Modelling Argument Convincingness using Bayesian Preference Learning

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

- Motivation: finding well-written and convincing arguments to enable better decision-making, analyse public opinion
- Need tools that can learn to identify convincing arguments in new topics and domains from small amounts of data
- Pairwise preferences provide a way for multiple people to communicate the relative convincingness of arguments
- Implicit preferences can be elicited from user actions, such as selecting a document from a list given its summary to read in more detail
- We propose a scalable approach to preference learning using stochastic variational inference to learn a Gaussian process model
- We show how Bayesian approach to preference learning can be applied to high-dimensional text problems
- The approach outperforms SVMs and neural networks when data is sparse or noisy
- ...and achieves state-of-the art performance for predicting argument convincingness given sufficient data
- We show how our Bayesian approach enables more effective active learning, thereby reducing the amount of data required to model argument convincingness in new domains
- We develop a gradient-based approach to automatic relevance determination that enables us to identify the most informative features for predicting arguments when the input space contains tens of thousands of dimensions
- Our analysis shows that word embeddings and linguistic features contain complementary information that boost performance when both types of features are used.

1 Introduction

Motivations for modelling argument convincingness:

- Learning about a controversial topic often requires reading large amounts of text, often with much duplicate information, in order to understand different points of view
- Points of view on controversial topics are often presented as arguments for or against a particular position
- Finding well-written arguments could allow better understanding of why people hold particular opinions
- Identifying arguments that are considered convincing to a particular group of people helps understand who holds which point of view
- Tools that identify convincing arguments could therefore assist in making better decisions and analysing public opinion

Pairwise preferences provide a way for multiple people to communicate the relative convincingness of arguments:

- Implicit preferences can be elicited from user actions, such as selecting a document from a list given its summary to read in more detail
- Explicit preferences can be easier to provide than ratings or classifications
- Since preferences are relative, they avoid calibration issues caused by multiple people using ratings (e.g. Mary rarely awards 5 stars to movies she likes, John frequently awards 5 stars)

Previous work on predicting convincingness:

- (Habernal and Gurevych, 2016) shows that it is possible to predict convincingness given linguistic features or word embeddings
- They show that models can be transferred reasonably well between topics in online debates (78% accuracy)
- New types of text, new domains, and new users with different preferences means we may face situations in practice where models trained on existing corpora are less effective, but data for the new task is limited (sparse)
- Use two different pipelines consisting of multiple steps: combining crowdsourced data, removing inconsistencies, classification; combining crowdsourced data, ranking using PageRank, regression
- In low-data situations, these approaches may underperform, since model uncertainty is not accounted for between each stage, nor in the final predictions, and errors propagate along the pipeline
- Training data may also contain errors, which would be propagated through the pipeline (this was avoided in previous work by combining labels from multiple crowdworkers; we should be able to handle the case where this is not possible).
- Feature space becomes very large when working with textual features – can we narrow it down automatically to improve scalability and improve performance?

In contrast to previous work, we propose to directly model the relationship between preferences for convincing arguments and textual features (incl. word embeddings) to predict which arguments a person will prefer:

- Bayesian approaches have been shown to successfully handle situations with small amounts of data, allow transfer of background knowledge through priors, and provide a good basis for actively selecting data
- Confidence estimates from Bayesian models account for sparsity and noise in data, as well as uncertainty in the model. This means they do not make overly-confident predictions when training data is small (they know when they don't know).
- Bayesian preference learning methods have been proposed but scalable implementations were not developed and models have not been applied to text with large numbers of features
- We address the limitations above by adapting Bayesian preference learning approach to argumentation
- Introduce stochastic variational inference (SVI) to train the model on large numbers of preferences and documents
- Develop gradient-based ARD to identify relevant text features

We demonstrate how our preference learning approach can be used to model convincingness of arguments:

- Evaluate the performance of our approach against state-of-the art deep learning and SVM methods
- Show that Bayesian Gaussian process (GP) models are applicable to performing preference learning over text (existing evaluation of GPs for text is very limited, although they have been used extensively with great success in domains such as Physics, finance, Biology. This is possibly because GPs were seen as more difficult to implement and could not be scaled up until recent advances such as SVI)

- Evaluate the ability of each method to handle noisy and sparse data, showing improved performance using our method in the presence of noise and data sparsity

- Analyse the features that are most informative when determining convincingness, providing insight into what makes a convincing argument

Structure:

- Review related work in more detail: on argumentation and persuasion; examples of Bayesian methods for NLP

- Method background: preference learning with GPs; scalability of GPs; related work on SVI

- Method part 1: proposed approach to scalable Bayesian preference learning

- Method part 2: automatic relevance determination: background; our proposed gradient-based method

- Experiments 1: comparison with state-of-the art on predicting preference in online debates; error analysis focussing on differences between each method

- Experiments 2: study of performance with noisy and sparse data; error analysis highlighting differences in each method

- Experiments 3: analysis of informative features for argumentation

- Conclusions and future work

2 Related Work

Related work on argumentation and persuasion. Related work on finding reasons for argument convincingness (cite Ivan). Related work on choosing the best argument in sequence (Rosenfeld and Kraus, 2016)(Monteserin and Amandi, 2013).

Preference learning from pairwise preferences is effective because it removes the need for humans to provide scores or classifications and allows them to make relevance judgements, which have been shown to be easier for human annotators in many cases(Brochu et al., 2008). Pairwise comparisons also occur in implicit feedback, for example, when a user chooses to click on link from a list of several. They are therefore a useful tool for practical learning from end users. However, the pairwise comparisons we observe may not be a perfect representation of their preferences as they may contain noise, leading to inconsistencies where items cannot be ranked in such a way that the ranking agrees with all the observed comparisons. Bayesian approaches are suited to handling these problems of data sparsity, noise and bias, The Gaussian process (GP) preference learning approach of (Chu and Ghahramani, 2005) resolves inconsistencies between preferences and provides a way to predict rankings or preferences for items for which we have not observed any pairwise comparisons based on the item's features. This model assumes that preferences are noisy, i.e. contain some erroneous labels. particularly as the modular nature of inference algorithms such as Gibb's sampling and variational approximation is suited to extending the model to handle different types of feedback that give indications of some underlying preferences.

The GP methods require $\mathcal{O}(P_n)$ steps, where P_n is the number of pairs for user n. We use SVI to address scalability in a variational Bayesian framework. The modular nature of VB allows us to take advantage of models for feedback of different types where the input values for each type of feedback do not directly correspond (e.g. explicit user ratings and number of clicks may have different values). By using SVI, we provide a formal way to deal with scalability that comes with guarantees(Hoffman et al., 2013). We also estimate the output scale of the GPs, the latent factors, and item bias as part of the variational approximation.

In most scenarios where we wish to make predictions about arguments, there is a very large number of input variables potentially associated with each argument in the dataset, but very sparse observations of these variables. To illustrate this, consider a simple bag-of-words representation of the argument text, and a set of click-data recording which actions each user took when presented with a choice between different pieces of text. Given a large vocabulary, the words present in an argument will be a very small subset of possible words. Users will likely see a subset of texts and the recorded choices will be a much smaller subset of the possible combinations of texts. To make predictions about unobserved preferences when presented with a new text with sparse data, we require an abstraction from the raw input data, and thus seek a way to embed the texts into a space where texts with similar properties are placed close together. In the case of arguments, one property that may determine whether texts should be close together is that they have similar levels of convincingness to similar types of people, in similar contexts. Our proposal therefore produces a form of argument embedding, driven by convincingness. A similar approach to learning latent features, VBMDS, is proposed by (Soh, 2016) for learning embeddings using approximate Bayesian techniques, but does not use the embeddings for preference learning to find separate person and item embeddings and does not apply this to NLP problems. Their proposal does, however, show how to combine points with and without side information – our input features – to make predictions about low-dimensional embeddings for unseen data. The kernelized probabilistic matrix factorization (KPMF) (Zhou et al., 2012) proposes a similar approach to VBMDS using GP priors over latent dimensions, but with a simpler MAP inference scheme, and different likelihood and distance functions.

An important aspect of convincingness is the context in which an argument is made, particularly as part of a dialogue. The sequence of arguments strongly correlates with their ability to change the audience's opinions (Tan et al., 2016), as does the prior stance of the audience(Lukin et al., 2017). In our approach, this context can be represented as input variables that affect the item and person embeddings, where the variables encapsulate the previously seen arguments. While out-of-scope of the present investigation, future work may investigate the best way to determine novelty of an argument given a small number of variables representing previously seen arguments. Another related avenue of improvement is to consider the structure of arguments to select argument components – it may be important to consider not just novelty, but whether claims have sufficient support and premises are clearly linked to the claims they support or attack. Embedding this structure may require complex graph structures of claims and premises to be represented as short vectors, and may therefore be a topic of future study.

Using textual data as inputs to a Gaussian process presents some challenge. Firstly, large vocabulary sizes lead to a large number of dimensions, which present problems when performing automatic relevance determination (ARD) to optimize the model to the most important features for predicting the target variables. Secondly, kernel functions are not typically learned or adapted to the data, which means that points with different features that commonly co-occur are not assigned high covariance, whereas it would be desirable to learn that commonly co-occurring features indicate similar target values. A solution to this problem is to represent input features such as words using vectors of continuous values, i.e. word embeddings. This approach was proposed for performing GP regression on text data by (Yoshikawa et al., 2015), who showed how to learn the word embeddings and map document distributions over word embeddings to points in a reproducing kernel Hilbert space. This approach can be used to obtain document embeddings from word embeddings.

The latent features allow us to interpolate between items and people in a low-dimensional embedding space. A key question in this latent feature approach is how to model the deviation of individual preferences from that predicted by latent features common to multiple people (item deviations can be modelled through an item mean function). This deviation occurs when there is still entropy in a user's preferences given the latent features because the latent features only describe patterns that are common to multiple users. A simple approach is to allow additional noise with uniform variance at each data point, so that all preference patterns are represented by the latent feature vectors of items and people. However, any individual preference patterns particular to one user must then be represented by additional latent features that are not activated for any other users. An alternative is to use a personal model of preference deviation for each person. Given the input features of the items and any state variables relating to the person, this model can capture correlations in the deviation for different items for the same person. Both the latent person features and the individual noise model can also include any input features of the person that change over time, e.g. representing their state and the arguments they have previously seen. This individual noise model allows us to differentiate preference patterns that are specific to one user, when the input features may not otherwise be sufficient to distinguish these users.

3 Identifying Common Patterns of Convincingness

4 Bayesian Preference Learning Model

The model introduced in (Houlsby et al., 2012) combines preference learning with matrix factorisation to identify latent features of items and users that affect their preferences. This allows for a collaborative filtering effect, whereby users with similar preferences on a set of observed items are assumed to have similar preferences for other items with similar features. This allows us to make better predictions about the unobserved preferences of a given user when we have seen preferences of a similar user.

The method presented in (Houlsby et al., 2012) uses a combination of expectation propagation (EP) and variational Bayes (VB). Since the inference steps require inverting a covariance matrix, this method scales with $\mathcal{O}(N^3)$ and is therefore impractical for large datasets. For our modified version of this method, we improve scalability by using stochastic variational inference to infer the complete model. The variational approximation to the posterior is given by...

The variational inference algorithm maximises a lower bound on the log marginal likelihood:

$$
\begin{aligned}
\mathcal{L} = \sum_{i=1}^{N} \mathbb{E}[\log p(t_i | x_{i,1}, x_{i,2}, \boldsymbol{f})] + \\
\sum_{u=1}^{U} \mathbb{E}\left[\log \frac{p(\boldsymbol{f}_u | \boldsymbol{w}\boldsymbol{y}_u, \boldsymbol{K}_{f,u}/s_{f,u})}{q(\boldsymbol{f}_u)}\right] + \\
\sum_{c=1}^{C} \mathbb{E}\left[\log \frac{p(\boldsymbol{w}_c | \boldsymbol{0}, \boldsymbol{K}_w/s_{w,c})}{q(\boldsymbol{w}_c)}\right] + \\
\sum_{c=1}^{C} \mathbb{E}\left[\log \frac{p(\boldsymbol{y}_c | \boldsymbol{0}, \boldsymbol{K}_y/s_{y,c})}{q(\boldsymbol{y}_c)}\right] + \\
\mathbb{E}\left[\log \frac{p(\boldsymbol{t} | \boldsymbol{\mu}, \boldsymbol{K}_t/s_t)}{q(\boldsymbol{t})}\right] + \\
\sum_{u=1}^{U} \mathbb{E}\left[\log \frac{p(s_{f,u} | a_{f,u}, b_{f,u})}{q(s_{f,u})}\right] + \\
\sum_{d=1}^{D} \mathbb{E}\left[\log \frac{p(s_{w,d} | a_{w,d}, b_{w,d})}{q(s_{w,d})}\right] + \\
\sum_{d=1}^{D} \mathbb{E}\left[\log \frac{p(s_{y,d} | a_{y,d}, b_{y,d})}{q(s_{y,d})}\right] \qquad (1)
\end{aligned}
$$

where t_i is the preference label for the ith pair,

To perform feature selection with large numbers of features, we introduce an automatic relevance determination (ARD) approach that uses the gradient of the lower bound on the log marginal likelihood to optimise the kernel length-scales using the L-BFGS-

B method(**?**). The gradient is given by:

$$\nabla\mathcal{L} = \left[\frac{\partial\mathcal{L}}{\partial l_{w,1}}, ..., \frac{\partial\mathcal{L}}{\partial l_{w,D_w}}, \frac{\partial\mathcal{L}}{\partial l_{y,1}}, ..., \frac{\partial\mathcal{L}}{\partial l_{y,D_y}}\right], \quad (2)$$

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial l_{w,d}} = \frac{\partial}{\partial l_{w,d}} \sum_{u=1}^{U} \mathbb{E}\left[\log\frac{p(\boldsymbol{f}_u|\boldsymbol{w}\boldsymbol{y}_u, \boldsymbol{K}_{f,u}/s_{f,u})}{q(\boldsymbol{f}_u)}\right] + \\ \sum_{c=1}^{C}\mathbb{E}\left[\log\frac{p(\boldsymbol{w}_c|\boldsymbol{0}, \boldsymbol{K}_w/s_{w,c})}{q(\boldsymbol{w}_c)}\right] - \\ \sum_{u=1}^{U}\mathbb{E}\left[\log q(s_{f,u})\right] - \sum_{d=1}^{D}\mathbb{E}\left[\log q(s_{w,d})\right] + \\ = 0.5(\hat{f}_u - wy_u)^T\boldsymbol{K}_{f,u}^{-1}\frac{\partial\boldsymbol{K}}{\partial\log l_{w,d}}\hat{s}_{f,u}\boldsymbol{K}_{f,u}^{-1}(\hat{f}_u - wy_u) \\ -0.5\mathrm{tr}\left((\boldsymbol{K}_{f,u}^{-1} - \frac{\boldsymbol{C}^{-1}}{\hat{s}_{f,u}})\frac{\partial\boldsymbol{K}_{f,u}}{\partial\log l_{w,d}}\right) \\ \frac{\partial\mathcal{L}}{\partial l_{y,d}} = \quad (3)\end{aligned}$$

$$(4)$$

where $l_{w,d}$ is a length-scale used for all the GPs over item features. The implicit terms are zero when the VB algorithm has converged.

5 Experiments

The first dataset contains a number of pairwise convincingness preference labels for a set of arguments from a crowd of workers. Each label is associated with two arguments and expresses whether a worker in the crowd found the first argument most convincing, the second argument, or had no preference.

The task is to train the models then predict the preference labels for held-out data. Each method can be assessed in terms of classification accuracy, since the labels have three possible values.

5.1 Hypotheses – needs to be reconciled with the above paragraph

Prior work on convincingness:

- (Habernal and Gurevych, 2016) shows how to predict convincingness of arguments by training a NN from crowdsourced annotations.
- (Lukin et al., 2017) shows that persuasion is correlated with personality traits.

We build on this to show...

- How we can rank arguments in terms of convincingness using preference learning, and resolve conflicts.
- Which input features are informative and how we can learn this using a Bayesian approach
- Uncertainty is modelled correctly in the Bayesian approach so that (a) we can filter out the uncertain decisions to improve accuracy; (b) the brier score/cross entropy error is lower;
- Uncertainty also means that simple active learning works better
- Bayesian approach means accuracy is better with sparse data, e.g. in the cold-start situation, esp. if we can use (a) and (c) to avoid acting on uncertain labels.

This is all useful because we can use the approach to determine which features are worth obtaining, make predictions when data is sparse, and obtain data from users efficiently.

Experiment story:

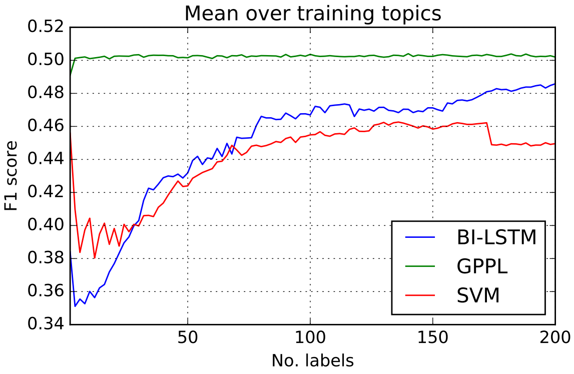
1. Compare GP preference learning (GPPL) to SVM using linguistic features. Result: better performance on ranking tasks.
2. Compare GPPL to Bi-LSTM. Result: GPPL is not as effective when using mean embeddings, but GPPL with linguistic features still outperforms Bi-LSTM on all metrics.
3. Do the embeddings and linguistic features provide complementary information? Run GPPL with both sets of features. Result: a small improvement on classification tasks, and larger improvement on ranking suggests that the feature sets contain complementary information. The computational cost (of kernel computation, which dominates the overall cost in these experiments) grows linearly with number of features.
4. How much does performance drop when we allow conflicts in the preference graph? Compare GPPL, SVM, Bi-LSTM. Result: all methods have a small drop in performance, but GPPL is affected least.

5. How much does performance drop if we use noisy crowdsourced labels, rather than the gold standard produced by MACE? Compare GPPL, SVM, Bi-LSTM. Result: as in previous experiment, GPPL copes best with the added noise.

6. Does GPPL improve ranking performance because of the way it resolves conflicting preferences, or the way it makes predictions? Compare against feature-free preference learning used to train an SVM regression model (PL+SVR). Result: GPPL improves classification slightly on noise-free dataset, and improves more over PL+SVR on ranking and noisy classification tasks. GPPL resolves conflicts at the same time as predicting scores using similarities between arguments in feature space, so therefore has more information to resolve erroneous labels than the feature-free PL.

7. GP methods are heavily affected by choice of kernel. The standard approach (using a product of kernels for each feature) is equivalent to taking the euclidean distance between points in feature space. These distances become very large when we have a large number of features. Each point needs to be close in all dimensions in order to be close overall. An alternative to this product kernel is to use a sum kernel, which will result in points having strong covariance if some (rather than all) features are similar. This may be suitable for high-dimensional settings where some features may have missing values. Compare the GPPL approach with product and sum kernels. Result: *needs rerunning due to bug in sum kernel*

8. Besides the choice of kernel itself, another important set of hyperparameters of the GPPL are the hyperparameters for the prior over the output scale of the kernel. The output scale controls the noise of the preferences, and needs to be large enough to allow the posterior to deviate substantially from the prior mean given only a small number of observations at each point. We test three different plausible settings of this hyperparameter to determine how sensitive the results are. Result: heavily informative values do not work well; very noninformative values are effective, although we observe a small boost when using an intermediate setting. This may be due to the data sparsity; the intermediate setting puts more weight onto each individual observation.



5.2 Alternative Application – Notes

The core contribution is to do preference learning with sparse observations with text data. There may be several other problems related to NLP and argumentation, more specifically, that could benefit from this approach. Argument cloze task? The model can be adapted to classification problems, regression, or mixed observation types by applying a different likelihood. The core of the method is the abstraction of a latent function over items and people, dependent on latent features of items and people, with the ability to include side information. The paper could therefore apply the model to multiple NLP tasks, and be a methodology paper. This needs us to discuss and distinguish the class of problems and novelty of the method. What are the alternative methods, e.g. if we were to use this for classification? One could supply all item and person data to a neural network and train it in a semi-supervised manner? Sticking to the idea of preference learning or ranking, what other tasks could be handled in this way?

5.3 Active Learning

Acknowledgments

References

E. Brochu, N. de Freitas, and A. Ghosh. 2008. Active preference learning with discrete choice data. In *Advances in Neural Information Processing Systems 20*, pages 409–416. MIT Press.

Dataset	Dataset properties	Hypothesis	Methods
1. UKPConvArgStrict	MACE output with confidence $\geq 95\%$; Discard arguments marked as equally convincing; Discard conflicting preferences.	Bayesian method is competitive with previous methods at predicting clean preference pairs from a clean dataset.	GP Preference learning + linguistic features + embeddings.
2. UKPConvArgAll	MACE output with confidence $\geq 95\%$; No further filtering.	Bayesian method is competitive with previous methods if filtering step is removed.	GP Preference learning + linguistic features + embeddings.
3. UKPConvArgRank	MACE output with confidence $\geq 95\%$; Equal arguments included; PageRank used to rank arguments.	Bayesian method is competitive with previous methods at ranking arguments and can perform ranking given pairs rather than rank scores.	GP regression with preference learning output + linguistic features + embeddings (trained on rank scores); GP Preference learning + linguistic features + embeddings (trained on pairs).
4. UKPConvArgCrowd	No filtering, all pairs from original workers are provided.	Bayesian method can predict argument pairs for individual annotators with competitive performance to rival methods on clean, combined data; There are patterns of common agreement/disagreement among workers.	Bayesian Preference Components + linguistic features + embeddings (pair prediction).
5. UKPConvArgCrowdR	No filtering, all pairs from original workers are provided; PageRank used to produce gold-standard ranking.	Bayesian model can predict individual argument rankings; Significant differences between individual rankings and gold-standard ranking.	Bayesian Preference Components + linguistic features + embeddings (ranking output).

Table 1: The datasets and hypotheses in each experiment.

Wei Chu and Zoubin Ghahramani. 2005. Preference learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 137–144. ACM.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Matthew D Hoffman, David M Blei, Chong Wang, and John William Paisley. 2013. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347.

Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Jose M Hernández-Lobato. 2012. Collaborative gaussian processes for preference learning. In *Advances in Neural Information Processing Systems*, pages 2096–2104.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *15th European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics.

Ariel Monteserin and Analía Amandi. 2013. A reinforcement learning approach to improve the argument selection effectiveness in argumentation-based negotiation. *Expert Systems with Applications*, 40(6):2182–2188.

Ariel Rosenfeld and Sarit Kraus. 2016. Providing arguments in discussions on the basis of the prediction of human argumentative behavior. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(4):30.

Harold Soh. 2016. Distance-preserving probabilistic embeddings with side information: Variational bayesian multidimensional scaling gaussian process. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, pages 613–624. International World Wide Web Conferences Steering Committee.

Yuya Yoshikawa, Tomoharu Iwata, and Hiroshi Sawada. 2015. Non-linear regression for bag-of-words data via gaussian process latent variable set model. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 3129–3135. AAAI Press.

Tinghui Zhou, Hanhuai Shan, Arindam Banerjee, and Guillermo Sapiro. 2012. Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In *Proceedings of the 2012 SIAM international Conference on Data mining*, pages 403–414. SIAM.