FACTORORIZATION MACHINES AS A TOOL FOR HEALTHCARE
CASE STUDY ON TYPE 2 DIABETES DETECTION

Ioakeim (Kimis) Perros and Jimeng Sun
perros@gatech.edu, jsun@cc.gatech.edu
Computational Science and Engineering
Georgia Institute of Technology

Artificial Intelligence in Medicine 2015
Workshop on Matrix Computations for Biomedical Informatics
Talk Outline
Talk Outline

THE PROBLEM:
PREDICT TYPE 2 DIABETES
Talk Outline

THE PROBLEM:
PREDICT TYPE 2 DIABETES

USING ELECTRONIC HEALTH RECORDS
Talk Outline

The Problem:
Predict Type 2 Diabetes

Using Electronic Health Records

The Method:
Factorization Machines
Healthcare Data Challenges
Healthcare Data Challenges

Heterogeneous, High-Dimensional

Diagnoses

Medications

Genomic
Healthcare Data Challenges

Heterogeneous, High-Dimensional, Noisy
Healthcare Data Challenges

Heterogeneous, High-Dimensional

Noisy

Sparse
Type 2 Diabetes Mellitus

Type 2 Diabetes
The Silent Pandemic

By 2050, 1 out of every 3 Americans will have diabetes.

FED UP
In Theaters May 9

[Image of a crowd with the text "Type 2 Diabetes The Silent Pandemic" and an advertisement for the movie "FED UP".]
Type 2 Diabetes Mellitus

Currently: More Than 3M Cases per Year in the US
Type 2 Diabetes Mellitus

Currently: More than 3M cases per year in the US

Early detection and treatment decreases the risk of developing diabetes’ complications.
Type 2 Diabetes Mellitus

Currently: More than 3M cases per year in the US

Target of this work: Accurate classification of diabetic patients, given their clinical status
Model learning from sparse data

**Linear Regression**

\[ y(x) = w_0 + \sum_{i=1}^{p} w_i x_i \]

\[ w_0 \in \mathbb{R}, \ w \in \mathbb{R}^p \]
Model learning from sparse data

LINEAR REGRESSION

Often interested in feature interactions (e.g. how diagnostic conditions JOINTLY affect the target prediction)

\[ y(x) = w_0 + \sum_{i=1}^{p} w_i x_i \]

\( w_0 \in \mathbb{R}, \ w \in \mathbb{R}^p \)
Model learning from sparse data

**POLYNOMIAL REGRESSION (d=2)**

Often interested in feature interactions (e.g. how diagnostic conditions **JOINTLY** affect the target prediction)

\[
y(x) = w_0 + \sum_{i=1}^{p} w_i x_i + \sum_{i=1}^{p} \sum_{j \geq i}^{p} W_{i,j} x_i x_j
\]

- **linear terms**
- **non-linear terms**

\[
w_0 \in \mathbb{R}, \ w \in \mathbb{R}^p, \ W \in \mathbb{R}^{p \times p}
\]
Model learning from sparse data

**POLYNOMIAL REGRESSION** (d=2)

**OFTEN INTERESTED IN FEATURE INTERACTIONS**
(e.g. how diagnostic conditions **JOINTLY** affect the target prediction)

\[
y(x) = w_0 + \sum_{i=1}^{p} w_i x_i + \sum_{i=1}^{p} \sum_{j \geq i} W_{i,j} x_i x_j
\]

\[\begin{align*}
\text{linear terms} & \\
\text{non-linear terms} &
\end{align*}\]

\[w_0 \in \mathbb{R}, \quad w \in \mathbb{R}^p, \quad W \in \mathbb{R}^{p \times p}\]

1) **VERY LARGE NUMBER OF PARAMETERS**
Model learning from sparse data

**Polynomial Regression (d=2)**

Often interested in feature interactions (e.g., how diagnostic conditions *jointly* affect the target prediction)

\[
y(x) = w_0 + \sum_{i=1}^{p} w_i x_i + \sum_{i=1}^{p} \sum_{j \geq i}^{p} W_{i,j} x_i x_j
\]

\[
\begin{align*}
\text{linear terms} & \\
\text{non-linear terms}
\end{align*}
\]

\[
\begin{align*}
w_0 & \in \mathbb{R}, \ w \in \mathbb{R}^p, \ W \in \mathbb{R}^{p \times p}
\end{align*}
\]

1) **Very large number of parameters**

e.g. 1000 features => more than 1Mil parameters!
Model learning from sparse data

**POLYNOMIAL REGRESSION** (d=2)

**Often interested in feature interactions**
(e.g. how diagnostic conditions **JOINTLY** affect the target prediction)

\[ y(\mathbf{x}) = w_0 + \sum_{i=1}^{p} w_i x_i + \sum_{i=1}^{p} \sum_{j \geq i}^{p} W_{i,j} x_i x_j \]

\[ w_0 \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^p, \mathbf{W} \in \mathbb{R}^{p \times p} \]

2) **Sparse dataset** (=> **few available features / patient**)
Model learning from sparse data

**Polynomial Regression** (d=2)

- Often interested in feature interactions (e.g. how diagnostic conditions **jointly** affect the target prediction)

\[ y(x) = w_0 + \sum_{i=1}^{p} w_i x_i + \sum_{i=1}^{p} \sum_{j \geq i}^{p} W_{i,j} x_i x_j \]

- Linear terms
- Non-linear terms

\[ w_0 \in \mathbb{R}, \ w \in \mathbb{R}^p, \ W \in \mathbb{R}^{p \times p} \]

2) Sparse dataset (=> Few available features / patient)

Estimate each interaction parameter independently from limited data
Model learning from sparse data

**Polynomial Regression (d=2)**

Often interested in feature interactions (e.g. how diagnostic conditions jointly affect the target prediction)

\[
y(x) = w_0 + \sum_{i=1}^{p} w_i x_i + \sum_{i=1}^{p} \sum_{j \geq i}^{p} W_{i,j} x_i x_j
\]

- Linear terms
- Non-linear terms

\[
w_0 \in \mathbb{R}, w \in \mathbb{R}^p, W \in \mathbb{R}^{p \times p}
\]

2) Sparse dataset (=> Few available features / patient)

Estimate each interaction parameter independently from limited data
Factorization Machines [Rendle 2010, 2012]
Factorization Machines [Rendle 2010, 2012]

Very successful in predicting your next movie ratings!
Factorization Machines (d=2) [Rendle 2010, 2012]

\[ y(x) = w_0 + \sum_{i=1}^{p} w_i x_i + \sum_{i=1}^{p} \sum_{j>i}^{p} \langle V_i, V_j \rangle x_i x_j \]

- Represent each feature via a vector (length k) => low-rank matrix V
- Interaction between features i, j expressed through: \( \langle V_i, V_j \rangle \)
- Weights are now dependent: \( \langle V_i, V_j \rangle, \langle V_i, V_l \rangle \) share \( V_i \)
- Thus: estimation of one interaction aids in estimating related interactions
Factorization Machines (d=2) [Rendle 2010, 2012]

\[
w_0 \in \mathbb{R}, \; \mathbf{w} \in \mathbb{R}^p, \; \mathbf{V} \in \mathbb{R}^{p \times k}
\]

\[
y(\mathbf{x}) = w_0 + \sum_{i=1}^{p} w_i x_i + \frac{1}{2} \sum_{f=1}^{k} \left( \left( \sum_{i=1}^{p} V_{i,f} x_i \right)^2 - \sum_{i=1}^{p} V_{i,f}^2 x_i^2 \right)
\]

- Represent each feature via a vector (length k) => low-rank matrix \( \mathbf{V} \)
- Interaction between features \( i, j \) expressed through: \( \langle \mathbf{V}_i, \mathbf{V}_j \rangle \)
- Weights are now dependent: \( \langle \mathbf{V}_i, \mathbf{V}_j \rangle, \langle \mathbf{V}_i, \mathbf{V}_l \rangle \) share \( \mathbf{V}_i \)
- Thus: estimation of one interaction aids in estimating related interactions

- Flexible form allows **LINEAR** computation of model equation in both \( k \) and \( p \)
Factorization Machines (d=2) [Rendle 2010, 2012]

\[ w_0 \in \mathbb{R}, w \in \mathbb{R}^p, V \in \mathbb{R}^{p \times k} \]

\[ y(x) = w_0 + \sum_{i=1}^p w_i x_i + \frac{1}{2} \sum_{f=1}^k \left( \left( \sum_{i=1}^p V_{i,f} x_i \right)^2 - \sum_{i=1}^p V_{i,f}^2 x_i^2 \right) \]

- Represent each feature via a vector (length k) \( \Rightarrow \) low-rank matrix \( V \)
- Interaction between features \( i, j \) expressed through: \( \langle V_i, V_j \rangle \)
- Weights are now dependent: \( \langle V_i, V_j \rangle, \langle V_i, V_l \rangle \) share \( V_i \)
- Thus: estimation of one interaction aids in estimating related interactions

- Flexible form allows \textbf{LINEAR} computation of model equation in both \( k \) and \( p \)
- By construction: \#params \( \ll p^2 \) of Polynomial Regression!
Experiments

- Dataset released from Practice Fusion
- Task: diabetes classification
- 9948 patients, 1904 of them are diabetic
- 702 FEATURES: Diagnoses (first-3 ICD9 digits), sex, year of birth
- Sparse dataset: 1% non-zero values

- libFM package for Factorization Machines
- MCMC as the learning method (requires the least number of hyperparameters)
- Competing methods (Matlab built-in): SVM with gaussian kernel, Random Forests
- All parameters selected via cross-validation
## Experiments - Prediction Accuracy

<table>
<thead>
<tr>
<th></th>
<th>F1-Score</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factorization Machines</td>
<td>0.3302</td>
<td>0.8187</td>
<td>0.5656</td>
<td>0.2334</td>
<td>0.8048</td>
</tr>
<tr>
<td>SVM</td>
<td>0.2046</td>
<td>0.8125</td>
<td>0.5438</td>
<td>0.1263</td>
<td>0.7598</td>
</tr>
<tr>
<td>Random Forests</td>
<td>0.2069</td>
<td>0.8178</td>
<td>0.6232</td>
<td>0.1242</td>
<td>0.8048</td>
</tr>
</tbody>
</table>

Results on detecting Type 2 Diabetes Mellitus, 5-fold CV, low rank of FMs k=4

- **FM**s outperform state-of-the-art classifiers in terms of F1-Score
- Approx. same performance with best alternative over Accuracy, AUC
Average precision of FM of increasing model complexity (varying $k = \{1, 2, 4, 8, 16, 32, 64\}$)
Experiments - Scalability to the model dimensionality

Effect of increasing model complexity (varying $k = \{1, 2, 4, 8, 16, 32, 64\}$) on the FM training time
Experiments - Scalability to the sample size

Effect of increasing dataset size on the FM training time (fixed \( k = 4 \) )
Concluding remarks - Future work

- **Factorization Machines**: efficient model learning method for **sparse** EHR datasets
- **Healthcare data are inherently sparse**: patients have to visit physician so that a diagnosis is recorded

- **Next**: incorporation of more data sources
- **Take domain knowledge into account**