Data Mining

CS57300
Purdue University

January 17, 2018
Goals

• Introduce a variety of Data Mining applications

• Explain some of the principles behind today’s Web
The Anatomy of the Today’s Web

User interaction

These days, even walking around is a form of interaction

Data

Predictions/Descriptions

Sells

Mining

insurance, advertisers, etc.
Descriptive vs. predictive modeling

• Descriptive (generative) models **summarize** the data
  • Provide insights into the domain
  • Focus on modeling joint distribution $P(X)$ or $P(X, Y)$
  • May be used for classification, but prediction is not the primary goal

• Predictive models **predict** the value of one variable of interest given known values of other variables
  • Focus on modeling the conditional distribution $P(Y | X)$ or on modeling the decision boundary for $Y$
Example: SPAM

- I was reading a little more about Tsalling entropy and trying to figure out whether it would be appropriate for relational learning problems. One possibility is to use it for exponential random graph models, which have features like the number of triangles in the graph. Since these grow with graph size, it seems to be an "extensive" property that the Tsalling entropy is trying to model...

- Don't Be Silly To Pay Hundred$ Or Thousand$ You Can Have The Exactly Same Licensed Software At 5%-10% Of The Retail Price : All popular softwares for PC & MAC : Language available: English, Deutsch, French, Italian, Spanish : Buy & start downloading right after you paid : You will be given your dedicated PERSONAL LICENSE right after you paid....
Data representation

- Class label: isSpam {+,-}

- Attributes?
  - Convert email text into a set of attributes

<table>
<thead>
<tr>
<th>isSpam</th>
<th>word\textsubscript{1}</th>
<th>word\textsubscript{2}</th>
<th>word\textsubscript{3}</th>
<th>...</th>
<th>word\textsubscript{n}</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Predictive modeling

• Data representation:
  • Training set: Paired attribute vectors and class labels \((x_i, y_i)\)
    or \(n \times p\) tabular data with class label \(y\) and \(p-1\) attributes \(x\)

• Task: estimate a predictive function \(f(x; W) = y\)
  • Assume that there is a parametric function \(y = f(x; W)\) that maps data instances \(X\) to class labels \(Y\) using function parameters \(W\)
  • Construct a model that approximates the mapping
    • Classification: if \(y\) is categorical
    • Regression: if \(y\) is real-valued
Classification

• In its simplest form, a classification model defines a decision boundary \( (h) \) and labels for each side of the boundary

• Input: \( x = \{ x_1, x_2, ..., x_n \} \) is a set of attributes, function \( f \) assigns a label \( y \) to input \( x \), where \( y \) is a discrete variable with a finite number of values
Classification output

- Different classification tasks can require different kinds of output
  - Each requires progressively more accurate models (e.g., a poor probability estimator can still produce an accurate ranking)

- Class labels — Each instance is assigned a single label
  - *Model only need to decide on crisp class boundaries*

- Ranking — Instances are ranked according to their likelihood of belonging to a particular class
  - *Model implicitly explores many potential class boundaries*

- Probabilities — Instances are assigned class probabilities $p(y|x)$
  - *Allows for more refined reasoning about sets of instances*
**Discriminative classification**

- Model the decision boundary directly
- Direct mapping from inputs $\mathbf{x}$ to class label $y$
- No attempt to model probability distributions
- May seek a discriminant function $f(\mathbf{x}; \mathbf{W})$ that maximizes measure of separation between classes
- Examples:
  - Perceptrons, nearest neighbor classifiers, support vector machines, decision trees
Probabilistic classification

- Model the underlying probability distributions
  - Posterior class probabilities: $p(y|x)$
  - Class-conditional and class prior: $p(x|y)$ and $p(y)$
- Maps from inputs $x$ to class label $y$ indirectly through posterior class distribution $p(y|x)$
- Examples:
  - Naive Bayes classifier, logistic regression, linear regression, most neural networks classification/regression tasks
Examples: Predictive/Descriptive(Generative) Models

• **Classification/Regression task**
  • Given an example \((x_i, y_i)\)
  • Wants to learn relationship between \(X\) and \(Y\) (often a probability distribution \(p(Y | X)\))
  • To use it, we need \(x_i\) and the output is the predicted class:
    \[
    \hat{y}_i = \arg \max_y p(y | x_i)
    \]

• **Generation (descriptive) task** (“closer to real intelligence”)
  • Given an example \((x_i, y_i)\)
  • Wants to learn joint probability \(p(y_i, x_i)\)
  • To use it, we sample another example from \(p(y, x)\), the output is an entirely new example (it understands “behavior” and can simulate it)
Example Classification Task: Retail & Healthcare

• Classification (drug safe or not safe, user buys or does not buy)

DrugBank dataset: http://www.drugbank.ca/downloads

• Descriptive vs Predictive
  • Tylenol and Advil are good for pain (descriptive)
  • Drug X will reduce fever Y by at least 5% (predictive)
Click-through Prediction Task:
Google News predicts probability you will click to read the news

- Google News
  - Ranked list of news from highest probability of clicking to lowest
  - “Filter bubble”

Ripple co-founder loses $44 billion on paper during cryptocurrency crash
CNBC • 4h ago

Related Coverage
China Escalates Crackdown on Cryptocurrency Trading
Highly Cited • Bloomberg • Jan 15, 2018

Woe Is Me? Dow Gains 320 Points, Nasdaq Surpasses Dot-Com High as Melt Up Resumes
Barron's • 37m ago

Related Coverage
Markets Live: ASX steps higher after Wall Street rally
Live Updating • The Sydney Morning Herald • 27m ago

Report: GM and Waymo lead driverless car race; Tesla lags far behind
Ars Technica • 3h ago
Example Regression Task:

• Obama’s 2008 campaign online effort first of its kind

• Try different strategies to get more donations

• Given website layout, predict wording that gives largest average campaign contribution

Ack: Dan Siroker
optimizely.com
Regression Task (cont)

- In 2016 candidates did similar experiments

- Politician statements are similarly tested
  - No surprises on the impact on likely voters
All Predictive Models Must be Tested

- **Hypotheses** are tentative statements of the expected relationships between two or more variables

- Formulate null and alternative hypothesis
  - $H_0$: Angry Trump donations $=$ Calm Trump donations
  - $H_1$: Angry Trump donations $\neq$ Calm Trump donations

- Gather a sample statistic (e.g., $\mu$=estimate of Angry Trump donations)

- Determine the sampling distribution for the statistic under the null hypothesis

- Use the sampling distribution to calculate the probability of obtaining the observed value of $\mu$, given $H_0$
  - If the probability is low, reject $H_0$ in favor of $H_1$
Two Types of Hypothesis (Model) Testing

- **Offline Testing**
  - Test if predictions are accurate over held out data (data not used to train the prediction model)
  - Why can’t we use the original (training) data, from which we constructed our prediction model?
    - Most prediction methods are not robust to overfitting
  - Examples of techniques:
    - Boosting
    - Cross-validation

- **Online Testing**
  - We “test” predictions in the live system
  - It is not quite a test, because we don’t really care which hypothesis (model) is more accurate
  - Examples:
    - Reinforcement learning
    - Multi-armed Bandits
Select 500 users to see headline chosen by model A

- Titanic Sinks

Select 500 users to see headline chosen by model B

- Ship Sinks Killing Thousands

We often refer to decision of choosing A or B as choosing an action (or arm)

Do people click more on headline of models A or B?

- If action A much better than action B, we are wasting users 500 users on a bad model... can we do better?
Truth is...

- Sometimes we don’t only want to quickly find whether hypothesis (model) A is better than hypothesis (model) B
- We really want to use the best-looking hypothesis (model) at any point in time
- Deciding if $H_0$ should be rejected is irrelevant
Real-world Problem

• Websites in perpetual state of testing

• Goal:
  Acquire just enough information about suboptimal action (headline) to ensure they are suboptimal. Looking for action (headline) $i$ of user $k$ that maximizes $E[X^{(i)}_k]$

$$X^{(i)}_k = \begin{cases} 
  1 & \text{, if } k\text{-th user seeing headline } i \text{ clicks} \\
  0 & \text{, otherwise}
\end{cases}$$

(A) Titanic Sinks

$$X^{(1)}_k = \begin{cases} 
  1 & \text{, with probability } p_1 \\
  0 & \text{, otherwise}
\end{cases}$$

(B) Ship Sinks Killing Thousands

$$X^{(2)}_k = \begin{cases} 
  1 & \text{, with probability } p_2 \\
  0 & \text{, otherwise}
\end{cases}$$

© 2016 Bruno Ribeiro
Role of Dependencies in Data
Relationship in Data

- Data is rich and sometimes dependent
  - Examples \((x_i, y_i)\) and \((x_j, y_j)\) may have dependencies
    - Say, i and j are friends on Facebook. X is a set of observable online behaviors and Y is whether they vote for the same party
  - Many very important prediction tasks using dependent data are related to graphs
Product Recommendation

Example of predicting links in a bipartite graph
Twitter’s Who to Follow

Another example of link prediction application
Graph-based Prediction Tasks also Find Fraud

- Detecting Fraud

Review Fraud?

*Buy Amazon Reviews*

**Never has it been easier to get multiple 4 and 5 star reviews on your Amazon product page. We provide real reviews from aged accounts with real buying activity. Most products in the Amazon marketplace will never even be seen. The more positive reviews you have the better your chances are.**

**Outrank the Competition**

Amazon is a highly competitive marketplace. Standing out from the crowd is the only way you are going to sell products. When you buy Amazon reviews you can rest assured that your product listing will rank higher in searches, get more attention and most importantly make more sales.

**We Review Any Product**

We utilize the talents of a professional and diverse writing team. You can buy Amazon reviews for any type of product. We write reviews for music, eBooks, supplements, cosmetics etc. We won't just copy reviews from elsewhere and rewrite them. Your reviews will be 100% unique.