Microsoft Research

UCL

Collective Noise Contrastive Estimation for Policy Transfer Learning

Weinan Zhang, University College London

Ulrich Paquet, Katja Hofmann, Microsoft Research

Xbox Radio Track Playing Problem

For each session:

- **Seed:** the user sets a seed artist
- **Action:** the policy plays tracks one by one to the user
- **Feedback**: if the user does not like the current track, he can push the 'skip' button

Problem:

- Design a policy to maximise user satisfaction, quantified as policy reward score

Two data sources:

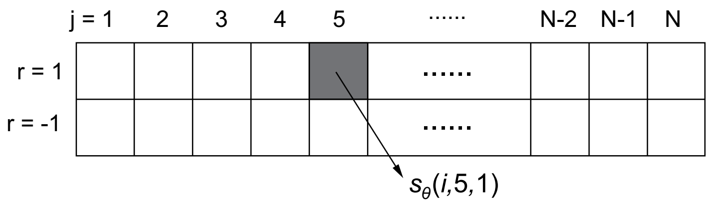
- Auxiliary data: user self-generated playlists (can be regarded as positive examples)
- Target data: user feedback (skip or listen) given the historic radio playing

Model

Notations:

- Context *i,* selected artist *j,* reward *r*
- Score function $s_\theta(i,j,r) = r \cdot w_i^T w_j + b_j$

Softmax-based Stochastic Policy:



- Conditional probability

  $P_\theta(j,r|i) = e^{s_\theta(i,j,r)} \,/\, \sum_{r'}\sum_{j'} e^{s_\theta(i,j',r')}$

- Radio selection

  $P_\theta(j|i,r=1) = e^{s_\theta(i,j,1)} \,/\, \sum_{j'} e^{s_\theta(i,j',1)}$

Objective 1: Maximising data generation likelihood

- Playlist data likelihood: $\mathcal{L}_P(P_\theta) = \prod_{(i,j,1)\in D_P} P_\theta(j|i,r=1)$
- Radio data likelihood: $\mathcal{L}_R(P_\theta) = \prod_{(i,j,r)\in D_R} P_\theta(j,r|i)$
- Joint optimisation:

$$\max_\theta \frac{\alpha}{|D_R|}\log\mathcal{L}_R(P_\theta) + \frac{1-\alpha}{|D_P|}\log\mathcal{L}_P(P_\theta)$$

$$= \max_\theta \frac{\alpha}{|D_R|}\sum_{(i,j,r)\in D_R}\log\frac{e^{s_\theta(i,j,r)}}{\sum_{r'}\sum_{j'} e^{s_\theta(i,j',r')}} + \frac{1-\alpha}{|D_P|}\sum_{(i,j,1)\in D_P}\log\frac{e^{s_\theta(i,j,1)}}{\sum_{j'} e^{s_\theta(i,j',1)}}$$

Objective 2: Maximising inverse propensity score (IPS) based policy value

- Expected reward of a policy: $\mathbb{E}_i[\mathbb{E}_{P_\theta(j|i,r'=1)}[\overrightarrow{r_i}[j]]]$
- IPS policy value on radio data:

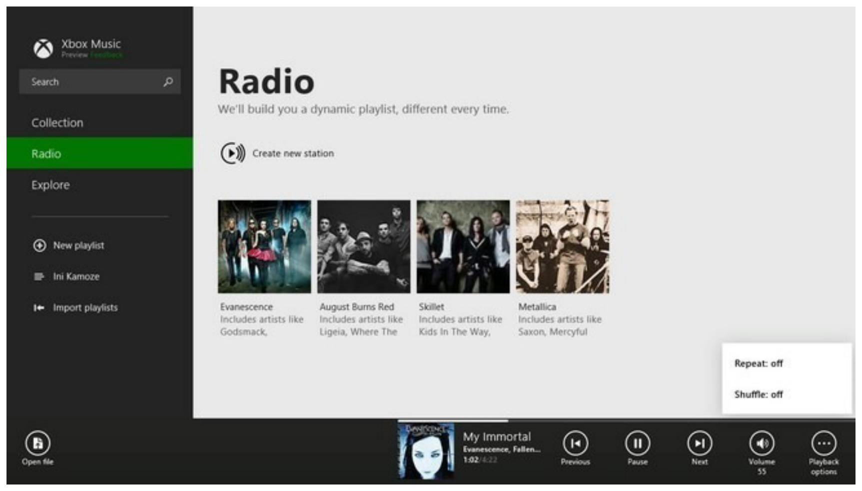
$$\hat{V}_{\text{ips}}(P_\theta) = \frac{1}{|D_R|}\sum_{(i,j,r)\in D_R}\frac{rP_\theta(j|i,r'=1)}{P_D(j|i)} = \frac{1}{|D_R|}\sum_{(i,j,r)\in D_R}\frac{r}{P_D(j|i)}\frac{e^{s_\theta(i,j,1)}}{\sum_{j'} e^{s_\theta(i,j',1)}}$$

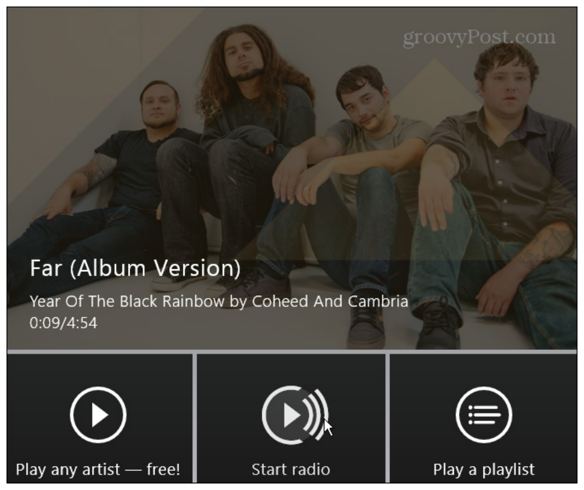
$$> \frac{1}{|D_R|}\sum_{(i,j,r)\in D_R}\frac{r}{P_D(j|i)}\log\frac{e^{s_\theta(i,j,1)}}{\sum_{j'} e^{s_\theta(i,j',1)}} = \tilde{V}_{\text{ips}}(P_\theta) \quad \text{[lower bound]}$$

- Joint optimisation:

$$\max_\theta \alpha\tilde{V}_{\text{ips}}(P_\theta) + \frac{1-\alpha}{|D_P|}\log\mathcal{L}_P(P_\theta)$$

$$= \max_\theta \frac{\alpha}{|D_R|}\sum_{(i,j,r)\in D_R}\frac{r}{P_D(j|i)}\log\frac{e^{s_\theta(i,j,1)}}{\sum_{j'} e^{s_\theta(i,j',1)}} + \frac{1-\alpha}{|D_P|}\sum_{(i,j,1)\in D_P}\log\frac{e^{s_\theta(i,j,1)}}{\sum_{j'} e^{s_\theta(i,j',1)}}$$





Gradient calculation via noise contrastive estimation (NCE)

- Expensive gradient calculation on softmax function:

$$\frac{\partial}{\partial\boldsymbol{\theta}}\log\frac{e^{s_\theta(i,j,r)}}{\sum_{j'} e^{s_\theta(i,j',r)}} = \frac{\partial s_{\boldsymbol{\theta}}(i,j,r)}{\partial\boldsymbol{\theta}} - \mathbb{E}_{P_\theta(j'|i,r)}\Big[\frac{\partial s_{\boldsymbol{\theta}}(i,j',r)}{\partial\boldsymbol{\theta}}\Big]$$

- NCE idea: define a loss function to quantify how likely the policy will separate a data point from *k* noise data points generated from a known noise probabilistic distribution.

$$\mathcal{L}^{(i,j,r)}_{\text{NCE}}(\boldsymbol{\theta}) = \log\frac{P_\theta(j|i,r)}{P_\theta(j|i,r)+kP_n(j)} + \sum_{m=1}^{k}\log\frac{kP_n(j_m)}{P_\theta(j_m|i,r)+kP_n(j_m)}$$

$$\frac{\partial}{\partial\boldsymbol{\theta}}\mathcal{L}^{(i,j,r)}_{\text{NCE}}(\boldsymbol{\theta}) = \frac{kP_n(j)}{e^{s_{\boldsymbol{\theta}}(i,j,r)}+kP_n(j)}\frac{\partial s_{\boldsymbol{\theta}}(i,j,r)}{\partial\boldsymbol{\theta}} - \sum_{m=1}^{k}\frac{e^{s_{\boldsymbol{\theta}}(i,j_m,r)}}{e^{s_{\boldsymbol{\theta}}(i,j_m,r)}+kP_n(j_m)}\frac{\partial s_{\boldsymbol{\theta}}(i,j_m,r)}{\partial\boldsymbol{\theta}}$$

- when $k\to+\infty$, the gradient $\frac{\partial}{\partial\boldsymbol{\theta}}\mathcal{L}^{(i,j,r)}_{\text{NCE}}(\boldsymbol{\theta}) \to \frac{\partial}{\partial\boldsymbol{\theta}}\log\frac{e^{s_\theta(i,j,r)}}{\sum_{j'} e^{s_\theta(i,j',r)}}$.

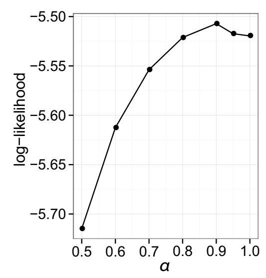
Experiment on Xbox Music Playlist & Radio Data

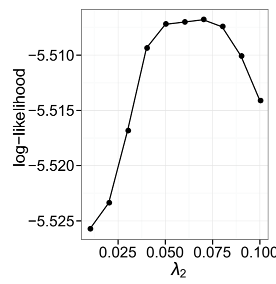
Datasets:

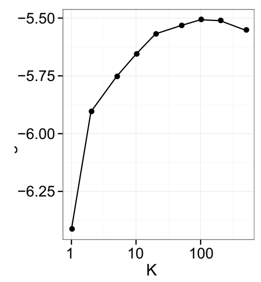
- **Playlist:** 20.3k playlists with 722.7k transitions on 1.81k artists
- **Radio:** 97.6k transation sequences on 1.44k artists (1.03k artists occur in playlists)

Performance with Objective 1:

Algorithm	log-likelihood
Random	-7.7932
Popularity	-5.8009
NCE-Playlist	-10.3978
NCE-Radio	-5.5197
NCE-Collective	**-5.5072**

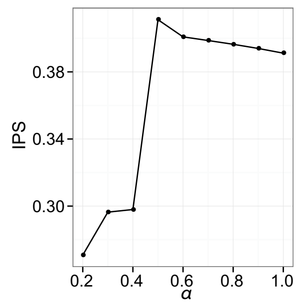


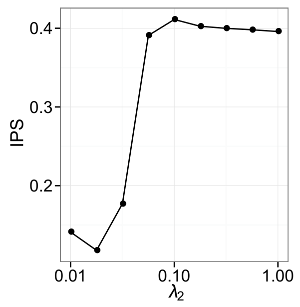


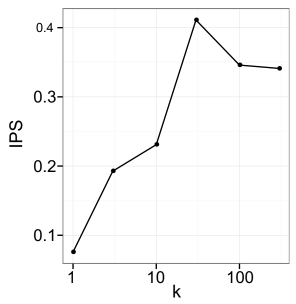


Performance with Objective 2:

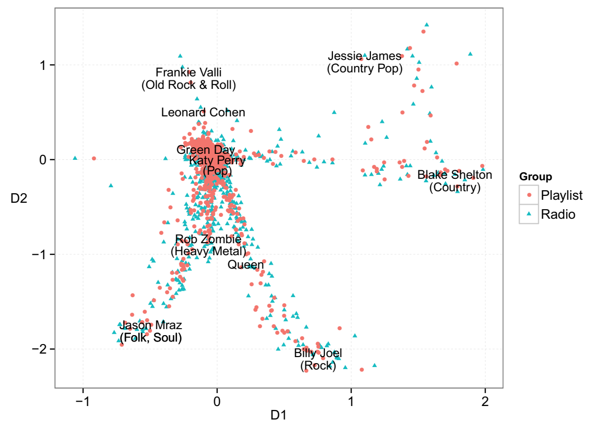
Algorithm	IPS Value
Random	0.0687
Popularity	0.0747
SameArtist	-0.3088
NCE-Playlist	0.0695
NCE-Radio	0.3912
NCE-Collective	**0.4111**







Case studies:



Seed	Queen	Blake Shelton	Billy Joel	Jessie James
1	Status Quo	Eric Church	Don Henley	Lila McCann
2	Uriah Heep	James Otto	Mr. Mister	Sara Evans
3	The Romantics	Steve Holy	Little River Band	Jamie O'Neal
4	David Gilmour	Miss Willie Brown	Peter Cetera	Chely Wright
5	Duane Allman	Bobby Pinson	Rita Coolidge	Lorrie Morgan
6	Ash Wednesday	Jason Blaine	Janis Ian	Tanya Tucker
7	Michael Bolton	Chad Brock	Karla Bonoff	K.T. Oslin
8	Lobo	Easton Corbin	Bruce Cockburn	Briston Latina
9	Nils Lofgren	Love And Theft	Orleans	Beat This Summer
10	David Knopfler	John Rich	Nils Lofgren	The Charlie Daniels Band
11	Orleans	Eli Young Band	David Knopfler	The Statler Brothers
12	Bruce Cockburn	Josh Turner	Bob Dylan	Roger Miller
13	Karla Bonoff	Billy Currington	Lobo	Steve Earle
14	Janis Ian	Darius Rucker	Gino Vannelli	June Carter Cash
15	Marc Cohn	Craig Morgan	Rick Springfield	Charlie Louvin