Empirical Bayes in Recommender System

HKUST Summer Program

Zihui Weng

Department of Mathematics, HKUST

2022.8.6

Starting Point

- Search for a good design of informative prior instead a non-informative prior
- Decouple the learning rate of different categories
- Solution to the trade-off between exploration and exploitation
- Provide a general solution in estimating empirical priors for a wide range of applications that are modeled as Bayesian bandits or involve Bayesian learning, esp. recommender system and MABs

Objective

Utilize early randomized data and Empirical Bayes(EB) method to construct an experiment-specific hierarchical informative prior

Assumption

- The kth category C_k has a distince hyperparameter meta-prior distribution(determined by experts' knowledge), here we assume it as Gaussian, $N(\nu_k, \tau_k^2)$
- One feature's true effect(i.e. coefficients) μ_i is drawn i.i.d from the corresponding category's meta-prior distribution: $\mu_i \sim N(\nu_k, \tau_k^2)$ and $\mathbb{E}[\mu_i] = \nu_k, \mathbb{V}[\mu_i] = \tau_k^2 \ \forall i \in C_k$
- $\tilde{\mu}_i$ and $\tilde{\sigma_i^2}$ is the estimators of $\mathbb{E}[\mu_i]$ and $\mathbb{V}[\mu_i]$ respectively.
- $\mathbb{E}\left[\tilde{\mu}_i \mid \mu_i\right] = \mu_i, \mathbb{V}\left[\tilde{\mu}_i \mid \mu_i\right] = \tilde{\sigma}_i^2, \mathbb{E}\left[\tilde{\mu}_i\right] = \nu_k, \quad \forall i \in C_k$
- Setting $\nu_k = 0$ to ensure the model is invariant to input feature sign changes.
- $y \in \{-1, 1\}$ denotes the response binary variable, x denotes the features

Derivation

- Variance Decomposition:
  $\mathbb{V}[\tilde{\mu}_i] = \mathbb{E}[\mathbb{V}[\tilde{\mu}_i \mid \mu_i]] + \mathbb{V}[\mathbb{E}[\tilde{\mu}_i \mid \mu_i]] = \mathbb{E}\left[\tilde{\sigma}_i^2\right] + \tau_k^2, \quad \forall i \in C_k$
- $\tau_k^2 = \mathbb{V}[\tilde{\mu}_i] - \mathbb{E}\left[\tilde{\sigma}_i^2\right]$
- $\hat{\tau}_{k,t}^2 = \widehat{\mathbb{V}[\tilde{\mu}_{i,t}]} - \widehat{\mathbb{E}\left[\tilde{\sigma}_{i,t}^2\right]} = \frac{\sum_{i \in C_k}\left(\tilde{\mu}_{i,t} - \hat{\nu}_{k,t}\right)^2}{N_k - 1} - \frac{\sum_{i \in C_k} \tilde{\sigma}_{i,t}^2}{N_k}$ as an estimator for the meta-prior variance at time t
- $\hat{\nu}_{k,t} = \frac{\sum_{i \in C_k} \tilde{\mu}_{i,t}}{N_k}, \quad \forall C_k$
- Setting $\nu_k = 0$ and obtain additional one degree of freedom
- $\hat{\tau}_{k,t}^2 = \frac{\sum_{i \in C_k}\left[\tilde{\mu}_{i,t}^2 - \tilde{\sigma}_{i,t}^2\right]}{N_k}, \quad \forall C_k$

Update Strategy

- Meta-prior $\mathcal{N}(0, \tau_k^2)$
- Bayesian Linear Probit Model
  $P(y \mid x, \tilde{\mu}) = \Phi\left(y \cdot \frac{\tilde{\mu}^T x}{\beta}\right) \quad \beta = 1$
- Component-wise[2] and Matrix-wise[3]

Component-wise

- A vector of means $\mu := (\mu_{1,1}, \ldots\ldots, \mu_{N,M_N})^T$ vector of variances $\sigma^2 := (\sigma^2_{1,1}, \ldots\ldots, \sigma^2_{N,M_N})^T$
- sample
  $x := (x_1^T, \ldots\ldots, x_N^T)$,$x_i := (x_{i,1}, \ldots\ldots, x_{i,M_i}), \sum_{j=1}^{M_i} x_{i,j} = 1$
- $\sum^2 := \beta^2 + x^T\sigma^2$
- $v(t) := \frac{\mathcal{N}(t;0,1)}{\Phi(t;0,1)} \quad w(t) := v(t)[v(t) + t]$
- $\tilde{\mu}_{i,j} = \mu_{i,j} + yx_{i,j}\frac{\sigma^2_{i,j}}{\Sigma} v(\frac{yx^T\mu}{\Sigma})$
- $\tilde{\sigma}^2_{i,j} = \sigma^2_{i,j}(1 - x_{i,j}\frac{\sigma^2_{i,j}}{\Sigma^2} w(\frac{yx^T\mu}{\Sigma}))$

Matrix-wise

- $Y_i^* = x_i^T \mu + \epsilon_i \sim^{i.i.d} \mathcal{N}(0,1)$ as latent variables
- $p(y_i|y_i^*) = \mathbf{1}_{y_i=0}\mathbf{1}_{y_i<0} + \mathbf{1}_{y_i=1}\mathbf{1}_{y_i=\geq 0}$
- The posterior function with augmented variables: $\pi(\mu, Y_i^*|y, X) \propto$ $\sum_{i=1}^{n}[p(y_i|y_i^*)] \times N_N(Y^*|X\mu, I_N) \times N_K(\mu|\mu_0, B_0)$
- Y_i^*'s full conditional distribution:
  - $Y_i^*|\mu, y, X \sim TN_{[0,\infty)}(x_i^T\mu, 1), \quad y_i = 1$
  - $Y_i^*|\mu, y, X \sim TN_{(-\infty,0)}(x_i^T\mu, 1), \quad y_i = 0$
- Update Formula:
  - $\mu|Y^*, X \sim N(\mu_n, B_n)$
  - $B_n = (B_0^{-1} + X^TX)^{-1}, \quad \mu_n = B_n(B_0^{-1}\mu_0 + X^TY^*)$

Adult Dataset

- One-hot encoding for first and second order features
- Variable selection through adaptive LASSO and two choice for selections:
  - select only the variables with non-zero weights
  - select the whole feature when there exists a non-zero weight interaction in it
- Problems about degenerative meta-prior variance
  - Speculation of low traffic of a training batch
  - Some solutions toward this: Bootstrapping, Ensembles, Epoch Training
  - Bootstrap for the first batch into several 5K instances batches, treat each bootstrapped set as a training epoch and compute the meta-prior variance through standard Gaussian prior after each epoch until obtaining the non-degenerative meta-prior variance

Proposed Models

- BLIP: Update the model in batch with day t data
- BLIPBayes: Specify the prior reset time t and upgrade the meta-prior variance using the bootstrapped data from the data observed until t
- BLIPTwice: update the model twice, first with the same bootstrapped data as BLIPBayes and second with all the data observed until day t
- Choose cross-entropy as criterion to compute the log loss of the testing set:
  $Logloss = -\frac{1}{n}\sum_{j=1}^{n} y_j log(P_j) + (1 - y_j) log(1 - P_j)$

Effect of First-Order Features

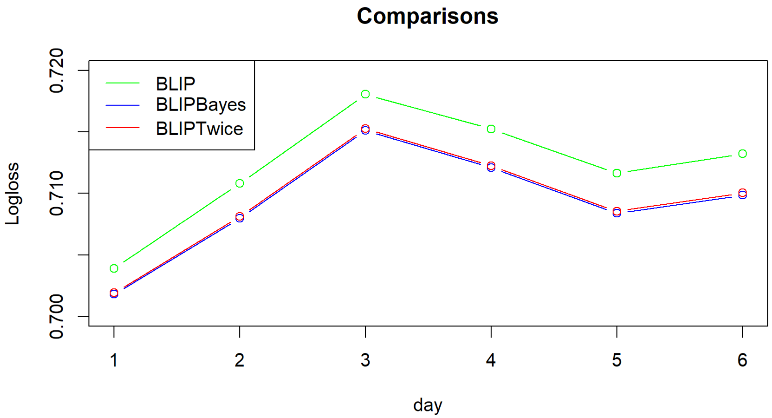
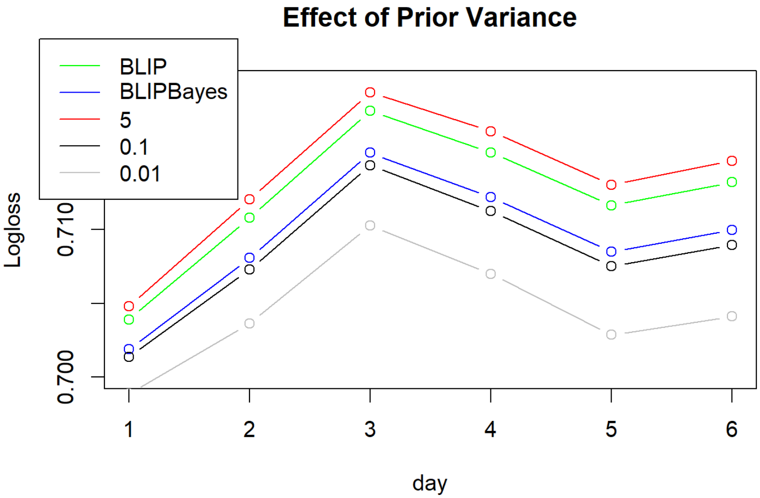


Figure 1: All First order

Effect of First-Order Features

Figure 2: Selected First order

Effect of Prior Variance



Effect of Reset Times

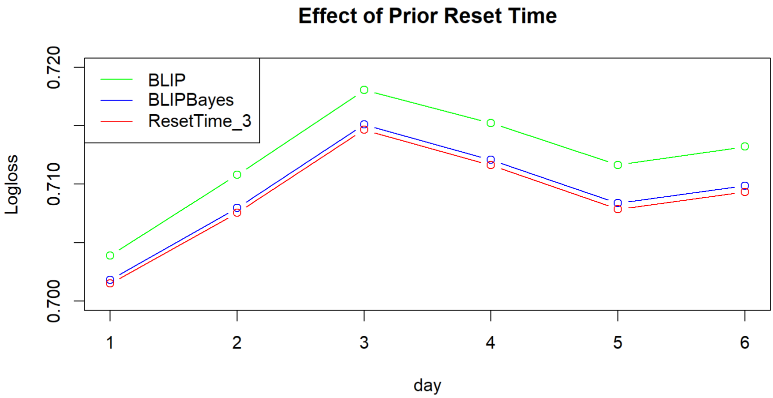


Figure 4: Effect of Reset Time

Effect of the Size of Batches

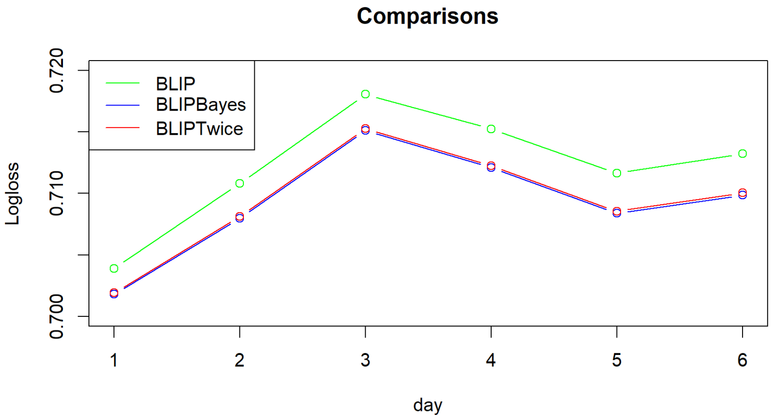
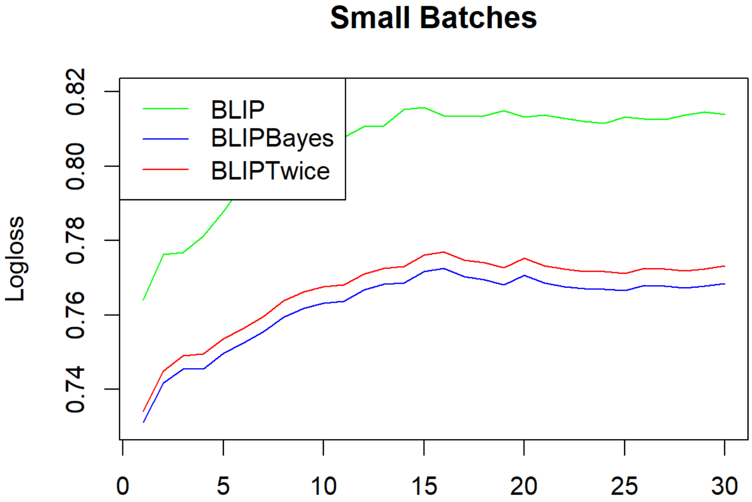


Figure 5: 6 Batches

Effect of the Size of Batches



Overview[5]

- Method: "Learning the experience from others"
- Examples: Robbins' Formula, The Missing Species Problem, James-Stein Estimator, False Discovery Rate......
- f-modeling and g-modeling

Recent research on Empirical Bayes

EB method when known F and unknown G

EB method with unknown F and G

EB method with unknown F and G

- Order statistic regression on replicated data[6]
- Driven Force: Replication makes it possible to estimate μ_i with no assumptions on F and G with the aim of matching the risk of Bayes rule
- Target: Point estimation of the posterior mean $\mathbb{E}_F[Z_{ij}|\mu_i, \alpha_i]$
- Key Method: The key insight is that the conditional mean $\mathbb{E}_{F,G}[Z_{ij}|X_i]$ is (almost surely) identical to the posterior mean $\mathbb{E}_{F,G}[\mu_i|X_i]$, which represents that Bayes rule could be estimated via $\mathbb{E}_{F,G}[\mu_i|X_i]$ and simply regress Y_i on X_i under any black-box predictive model.

Algorithms

Aurora: "Averages of Units by Regressing on Ordered Replicates Adaptively."

For $j \in \{1, \ldots, K\}$

1. Split the replicates for each unit, $\boldsymbol{Z}_i$, into $\boldsymbol{X}_i := (Z_{i1}, \ldots, Z_{i(j-1)}, Z_{i(j+1)}, \ldots, Z_{iK})$ and $Y_i := Z_{ij}$, as in (5).
2. For each $\boldsymbol{X}_i$, order the values to obtain $\boldsymbol{X}_i^{(\cdot)}$.
3. Regress Y_i on $\boldsymbol{X}_i^{(\cdot)}$ using any black-box predictive model. Let $\hat{m}_j$ be the fitted regression function.
4. Let $\hat{\mu}_{i,j}^{\text{Aur}} := \hat{m}_j(\boldsymbol{X}_i^{(\cdot)})$.

end

Estimate each μ_i by $\hat{\mu}_i^{\text{Aur}} := \frac{1}{K} \sum_{j=1}^{K} \hat{\mu}_{i,j}^{\text{Aur}}$.

Figure 7: Aurora

Confidence Intervals for Nonparametric Empirical Bayes Analysis[7]

- Starting Point: The importance of measuring the uncertainty of the estimators using empirical bayes methods
- Two Approaches:
  - F-localization
  - AMARI
- Both two approches will finally be transformed into an optimization problem using the Charnes and Cooper transformation for linear-fractional programming.
- Difference between two approaches is that F-localization constructs a simultaneous interval while AMARI constructs pointwise interval.

Intro to MABs

- Find strategies for exploration-exploitation and cold start problem.
- Information Ratio, Regret Bound
- Algorithms: non-Bayesian: ϵ-greedy, UCB,etc; Bayesian: Thompson Sampling.

Bayes in Bandits[8]

- Original Thompson Sampling: Conjucate Prior (Beta) in Bernoulli bandit settings
- What if the distribution behind the prior and likelihood function is not conjucate(inconsistent) to the bandits settings? For example, Gaussian in Bernoulli?
- Gaussian Imagination(i.e. regard the observed data from true belief as generation of Gaussian prior)

Bayes in Bandits

- Assumptions:
    - For all $t \in \mathbb{Z}_{++}$, $\mathbb{E}\left[\mathbb{E}\left[\tilde{R}_* \mid \tilde{H}_t \leftarrow H_t\right]\right] \geq \mathbb{E}\left[R_*\right]$
    - The imaginary learning target $\tilde{\chi}$ and the imaginary mean reward $\tilde{\theta}$ are jointly Gaussian.
- Regret Bound:
  $\mathcal{R}(T) \leq \sqrt{\mathbb{I}(\tilde{\chi}, \tilde{\mathcal{E}})\tilde{\Gamma}_{\tilde{\chi},\epsilon} T} + \epsilon T + \gamma\sqrt{2d_{KL}(\mathbb{P}(\theta \in \cdot)||\mathbb{P}(\tilde{\theta} \in \cdot))T}$
- Relation to the robust estimation

1. Background[1]
2. Method and Derivations
3. Simulations and Results
4. Literature Review and Extensions
   - Empirical Bayes
   - Bayesian in MABs
   - Meta-Learning v.s. Hierarchical Bayesian Model[4]
   - Development of Bayesian Linear Probit Model
5. References

Meta-Learning

- High requirement of generalization ability in multiple tasks $\rightarrow$ Meta-Learning
- Two methods of the construction of meta-learning:
    - Gradient-based hyperparameter optimization
    - Probabilistic inference in a hierarchical Bayesian model

Type 1 Construction

- Provides a gradient-based meta-learning procedure that employs a single additional parameter (the meta-learning rate) and operates on the same parameter space for both meta-learning and fast adaptation.
- The objective of MAML: $\mathcal{L}(\theta) = \frac{1}{\mathcal{J}} \sum_j [\frac{1}{\mathcal{M}} \sum_m -logp(x_{j_{N+m}} | \theta - \alpha \nabla_\theta \frac{1}{\mathcal{N}} \sum_n -logp(x_{j_N} | \theta))]$
- In particular, in the case of meta-learning, each task-specific parameter ϕ_j is distinct from but should influence the estimation of the parameters from other tasks.

Type 2 Construction

- Objective: $p(X|\theta) = \prod_j (\int p(x_{j_1}, ..., x_{j_N}|\theta_j) p(\phi_j|\theta) d\phi_j)$
- We maximize this objective as a function of θ and apply empirical bayes to estimate the prior parameters.

Connection between Two Methods

- The task-specific parameter ϕ_j's exactly marginal distribution in the objective of hierarchical bayesian model is not tractable to obtain
- Consider an approximation of that objective using $\hat{\phi}_j$: $-logp(x|\theta) \approx \sum_j [-logp(x_{j_{N+1}}, ..., x_{j_{N+M}} | \hat{\phi}_j)]$
- Setting $\hat{\phi}_j = \theta + \alpha \nabla_\theta logp(x_{j_1}, ..., x_{j_N} | \theta)$ we could obtain the unscaled form of objective in gradient-based hyperparameter optimization.
- Trade-off between optimizing the objective and staying close to θ.

Conclusion: MAML can be understood as empirical Bayes in a hierarchical probabilistic model!

1. Background[1]
2. Method and Derivations
3. Simulations and Results
4. Literature Review and Extensions
    - Empirical Bayes
    - Bayesian in MABs
    - Meta-Learning v.s. Hierarchical Bayesian Model[4]
    - Development of Bayesian Linear Probit Model
5. References

History

- Assumed Density Filtering Expectation Propogation [9]
- Factor Graph Sum Product Algorithm [10]
- TrueSkill [11]
- Bayesian Linear Probit Model

[1] S. Nabi, H. Nassif, J. Hong, H. Mamani, and G. Imbens, "Bayesian meta-prior learning using empirical bayes," 2020.

[2] T. Graepel, J. Quiñonero Candela, T. Borchert, and R. Herbrich, "Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine," in *Proceedings of the 27th International Conference on Machine Learning ICML 2010, Invited Applications Track (unreviewed, to appear)*, June 2010.

[3] J. H. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 669–679, 1993.

[4] E. Grant, C. Finn, S. Levine, T. Darrell, and T. Griffiths, "Recasting gradient-based meta-learning as hierarchical bayes," 2018.

[5] B. Efron, "Empirical bayes: Concepts and methods," 2010.

[6] N. Ignatiadis, S. Saha, D. L. Sun, and O. Muralidharan, "Empirical bayes mean estimation with nonparametric errors via order statistic regression on replicated data," 2019.

[7] N. Ignatiadis and S. Wager, "Confidence intervals for nonparametric empirical bayes analysis," 2019.

[8] Y. Liu, A. M. Devraj, B. Van Roy, and K. Xu, "Gaussian imagination in bandit learning," 2022.

[9] D. Khashabi, "Expectation propagation for bayesian inference," 2010.

[10] F. Kschischang, B. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[11] R. Herbrich, T. Minka, and T. Graepel, "Trueskill™: A bayesian skill rating system," in *Advances in Neural Information Processing Systems* (B. Schölkopf, J. Platt, and T. Hoffman, eds.), vol. 19, MIT Press, 2006.

[12] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4.
Springer, 2006.

[13] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, and J. Q. n. Candela, "Practical lessons from predicting clicks on ads at facebook," in *Association for Computing Machinery*, Association for Computing Machinery, 2014.

Thank You