Less Naïve Bayes for Recommender Systems

Anonymous Author(s)

Abstract

Recommender systems are an import part of modern e-commerce as they provide a means to find content of interest to users specific to their preferences. The Naive Bayes Classifier continues to see application as a component of recommender systems on the account of its efficiency and simplicity. However, the conditional independence assumption that Naive Bayes hinges on is questionable in the recommendation setting. An approach similar in design to Naive Bayes is present that seeks to dull this conditional assumption by blending the predictions that result from the different possible independence assumptions that can be made. Also considered is the importance of an intelligent prior on pair wise joint probabilities, the effect of including prediction for the opposite of a user’s preferences, and the application of classification from both the user and item dimension perspectives. Comparison is performed on the recommendation task of classifying content as being liked or disliked by a user and measuring performance via precision and recall.

1 Introduction

Recommender systems (RS) serve an increasingly important role in the aid of discovery of personalized content. Although search engines can help with the filtering of content it places the burden of what to query for on the user. In contrast to passively returning requested content RS’s seek to actively push content that is expect to be of interest elevating the need to know what query to ask. This is achieved by identifying users’ personal preferences from feedback provided on the content in the system. Collaborative filtering RS’s avoid needing domain specific knowledge by base recommendations entirely on preference dependencies that can be extracted from user feedback.

Of important for a RS to be effective is that when it presents content (movies, books, etc.) it has reason to be confident that there is a significant chance that this content will interest the user it is present to. In competition is the desire to identify novel content that may be liked by the user but the user may not be aware of. In the framing of recommendation as a classification problem this trade off is that of aiming to achieve high recall of relevant content (avoid missing content that would be liked) while maintaining high precision (avoiding recommending content that would not be liked).

For the task of classification when dealing with significant amounts of data, as is the case with the feedback handled by RS’s, the Naive Bayes Classifier is a commonly applied due in part to its scalability and simplicity. However, to achieve this the Naive Bayes model makes the assumption that the preferences being used to predict a target preference are conditionally independent given the target preference. This assumption of conditional independence is questionable in the recommendation setting. Unlike in other domains where the class that is being predicted is distinguished in someway (ex. spam or sentiment classification) in the
recommendation setting preferences are being predicted according to other preferences. As such, there is no comparable asymmetry between the class being predicted and the attributes being used in the prediction. The intuitively weaker significance of the class of “will this content be liked or not?” being predicted leaves in doubt whether it is reasonable to assert that when this one preference is known all of the other content preferences used in its prediction become independent of each other.

Presented in this paper is a method that takes a similar approach as Naive Bayes by making conditional independence assumptions so to approximately decompose probabilities involving many variables down to a combination of probabilities involving one or two variables. Instead of assuming conditional independence from only the perspective of the class being predicted the aim is to consider the predictions that result from the possible combinations of conditional independence assumptions that result in the simplification of probability down to the set of known probabilities. By blending the predictions that result from these different possible independence assumptions the aim is to weaken the significance of any one the assumptions so to make the violation of anyone of them less detrimental to prediction performance.

When computing the pairwise probabilities, used in both Naive Bayes and the proposed approach, from the feedback data the use of the single variable probabilities to placing a priors on pair wise probabilities is compared against a uniform prior. Also considered is whether any improvement to predictions can be attained by predicting using the opposite of user’s preferences and avoiding recommending content that would be recommended to the user’s opposite. A final natural extension is instead of performing only prediction according to dependencies between items the same methods can be applied considering dependencies on the user dimension.

The rest of this paper is organized as follows: background and further motivation is presented in Section 2. The proposed approach is explained in Section 3. In Section 4 the additional enhancements considered are described in more detail. The experimental evaluation performed is outlined and result shown in Section 5. A discussion of the experimental findings is in Section 6. Finally, the paper is concluded with the highlights as well as remaining open considerations.

2 Background and Motivation

For the task of classification the probabilistic framework provided by the Bayesian classifier can be applied. The problem is given an example with $N$ attributes ($A_1, A_2, ... A_N$) the aim is to predict the class $C_k$ by way of finding the $C_k$ that maximizes $P(C_k | A_1, A_2, ... A_N)$, or alternatively by Bayes’ theorem $P(C_k | A_1, A_2, ... A_N) \propto P(A_1, A_2, ... A_N | C_k) P(C_k)$. To estimate $P(A_1, A_2, ... A_N | C_k)$ the Naive Bayes Classifier makes the assumption that the attributes are independent given the class. This assumption gives $P(A_1, A_2, ... A_N | C_k) = P(A_1 | C_k) P(A_2 | C_k) ... P(A_N | C_k)$. The benefit being that each of these smaller conditional probabilities can be approximated based on the empirical co-occurrence of $C_k$ and $A_i$ in the training data, which together enables the estimation of the full conditional probability.

The validity of the conditional independence assumption is what the approximation quality rests on. Despite this assumption being questionable in the recommendation setting this hasn’t stopped the Naive Bayes Classifier being incorporated into recommender systems [5] [2]. An approach taken by [4] to improve the applicability of Naive Bayes to the recommendation task is to only use a subset of known preferences that each individually give high information on the target preference. This form of truncated Naive Bayes achieving improved prediction quality suggests there is substantial error cause by the assumption not holding. While this solution reduces the issue, ignoring some preferences is not a particularly desirable solution.

Without making a conditional independence assumption computing $P(A_1, A_2, ... A_N | C_k)$ can be handled by, first approximating the dependencies between all of the attributes $A_i$ and the class, $C_k$. This is generally referred to as structure learning, which determines parameters that specify a Bayesian Network or a Markov Random Field dependency structure. Then
inference on the resulting structure can be performed to make predictions. For instance in the Pairwise Markov Random Field (PMRF) considered by [1] nodes representing content are linked via pairwise potentials according to preference similarity. However, the restriction to only pairs limits the dependencies that can be captured. Although there has been work on learning potentials connecting more than just pairs of nodes in other settings [6] it remains prohibitively expensive computationally to learn such potentials on the scale that a recommender system would need.

Instead of learning directly dependencies between users or content an alternative perspective that has proven to be effective is treating preferences as depend only on latent user and content factors. In particular matrix factorization to learn such latent factors as used by [3] achieved notoriety by its success on the Netflix challenge. Although successful in terms of prediction performance, by side stepping the challenge of working with dependencies directly between preferences it becomes problematic to isolate why a user is predicted to like content or not. This is due to all the individual’s preferences contributing to the user’s latent factors obscures which reported preferences are the reason for a particular recommendation. As a result work continues on methods that avoid operating in latent features spaces despite their success with respect to prediction quality.

3 Less Naïve Bayes

In Naïve Bayes conditional independence of attributes used in prediction is assumed given the class being predicted. While reasonable in some applications it is a rather dubious assertion in the recommender setting. Instead of making conditional independence assumptions specific to the class being predicted for the following method seeks to blending the predictions that result from the possible different combinations of conditional independence assumptions that suffice to approximately decompose the full conditional probability.

By the definition of condition probability the condition probability needed, \(P(A_1, A_2, ..., A_N|C_k)P(C_k)\), is equivalent to the full joint probability \(P(A_1, A_2, ..., A_N, C_k)\). Approximation of the full joint distribution is considered as it removes the uniqueness placed on the class \((A_i \text{ and } C_k \text{ are merged into } I_j)\) being predicted that seems out of place in the recommender setting where both the attributes and the class are user-content preferences. To approximate this joint probability it needs to be decomposed down to only probabilities involving a single or pairs of preferences.

The weakest conditional independence assumption that can be made not knowing the true dependencies is given all other preferences a particular preference is conditionally independent of a particular other preference, as shown in Equation 1. What this amounts to is dropping a single dependency from an initially fully connected dependency network.

\[
P(I_1, I_2, ..., I_{k-1}, I_{k+1}, ..., I_m|I_k)P(I_k)
= P(I_1, ..., I_m)
≈ P(I_1, ..., I_{i-1}, I_{i+1}, ..., I_m)P(I_i|I_i, ..., I_{i-1}, I_{i+1}, ..., I_{j-1}, I_{j+1}, ..., I_m)
≈ \frac{P(I_1, ..., I_{i-1}, I_{i+1}, ..., I_m)P(I_1, ..., I_{j-1}, I_{j+1}, ..., I_m)}{P(I_1, ..., I_i, I_{i+1}, ..., I_{j-1}, I_{j+1}, ..., I_m)}
\]

In the fully connected dependency network of a join probability of \(m\) variables there are \(m(m-1)/2\) possible such conditional independence assumptions. Working with no knowledge on which is more valid the predictions that result from each can be blended together. To accomplish this the geometric mean of all of these predictions is used as shown in Equation 2. What this accomplishes is that the full joint probability can now be approximated using joint probabilities that involve at most one few variables.

\[
P(I_1, ..., I_m) \approx \left(\prod_{i=1}^{m} \prod_{j=i+1}^{m} \frac{P(I_1, ..., I_{i-1}, I_{i+1}, ..., I_m)P(I_1, ..., I_{j-1}, I_{j+1}, ..., I_m)}{P(I_1, ..., I_{i-1}, I_{i+1}, ..., I_{j-1}, I_{j+1}, ..., I_m)}\right)^{\frac{2}{m(m-1)}}
\]
For example prediction of whether a user likes content A given that they have provided preferences for content B and content C under Naive Bayes is predicted by equation 3. Where as under the proposed method the resulting prediction after term cancellation is that shown in Equation 4. The dependency structure used is also compared by Figure 1. Note the difference being the powers to which the single and pair probabilities are raised.

$$
$$

(3)

$$
\frac{P(A = 1|B, C)}{P(A = -1|B, C)} \approx \left(\frac{P(B, A = 1)P(C, A = 1)}{P(B, A = -1)P(C, A = -1)}\right)^\frac{1}{3} \left(\frac{P(A = -1)}{P(A = 1)}\right)^\frac{1}{3}
$$

(4)

\[\text{Figure 1: Comparision of Naive Bayes and Blended dependency structure}\]

To be able to generalize this approximation for a joint probability with arbitrarily many variables recursive approximate decomposition of joint probabilities that are not pairs or single variable probabilities is needed. This is where blending using the geometric mean enables such recursive approximation to computed efficiently. Following every decomposition the approximation is composed of joint probabilities of \(i\) variables in the numerator and joint probabilities of \(i-1\) variables in the denominator. Importantly, all of the probability terms in the numerator will have the same power by symmetry and similarly all terms in denominator have the same power.

Let \(T(i)\) and \(B(i)\) be the powers for the top and bottom terms respectively in an approximation where \(i\) is the layer of the approximation and corresponds to using joint probabilities of size of at most \(i\) variables. For a joint probability of full size \(m\) when moving down to layer \(i\) in the approximation we need to compute the power for numerator and denominator for the layer using those of the previous layer. There are \(mC_{i+1}\) numerator terms in layer above layer \(i\) each with power \(T(i+1)\). When decomposed each creates 2 terms for the numerator of layer \(i\). Since there are \(mC_{i}\) terms in layer \(i\) and power is evenly distributed the added power that results for these terms is \(\frac{2mC_{i+1}}{mC_{i}}T(i+1)\). The pre-existing power for terms of this size is \(B(i+1)\). However, there these terms are in the denominator of the layer and so this power is subtracted. The denominator power is computed similarly according to number of terms in layer above, \(mC_{i+1}\), their power \(T(i+1)\) and number of terms a layer down, \(mC_{i-1}\), except each numerator term decomposes to produce only one such denominator term. Mathematically the update calculations are show in Equation 5.

$$
T(i) \leftarrow \frac{2 \cdot mC_{i+1}}{mC_{i}}T(i+1) - B(i+1) \quad , \quad B(i) \leftarrow \frac{mC_{i+1}}{mC_{i-1}}T(i+1)
$$

(5)
Simplifying the combinatorials gives the layer update used in Algorithm 1 for computing the numerator and denominator powers. To limit numeric error the powers are computed in log form but in the interest of clarity the algorithm is presented in non-log form.

Algorithm 1 ComputeLayerPowers

Input: $m, k$
Output: $T(k), B(k)$

1. $i \leftarrow m$
2. $T(i) \leftarrow 1$
3. $B(i) \leftarrow 0$
4. while $i > k$ do
5. \hspace{1em} $i \leftarrow i - 1$
6. \hspace{2em} $T(i) \leftarrow 2(m - i + 1)/i \cdot T(i) - B(i)$
7. \hspace{2em} $B(i) \leftarrow (m - i + 2)(m - i + 1)/(i(i - 1)) \cdot T(i)$

4.1 Priors

For both Naive Bayes and the proposed method single and pairwise probabilities are needed. These can be estimated from data by frequency/co-occurrence but due to limited data it is expected to be important to apply a prior. For single probabilities a uniform prior is applied by adding a pseudo count of 1 to the number of times content has been like and disliked.

$$P(C_k = 1 | A_1, \ldots, A_N) \approx \prod_{i=1}^N \left( \frac{P(A_i, C_k = 1)}{P(A_i, C_k = -1)} \right)^{T(2)} \left( \frac{P(C_k = -1)}{P(C_k = 1)} \right)^{B(2)}$$

As was hinted at by the example the end result is identical in form to that of Naive Bayes in terms of the probabilities that it uses however the powers weight the pairs and class prior differently.

4.2 Flipped Preferences

When making prediction a user’s preferences are used to determine whether it more likely for target content to be liked or not. It may also be worth considering if the user’s preferences were flipped what would be recommended to this made-up user with opposite preferences.
By reversal of preferences it is expected that if the opposite user would like particular content there is reason to believe the true user would not.

An interesting effect when combining the prediction ratio according to a user’s preferences with the inverse of the predication ratio for the users’ flipped ratings is that the class probability cancels. As such the resulting prediction operates only on the users’ preferences not on content priors. Due to the absence of a content prior it is expected that these prediction are unlikely to be as effective. However, of interest is whether combining these prediction with either Naive Bayes or the proposed method achieves any gain.

4.3 Double Sided Predictions

In addition to the predictions made by using a user’s preference vector it is also possible to use the item’s preference vector to predict the target user-content preference using the same methods. Combining these two predictions may achieve improvement by the use of an alternative perspective. These two can be seen as one leveraging item dependencies (e.g. if item 1 is liked by a user then item 2 is probably liked by that user) to predict where as the other using user dependences (e.g. if user 1 likes particular content then user 2 will also probably like that content).

5 Experimental Evaluation

Evaluation is perform on a MovieLens data set that contains 6040 users and 3952 movies with approximately 1M ratings. The ratings are on a scale of 1 to 5. Ratings 4 and above are converted to like, 3 and below are converted to dislike. This gives 575,281 likes and 424,928 dislikes. A random sample of 10% of these preferences were selected for testing and the remaining 90% used in training.

The methods compared are as follows:

- **NBsimprior**: Naive Bayes using uniform priors on pairwise probabilities.
- **NB** and **NB+flip**: Naive Bayes and Naive Bayes combined with flip respectively.
- **flip**: Inclusion of inverted preferences as discussed in Section 4.2.
- **LNB** and **LNB+flip**: Proposed method and it combined with flip respectively.
- **doubleLNB+flip**: LNB+flip applied on both user and item dimensions.

All methods other than **NBsimprior** are using priors as shown in Section 4.1. All prediction ratio combinations are performed using product. Methods other than **doubleLNB+flip** are applied on the item dimension and learn dependencies between items.

Using the prediction ratios content is recommend if the ratio is larger than a set threshold. Setting the threshold trades between precision and recall. Increasing the threshold results in fewer recommendations likely reducing recall but precision is generally increased as those that are recommended the system is more confident the user will like. Decreasing the threshold accomplishes the opposite.

\[
\text{Precision} = \frac{|\{\text{positive user-content preferences}\} \cap \{\text{user-content recommendations}\}|}{|\{\text{user-content recommendations}\}|}
\]

\[
\text{Recall} = \frac{|\{\text{positive user-content preferences}\} \cap \{\text{user-content recommendations}\}|}{|\{\text{positive user-content preferences}\}|}
\]

Shown if Figure 2 is the ROC curve for the considered methods. This compares the trade-off between precision and recall as the threshold is varied for each of the methods. Better performance is indicated by a curve that is stretched towards the top left corner, which corresponds to perfect precision of 1 and perfect recall of 1.

---

3Available at: http://www.grouplens.org/node/73
The F1-score evaluates a classifier’s performance by combining its precision and recall into a single performance metric. For good performance to be achieved both precision and recall should be high and balanced.

\[
F1\text{-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]

Table 1 shows the precision, recall, and F1-score that each method obtains when the threshold is chosen to correspond to the maximum F1-score.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>doubleLNB+flip</td>
<td>0.7011</td>
<td>0.9021</td>
<td>0.7890</td>
</tr>
<tr>
<td>LNB+flip</td>
<td>0.7113</td>
<td>0.8679</td>
<td>0.7818</td>
</tr>
<tr>
<td>LNB</td>
<td>0.6774</td>
<td>0.8950</td>
<td>0.7712</td>
</tr>
<tr>
<td>NB+flip</td>
<td>0.6694</td>
<td>0.8867</td>
<td>0.7629</td>
</tr>
<tr>
<td>NB</td>
<td>0.6497</td>
<td>0.9135</td>
<td>0.7593</td>
</tr>
<tr>
<td>flip</td>
<td>0.6404</td>
<td>0.9079</td>
<td>0.7510</td>
</tr>
<tr>
<td>NBsimprior</td>
<td>0.5858</td>
<td>0.9760</td>
<td>0.7321</td>
</tr>
</tbody>
</table>

Table 1: Best F1-scores for each method and associate Precision and Recall

6 Discussion

Although the proposed method achieves a higher F1-score than Naive Bayes it is important to note that there is a clear cross over where at high recall LNB has higher precision than NB for the same recall and the reverse is true for recall lower than about 0.66. This could possibly be a consequence of LNB placing reduced weight on the prior enabling greater recall by more aggressively adjusting recommendations according to a user’s preferences.
but at the consequence of also being more prone to error causing it to suffer at higher
precisions.

The poor performance of \( NB \text{simprior} \) with a uniform prior over pairs gives a clear indication
of the importance an intelligent choice for the prior. There is possibly room for improvement
by further tuning of these priors. For instance the prior on single probabilities could possibly
be improved by using the average probability of content to be liked as a prior in place of
uniform. The Failure of \( \text{flip} \) to perform well indicates the importance of inclusion at least
some prior. While \( \text{flip} \) performs ineffectively on its own substantial improvement is seen
for both \( NB \) of \( LNB \) when combining \( \text{flip} \) with them. \( \text{flip} \) giving a greater improvement
to \( LNB \) could be due to the stronger prior used by \( NB \) resisting recommendations that
\( \text{flip} \) would like to give. The double-sided consideration gives noteworthy improvements to
precision outside of the 0.7 to 0.85 recall range. This suggests that there are important
dependencies between the preferences of users in addition to the dependencies between
items.

The method of computing powers presented Section 3 is likely overkill for computing the
powers and there is likely a simpler means to arrive and the needed powers. However
its potential is in being applicable for determining weights to use when joint probabilities
larger than pairs are to be used. Of further interest is whether incorporating such larger
probabilities can yield significant prediction improvements.

7 Conclusions and Future Work

Proposed is a novel method that uses the geometric mean to combine the predictions that
result from the possible difference conditional independency assumptions that decompose
and approximate a joint probability of many variables so to more robustly approximate said
joint probability. According to empirical evaluation on a sizable MovieLens data set the
proposed method is shown to achieve gains over the Naives Bayes Classifier with respect
to F1-score. In addition, the consideration of predictions of a user’s flipped preferences is
found to be beneficial to both Naive Bayes and the proposed method. Performing predic-
tion with respect to the user dimension in addition to the item dimension also found to
yield performance improvements. Though by design the proposed method can make use of
known probabilities larger than pairs it remains to be evaluated whether incorporating such probabilities can yield prediction improvements.

References

belief propagation over pairwise markov random fields. In Communication, Control,

system using naive bayes classifier and collaborative filtering. In The 2010 IAENG
International Conference on Data Mining and Applications, April 2010.


In Proceedings of the 6th Pacific Rim International Conference on Artificial Intelligence,

2003 IEEE/WIC International Conference on Web Intelligence, WI ’03, pages 168–,

pairwise potentials. In Proceedings of International Workshop on Artificial Intelligence