Deep Bayesian Bandits: Exploring in Online Personalized Recommendations

Dalin Guo*, Sofia Ira Ktena, Ferenc Huszar, Pranay Kumar Myana, Michael Kneier, Sourav Das, Wenzhe Shi, Alykhan Tejani
***dag082@ucsd.edu, UC San Diego; {first name initial}{surname}@twitter.com, Twitter**

#TheProblem

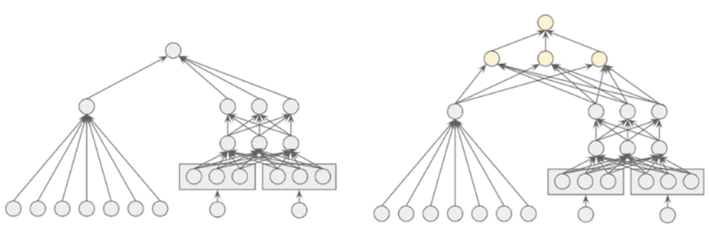
- **Continuous training:** trained with newly collected data continuously due to non-stationarity of feature distribution
- **Algorithmic bias:** data selected by the model (a feedback loop)
- **Exploration-Exploitation tradeoff**: explore: information (high accuracy) vs. exploit: reward (revenue)
- **Ads prediction:** Predict the **p**robability of **C**lick **T**hrough **R**ate (pCTR) given user, ads and other contextual features

#KeyIdeas

- Model as a Bernoulli contextual bandit problem
- Utilize neural network to generalize across users and items
- Obtain uncertainty estimation from neural network for bandit

#Models

- pCTR model: 3-layer feedforward neural network (offline), Wide-and-Deep neural network [1] (online)
- Bandit algorithms: ε-greedy, Thompson sampling (TS), Upper Confidence Bound (UCB)
- Posterior approximation methods: Dropout, Bootstrapping [2,3,4]
  - **Multihead**: bottom network shared [4]
  - **Hybrid model**: Dropout units on the second to last layer only



Original Wide & Deep Models

Multi-head Wide & Deep Models

#Dataset

- Offline: ADS-16 [5]: 120 users rate 300 ads (full observation)
- Online: Twitter ads traffic
- Metrics: PR-AUC on a test set & Accumulated averaged CTR

#Results

Offline Simulation

- trade-off between CTR and PR-AUC
- trade-off between CTR and computational cost
- UCB > TS

Table 1: Offline simulation: performance comparison

Model	CTR (+%)	PR-AUC
Random	0	0.5
Greedy	91.77	0.6565
ϵ-greedy	91.94	0.6501
Dropout TS	94.60	0.6421
Dropout UCB	97.16	0.5236
Bootstrap TS	94.83	0.5519
Bootstrap UCB	139.03	0.5307
SGD UCB	127.95	0.5335
Multihead UCB	112.79	0.5279
Multihead SGD UCB	96.30	0.5218
Hybrid TS	67.56	0.6311
Hybrid UCB	82.44	0.5165

Warm-start hybrid model

- Better performance with longer training epochs

Table 2: Offline simulation: Warm-start the hybrid model

Model (# epochs)	train PR-AUC	CTR (+%)	test PR-AUC
Random	0.5	0	0.5
ϵ-greedy (100)	0.5951	94.30	0.6692
Hybrid (100)	0.5001	85.99	0.5108
Hybrid (200)	0.5584	60.51	0.5165
Hybrid (500)	0.5895	128.66	0.5294

Online Model Performance

- Similar predictive performance as production (-0.0253 RCE)
- +2% impressions with a flat revenue (no significant +/-)
- no significant decrease in training and serving speed
- no direct improvement in product metrics
- no increase in negative engagement rate
  - vs. ε-greedy: 100% increase in negative engagement rate
- a higher RCE and ROC-AUC of trained model than production

Table 3: Predictive performance of models self-trained in online A/B test

Model	RCE	ROC-AUC(%)
Hybrid	**8.12**	**68.37**
Control	7.95	67.13

#Conclusion

- Bandit algorithms + neural network + uncertainty approximation
- A hybrid method: dropout units in second-to-last layer
- Offline simulation + online AB testing
- Efficiency + effectiveness

#References

[1] Cheng, Heng-Tze, et al. "Wide & deep learning for recommender systems." Proceedings of the 1st workshop on deep learning for recommender systems. 2016.
[2] Riquelme C et al, “Deep Bayesian Bandits Showdown: An Empirical Comparison of Bayesian Deep Networks for Thompson Sampling”, ICLR, 2018.
[3] Gal, Y et al, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning.” ICML, 2016.
[4] Osband, I et al, “Deep exploration via bootstrapped DQN.”, Neurips, 2016.
[5] Roffo, G et al, “Personality in computational advertising: A benchmark”, EMPIRE, 2016