A PREDICTIVE MODEL OF MUSIC PREFERENCE USING PAIRWISE COMPARISONS

*Bjørn Sand Jensen, Javier Saez Gallego & Jan Larsen**

Technical University of Denmark, Department of Informatics and
Mathematical Modeling, Richard Petersens Plads B321
2800 Lyngby, Denmark

ABSTRACT

Music recommendation is an important aspect of many streaming services and multi-media systems, however, it is typically based on so-called collaborative filtering methods. In this paper we consider the recommendation task from a personal viewpoint and examine to which degree music preference can be elicited and predicted using simple and robust queries such as pairwise comparisons. We propose to model - and in turn predict - the pairwise music preference using a very flexible model based on Gaussian Process priors for which we describe the required inference. We further propose a specific covariance function and evaluate the predictive performance on a novel dataset. In a recommendation style setting we obtain a leave-one-out accuracy of 74% compared to 50% with random predictions, showing potential for further refinement and evaluation.

***Index Terms*—** Music Preference, Kernel Methods, Gaussian Process Priors, Recommendation

1. INTRODUCTION

Methods for music recommendation has received a great deal of attention the last decade with most approaches typically being classified as collaborative filtering (*top-down*) or content-based (*bottom-up)* methods, with hybrid methods (see e.g. [1]) comprising both. Such hybrid systems exploits both ratings and contents to make recommendations, but the focus is still on the recommendation itself and not the basic questions of preference. Although from a fundamental point of view it is also interesting how well human preference can be elicited and represented without relying on the help of others. This also includes the aim to answer basic questions such as which properties of music determines the persons music preference. Obviously the potential power of collaborative filtering should not be discarded, but exploited in a principled manner in order to answer basic questions and hopefully provide an even better predictive model of individual music preference.

Based on these observations we consider music preference in a personalized setting by applying a Gaussian Process regression model which takes into account both human ratings and audio features. In contrast to many audio rating systems it is not based on absolute ratings of a single track, but on a pairwise comparisons between tracks, which is typically considered robust and have a low cognitive load (see e.g. [2]).

We furthermore propose to use a covariance function motivated from a generative view of audio features with a potential multi-task part which lead to similar capabilities as standard collaborative filtering, but with the added information level provided by subject features. Posterior inference in the resulting non-parametric Bayesian regression model is performed using a Laplace approximation of the otherwise intractable distribution. Any hyperparameters in the model can be learned using an empirical Bayes approach.

We evaluate the resulting model by its predictive power on a small scale, public available dataset [3] where 10 subjects evaluate 30 tracks in 3 genres. We report and discuss a number of aspects of the performance such as the learning curves as a function of the number of pairwise comparisons and learning curves when leaving out a track as test set.

2. METHODS

In this work we focus on modeling preference elicited by pairwise queries, i.e., given two inputs tracks u and v we obtain a response, $y \in \{-1, 1\}$, where $y = -1$ corresponds to a preference for u, and $+1$ corresponds to a preference for v. We consider n distinct input tracks $x_i \in \mathcal{X}$ denoted $\mathcal{X} = \{x_i | i = 1, ..., n\}$, and a set of m responses on pairwise comparisons between any two inputs in $\mathcal{X}$, denoted by

$$\mathcal{Y} = \{(y_k; u_k, v_k) | k = 1, ..., m\},$$

where $y_k \in \{-1, 1\}$. $u_k \in \mathcal{X}$ and $v_k \in \mathcal{X}$ are option one and two in the k'th pairwise comparison.

We consider y_k as a stochastic variable and we can then formulate the likelihood of observing a given response as cu-

*This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views.
** This work was supported in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328.

mulative normal distribution.

$$p\left(y_k|\mathbf{f}_k,\boldsymbol{\theta}_{\mathcal{L}}\right)=\Phi\left(y_k\frac{f\left(v_k\right)-f\left(u_k\right)}{\sqrt{2}\sigma}\right), \tag{1}$$

with $\mathbf{f}_k=\left[f\left(u_k\right),f\left(v_k\right)\right]$, $\Phi(x)$ defines a cumulative Gaussian (with zero mean and unity variance) and $\boldsymbol{\theta}_{\mathcal{L}}=\{\sigma\}$. This is in turn the well known Probit classification model, where the argument is the difference between two latent variables (functional values) and not just a single latent variable. This in effect implies that the $f(\cdot)$ encodes an internal, but latent preference function which can be elicited by pairwise comparisons via the likelihood model in Eq.(1). This idea was already considered by [4], but recently suggested in a Gaussian Process context by [5].

2.1. Gaussian Process Prior

The real question remains, namely how f is modelled. We will follow the principle suggested by [5] in which f is considered an abstract function and we can in turn place a prior distribution over it. A natural prior is a Gaussian Process (GP) defined as "a collection of random variables, any finite number of which have a (consistent) joint Gaussian distribution" [6]. Following [6] we denote a function drawn from a GP as $f\left(x\right)\sim\mathcal{GP}\left(\mathbf{0},k(\cdot,\cdot)_{\boldsymbol{\theta}_c}\right)$ with a zero mean function, and $k(\cdot,\cdot)_{\boldsymbol{\theta}_c}$ referring to the covariance function with hyperparameters $\boldsymbol{\theta}_c$, which defines the covariance between the random variables as a function of the inputs $\mathcal{X}$. The consequence of this formulation is that the GP can be considered a distribution over functions, i.e., $p\left(\mathbf{f}|\mathcal{X},\boldsymbol{\theta}_c\right)$, with hyper-parameters $\boldsymbol{\theta}_c$ and $\mathbf{f}=[f(x_1),f(x_2),...,f(x_n)]^T$.

In a Bayesian setting we can directly place the GP as a prior on the function defining the likelihood. This leads us directly to a formulation given Bayes relation with $\boldsymbol{\theta}=\{\boldsymbol{\theta}_{\mathcal{L}},\boldsymbol{\theta}_c\}$

$$p\left(\mathbf{f}|\mathcal{Y},\mathcal{X},\boldsymbol{\theta}\right)=\frac{p\left(\mathcal{Y}|\mathbf{f},\boldsymbol{\theta}_{\mathcal{L}}\right)p(\mathbf{f}|\mathcal{X},\boldsymbol{\theta}_c)}{p\left(\mathcal{Y}|\boldsymbol{\theta},\mathcal{X}\right)}. \tag{2}$$

The prior $p(\mathbf{f}|\mathcal{X},\boldsymbol{\theta}_c)$ is given by the GP and the likelihood $p\left(\mathcal{Y}|\mathbf{f},\boldsymbol{\theta}_{\mathcal{L}}\right)$ is the two likelihood defined previously, with the usual assumption that the likelihood factorizes, i.e., $p\left(\mathcal{Y}|\mathbf{f},\boldsymbol{\theta}_{\mathcal{L}}\right)=\prod_{k=1:m}p\left(y_k|f(u_k),f(v_k),\boldsymbol{\theta}_{\mathcal{L}}\right)$

The posterior of interest, $p\left(\mathbf{f}|\mathcal{Y},\mathcal{X},\boldsymbol{\theta}\right)$, is defined when equipped with the likelihood and the prior, but it is unfortunately not of any known analytical form, thus we rely on the Laplace approximation.

2.2. Inference & Hyperparameters

We apply the Laplace approximation and approximate the posterior by a multivariate Gaussian distribution, such that $p\left(\mathbf{f}|\mathcal{Y}\right)\approx\mathcal{N}(\mathbf{f}|\hat{\mathbf{f}},\mathbf{A}^{-1})$. Where $\hat{\mathbf{f}}$ is the mode of the posterior and $\mathbf{A}$ is the Hessian of the negative log-likelihood at the mode.

The mode is found as $\hat{\mathbf{f}}=\arg\max_{\mathbf{f}}\,p\left(\mathcal{Y}|\mathbf{f}\right)p\left(\mathbf{f}\right)$. We solve the problem by considering the unnormalized log-posterior and the resulting cost function which is to be maximized, is given by

$$\begin{aligned}\psi\left(\mathbf{f}|\mathcal{Y},\mathcal{X},\boldsymbol{\theta}\right)=&\log p\left(\mathcal{Y}|\mathbf{f},\mathcal{X},\boldsymbol{\theta}_{\mathcal{L}}\right)-\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f}\\&-\frac{1}{2}\log|\mathbf{K}|-\frac{N}{2}\log 2\pi.\end{aligned} \tag{3}$$

where $\mathbf{K}_{i,j}=k(x_i,x_j)_{\boldsymbol{\theta}_c}$. We use a damped Newton method with soft linesearch to maximize Eq. (3). In our case the basic damped Newton step (with adaptive damping factor λ) can be calculated without inversion of the Hessian (see [7])

$$\begin{aligned}\mathbf{f}^{new}=&\left(\mathbf{K}^{-1}+\mathbf{W}-\lambda\mathbf{I}\right)^{-1}\\&\cdot\left[\left(\mathbf{W}-\lambda\mathbf{I}\right)-\mathbf{f}+\nabla\log p(\mathcal{Y}|\mathbf{f},\mathcal{X},\boldsymbol{\theta}_{\mathcal{L}})\right],\end{aligned} \tag{4}$$

Using the notation $\nabla\nabla_{i,j}=\frac{\partial^2}{\partial f(x_i)\partial f(x_j)}$ we apply the definition $\mathbf{W}_{i,j}=-\sum_k\nabla\nabla_{i,j}\log p(y_k|\mathbf{f}_k,\boldsymbol{\theta}_{\mathcal{L}})$. We note that the term $\nabla\nabla_{i,j}\log p(y_k|\mathbf{f}_k,\boldsymbol{\theta}_{\mathcal{L}})$ is only nonzero when both x_i and x_j occur as either v_k or u_k in $\mathbf{f}_k$. In contrast to standard binary GP classification, the negative Hessian, $\mathbf{W}$ is not diagonal, which makes the approximation slightly more involved. When converged, the resulting approximation is

$$p\left(\mathbf{f}|\mathcal{Y},\mathcal{X},\boldsymbol{\theta}\right)\approx\mathcal{N}\left(\mathbf{f}|\hat{\mathbf{f}},\left(\mathbf{W}+\mathbf{K}^{-1}\right)^{-1}\right). \tag{5}$$

We refer to [7] for a full derivation and for the required derivatives as first outlined in [5]. Parameters in the likelihood and covariance function, collected in θ, are found by evidence optimization using a standard BFGS method.

2.3. Predictions & Evaluations

Given the model, in essence defined by f, we wish to make predictions of the observed variable y for a pair of test inputs $r\in\mathcal{X}_t$ and $s\in\mathcal{X}_t$. We are especially interested in the discrete decision, i.e., whether r is preferred over s denoted by $r\succ s$, or vice versa. Omitting the conditioning on $\mathcal{X}$ and $\mathcal{X}_t$, we can write the joint prior distribution between $\mathbf{f}\sim p\left(\mathbf{f}|\mathcal{Y},\boldsymbol{\theta}\right)$ and the test variables $\mathbf{f}_t=\left[f\left(r\right),f\left(s\right)\right]^T$ as

$$\begin{bmatrix}\mathbf{f}\\\mathbf{f}_t\end{bmatrix}=\mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\\mathbf{0}\end{bmatrix},\begin{bmatrix}\mathbf{K}&\mathbf{k}_t\\\mathbf{k}_t^T&\mathbf{K}_t\end{bmatrix}\right), \tag{6}$$

where $\mathbf{k}_t$ is a matrix with elements $\mathbf{k}_{2,i}=k(s,x_i)_{\boldsymbol{\theta}_c}$ and $\mathbf{k}_{1,i}=k(r,x_i)_{\boldsymbol{\theta}_c}$ with x_i being a training input. The conditional $p\left(\mathbf{f}_t|\mathbf{f}\right)$ is obviously Gaussian as well and can be obtained directly from Eq. (6). The predictive distribution is given as $p\left(\mathbf{f}_t|\mathcal{Y},\boldsymbol{\theta}\right)=\int p\left(\mathbf{f}_t|\mathbf{f}\right)p\left(\mathbf{f}|\mathcal{Y},\boldsymbol{\theta}\right)d\mathbf{f}$. With the posterior approximated with the Gaussian from the Laplace approximation, then $p\left(\mathbf{f}_t|\mathcal{Y},\boldsymbol{\theta}\right)$ will be Gaussian too and is given as $\mathcal{N}\left(\mathbf{f}_t|\mu^*,\mathbf{K}^*\right)$ with $\mu^*=[\mu_r^*,\mu_s^*]^T=\mathbf{k}_t\mathbf{K}^{-1}\hat{\mathbf{f}}$ and

$$\mathbf{K}^*=\begin{bmatrix}\mathbf{K}_{rr}^*&\mathbf{K}_{rs}^*\\\mathbf{K}_{sr}^*&\mathbf{K}_{ss}^*\end{bmatrix}=\mathbf{K}_t-\mathbf{k}_t^T\left(\mathbf{I}+\mathbf{W}\mathbf{K}\right)\mathbf{k}_t,$$

where $\hat{\mathbf{f}}$ and $\mathbf{W}$ are obtained from Eq. (5). With the predictive distribution for $\mathbf{f}_t$, the final prediction of the observed variable is available from

$$p\left(y_t|\mathcal{Y},\boldsymbol{\theta}\right) = \int p\left(y_t|\mathbf{f}_t,\boldsymbol{\theta}_{\mathcal{L}}\right) p\left(\mathbf{f}_t|\mathcal{Y},\boldsymbol{\theta}\right) d\mathbf{f}_t \tag{7}$$

If the likelihood is an odd function, as in our case, the binary preference decision between r and s can be made directly from $p\left(\mathbf{f}_t|\mathcal{Y}\right)$.

If $p\left(\mathbf{f}_t|\mathcal{Y},\boldsymbol{\theta}\right)$ is Gaussian and we consider the Probit likelihood, the integral in Eq. (7) can be evaluated in closed form as a modified Probit function given by [5]

$$P\left(r \succ s|\mathcal{Y}\right) = \Phi\left(\left(\mu_r^* - \mu_s^*\right)/\sigma^*\right) \tag{8}$$

with $\left(\sigma^*\right)^2 = 2\sigma^2 + \mathbf{K}_{rr}^* + \mathbf{K}_{ss}^* - \mathbf{K}_{rs}^* - \mathbf{K}_{sr}^*$

2.4. Kernels for Audio Preference

We suggest a general purpose covariance function for audio modeling tasks with GPs. It can easily integrate different modalities and meta-data types, such as audio features, tags, lyrics and subject features. The general covariance function is defined as

$$k\left(x,x'\right) = \left(\sum\nolimits_{i=1}^{N_a} k_i\left(x_a,{x_a}'\right)\right) k_u\left(x_u,{x_u}'\right) \tag{9}$$

where the first factor is the sum of all the N_a covariance functions defining the correlation structure of the audio part, x_a, of the complete instance, x. The second factor, or multi-task part, is the covariance function defining the covariance structure of the subject meta-data part, x_u. The practical evaluation is limited to the a individualized setting using only x_a, thus $k\left(x,x'\right) = k\left(x_a,{x_a}'\right)$, where we apply the probability product kernel formulation [8]. The probability product kernel is defined directly as an inner product, i.e., $k\left(x_a,{x_a}'\right) = \int \left[p\left(x_a\right) p\left({x_a}'\right)\right]^q dx$, where $p\left(x_a\right)$ is a density estimate of each audio track feature distribution. In this evaluation we fix $q = 1/2$, leading to the Hellinger divergence [8]. As custom in the audio community, see e.g. [9], we will resort to a (finite) Gaussian Mixture Model (GMM) in order to model the feature distribution. So $p(x)$ is in general given by $p\left(x\right) = \sum_{z=1}^{N_z} p\left(z\right) p\left(x|z\right)$, where $p\left(x|z\right) = \mathcal{N}\left(x|\mu_z,\sigma_z\right)$ is a standard Gaussian distribution. The kernel can be calculated in closed form [8] as.

$$k\left(p_a\left(x\right),p_a\left(x\right)\right) = \sum_z \sum_{z'} \left(p_a\left(z\right) p_{a'}\left(z'\right)\right)^q \tilde{k}\left(p\left(x|\theta_z\right),p\left(x|\theta_{z'}\right)\right) \tag{10}$$

where $\tilde{k}\left(p\left(x|\theta_z\right),p\left(x|\theta_{z'}\right)\right)$ is the probability product kernel between two single components, which is also available in closed form [8].

3. EXPERIMENT

In order to evaluate the model proposed in section in 2, we consider a small-scale dataset which is publicly available [3]. Specifically it consist of 10 test subjects, 30 audio tracks and 10 audio tracks per genre. The genres are Classical, Heavy Metal and Rock/Pop.

The experiment is based on a partial, full pairwise design, so that 155 out of the 420 combinations was evaluated by each of the 10 subjects. We extract standard audio features from the audio tracks, namely MFCCs (26 incl. delta coefficients). A GMM was fitted to each track distribution with a fixed model complexity of $N_z = 2$ and each components restricted to a diagonal covariance structure. Parameters where fitted using a standard maximum likelihood based EM algorithm using K-means initialization.

The experiment itself was conducted using a Matlab interface in a 2-Alternative-Forced-Choice setup inline with the model. The interface allowed subjects to listen to the two presented tracks as many times they wanted before making a choice between them. A questionnaire gathered subject meta-data such as, age, musical training, context and a priori genre preference. This data is, however, not used in this individualized evaluation, but can easily be applied in the multi-task kernel suggested in Sec. 2.4.

In the evaluation we are primarily interested in two aspects. The first, and main result, is an estimate of the generalization error on new unseen tracks, e.g., relevant for recommendation purposes. In order to evaluate this, we make an extensive cross-validation using a 30-fold cross-validation in which each track (incl. all connected comparisons) is left out once; the model with $\sigma = 1$ is then trained on 10 random subsets of tracks for each training set size, which results in an estimated of the average test error. The resulting learning curve is shown in Fig. 1 with the box plot illustrating the distribution of the average subject performance. When considering $N_{\text{tracks}} = 29$ we obtain an average prediction performance of 74.2% , which is the main result in a typical (individual) recommendation scenario.

Secondly, we investigate how many pairwise comparisons the model requires in order to learn the individual preferences. This is evaluated using a 10-fold cross-validation over the comparisons which gives the learning curve in Fig. 2. We notice that on average we only require approximately 40% or 56 comparisons in order to reach the 25% level, corresponding to approximately two comparisons per track.

4. DISCUSSION & CONCLUSION

We have outlined a pairwise regression model based on Gaussian Process priors for modeling and predicting the pairwise preference of music. We proposed an appropriate covariance structure suitable for audio features (such as MFCCs) based on generative models of audio features. The general version

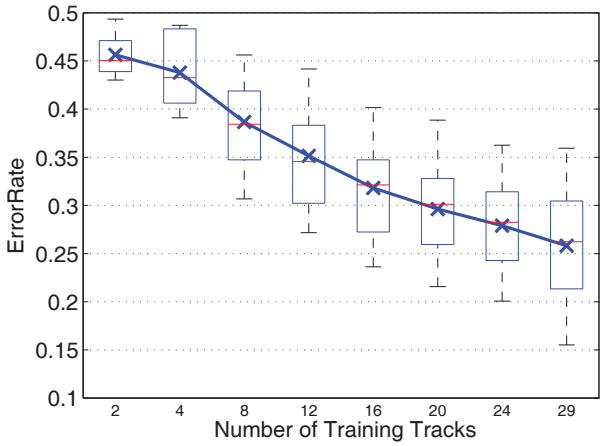

Fig. 1. Mean learning curve (blue line) and box plot over subjects. Leave-one-(track)-out test error as a function of the number of tracks in the training set. Thus, there can maximum be 29 tracks in the training set to predict the preference between the left out track and the rest. The baseline is 0.5 corresponding to random guessing.

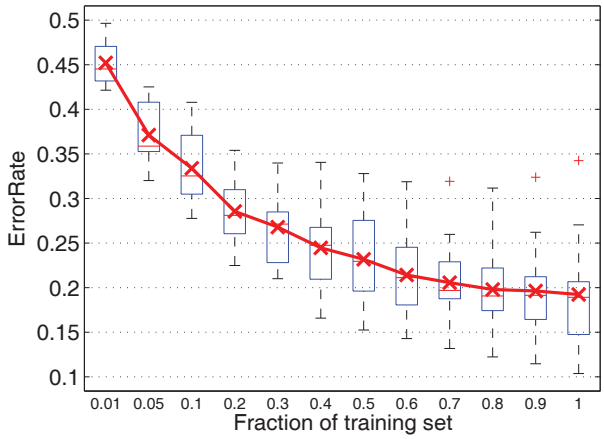

Fig. 2. Mean learning curve (red line) and box plot over the subjects mean performance. Test error rate as a function of the number of pairwise comparisons in the the training set. Notice that a fraction of one corresponds to $(155 \cdot 90\%)/420 \sim 33.2\%$ of all possible pairwise experiments.

of the covariance function allows for multi-task scenarios and feature integration. We evaluated the setup in a individual scenario in which we showed a 74% average accuracy. This indicates that there might very well be a promising upper bound on the number of required pairwise comparisons in this music setting, in effect implying that the specified correlation structure makes sense. This will ensure that the required number of pairwise comparisons does not scale quadratically when including more tracks.

We furthermore observe a difference among the different subjects indicating that some subjects may have a very consistent preference, possibly aligning well with the applied covariance function, while others seem very difficult to predict (observed as outliers in the box plot). We speculate that the pairwise approach to music preference is only possible for certain groups of subjects and/or in special contexts, which is to be investigated in future research.

The current model is intended for modeling personal preferences over a small/medium size dataset. For large datasets with millions of tracks, we see sparse techniques using pseudo-inputs and sequential selection as a powerful combination to scale the model and only use informative comparisons. Furthermore, a direct comparison between classic collaborative filtering with absolute ratings is obvious when a suitable dataset supporting is available.

In conclusion we have proposed a novel rating and modeling paradigm for eliciting music preference using pairwise comparisons. We conducted a preliminary evaluation of the performance on a small dataset and find the results promising for robust elicitation of music and audio preference in general.

5. REFERENCES

[1] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H.G. Okuno, "An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 435–447, 2008.

[2] R.D. Bock and J.V. Jones, "The measurement and prediction of judgment and choice," 1968.

[3] B. S. Jensen, J. S. Gallego, and J. Larsen, "A Predictive Model of Music Preference using Pairwise Comparisons - Supporting Material and Dataset," www.imm.dtu.dk/pubdb/p.php?6143.

[4] L L Thurstone, "A law of comparative judgement," *Psychological Review*, vol. 34, 1927.

[5] W. Chu and Z. Ghahramani, "Preference learning with Gaussian processes," *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, pp. 137–144, 2005.

[6] C.E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[7] B. S. Jensen and J. B. Nielsen, "Pairwise Judgements and Absolute Ratings with Gaussian Process priors," Techinal Report, DTU Informatics, September 2011.

[8] T. Jebara and A. Howard, "Probability Product Kernels," *Journal of Machine Learning Research*, vol. 5, pp. 819–844, 2004.

[9] A. Meng and J. Shawe-Taylor, "An investigation of feature models for music genre classification using the support vector classifier," in *International Conference on Music Information Retrieval*, 2005, pp. 604–609.