The Application of Deep Learning in Collaborative Filtering

Kai Lu

June-5th-2013
Outline

• Background Introduction
• RBM model introduction
• RBM in Collaborative Filtering
• Conditional RBM model
• SVD model
• Experimental Result
Introduction

• Users are facing more and more severe “Information overload” problem in the Internet.

• Collaborative filtering (CF) is an effective way to provide personalized recommendation for users.

• CF methods usually utilize user’s histories (rating, watching, purchasing) to build model, and make recommendations.
Introduction

We know what you ought to be watching!!!
Deep Learning Introduction

• DL is a new area of ML research, which has been introduced with the objective of moving ML closer to one of its original goals: Artificial Intelligence.

• Restricted Boltzmann Machines is one of the deep learning models.

• The main goal of this project is to explore the performance of RBM on the collaborative filtering task.
RBM Introduction

- RBMs are a specific type of undirected graphical model consisting of two layers of binary variables-hidden (H) and visible (V) with no intra-layer connections:

\[
\text{Visible units} \\
\text{Weights} \\
\text{Hidden units}
\]

- The joint over the visible and hidden variables is given by the Gibbs Distribution: 
  \[
P(V, H) = \frac{1}{Z} \exp(-E(V, H))
\]

- \(E(V, H)\) is an energy function defined as:
  \[
  E(V, H) = -\sum_{i=1}^{N_v} \sum_{j=1}^{N_h} v_i h_j w_{ij} - \sum_{i=1}^{N_v} v_i b_i - \sum_{j=1}^{N_h} h_j b_j
  \]
RBM introduction

- Probability of observing a single \( N_v \)-dimensional data point \( V \) is given by marginal:

\[
P^\theta(V) = \sum_h P^\theta(V, H) = \frac{1}{Z^\theta} \sum_h \exp(-E^\theta(V, H))
\]

- The log-likelihood can be written as:

\[
l(\theta) = \frac{1}{N} \sum_{n=1}^{N} \ln P^\theta(V_n) = l(\theta)^+ - l(\theta)^-
\]

\[
l(\theta)^+ = \frac{1}{N} \sum_{n=1}^{N} \ln \sum_h \exp(-E^\theta(V_n, H))
\]

\[
l(\theta)^- = \ln(Z^\theta) = \ln \sum_{h,v} \exp(-E^\theta(V, H))
\]

- Perform Maximum Likelihood, the positive term gradient:

\[
\frac{\partial l(\theta)^+}{\partial \theta} = \frac{1}{N} \sum_{n=1}^{N} \frac{\partial}{\partial \theta} \ln \sum_h \exp(-E^\theta(V_n, H)) = \frac{1}{N} \sum_{n=1}^{N} \sum_h P^\theta(H | V_n) \frac{\partial(-E^\theta(V_n, H))}{\partial \theta}
\]

- The negative term gradient:

\[
\frac{\partial l(\theta)^-}{\partial \theta} = \frac{\partial}{\partial \theta} \ln \sum_{h,v} \exp(-E^\theta(V, H)) = \sum_{h,v} P^\theta(H, V) \frac{\partial(-E^\theta(V, H))}{\partial \theta}
\]
RBM Introduction-Learning

- **Full gradient:**

\[
\Delta w_{ij} = \frac{\partial \log P^\theta(V)}{\partial w_{ij}} = \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}}
\]

- **Evaluate** \( P^\theta(H \mid V) \) and \( P^\theta(H,V) \):

\[
P^\theta(h_j = 1 \mid V) = \frac{P^\theta(h_j = 1,V)}{P^\theta(h_j = 0,V) + P^\theta(h_j = 1,V)} = \sigma(b_j + \sum_{i=1}^{N_v} w_{ij} v_i)
\]

\[
P^\theta(v_i \mid H) = \sigma(b_i + \sum_{j=1}^{N_h} w_{ij} h_j)
\]

- **Approximate** \( P^\theta(v_i,H) \) by Gibbs samples alternate between:

\[
h_j^{(t)} \sim P(h_j \mid V^{(t-1)})
\]

- and

\[
v_i^{(t)} \sim P(v_i \mid H^{(t)})
\]
RBM in Collaborative Filtering

• The CF problem is presented with an $N \times M$ ratings matrix:
  – $N$ is user number, $M$ is user number,
  – Rating is integer values from 1 to $K$.

• The CF RBM is as follows:
RBM in CF

• In CF RBM, the conditional $P^q(v_i | H)$ take the form of ‘softmax’ function:

$$p(v_i^k = 1 | h) = \frac{\exp(b_i^k + \sum_{j=1}^{F} h_j W_{ij}^k)}{\sum_{l=1}^{K} \exp(b_l^j + \sum_{j=1}^{Nh} h_j w_{ij}^l)}$$

$$v_i^k = 1 \rightarrow v_i = k$$

• The new $p(h_j = 1 | V)$ is as follows:

$$p(h_j = 1 | V) = \sigma(b_j + \sum_{i=1}^{m} \sum_{k=1}^{K} v_i^k W_{ij}^k)$$

• The marginal distribution over the visible ratings $V$ is:

$$p(V) = \sum_h \frac{\exp(-E(V, h))}{\sum_{V', h'} \exp(-E(V', h'))}$$

• Energy is:

$$E(V, h) = -\sum_{i=1}^{m} \sum_{j=1}^{F} \sum_{k=1}^{K} W_{ij}^k h_j v_i^k + \sum_{i=1}^{m} \log Z_i - \sum_{i=1}^{m} \sum_{k=1}^{K} v_i^k b_i^k - \sum_{j=1}^{F} h_j b_j$$

$$Z_i = \sum_{k=1}^{K} \exp(b_i^j + \sum_{j=1}^{h} h_j W_{ij}^l)$$
Learning and Prediction

- **Parameter Learning:**
  \[
  \Delta W_{ij}^k = \varepsilon \frac{\partial \log p(V)}{\partial W_{ij}^k} = \varepsilon(<v_i^k h_j^k>_{data} - <v_i^k h_j^k>_{model})
  \]
  \[
  \Delta b_i^k = \varepsilon(<v_i^k>_{data} - <v_i^k>_{model}) \quad \Delta b_j = \varepsilon(<h_j>_{data} - <h_j>_{model})
  \]

- **Efficient Learning:** Contrastive Divergence (CD) (Hinton, 2002)
  \[
  \Delta W_{ij}^k = \varepsilon(<v_i^k h_j^k>_{data} - <v_i^k h_j^k>_{T})
  \]

- **Making Prediction:**
  \[
  p(v_q^k = 1|V) \propto \sum_{h_1,...,h_p} \exp(-E(v_q^k, v, h)) \propto \exp(v_q^k b_q^k) \prod_{j=1}^{F} (1 + \exp(\sum_{ij} v_q^k W_{qj}^k + b_j))
  \]

- **The expectation** is then computed by normalizing over all K possible ratings.
Conditional RBM’s

• We can use the user/movies of the test data into RBMs: movies were viewed but rating is unknown.
• Conditional RBM model takes this extra information into account:

\[
p(v_i^k = 1 | h) = \frac{\exp(b_i^k + \sum_{j=1}^{F} h_j W_{ij}^k)}{\sum_{i=1}^{V} \exp(b_i' + \sum_{j=1}^{N_h} h_j W_{ij}')}
\]

\[
p(h_j = 1 | V, r) = \sigma(b_j + \sum_{i=1}^{m} \sum_{k=1}^{K} v_i^k W_{ij}^k + \sum_{i=1}^{M} r_i D_{ij})
\]

• The idea is to define a joint distribution over \((V, h)\) conditional on \(r\), \(r\) will affect the states of hidden units:

![Diagram of Conditional RBM](image-url)
SVD Model

• SVD seeks a low-rank matrix:

\[ X = UV' \quad U \in \mathbb{R}^{N \times C}, V \in \mathbb{R}^{M \times C} \]

• The optimization function is as follows:

\[
\begin{align*}
    f &= \sum_{i=1}^{N} \sum_{j=1}^{M} I_{ij} (u_i v_j' - Y_{ij})^2 + \lambda \sum_{ij} I_{ij} (\| u_i \|_{Fro}^2 + \| v_j \|_{Fro}^2) \\
    (U^*, V^*) &= \arg \min_{(U,V)} f \\
    I_{ij} &= \begin{cases}
        1 & \text{if } v_{ij} \neq \text{null} \\
        0 & \text{otherwise}
    \end{cases}
\end{align*}
\]

• The gradient is as follows:

\[
\begin{align*}
    \frac{\partial}{\partial u_{ik}} &= -e_{ij} v_{kj} + \lambda u_{ik}, \\
    \frac{\partial}{\partial v_{jk}} &= -e_{ij} u_{ik} + \lambda v_{kj}
\end{align*}
\]

\[
\begin{align*}
    u_{ik}' &= u_{ik} + \eta (e_{ij} v_{kj} - \lambda u_{ik}), \\
    v_{kj}' &= v_{kj} + \eta (e_{ij} u_{ik} - \lambda v_{kj})
\end{align*}
\]

• Prediction:

\[
\hat{r}_{ij} = \sum_{k=1}^{C} u_{ik} \cdot v_{kj}
\]
Experiments

• Dataset:

<table>
<thead>
<tr>
<th>#User</th>
<th>#item</th>
<th>rating</th>
<th>#train</th>
<th>#validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>480,189</td>
<td>17,770</td>
<td>[1,5]</td>
<td>99,072,112</td>
<td>1,408,395</td>
</tr>
</tbody>
</table>

• Evaluation Metric:

\[
RMSE(T) = \sqrt{\frac{1}{|T|} \sum_{(i,j) \in T} (\hat{r}_{ij} - r_{ij})^2}
\]
Experimental

- RBM VS MF: Factor SIZE = 40
Experimental Results

• Comparison with varying factor size:
Reference:

• [R. Salakhutdinov, G. Hinton] Restricted Boltzmann Machines for Collaborative Filtering. ICML07
• [G. Hinton] A practical guide to Training Restricted Boltzmann Machines; 2010
• [Asja Fischer] An Introduction to Restricted Boltzmann Machines.