# CS6784 - Advanced Topics in Machine Learning

**Spring 2010**

**Thorsten Joachims**
**Cornell University**

## Outline of Today

- **Introduction**
- **Overview of Class Topics**
  - Structured Prediction
  - Learning with Humans in the Loop
  - Understanding Archives
- **Administrivia**
  - Pre-Requisites
  - Credit Options and Format
  - Project
  - Course Material
  - Office Hours

## Topic 1

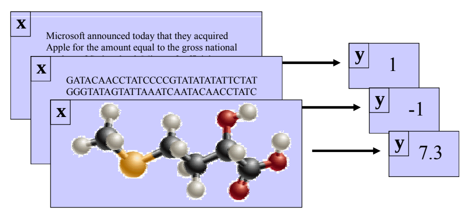
**Structured Output Prediction**

## Conventional Supervised Learning

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

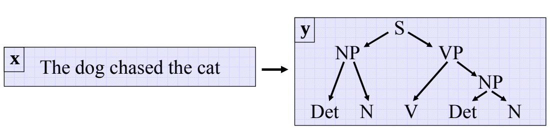
$$h : X \longrightarrow Y$$

**such that the prediction error is low.**



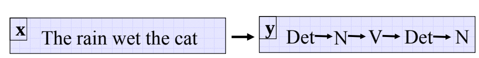

## Examples of Complex Output Spaces

- **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.




## 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.




→ 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.

**x**

```
s=(ABJLHBNJYAUGAI)
t=(BHJKBNYGU)
```

→

**y**

```
AB-JLHBNJYAUGAI
 |  ||  ||  |||
 BHJK-BN-YGU
```

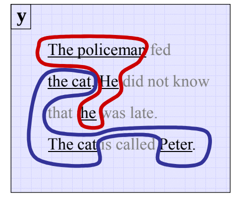
## Examples of Complex Output Spaces

- **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.

**x**

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

→




## Examples of Complex Output Spaces

- **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)

**x**

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

→

**y**

| | |
|---|---|
| -1 | antarctica |
| -1 | benelux |
| -1 | germany |
| +1 | iraq |
| +1 | oil |
| -1 | coal |
| -1 | trade |
| -1 | acquisitions |

## 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**

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

…

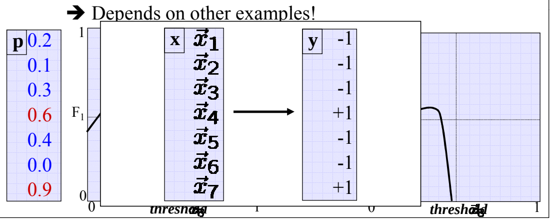
## Examples of Complex Output Spaces

- **Non-Standard Performance Measures (e.g. $F_1$-score, Lift)**
  - $F_1$-score: harmonic average of precision and recall

$$F_1 = \frac{2\,Prec\,Rec}{Prec + Rec}$$

  - New example vector $\vec{x}_8$. Predict $y_8=1$, if $P(y_8=1|\vec{x}_8)=0.4$?

    ➔ Depends on other examples!



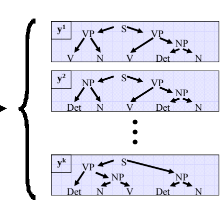

## Why is Structured Output Prediction Interesting?

- **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!

**x**

The dog chased the cat

→



## Overview: Structured Output Prediction

- **Definition of Problem**
- **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

**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

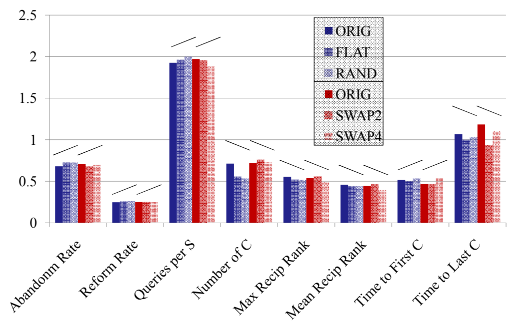
[Radlinski

## Absolute Metrics: Metrics

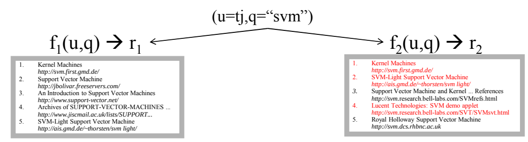
| Name | Description | Aggre-gation | Hypothesized Change with Decreased Quality |
|---|---|---|---|
| Abandonment Rate | % of queries with no click | N/A | Increase |
| Reformulation Rate | % of queries that are followed by reformulation | N/A | Increase |
| Queries per Session | Session = no interruption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

(*) only queries with at least one click count

## Absolute Metrics: Results

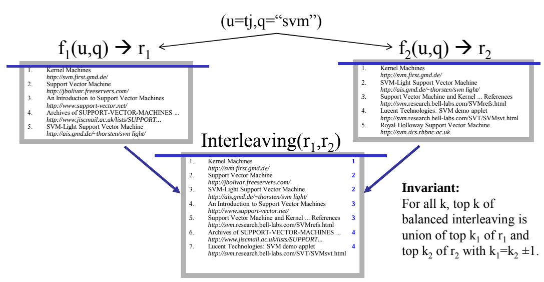



## Paired Comparisons: What to Measure?



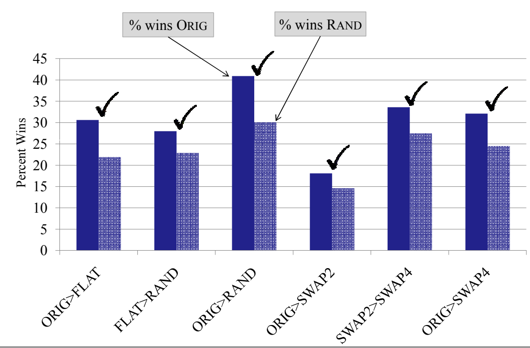

**Interpretation:** $(r_1 \succ r_2) \leftrightarrow \text{clicks}(r_1) > \text{clicks}(r_2)$

## Paired Comparisons: Balanced Interleaving




**Interpretation:** $(r_1 \succ r_2) \leftrightarrow \text{clicks}(\text{topk}(r_1)) > \text{clicks}(\text{topk}(r_2))$

## Balanced Interleaving: Results




## 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

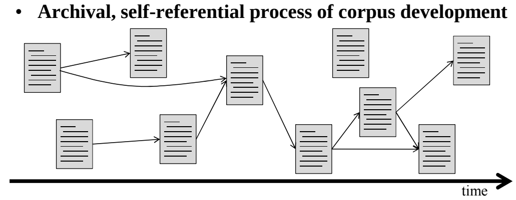
## Topic 3

**Understanding Archives**

## Archives

**Motivation: We now have more then >10 years of online**
- Newspaper archives
- Conference proceeding
- Personal email and photos
- Etc.

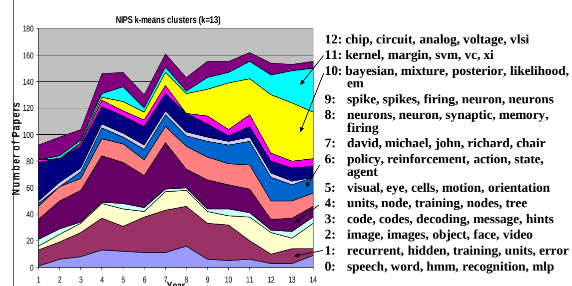
- **Archival, self-referential process of corpus development**




## ML Task: Information Genealogy

- **Task: Understand where information originates, how it spreads, and how information streams evolve over time**
  - How did the topics in the NIPS conference evolve and who were the most influential authors driving the change?
  - Did one news article influence another article?
  - Who are the bloggers that are ahead of the curve?
  - An automatic personal diary from email and photos.
  - Etc.

## Summarizing Temporal Development: Neural Information Processing Systems (NIPS) 1987 - 2000




[Benyah Shaparenko]

## Identifying Dependencies and Influence

Which papers were influenced by "Shrinking the Tube: a New Support Vector Regression Algorithm" written by B. Schoelkopf et al.?
- Assume unigram word distribution is mixture of past papers
- Likelihood ratio test for non-zero mixture weight (convex program)

| log(Λ(d)) | Cite? | Title and Authors |
|---|---|---|
| 321.2 | No | "Support Vector Method for Novelty Detection", B. Schoelkopf, R. Williamson, A. Smola, J. Shawe-Taylor, J. Platt. |
| 221.8 | Yes | "An Improved Decomposition Algorithm for Regression Support Vector Machines", Pavel Laskov. |
| 219.9 | Yes | "ν-arc: Ensemble Learning in the Presence of Outliers", G. Raetsch, B. Schoelkopf, A. Smola, K. Miller, T. Onoda, S. Mims. |
| 184.6 | No | "Fast Training of Support Vector Classifiers", F. Perez-Cruz, P. Alarcon-Diana, A. Navia-Vazquez, A. Artes-Rodriguez. |
| 168.9 | Yes | "Uniqueness of the SVM Solution", C. Burges, D. Crisp. |

[ShapaJo07]

## Identifying Key Documents: NIPS

| Score | Year | Cites | Paper Title and Authors |
|---|---|---|---|
| 1.167 | 1996 | 128 | "improving the accuracy and speed of support vector machines" by chris j.c. burges, b. schoelkopf |
| 1.128 | 1999 | 17 (466) | "using analytic qp and sparseness to speed training of support vector machines" by john c. platt |
| 0.986 | 1999 | 18 | "regularizing adaboost" by gunnar raetsch, takashi onoda, klaus-robert mueller |
| 0.953 | 1996 | 41 (3711) | "support vector method for function approximation, regression, and signal processing" by v. vapnik, s. golowich, a. smola |
| 0.945 | 1998 | 27 | "training methods for adaptive boosting of neural networks" by holger schwenk, yoshua bengio |
| 0.945 | 1997 | 3 | "modeling complex cells in an awake macaque during natural image viewing" by william e. vinje, jack l. gallant |
| 0.934 | 1998 | 17 | "em optimization of latent-variable density models" by chris bishop, markus svensen, chris william |
| 0.934 | 1995 | 584 | "a new learning algorithm for blind signal separation" by s. amari, a. cichocki, h. h. yang |

[ShapaJo07]

## Overview: Understanding Archives

- **Idea flow**
  - Dependencies between documents and authors
- **Temporal development of content**
  - Bursts and topic drift
- **Meta data and access data**
  - Using temporally grown link structure
  - Using access logs to identify relationships
- **Personal information management**
  - Desktop search
  - Photo archives

## Outline of Today

- **Introduction**
- **Overview of Class Topics**
  - Structured Prediction
  - Learning with Humans in the Loop
  - Understanding Archives
- **Administrivia**
  - Pre-Requisites
  - Credit Options and Format
  - Project
  - Course Material
  - Office Hours

## 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**
- **Research papers**
  - Everybody reads the paper in preparation for class
  - Some assignment (e.g. quiz, review, critique) about each paper
  - One student presents the paper in class
    - Slide presentation
    - Create examples, demo software, experiments etc. that help understand the paper
    - Prepare discussion topics

  → I'll give you feedback before your presentation
- **Project**

## Project

- **Full Semester Project**
  - Topic of your choice that relates to CS6784
  - Undergrad/MEng students: groups of 3-4
  - Ph.D. students: group or individual
- **Timeline**
  - 2/11: Proposal (10 %)
  - 3/18: First status report (10 %)
  - 4/20: Second status report (10 %)
  - 5/4-6: Project presentation (20 %)
  - 5/16: Final project report (50 %)

## Credit Options and Grades

- **Letter grade:**
  - project (50%)
  - paper presentation (25%)
  - assignments (15%)
  - discussion (10%)
- **Pass/Fail:**
  - paper presentation (50%)
  - assignments (30%)
  - discussion (20%)
- **Audit:**
  - not allowed, unless you have very good arguments

## Course Material

- **Background Reading**
  - T. Mitchell, "Machine Learning", McGraw Hill, 1997.
  - B. Schoelkopf, A. Smola, "Learning with Kernels", MIT Press, 2001. (online)
  - C. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.
  - R. Duda, P. Hart, D. Stork, "Pattern Classification", Wiley, 2001.
  - T. Hastie, R. Tishirani, and J. Friedman, "The Elements of Statistical Learning", Springer, 2001.
  - N. Cristianini, J. Shawe-Taylor, "Introduction to Support Vector Machines", Cambridge University Press, 2000. (online)
  - Ethem Alpaydin, "Introduction to Machine Learning", MIT Press, 2004.
- **Slides, Notes and Papers**
  - Slides available on course homepage
  - Papers on course homepage

## How to Get in Touch

- **Course Web Page**
  - http://www.cs.cornell.edu/Courses/cs6784/2010sp/
- **Email**
  - Thorsten Joachims: tj@cs.cornell.edu
- **Office Hours**
  - Tuesdays 4:00pm – 5:00pm, 4153 Upson Hall