Core Lecture 4: Big Educational Data

TECHNOLOGIES FOR LEARNING

Sergey Sosnovsky
Big Data is Big

- Over the last 2 years we have created more data than over the entire human history before that
- 10 zettabytes ($10^{22}$ bytes) of data were created in 2015
- Market prediction:
  - $203$ billion in revenue in 2020
  - In 2018, shortage of $\sim 150,000$ data analysts in US alone

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.
Big Educational Data
Learning Analytics: Data Sources

- Student-generated data
  - Browsing learning material
  - Taking tests
  - Forum discussions,
  - HWs, videos,…

- Context of the student
  - Location, type of access, calendar, email exchange, …

- Overall academic profile
  - Demographics (gender, age, …), program of study, academic records

- Evaluation Data
  - Course evaluation, other surveys and questionnaires

- Course meta-data
  - Structure, goals, learning resources and their parameters

- Performance, Assessment and Grades
<table>
<thead>
<tr>
<th>What is it good for</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery of items, users and usage patterns...</td>
</tr>
<tr>
<td>Prediction of events, behavior...</td>
</tr>
<tr>
<td>Detections of parameters, categories, traits...</td>
</tr>
<tr>
<td>Evaluation of systems, content...</td>
</tr>
<tr>
<td>Optimization of models, designs...</td>
</tr>
<tr>
<td>Personalization of learning experience...</td>
</tr>
<tr>
<td>Visualization of ... data</td>
</tr>
</tbody>
</table>
Learning Analytics: Users

- Teachers
- Administration
- Parents
- Learners
Early Warning Systems

• Identifying at-risk students
  ◦ Failing a course
  ◦ Dropping from a school
  ◦ Not finishing a MOOC

• Mostly rely on data from LMS

• Classification
  ◦ Analyze existing logs of filed students
  ◦ Train classifier
  ◦ Use it to predict failure of current students
Learning Dashboards

See where students are struggling.

See which students are practicing.
Course analysis

- Hard vs. Easy course
- Guys vs. Gals

Introductory STEM courses
Course analysis for students
Natural Language Processing

- Out of the box: Sentiment, bag-of-words groupings, reading level
- Pairs, triads, sentence structures
- Topic modeling
- Custom dictionaries by discipline, content area, etc.
Learning Analytics: Ethical Concerns

- Privacy
  - do all student appreciate their data to be collected
  - do they agree that it is manipulated, reasoned upon and used to draw conclusions about them

- Transparency
  - Who can see the data
  - How much students need (have to) know

- Consent
  - What kind of consent students need to give
  - From a legal/ethical/pragmatic perspective

- Location and security of data
  - Where the data/model can/should be stored and how they should be protected
  - Organization/Country?
  - Ethic vs. Legal vs. Capacity
Learning Analytics: Ethical Concerns

• Ownership
  ◦ Who owns the data? The models?
  ◦ Who decides how they can be used?
  ◦ Can students require to be “forgotten”? Is it technically possible?
  ◦ Can data sets be shared across organizations?
  ◦ Can they be used for public research?

• Misinterpretation and Accountability
  ◦ What if the data is misinterpreted by a teacher
  ◦ What if a model makes a wrong prediction
  ◦ What if the model output is misinterpreted by a student

• Obligation to act (from another perspective)
  ◦ Once we have information, are we obliged to act?
  ◦ Or even, once we have the capacity to collect data, interpret it, make predictions, are we obliged to do so?
<table>
<thead>
<tr>
<th></th>
<th>LA(K)</th>
<th>EDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery</td>
<td>Leveraging human judgement is key; automated discovery is a tool to</td>
<td>Automated discovery is key; leveraging human judgment is a</td>
</tr>
<tr>
<td></td>
<td>accomplish this goal</td>
<td>tool to accomplish this goal</td>
</tr>
<tr>
<td>Reduction &amp;</td>
<td>emphasis on understanding systems as wholes, in their full complexity</td>
<td>emphasis on reducing to components and analyzing individual</td>
</tr>
<tr>
<td>Holism</td>
<td></td>
<td>components and relationships between them</td>
</tr>
<tr>
<td>Origins</td>
<td>stronger origins in semantic web, information visualization, practical</td>
<td>strong origins in educational software and student modeling</td>
</tr>
<tr>
<td></td>
<td>pedagogy</td>
<td></td>
</tr>
<tr>
<td>Adaptation &amp;</td>
<td>Greater focus on informing and empowering instructors and learners</td>
<td>Greater focus on automated adaption (by the computer with</td>
</tr>
<tr>
<td>Personalization</td>
<td></td>
<td>no human in the loop)</td>
</tr>
<tr>
<td>Techniques &amp;</td>
<td>Social network analysis, sentiment analysis, discourse analysis,</td>
<td>Classification, clustering, Bayesian modeling, relationship</td>
</tr>
<tr>
<td>Methods</td>
<td>learner success prediction, sense-making models</td>
<td>mining</td>
</tr>
</tbody>
</table>
Educational Data Mining - Prediction

**Prediction**

- **Classification**
  - predicted variable is binary or categorical
  - Predict the moment of learning

- **Regression**
  - predicted variable is continuous
  - Predict student’s course grade based on their work with an ITS

- **Latent knowledge estimation**
  - Special type of a classifier
  - (example later)

- Develop a model which can infer a single aspect of the data
- Model created for a small data set and then applied on larger scale
Educational Data Mining – Structure Discovery

**Structure discovery**

- Look for structure in the data without an a priori idea
- No specific variable of interest (different from prediction)

**Clustering**
- Group data points into a set of categories
  - Discover groups of students
  - Categorize student behavior into successful/unsuccessful strategies

**Factor analysis**
- Split a set of variables based on hidden factors
  - Dimensionality reduction, e.g. identify key successful design choices made by ITS developers

**Social network analysis**
- Model relations & interactions between individual actors
  - Differences between effective and ineffective project groups
  - Changes in communication behavior over time

**Domain structure discovery**
- Determine the structure of a domain model
  - How specific content maps to specific knowledge components
Educational Data Mining – Relationships Mining

**Relationship Mining**

- **Association Rule Mining**
  - Discover If-then rules of variables co-occurrence
  - e.g. patterns of successful performance in an engineering simulation to make suggestions to students having problems

- **Correlation Mining**
  - find positive or negative linear correlations between variables
  - e.g. correlations between features of the design of ITS lessons and students’ prevalence of gaming the system

- **Sequential Pattern Mining**
  - find temporal associations between events
  - determine what path of student collaboration behaviors leads to a more successful eventual group project

- **Causal Data Mining**
  - find whether one event was the cause of another
  - e.g. determine which factors will lead a student to do poorly in a class

**Discover connections between variables in a data set with a large number of variables**

**Most popular subfield**
EDM: Learning Curve Analysis

• Learning curve / power law of practice
  ◦ Originally for motor skills: The logarithm of the reaction time for a particular task decreases linearly with the logarithm of the number of practice trials taken (motor tasks)
  ◦ Extended to cognitive skills: The probability of making a mistake in applying a skill decreases as a power function of the number of attempts to apply this skill
EDM: Learning Curve Analysis

- **Method:** build learning curves from system logs to examine the underlying domain model

- **Evaluation of the quality of the domain model:**
  - Refining knowledge components

- **Evaluation of the quality of the learning tasks:**
  - Filter out tasks that introduce noise (too difficult / too easy)

- **Evaluation of the quality of the instructional design:**
  - Optimize access to learning tasks based on their difficulty
EDM: Student Model Predictive Validity

• Student modeling as a classification task
  ◦ SM should be able to predict correct/incorrect application of knowledge components, i.e. classify attempts into successes and failures

• Truth table

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>True positive</td>
</tr>
<tr>
<td>Negative</td>
<td>False negative</td>
</tr>
</tbody>
</table>

• Precision = TP/ (TP+FP)
• Recall(Sensitivity, TPR) = TP/(TP+FN)
• Fall-out (1-specificity, FPR) = FP / (FP+TN)
• Accuracy = (TP + TN) / (TP + TN + FP + FN)
• F-measure = ( Precision * Recall ) / ( Precision + Recall )
EDM: Student Model Predictive Validity

• Method:
  ◦ compare at each learning step the SM prediction with an actual outcome to populate truth table
  ◦ compute accuracy measures
  ◦ compare with alternatives