theta}^{(Kt)}, [\frac{T}{2K}] \leq t \leq [\frac{T}{K}]\}$
- No computational overhead for training and inference

Discrete regression dataset

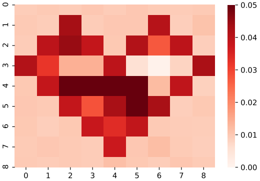


(a) True Data Unc.

Discrete regression dataset

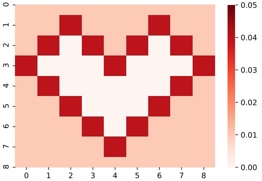


(a) True Data Unc.

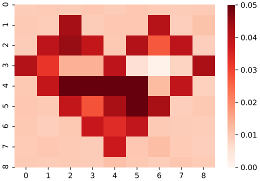


(b) Estimated Data Unc.

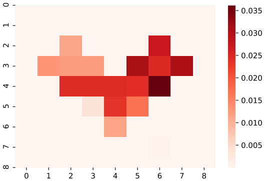
Discrete regression dataset



(a) True Data Unc.

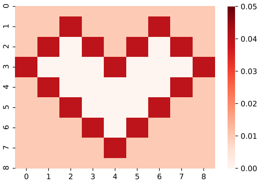


(b) Estimated Data Unc.

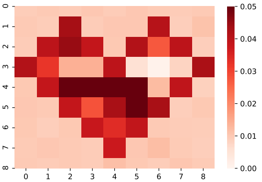


(c) Knowledge Unc. (SGLB)

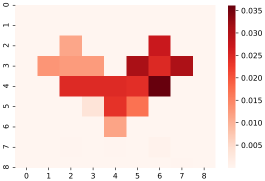
Discrete regression dataset



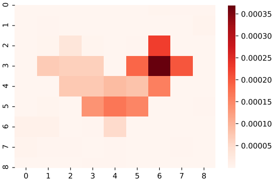
(a) True Data Unc.



(b) Estimated Data Unc.



(c) Knowledge Unc. (SGLB)



(d) Knowledge Unc. (vSGLB)

Out-of-domain detection

Table: Classification % AUC-ROC (↑)

Dataset		Single		Ensemble		
		SGB	SGLB	SGB	SGLB	vSGLB
Adult	TU	53	50	52	51	51
	KU	—	—	89	89	85
Amazon	TU	86	87	86	86	86
	KU	—	—	88	74	67
Click	TU	61	67	64	64	68
	KU	—	—	91	92	90
Internet	TU	67	68	70	69	68
	KU	—	—	87	89	81
KDD-Appetency	TU	29	48	47	50	52
	KU	—	—	90	91	93
KDD-Upselling	TU	53	51	62	60	47
	KU	—	—	97	97	78
Kick	TU	45	37	52	58	38
	KU	—	—	98	98	89

To sum up

- Ensembles of GBDT models allow to estimate data and knowledge uncertainty
- Virtual ensembles can be used as cheaper alternative to the true ones
- The proposed methods are implemented in CatBoost https://catboost.ai
- Data and experiments can be found here:
  https://github.com/yandex-research/GBDT-uncertainty

References

[Ustimenko et al., 2020] Ustimenko A., Prokhorenkova L. "SGLB: Stochastic Gradient Langevin Boosting", arXiv preprint arXiv:2001.07248, 2020.

[Duan et al., 2020] Duan T., Avati A., Ding D. Y., Basu S., Ng A. Y., Schuler A. "NGBoost: Natural Gradient Boosting for Probabilistic Prediction", ICML, 2020.

[Lakshminarayanan et al., 2020] Lakshminarayanan B., Pritzel A., Blundell C. "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles", NIPS, 2017.

Thank you!