Web Mining and Recommender Systems

Temporal data mining: Regression for Sequence Data
• Discuss how to use regression to predict temporally evolving data
Temporal models

This topic will look back on some of the topics already covered in this class, and see how they can be adapted to make use of temporal information

1. Regression – sliding windows and autoregression
2. Social networks – densification over time
3. Text mining – “Topics over Time”
4. Recommender systems – some results from Koren
Previously – Regression

Given **labeled training data** of the form

\[(data_1, label_1), \ldots, (data_n, label_n)\]

Infer the function

\[f(data) \rightarrow labels\]
Here, we’d like to predict sequences of real-valued events as accurately as possible.
Here, we’d like to predict sequences of **real-valued** events as accurately as possible.

Given: a time series:

\[(x_1, \ldots, x_N) \in \mathbb{R}^N\]

Suppose we’d like to minimize the MSE (as usual!) of the final part of some continuous portion of the sequence

\[
\frac{1}{u-v+1} \sum_{t=u}^{v} (f_t(x_1, \ldots, x_{u-1}) - x_t)^2
\]
Method 1: maintain a "moving average" using a window of some fixed length

\[ f(x_1, \ldots, x_m) = \frac{\sum_{k=0}^{K-1} x_{n-k}}{K} \]

\[ \mathcal{O}(n/K) \]
Method 1: maintain a “moving average” using a window of some fixed length

- This can be computed efficiently via dynamic programming:

\[ f(x_1, \ldots, x_{m+1}) = \frac{K \sum_{i=1}^{p} (x_{n-i} - x_{n-k}) - x_{n-k+1} + x_{n+1}}{K} \]
Time-series regression

**Method 1:** maintain a “moving average” using a window of some fixed length

\[
f(x_1, \ldots, x_m) = \frac{1}{K} \sum_{k=0}^{K-1} x_{m-k}
\]

- This can be computed efficiently via dynamic programming:

\[
f(x_1, \ldots, x_{m+1}) = \frac{1}{K} \left( K \cdot f(x_1, \ldots, x_m) - x_{m-k} + x_{m+1} \right)
\]

"peel-off" the oldest point  
add the newest point
Time-series regression

Also useful to plot data:

BeerAdvocate, ratings over time

Scatterplot

Sliding window (K=10000)

long-term trends

seasonal effects

Code on course webpage
Method 2: weight the points in the moving average by age

\[ f(x_1, \ldots, x_m) = \frac{\sum_{k=0}^{K-1} \frac{(K-k)}{K} x_{n-k}}{1 + 2 + \ldots + K} \]
**Method 2:** weight the points in the moving average by age

newest points have the highest weight
weight decays to zero after $K$ points

$$f(x_1, \ldots, x_m) = \sum_{k=0}^{K-1} \binom{K-k}{K} x_{m-k}$$
**Method 3**: weight the most recent points exponentially higher

\[
f(x_1) = x_1
\]

\[
f(x_1, \ldots, x_m) = \alpha f(x_1, \ldots, x_{m-1}) + (1-\alpha)x_m
\]
Methods 1, 2, 3

Method 1: Sliding window
Method 2: Linear decay
Method 3: Exponential decay
Method 4: all of these models are assigning weights to previous values using some predefined scheme, why not just learn the weights?

\[ f(x_1, \ldots, x_m) = \sum_{k=0}^{K-1} \theta_k x_n - k \]

\[ \sum_{m} \left( \sum_{k=0}^{K-1} f(x_k, \ldots, x_n) - x_{n-1} \right)^2 \]
Method 4: all of these models are assigning weights to previous values using some predefined scheme, why not just learn the weights?

- We can now fit this model using least-squares
- This procedure is known as autoregression
- Using this model, we can capture periodic effects, e.g. that the traffic of a website is most similar to its traffic 7 days ago
Learning Outcomes

• Introduced several schemes to predict values in sequences
• Introduced autoregression
Web Mining and Recommender Systems

Temporal dynamics in social networks
Learning Goals

• Discuss how social networks change over time
How can we characterize, model, and reason about the structure of social networks?

1. Models of network structure
2. Power-laws and scale-free networks, “rich-get-richer” phenomena
3. Triadic closure and “the strength of weak ties”
4. Small-world phenomena
5. Hubs & Authorities; PageRank
Previously we saw some processes that model the generation of social and information networks

- Power-laws & small worlds
- Random graph models

These were all defined with a “static” network in mind. But if we observe the order in which edges were created, we can study how these phenomena change as a function of time.

First, let’s look at “microscopic” evolution, i.e., evolution in terms of individual nodes in the network.
Temporal dynamics of social networks

Q1: How do networks grow in terms of the number of nodes over time?

A: Doesn’t seem to be an obvious trend, so what do networks have in common as they evolve?

(from Leskovec, 2008 (CMU Thesis))
Q2: When do nodes create links?
- x-axis is the age of the nodes
- y-axis is the number of edges created at that age

A: In most networks there’s a “burst” of initial edge creation which gradually flattens out. Different behavior on LinkedIn?
Q3: How long do nodes “live”?  
• x-axis is the diff. between date of last and first edge creation  
• y-axis is the frequency  

A: Node lifetimes follow a power-law: many many nodes are shortlived, with a long-tail of older nodes
What about “macroscopic” evolution, i.e., how do global properties of networks change over time? **Q1:** How does the # of nodes relate to the # of edges?

- A few more networks: citations, authorship, and autonomous systems (and some others, not shown)
- **A:** Seems to be linear (on a log-log plot) but the number of edges grows faster than the number of nodes as a function of time
Q1: How does the # of nodes relate to the # of edges?

A: seems to behave like

\[ E(t) \propto N(t)^a \]

where

\[ 1 \leq a \leq 2 \]

- a = 1 would correspond to constant out-degree – which is what we might traditionally assume
- a = 2 would correspond to the graph being fully connected
- What seems to be the case from the previous examples is that a > 1 – the number of edges grows faster than the number of nodes
Q2: How does the degree change over time?

• A: The average out-degree **increases** over time.
Q3: If the network becomes denser, what happens to the (effective) diameter?

- A: The diameter seems to decrease.
- In other words, the network becomes more of a small world as the number of nodes increases.
Q4: Is this something that must happen – i.e., if the number of edges increases faster than the number of nodes, does that mean that the diameter must decrease?

A: Let’s construct random graphs (with $a > 1$) to test this:
So, a decreasing diameter is not a “rule” of a network whose number of edges grows faster than its number of nodes, though it is consistent with a preferential attachment model.

**Q5:** is the degree distribution of the nodes sufficient to explain the observed phenomenon?

**A:** Let’s perform **random rewiring** to test this.

Random rewiring preserves the degree distribution, and randomly samples amongst networks with observed degree distribution.
So, a decreasing diameter is **not** a “rule” of a network whose number of edges grows faster than its number of nodes, though it is consistent with a preferential attachment model.

**Q5:** is the degree distribution of the nodes sufficient to explain the observed phenomenon?
Temporal dynamics of social networks

So, a decreasing diameter is not a “rule” of a network whose number of edges grows faster than its number of nodes, though it is consistent with a preferential attachment model.

Q5: Is the degree distribution of the nodes sufficient to explain the observed phenomenon?

A: Yes! The fact that real-world networks seem to have decreasing diameter over time can be explained as a result of their degree distribution and the fact that the number of edges grows faster than the number of nodes.
Temporal dynamics of social networks

Other interesting topics...

“memetracker”
Temporal dynamics of social networks

Other interesting topics...

Aligning query data with disease data – Google flu trends: https://www.google.org/flutrends/us/#US

Learning Outcomes

- Discussed how social networks change over time
- Described some mechanisms to explain this phenomenon
Further reading:
“Dynamics of Large Networks” (most plots from here)
Jure Leskovec, 2008

“Microscopic Evolution of Social Networks”
Leskovec et al. 2008

“Graph Evolution: Densification and Shrinking Diameters”
Leskovec et al. 2007
Web Mining and Recommender Systems

Temporal dynamics of text
Learning Goals

• Discuss how text can change over time
Previously...

Bag-of-Words representations of text:

F_{text} = [150, 0, 0, 0, 0, 0, 0, ... , 0]

- a
- aardvark
- zoetrope
Previously, we tried to develop low-dimensional representations of documents:

What we would like:

87 of 102 people found the following review helpful

⭐⭐⭐⭐⭐ You keep what you kill, December 27, 2004
By Schlinky “Schlinky” (Washington State) - See all my reviews

This review is from: The Chronicles of Riddick (Widescreen Unrated Director's Cut) (DVD)
Even if I have to apologize to my Friends and Favorites, and my family, I have to admit that I really liked this movie. It's a Sci-Fi movie with a "Mad Maxx" appeal that, while changing many things, left Riddick from 'Pitch Black' to be just Riddick. They did not change his attitude or soften him up or bring him out of his original character, which was very pleasing to 'Pitch Black' fans like myself.

First off, let me say that when playing the DVD, the first selection to come up is Convert or Fight, and no explanation of the choices. This confused me at first, so I will mention off the bat that they are simply different menu formats, that each menu has the very same options, simply different background visuals. Select either one and continue with the movie.

(review of “The Chronicles of Riddick”)
We saw how **LDA** can be used to describe documents in terms of **topics**

- Each document has a **topic vector** (a stochastic vector describing the fraction of words that discuss each topic)
- Each topic has a **word vector** (a stochastic vector describing how often a particular word is used in that topic)
Topics and documents are **both** described using stochastic vectors:

Each document has a **topic distribution** which is a mixture over the topics it discusses

$$\theta_d \in \Delta^K \text{ i.e., } \forall_d \sum_k \theta_{d,k} = 1$$

Each topic has a **word distribution** which is a mixture over the words it discusses

$$\phi_k \in \Delta^D \text{ i.e., } \forall_k \sum_w \phi_{k,w} = 1$$
Latent Dirichlet Allocation

Topics over Time (Wang & McCallum, 2006) is an approach to incorporate temporal information into topic models.

E.g.

- The topics discussed in conference proceedings progressed from neural networks, towards SVMs and structured prediction (and back to neural networks).
- The topics used in political discourse now cover science and technology more than they did in the 1700s.
- With in an institution, e-mails will discuss different topics (e.g. recruiting, conference deadlines) at different times of the year.
Topics over Time (Wang & McCallum, 2006) is an approach to incorporate temporal information into topic models.

The ToT model is similar to LDA with one addition:

1. For each topic $K$, draw a word vector $\phi_k$ from $\text{Dir}(\beta)$
2. For each document $d$, draw a topic vector $\theta_d$ from $\text{Dir}(\alpha)$
3. For each word position $i$:
   1. draw a topic $z_{di}$ from multinomial $\theta_d$
   2. draw a word $w_{di}$ from multinomial $\phi_{z_{di}}$
   3. **draw a timestamp** $t_{di}$ **from Beta($\psi_{z_{di}}$)**
Latent Dirichlet Allocation

Topics over Time (Wang & McCallum, 2006) is an approach to incorporate temporal information into topic models.

3.3. draw a timestamp $t_{\{di\}}$ from Beta($\psi_{\{z\{di\}}}$)

- There is now one Beta distribution per topic.
- Inference is still done by Gibbs sampling, with an outer loop to update the Beta distribution parameters.

Beta distributions are a flexible family of distributions that can capture several types of behavior – e.g. gradual increase, gradual decline, or temporary “bursts”.

\[
\text{p.d.f.:} \quad \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}
\]
Latent Dirichlet Allocation

**Results:**
Political addresses – the model seems to capture realistic “bursty” and gradually emerging topics

<table>
<thead>
<tr>
<th>Topic</th>
<th>Probability</th>
<th>Topic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mexican War</td>
<td>0.02032</td>
<td>Panama Canal</td>
<td>0.01521</td>
</tr>
<tr>
<td>Mexico</td>
<td>0.01832</td>
<td>United States</td>
<td>0.01607</td>
</tr>
<tr>
<td>Government</td>
<td>0.01521</td>
<td>Islands</td>
<td>0.01167</td>
</tr>
<tr>
<td>War</td>
<td>0.01059</td>
<td>Canal</td>
<td>0.01014</td>
</tr>
<tr>
<td>Congress</td>
<td>0.00951</td>
<td>American</td>
<td>0.00872</td>
</tr>
<tr>
<td>Country</td>
<td>0.00906</td>
<td>Cuba</td>
<td>0.00834</td>
</tr>
<tr>
<td>Texas</td>
<td>0.00852</td>
<td>Made</td>
<td>0.00747</td>
</tr>
<tr>
<td>Made</td>
<td>0.00727</td>
<td>General</td>
<td>0.00731</td>
</tr>
<tr>
<td>Great</td>
<td>0.00611</td>
<td>War</td>
<td>0.00666</td>
</tr>
<tr>
<td>World</td>
<td>0.01875</td>
<td>Energy</td>
<td>0.03902</td>
</tr>
<tr>
<td>States</td>
<td>0.02132</td>
<td>National</td>
<td>0.01534</td>
</tr>
<tr>
<td>Security</td>
<td>0.02067</td>
<td>Development</td>
<td>0.01448</td>
</tr>
<tr>
<td>Soviet</td>
<td>0.01664</td>
<td>Space</td>
<td>0.01436</td>
</tr>
<tr>
<td>United</td>
<td>0.01491</td>
<td>Science</td>
<td>0.01227</td>
</tr>
<tr>
<td>Nuclear</td>
<td>0.01454</td>
<td>Technology</td>
<td>0.01227</td>
</tr>
<tr>
<td>Peace</td>
<td>0.01408</td>
<td>Oil</td>
<td>0.01178</td>
</tr>
<tr>
<td>Nations</td>
<td>0.01069</td>
<td>Make</td>
<td>0.00994</td>
</tr>
<tr>
<td>International</td>
<td>0.01024</td>
<td>Effort</td>
<td>0.00969</td>
</tr>
<tr>
<td>America</td>
<td>0.00987</td>
<td>Administration</td>
<td>0.00957</td>
</tr>
</tbody>
</table>

Fitted Beta distribution assignments to this topic.
Latent Dirichlet Allocation

Results:
e-mails & conference proceedings

<table>
<thead>
<tr>
<th>Faculty Recruiting</th>
<th>ART Paper</th>
<th>MALLET</th>
<th>CVS Operations</th>
<th>Recurrent NN</th>
<th>Game Theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs 0.03572</td>
<td>xenui 0.02113</td>
<td>code 0.05668</td>
<td>clock 0.04473</td>
<td>state 0.05903</td>
<td>game 0.02850</td>
</tr>
<tr>
<td>april 0.02724</td>
<td>data 0.01814</td>
<td>files 0.04212</td>
<td>page 0.04970</td>
<td>recurrent 0.03765</td>
<td>strategy 0.02378</td>
</tr>
<tr>
<td>faculty 0.02341</td>
<td>word 0.01961</td>
<td>mail 0.04673</td>
<td>version 0.00328</td>
<td>sequence 0.03636</td>
<td>play 0.01690</td>
</tr>
<tr>
<td>david 0.02012</td>
<td>research 0.01408</td>
<td>java 0.03805</td>
<td>cvs 0.01557</td>
<td>sequences 0.02462</td>
<td>games 0.01173</td>
</tr>
<tr>
<td>lunch 0.01766</td>
<td>topic 0.01366</td>
<td>file 0.02847</td>
<td>add 0.03083</td>
<td>time 0.02402</td>
<td>player 0.01511</td>
</tr>
<tr>
<td>schedule 0.01656</td>
<td>model 0.01238</td>
<td>al 0.02479</td>
<td>update 0.02589</td>
<td>states 0.02657</td>
<td>agents 0.01346</td>
</tr>
<tr>
<td>candidate 0.01560</td>
<td>andres 0.01238</td>
<td>directory 0.02890</td>
<td>latest 0.02530</td>
<td>transition 0.01300</td>
<td>expert 0.01281</td>
</tr>
<tr>
<td>talk 0.01355</td>
<td>sample 0.01152</td>
<td>version 0.01664</td>
<td>updated 0.02317</td>
<td>finite 0.01242</td>
<td>strategies 0.01123</td>
</tr>
<tr>
<td>bruce 0.01273</td>
<td>enron 0.00967</td>
<td>pdf 0.01421</td>
<td>checkcd 0.02777</td>
<td>length 0.01154</td>
<td>opponent... 0.01088</td>
</tr>
<tr>
<td>visit 0.01232</td>
<td>dataset 0.00960</td>
<td>bug 0.01352</td>
<td>change 0.02156</td>
<td>strings 0.01013</td>
<td>nash 0.00648</td>
</tr>
</tbody>
</table>
Latent Dirichlet Allocation

**Results:**
conference proceedings (NIPS)

Relative weights of various topics in 17 years of NIPS proceedings
• Discussed how text can change over time
Further reading:
“Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends”
(Wang & McCallum, 2006)
http://people.cs.umass.edu/~mccallum/papers/tot-kdd06.pdf
Web Mining and Recommender Systems

Temporal recommender systems
• Discuss how temporal dynamics can be incorporated into recommender systems
Recommender Systems go beyond the methods we’ve seen so far by trying to model the relationships between people and the items they’re evaluating.

- Preference toward "action"
- Preference toward "special effects"
- Is the movie action-heavy?
- Are the special effects good?

Compatibility
Predict a user’s rating of an item according to:

\[ f(u, i) = \alpha + \beta_u + \beta_i + \gamma_u \cdot \gamma_i \]

By solving the optimization problem:

\[
\arg\min_{\alpha, \beta, \gamma} \sum_{u,i} (\alpha + \beta_u + \beta_i + \gamma_u \cdot \gamma_i - R_{u,i})^2 + \lambda \left( \sum_u \beta_u^2 + \sum_i \beta_i^2 + \sum_i \|\gamma_i\|_2^2 + \sum_u \|\gamma_u\|_2^2 \right)
\]

(error) \quad (\text{regularizer})

(e.g. using stochastic gradient descent)
Temporal latent-factor models

To build a reliable system (and to win the Netflix prize!) we need to account for **temporal dynamics**:

![Graphs showing Netflix ratings over time and by movie age.](Image)

Netflix ratings by movie age

So how was this actually done?

Figure from Koren: “Collaborative Filtering with Temporal Dynamics” (KDD 2009)
Temporal latent-factor models

To start with, let’s just assume that it’s only the bias terms that explain these types of temporal variation (which, for the examples on the previous slides, is potentially enough)

\[ b_{u,i}(t) = \alpha + \beta_u(t) + \beta_i(t) \]

Idea: temporal dynamics for items can be explained by long-term, gradual changes, whereas for users we’ll need a different model that allows for “bursty”, short-lived behavior
Temporal latent-factor models

temporal bias model:

\[ b_{u,i}(t) = \alpha + \beta_u(t) + \beta_i(t) \]

For item terms, just separate the dataset into (equally sized) bins:

\[ \beta_i(t) = \beta_i + \beta_i,\text{Bin}(t) \]

*in Koren's paper they suggested ~30 bins corresponding to about 10 weeks each for Netflix

or bins for periodic effects (e.g. the day of the week):

\[ \beta_i(t) = \beta_i + \beta_i,\text{Bin}(t) + \beta_i,\text{period}(t) \]

What about user terms?

• We need something much finer-grained

• **But** – for most users we have far too little data to fit very short term dynamics
Temporal latent-factor models

Start with a simple model of drifting dynamics for users:

$$\text{dev}_u(t) = \text{sign}(t - t_u) \cdot |t - t_u|^x$$

- **mean** rating date for user $u$
- before (-1) or after (1) the mean date
- days away from mean date
- hyperparameter (ended up as $x=0.4$ for Koren)
Temporal latent-factor models

Start with a simple model of drifting dynamics for users:

\[ \text{dev}_u(t) = \text{sign}(t - t_u) \cdot |t - t_u|^x \]

- **mean** rating date for user \( u \)
- hyperparameter (ended up as \( x=0.4 \) for Koren)
- days away from mean date
- before (-1) or after (1) the mean date

Time-dependent user bias can then be defined as:

\[ \beta^{(1)}_u(t) = \beta_u + \alpha_u \cdot \text{dev}_u(t) \]

- overall user bias
- sign and scale for deviation term
Temporal latent-factor models

Netflix ratings over time

Real data

Fitted model
Temporal latent-factor models

time-dependent user bias can then be defined as:

$\beta_u^{(1)}(t) = \beta_u + \alpha_u \cdot \text{dev}_u(t)$

- overall user bias
- sign and scale for deviation term

• Requires only two parameters per user and captures some notion of temporal “drift” (even if the model found through cross-validation is (to me) completely unintuitive)

• To develop a slightly more expressive model, we can interpolate smoothly between biases using splines
Temporal latent-factor models

\[ \beta_u^{(2)}(t) = \beta_u + \frac{\sum_{l=1}^{k_u} e^{-\gamma|t-t_l^u|} b_{u,t_l}}{\sum_{l=1}^{k_u} e^{-\gamma|t-t_l^u|}} \]

- number of control points for this user \((k_u = n_u^{0.25} \text{ in Koren})\)
- user bias associated with this control point
- time associated with control point (uniformly spaced)
Temporal latent-factor models

\[ \beta_u^{(2)}(t) = \beta_u + \sum_{l=1}^{k_u} \frac{e^{-\gamma|t - t^u_l|} b^u_{tl}}{\sum_{l=1}^{k_u} e^{-\gamma|t - t^u_l|}} \]

- This is now a reasonably flexible model, but still only captures \textit{gradual drift}, i.e., it can’t handle sudden changes (e.g. a user simply having a bad day)
Temporal latent-factor models

- Koren got around this just by adding a “per-day” user bias:

\[ \beta_{u,t} \]

bias for a particular day (or session)

- Of course, this is only useful for particular days in which users have a lot of (abnormal) activity

- The final (time-evolving bias) model then combines all of these factors:

\[ \beta_{u,i}(t) = \alpha + \beta_u + \alpha_u \cdot \text{dev}_u(t) + \beta_{u,t} + \beta_i + \beta_i,\text{Bin}(t) \]
Finally, we can add a time-dependent scaling factor:

$$\beta_{u,i}(t) = \alpha + \beta_u + \alpha_u \cdot \text{dev}_u(t) + \beta_{u,t} + (\beta_i + \beta_{i,Bin(t)}) \cdot c_u(t)$$

also defined as $c_u + c_{u,t}$

Latent factors can also be defined to evolve in the same way:

$$\gamma_{u,k}(t) = \gamma_{u,k} + \alpha_{u,k} \cdot \text{dev}_u(t) + \gamma_{u,k,t}$$

factor-dependent user drift
factor-dependent short-term effects
Summary

• Effective modeling of temporal factors was absolutely critical to this solution outperforming alternatives on Netflix’s data
• In fact, even with only temporally evolving bias terms, their solution was already ahead of Netflix’s previous (“Cinematch”) model

On the other hand...
• Many of the ideas here depend on dynamics that are quite specific to “Netflix-like” settings
• Some factors (e.g. short-term effects) depend on a high density of data per-user and per-item, which is not always available
Temporal latent-factor models

Summary

• Changing the setting, e.g. to model the stages of progression through the symptoms of a disease, or even to model the temporal progression of people’s opinions on beers, means that alternate temporal models are required.

rows: models of increasingly “experienced” users

columns: review timeline for one user
Learning Outcomes

- Discussed how temporal dynamics can be incorporated into recommender systems
- Discussed how this was useful for Netflix in particular
Further reading:
“Collaborative filtering with temporal dynamics”
Yehuda Koren, 2009