Outline of Today

• Introduction
  – Thorsten Joachims + Joshua Moore

• Overview of Class Topics
  – Structured Prediction
  – Machine Learning with Humans in the Loop
  – Learning Representations

• Administrivia
  – Goals for the Class
  – Pre-Requisites
  – Credit Options and Format
  – Project and Assignments
  – Course Material
  – Warm-up Assignment
  – Contact Info
Structured Output Prediction
Conventional Supervised Learning

- Find function from input space $X$ to output space $Y$

$$h: X \rightarrow Y$$

such that the prediction error is low.

Microsoft announced today that they acquired Apple for the amount equal to the gross national product of Switzerland. Microsoft officials stated that they first wanted to buy Switzerland, but eventually were turned off by the mountains and the snowy winters…
Examples of Complex Output
Spaces

• Natural Language Parsing
  – Given a sequence of words \( x \), predict the parse tree \( y \).
  – Dependencies from structural constraints, since \( y \) has to be a tree.

\[
\begin{align*}
\text{The dog chased the cat} &
\end{align*}
\]
Examples of Complex Output Spaces

- Part-of-Speech Tagging
  - Given a sequence of words $x$, predict sequence of tags $y$.
  - Dependencies from tag-tag transitions in Markov model.

$\begin{align*}
x & \quad \text{The rain wet the cat} \\
y & \quad \text{Det$\rightarrow$N$\rightarrow$V$\rightarrow$Det$\rightarrow$N}
\end{align*}$

$\rightarrow$ Similarly Named-Entity Recognition, Protein Intron Tagging, etc.
Examples of Complex Output Spaces

• Protein Sequence Alignment
  – Given two sequences \( x=(s,t) \), predict an alignment \( y \).
  – Structural dependencies, since prediction has to be a valid global/local alignment.

\[
\begin{align*}
x & = (ABJLHBNJYAUGAI) \\
st & = (BHJKBNYGU)
\end{align*}
\]

\[
\begin{align*}
y & = AB-JLHBNJYAUGAI \\
& \quad | \quad | \quad | \quad | \quad | \\
& \quad BHJK-BN-YGU
\end{align*}
\]
Examples of Complex Output
Spaces

• Noun-Phrase Co-reference
  – Given a set of noun phrases $x$, predict a clustering $y$.
  – Structural dependencies, since prediction has to be an equivalence relation.
  – Correlation dependencies from interactions.

The policeman fed the cat. He did not know that he was late.
The cat is called Peter.

The policeman fed the cat. He did not know that he was late.
The cat is called Peter.
Examples of Complex Output

- Multi-Label Classification
  - Given a (bag-of-words) document $x$, predict a set of labels $y$.
  - Dependencies between labels from correlations between labels (“iraq” and “oil” in newswire corpus)

Due to the continued violence in Baghdad, the oil price is expected to further increase. OPEC officials met with …

$y$

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>-1</th>
<th>-1</th>
<th>+1</th>
<th>+1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>antarctica</td>
<td>benelux</td>
<td>germany</td>
<td>iraq</td>
<td>oil</td>
</tr>
<tr>
<td></td>
<td>coal</td>
<td>trade</td>
<td>acquisitions</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Examples of Complex Output Spaces

- Information Retrieval
  - Given a query \( x \), predict a ranking \( y \).
  - Dependencies between results (e.g. avoid redundant hits)
  - Loss function over rankings (e.g. AvgPrec)

\[ x \xrightarrow{SVM} y \]

1. Kernel-Machines
2. SVM-Light
3. Learning with Kernels
4. SV Meppen Fan Club
5. Service Master & Co.
6. School of Volunteer Management
7. SV Mattersburg Online
...

Why is Structured Output Prediction Interesting?

- **Application Perspective**
  - Many interesting real-world problems have structure in outputs

- **Research Perspective**
  - Like a multi-class problem with exponentially many classes!
    - How to predict efficiently?
    - How to learn efficiently?
    - Potentially huge models!

\[ \begin{align*}
  y^1 & : \text{VP} \rightarrow \text{S} \rightarrow \text{VP} \rightarrow \text{NP} \\
  y^2 & : \text{NP} \rightarrow \text{S} \rightarrow \text{VP} \rightarrow \text{NP} \\
  y^k & : \text{VP} \rightarrow \text{NP} \rightarrow \text{NP} \rightarrow \text{NP}
\end{align*} \]

The dog chased the cat
Overview: Structured Output Prediction

• Existing methods and their properties / limitations
  – Generative models
  – Structural SVMs and other maximum margin methods
  – Conditional Random Fields
  – Search-based methods
  – Gaussian Processes
  – Kernel Dependency Estimation

• Applications
  – Search engines
  – Natural language processing
  – Reinforcement learning
  – Probabilistic reasoning
  – Computational biology
Topic 2

Machine Learning with Humans in the Loop
Interactive Learning Systems

• WHILE(forever)
  – “System” presents options to the user
  – User examines the “Options” and reacts to them
  – “System” observes the selection and learns from it
• “System” / “Options” =
  – Search engine / search results
  – Movie recommender system / recommended movies
  – Online shopping site / products to buy
  – GPS navigation software / route
  – Spelling correction in word processor / word
  – Social network extension / friend
  – Twitter / post
Implicit Feedback in Web Search

- Observable actions
  - Queries / reformulations
  - Clicks
  - Order, dwell time
  - Etc.

- Implicit feedback
  - Personalized
  - Democratic
  - Timely
  - Human intelligence
  - Cheap
  - Abundant
Does User Behavior Reflect Retrieval Quality?

User Study in ArXiv.org

- Natural user and query population.
- User in natural context, not lab.
- Live and operational search engine.
- Ground truth by construction

**Orig ⊃ Swap2 ⊃ Swap4**
- **Orig**: Hand-tuned fielded
- **Swap2**: Orig with 2 pairs swapped
- **Swap4**: Orig with 4 pairs swapped

**Orig ⊃ Flat ⊃ Rand**
- **Orig**: Hand-tuned fielded
- **Flat**: No field weights
- **Rand**: Top 10 of Flat shuffled

[Radlinski et al., 2008]
## Absolute Metrics: Metrics

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Aggregation</th>
<th>Hypothesized Change with Decreased Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abandonment Rate</td>
<td>% of queries with no click</td>
<td>N/A</td>
<td>Increase</td>
</tr>
<tr>
<td>Reformulation Rate</td>
<td>% of queries that are followed by reformulation</td>
<td>N/A</td>
<td>Increase</td>
</tr>
<tr>
<td>Queries per Session</td>
<td>Session = no interruption of more than 30 minutes</td>
<td>Mean</td>
<td>Increase</td>
</tr>
<tr>
<td>Clicks per Query</td>
<td>Number of clicks</td>
<td>Mean</td>
<td>Decrease</td>
</tr>
<tr>
<td>Max Reciprocal Rank*</td>
<td>1/rank for highest click</td>
<td>Mean</td>
<td>Decrease</td>
</tr>
<tr>
<td>Mean Reciprocal Rank*</td>
<td>Mean of 1/rank for all clicks</td>
<td>Mean</td>
<td>Decrease</td>
</tr>
<tr>
<td>Time to First Click*</td>
<td>Seconds before first click</td>
<td>Median</td>
<td>Increase</td>
</tr>
<tr>
<td>Time to Last Click*</td>
<td>Seconds before final click</td>
<td>Median</td>
<td>Decrease</td>
</tr>
</tbody>
</table>

(*) only queries with at least one click count
Paired Comparisons: What to Measure?

Interpretation: $(r_1 > r_2) \iff \text{clicks}(r_1) > \text{clicks}(r_2)$
Balanced Interleaving

\[ f_1(u, q) \rightarrow r_1 \leftarrow \text{Interleaving}(r_1, r_2) \rightarrow f_2(u, q) \rightarrow r_2 \]

(u=tj, q=“svm”)

Interpretation: \((r_1 \succ r_2) \iff \text{clicks(topk}(r_1)) > \text{clicks(topk}(r_2))\) → see also [Radlinski, Craswell, 2012] [Hofmann, 2012]

Invariant: For all \(k\), top \(k\) of balanced interleaving is union of top \(k_1\) of \(r_1\) and top \(k_2\) of \(r_2\) with \(k_1=k_2 \pm 1\).

Model of User:
Better retrieval functions is more likely to get more clicks.

[Joachims, 2001] [Radlinski et al., 2008]
Arxiv.org: Interleaving Results

% wins ORIG

% wins RAND

Percent Wins

ORIG>FLAT  FLAT>RAND  ORIG>RAND  ORIG>SWAP2  SWAP2>SWAP4  ORIG>SWAP4
Issues in Learning with Humans

• Presentation Bias
  – Get accurate training data out of biased feedback
  – Use randomization to collect unbiased data
  – Experiment design

• Online Learning
  – Exploration/exploitation trade-offs
  – Observational vs. experimental data
  – Ability to run interactive experiments with users

• Measuring User Satisfaction
  – Turning behavior into evaluation measure
Overview: Learning with Humans

• Methods
  – Online learning and multi-armed bandits
  – Methods for interpreting user behavior
  – Matrix decomposition methods for recommendation
  – Active learning

• Applications
  – Information retrieval
  – Recommender systems
  – Online shopping
  – Mechanical turk
  – Web server usage
Topic 3

Learning Representations
Learning about Music

• Collection of songs
  – \( S = \{s_1, s_2, s_3, \ldots, s_N\} \)

• Example playlists
  – \( p_1 = [s_9, s_{527}, s_{12}, \ldots] \)
  – \( p_2 = [s_{7192}, s_{67}, s_{726}, \ldots] \)

• Goals
  – Automatically generate new playlists
  – Understand semantic space of songs
    • Query by tag, similar song
  – Visualization
Learning about Products

• Collection of products
  – $P = \{p_1, p_2, p_3, \ldots, p_N\}$

• Example browsing sequences
  – $s_1 = [p_9, p_{527}, p_{12}, \ldots]$  
  – $s_2 = [p_{7192}, p_{67}, p_{726}, \ldots]$ 

• Goals
  – Automatically recommend other items based on session prefix
  – Understand semantic space of products
    • Query by keyword, similar product
  – Visualization
Challenges and Approach

• Challenges
  – Items (i.e. songs, products) don’t have good features
    → We would like to generate “style” features
  – Number of items is large
    → Traditional sequence models (e.g., NLP) do not scale

• Approach
  – Model sequences as (k-th order) Markov model
    • \( P(s_1, ..., s_k) = \prod P(s_i | s_{i-1}) \)
  – Find model for transitions \( P(s_i | s_{i-1}) \) that
    • does not require \( N^2 \) storage.
    • generalizes well beyond the observed data.
Logistic Markov Embedding

• Model
  – Distance in space ~ transition probability

\[
\Pr(p[i] | p[i-1]) = \frac{e^{-\|X(p[i]) - X(p[i-1])\|^2_2}}{\sum_j e^{-\|X(p[j]) - X(p[i-1])\|^2_2}}
\]

• Training
  – Maximum likelihood
  – Stochastic gradient
  – O(n) iteration complexity
  \[\rightarrow O(1) \text{ iteration complexity}\]
Learning Song Positions

- Prince
- Tupac
- Garth Brooks
- Iron Maiden
- Metallica
- Michael Jackson
Learning Song Positions

Tupac
Prince
Iron Maiden
Metallica
Michael Jackson
Learning Song Positions

Tupac
Garth Brooks
Metallica
Prince
Iron Maiden
Michael Jackson
Learning Song Positions

- Tupac
- Garth Brooks
- Metallica
- Prince
- Iron Maiden
- Michael Jackson

Diagram showing the positions of various artists in a learning context.
Result: Song Embedding

- Garth Brooks
- Bob Marley
- The Rolling Stones
- Michael Jackson
- Lady Gaga
- Metallica
- T.I.
- All
Extension: Tag Model

Tag 1: Pop Music
Tag 2: 1980’s
Tag 3: Male vocals

Actual position of “Billie Jean”
Overview:
Learning Representations

• Methods
  – Embeddings based on sequence data
  – Embeddings based on co-occurrence data
  – Embeddings for bipartite graphs
  – Matrix factorization for rating data
  – Modeling structured objects
  – Modeling compositionality

• Applications
  – Playlist modeling
  – Natural language processing
  – Image search
  – Modeling human behavior
Outline of Today

• Introduction
• Overview of Class Topics
  – Structured Prediction
  – Machine Learning with Humans in the Loop
  – Learning Representations
• Administrivia
  – Goals for this class
  – Pre-Requisites
  – Credit Options and Format
  – Project and Assignments
  – Course Material
  – Warm-up Assignment
  – Contact Info
Goals for this Class

• Deepen your knowledge in three active research areas of ML
• Enable and improve your thesis research
• Practice being a successful academic

→ Class targeted towards current (or soon to be) PhD students!
Pre-Requisites

- This is not an introductory Machine Learning class!
- You need to satisfy one of the following ML pre-reqs:
  - Successfully taken CS4780 “Machine Learning”
  - Successfully taken CS6780 “Advanced Machine Learning”
  - Successfully taken a comparable “Intro to ML” class (*)
  - Acquired the equivalent ML knowledge in some other way (e.g. strong background in Statistics + ML textbook) (*)
- Basic probability and linear algebra
- Programming skills required for many projects

(*) means talk to me
Format of Class

• Lectures (by TJ)
  – Background material on general ML and 3 topics
• Research paper presentations (by students)
  – Reach current state of the art in each of 3 topics
• Project
  – Semester long, original research project
• Mock funding proposals
  – Develop your own research ideas for the 3 topics
• Peer reviewing
Research Paper Presentations

• Pair of students present the paper in class
  – Slide presentation
  – Create critique, extended bibliography, examples, demo software, experiments etc. that help understand the paper
  – Prepare discussion topics / group activity
  – Prepare quiz
  – Do dry-run of presentation in my office before class (30% of the grade).

• Everybody reads the paper in preparation for class
  – Quiz about each paper

• All students give feedback afterwards.
Mock Funding Proposals

• Write short funding proposal
  – Practice to develop your own research ideas and research plan
  – Practice to justify your research
  – Practice to convince others of your ideas

• Individual or group

• Peer reviewed
Project

• Full Semester Project
  – Topic of your choice that relates to CS6784
  – Scoped to be a publishable paper
  – Individual or group

• Timeline
  – 2/9: Proposal (10 %)
  – 3/16: First status report (10 %)
  – 4/20: Second status report (10 %)
  – 5/1-6: Project presentation (20 %)
  – 5/12: Final project report (50 %)

• At each step peer review
  – 5/18: Peer reviews due for project reports
Credit Options and Grades

• Letter grade:
  – project (40%)
  – paper presentation (20%)
  – in-class assignments and participation (15%)
    • three lowest grades dropped
  – funding proposals (12%)
  – peer reviewing (10%)
  – warm-up assignment (3%)

• Pass/Fail:
  – not allowed, unless you have very good arguments

• Audit:
  – not allowed, unless you have very good arguments
Course Material

• Background Reading

• Slides, Notes and Papers
  – Slides available on course homepage
  – Papers on course homepage
Warm-up Assignment

• Read the paper:

• Write a short paper that
  – is at most 800 words long
  – is submitted by Tuesday Jan 28 at 11:59pm EST

and that addresses the following questions:
  – What are the main original contributions described in this paper? Briefly describe the top 3 and argue why those are top.
  – For each original contribution, briefly describe in how far related ideas were already present in earlier papers.

• Peer review
How to Get in Touch

• Course Web Page

• Email
  – Thorsten Joachims: tj@cs.cornell.edu
  – Joshua Moore: jlmo@cs.cornell.edu

• Office Hours
  – Thursdays 3:00pm – 4:00pm, 418 Gates Hall

• Piazza
  – https://piazza.com/cornell/spring2014/cs6784

• Peer reviewing platform
  – TBA