Structured Data

Learning, prediction, dependency, testing
Lecture 1

Florence d'Alché-Buc, Slim Essid, Zoltan Szabo, Arthur Tenenhaus

Contact: florence.dalche@telecom-paristech.fr,
M2 Data Science, Télécom ParisTech & Ecole Polytechnique, Université of Paris-Saclay, France

Table of contents

Overview

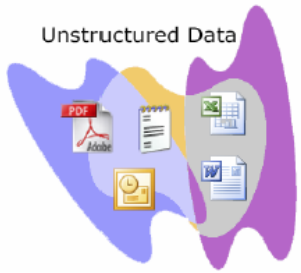
Motivation

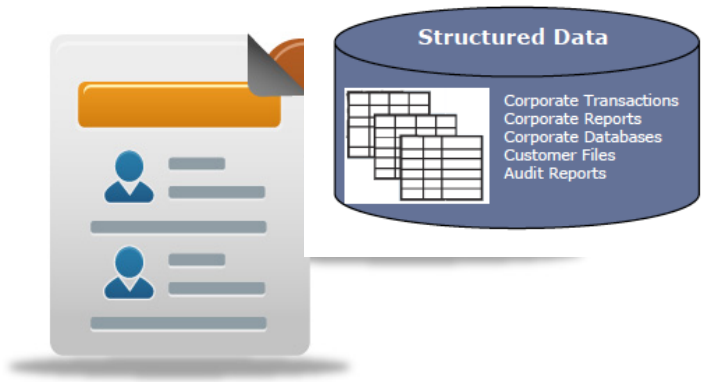
- Data as structured objects (sequences, trees, graphs) $\neq$ structured data in the sense of database
- Heterogeneous data coming from multiple sources
- A mix of structured and unstructured data
- Interdependent data

Patient records

- medical images
- biomedical signals
- results of medical exams
- symptoms measured by various sensors
- genotype
- transcriptomics

Client data

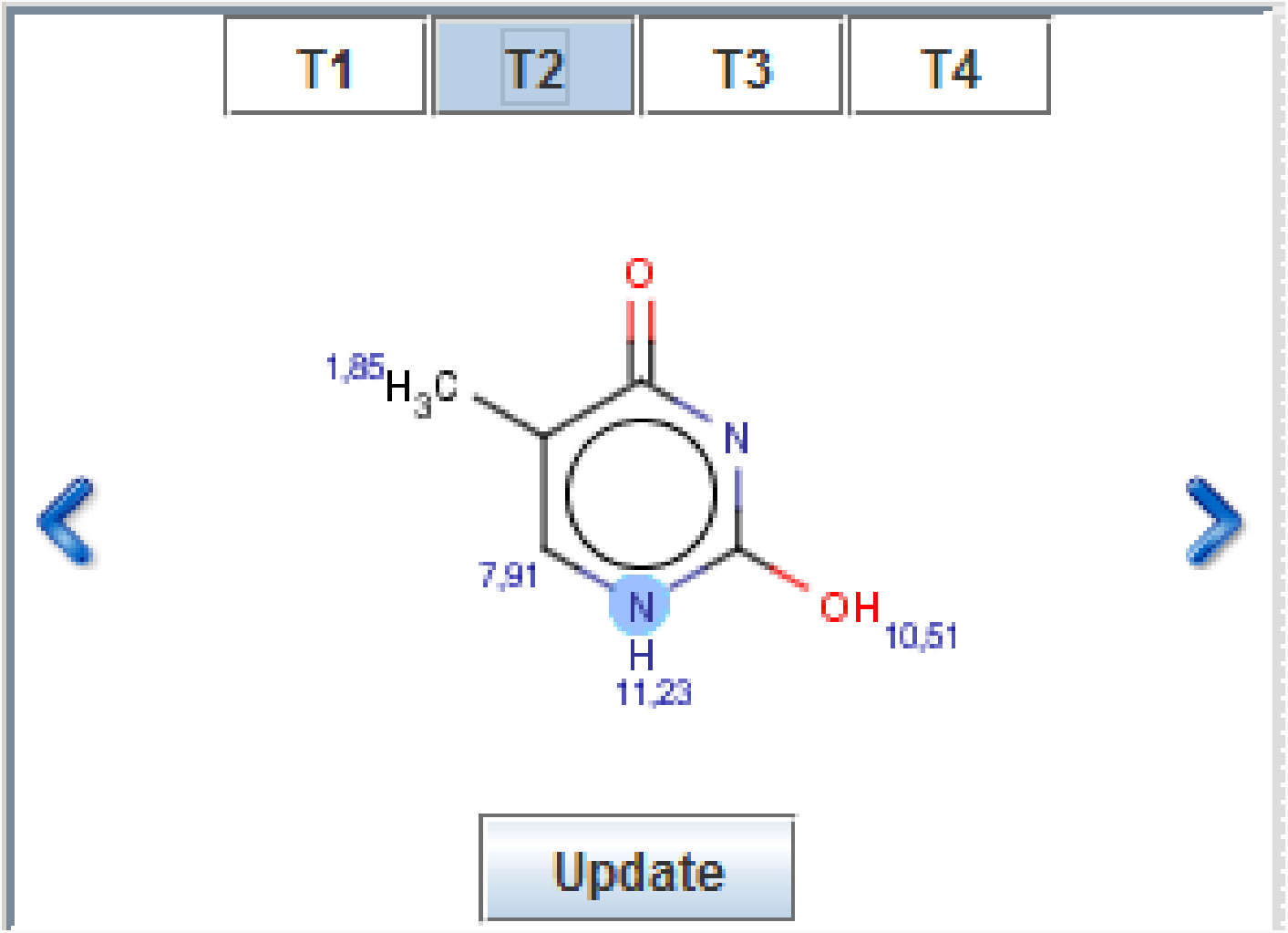




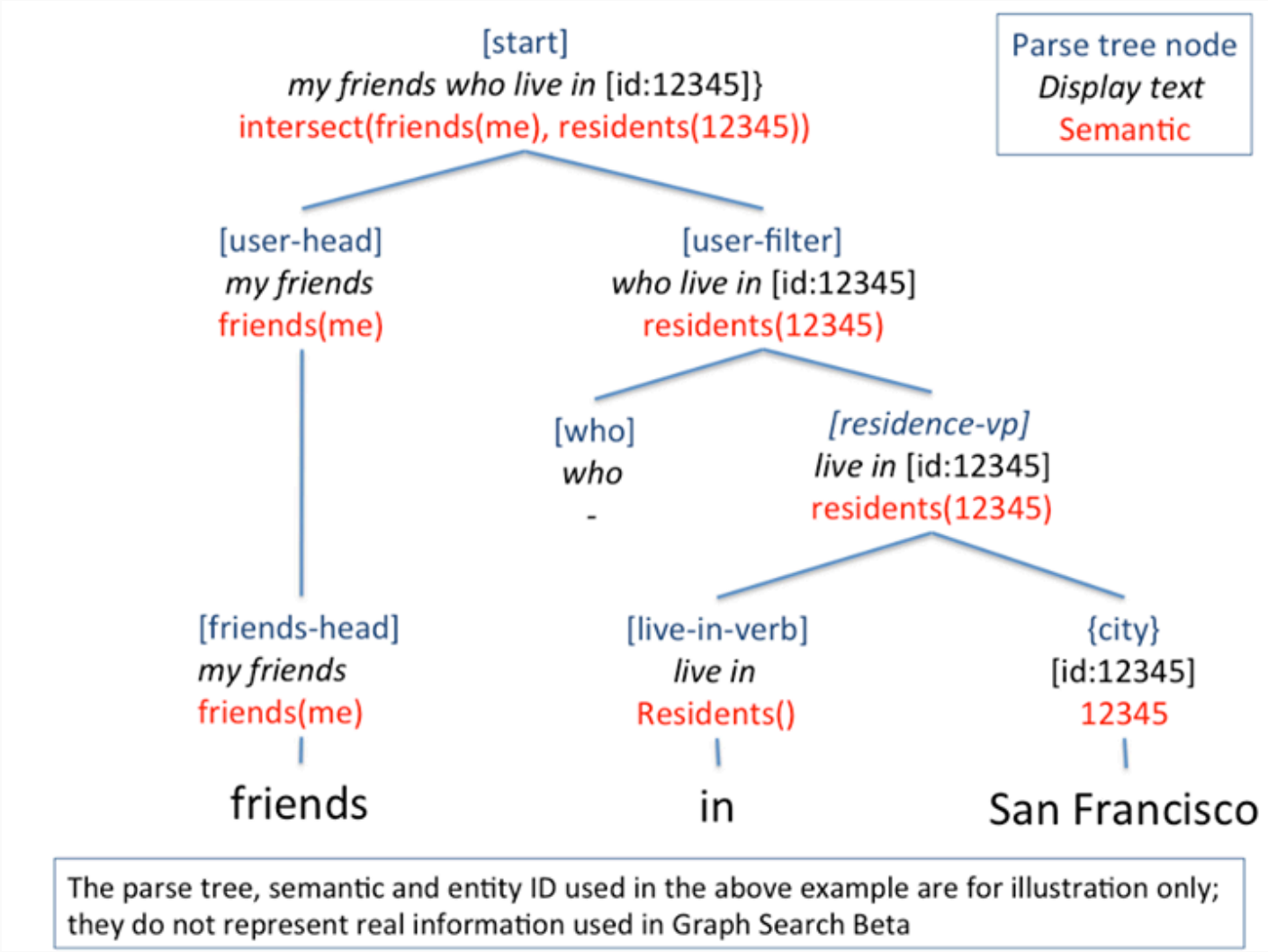
- documents, reports

Structured data

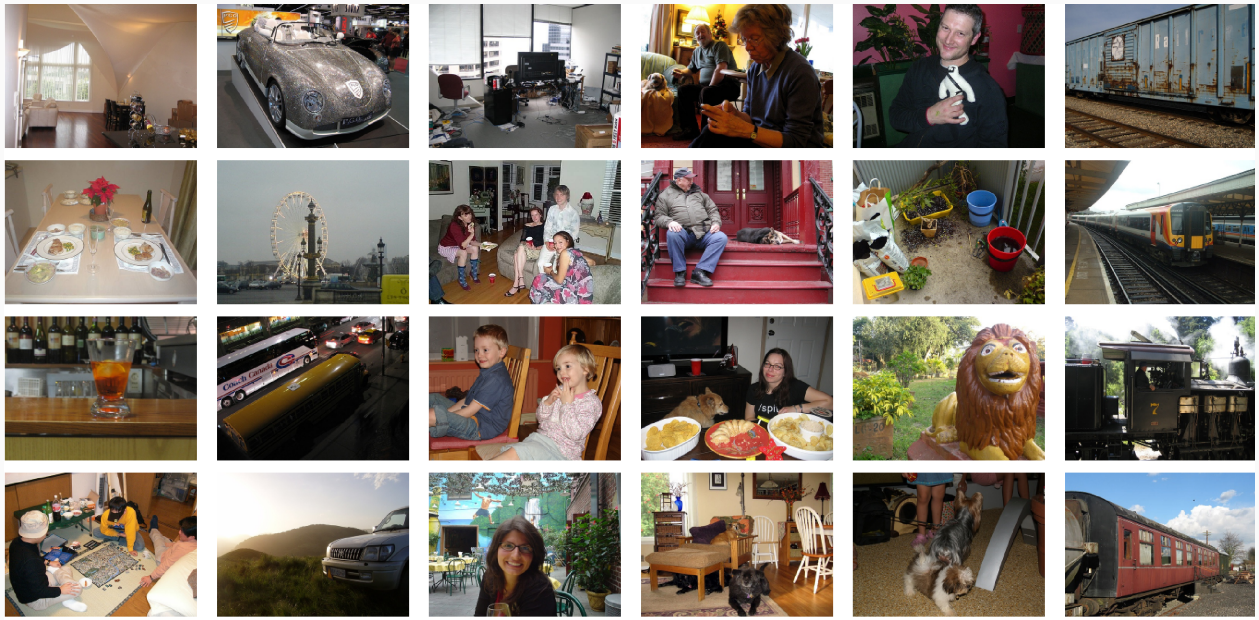
Structured data



Structured data



Unstructured data or … implicitely structured data



Unstructured data or … implicitely structured data

This method is naturally geared toward document-pivoted TC, since ranking the training documents for their similarity with the test document can be done once for all categories. For category-pivoted TC, one would need to store the document ranks for each test document, which is obviously clumsy; DPC is thus *de facto* the only reasonable way to use k-NN.

A number of different experiments (see Section 7.3) have shown k-NN to be quite effective. However, its most important drawback is its inefficiency at classification time: while, for example, with a linear classifier only a dot product needs to be computed to classify a test document, k-NN requires the entire training set to be ranked for similarity with the test document, which is much more expensive. This is a drawback of "lazy" learning methods, since they do not have a true training phase and thus defer all the computation to classification time.

6.9.1. Other Example-Based Techniques. Various example-based techniques have been used in the TC literature. For example, Cohen and Hirsh [1998] implemented an example-based classifier by extending standard relational DBMS technology with "similarity-based soft joins." In their WHIRL system they used the scoring function

$$CSV_i(d_j) = 1 - \prod_{d_z \in Tr_k(d_j)} (1 - RSV(d_j, d_z))^{\|\breve{\Phi}(d_z, c_i)\|}$$

as an alternative to (9), obtaining a small but statistically significant improvement over a version of WHIRL using (9). In their experiments this technique outperformed a number of other classifiers, such as a C4.5 decision tree classifier and the RIPPER CNF rule-based classifier.

A variant of the basic k-NN approach was proposed by Galavotti et al. [2000], who reinterpreted (9) by redefining

The difference from the original k-NN approach is that if a training document d_z similar to the test document d_j does not belong to c_i, this information is not discarded but weights *negatively* in the decision to classify d_j under c_i.

A combination of profile- and example-based methods was presented in Lam and Ho [1998]. In this work a k-NN system was fed *generalized instances* (GIs) in place of training documents. This approach may be seen as the result of

—clustering the training set, thus obtaining a set of clusters $K_i = \{k_{i1}, \ldots, k_{i|K_i|}\}$;

—building a profile $G(k_{iz})$ ("generalized instance") from the documents belonging to cluster k_{iz} by means of some algorithm for learning linear classifiers (e.g., Rocchio, WIDROW-HOFF);

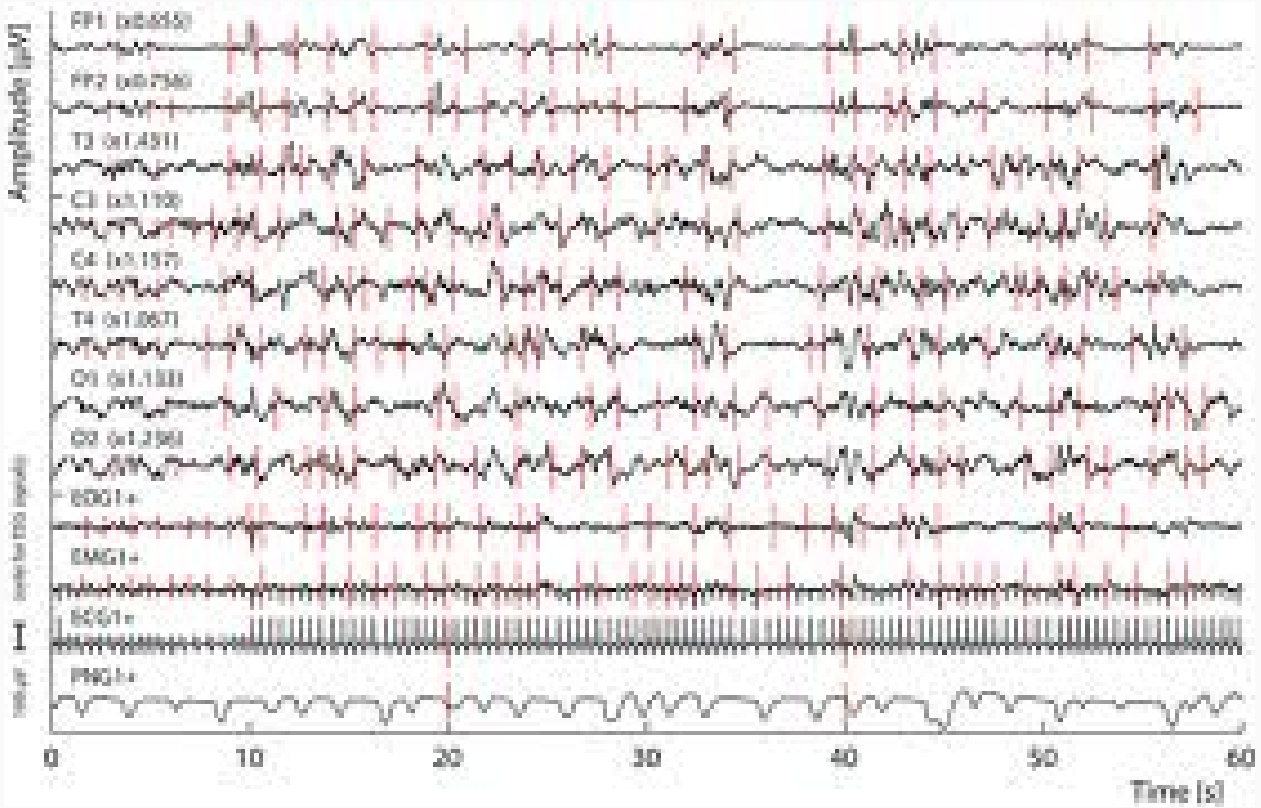
—applying k-NN with profiles in place of training documents, that is, computing

$$CSV_i(d_j) \stackrel{def}{=} \sum_{k_{iz} \in K_i} RSV(d_j, G(k_{iz})) \cdot \frac{|\{d_j \in k_{iz} \mid \breve{\Phi}(d_j, c_i) = T\}|}{|\{d_j \in k_{iz}\}|} \cdot \frac{|\{d_j \in k_{iz}\}|}{|Tr|} = \sum_{k_{iz} \in K_i} RSV(d_j, G(k_{iz})) \cdot \frac{|\{d_j \in k_{iz} \mid \breve{\Phi}(d_j, c_i) = T\}|}{|Tr|}, \quad (10)$$

where $\frac{|\{d_j \in k_{iz} \mid \breve{\Phi}(d_j, c_i) = T\}|}{|\{d_j \in k_{iz}\}|}$ represents the "degree" to which $G(k_{iz})$ is a positive instance of c_i, and $\frac{|\{d_j \in k_{iz}\}|}{|Tr|}$ represents its weight within the entire process.

This exploits the superior effectiveness (see Figure 3) of k-NN over linear classifiers while at the same time avoiding

Unstructured data or … implicitely structured data



Learning from structured data

- Prediction from structured data (not this course)
- Structured output prediction (this course)

Analyzing (structured) heterogeneous data

- How to extend Canonical Correlation Analysis to heterogeneous data ?
- How to extend independence test to heterogeneous data ?

in a single common framework ?

Outline of this course

- Part 1: Structured data: learning and prediction
  - Energy-based or scoring methods: Max margin methods, Conditional Random Fields, structured deep learning,...
  - Operator-valued Kernel Regression, Multi-task regression
  - Output Kernel Regression, pre-image problems
- Part 2: Structured data: Dependency and testing
  - Kernel canonical correlation analysis
  - Mean embedding, maximum mean discrepancy, integral probability metric, characteristic/universal kernel,
  - Hilbert-Schmidt independence criterion, covariance operator
  - Kernel based two-sample and independence tests.
  - Multiway data analysis, generalization of CCA and KCCA.

Outline of this course

- Part 1: Structured data: learning and prediction (Jan 9 to Feb 13)
- Part 2: Structured data: Dependency and testing (Feb 20 to March 20)

Schedule of this course

- Session 1 (9/01) - Introduction, beginning of maximum margin approaches, multi-class problems (FAB)
- Session 2 (16/01) - Conditional random fields (CRF; sequence labelling) (Slim Essid)
- Session 3 (23/01) - Lecture : End of maximum margin approaches, deep structured learning (FAB)
- Session 4 (30/01) - Datalab 1: Practice of CRF and M3N (Alexandre Garcia, Slim Essid)
- Session 5 (06/02) - Multi-task regression and operator-valued kernels for multi-task learning (FAB)
- Session 6 (13/02) - Output representations, pre-image problem (FAB)
- Session 7 (Datalab 2, not a Monday): operalib, structured variational autoencoder (Moussab Djerrab, Alexandre Garcia, Romain Brault)
- Session 8-9 (20/02, 27/02) - Kernel canonical correlation analysis, mean embedding, maximum mean discrepancy,

Concepts and important notions of this course

Present the key components used today to predict and analyze structured/heterogeneous data

- input/output feature vector: $\phi(x, y)$
- extension of the margin notion
- beyond likelihood maximization: cross-entropy, ...
- optimization for inference, optimization for prediction in graphical models
- extension of scalar-valued kernels
- kernel trick in the output space
- working in RKHS, distribution embedding in RKHS
- maximum mean discrepancy, Hilbert-Schmidt independence criterion
- ...

Evaluation of this course

- Project
  - 4 students per project (20 projects will be proposed, exceptionally you might propose your own project)
  - Target a paper among a list of proposed papers (often a pair of papers)
  - Give a critical analysis of the paper
  - Implement or improve existing code, apply on new datasets
  - Compare proposed methods with baselines
  - Paper analysis + Report (not a notebook) + code
  - Beginning: Jan 30
  - Deadline: March 31

How to prepare the project ?

- Attend courses and datalab
- Read carefully the proposed references
- import ans study the software libraries

Introduction to structured output prediction

Classic supervised machine learning / structured output learning

Classic machine learning:

$$\mathcal{X} \to \mathbb{R} \tag{1}$$

Structured output learning :

$$\mathcal{X} \to \mathcal{Y} \tag{2}$$

$\mathcal{Y}$: set of structured objects

Sequence labeling

- *INPUT:* a sequence of tokens
- *OUTPUT:* a sequence of labels (one par token)

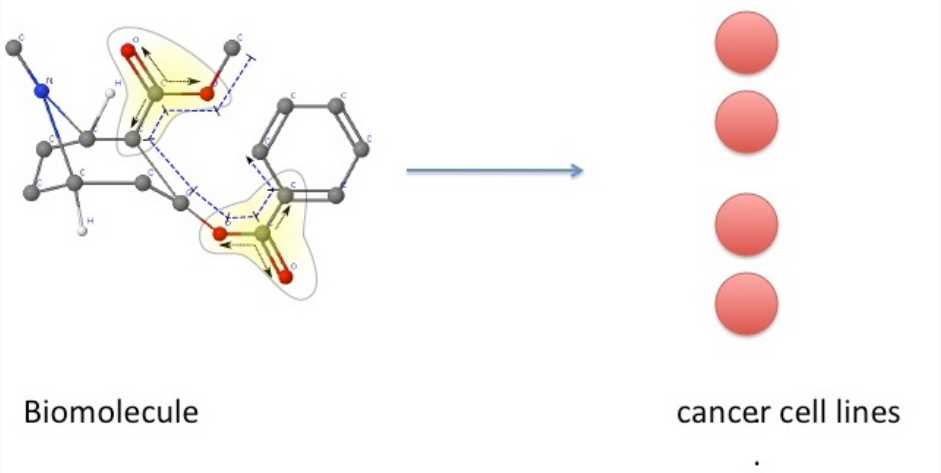
Example: Part-of-Speech tagging

Given a sentence, find parts-of-speech of all words.

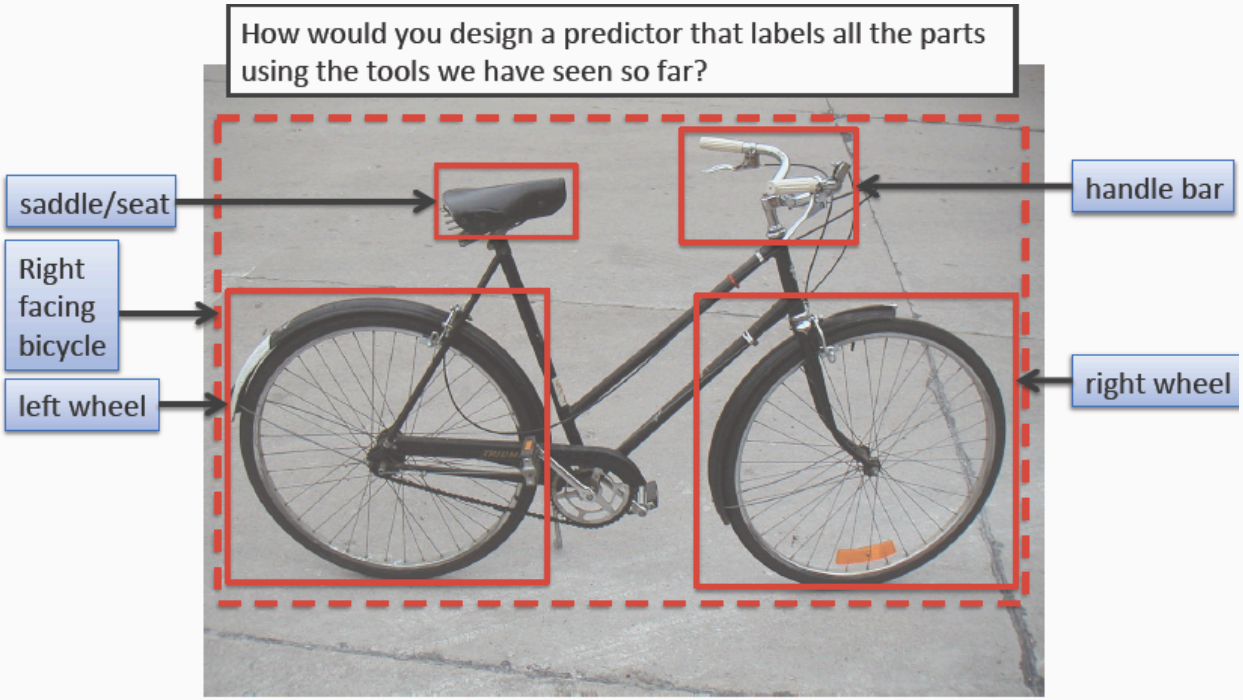
The	Fed	raises	interest	rates
Determiner	Noun	Verb	Noun	Noun
Other possible tags in different contexts,	Verb (I *fed* the dog)		Verb (Poems don't *interest* me)	Verb (He rates movies online)

Multi-task regression

- *INPUT:* an object
- *OUTPUT:* multiple interdependent outputs



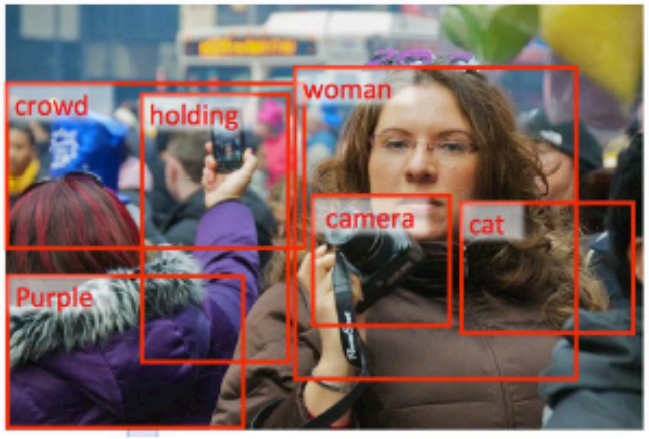
Object detection



- INPUT: a raw image
- OUTPUT : an object defined by several parts

Automatic image captioning

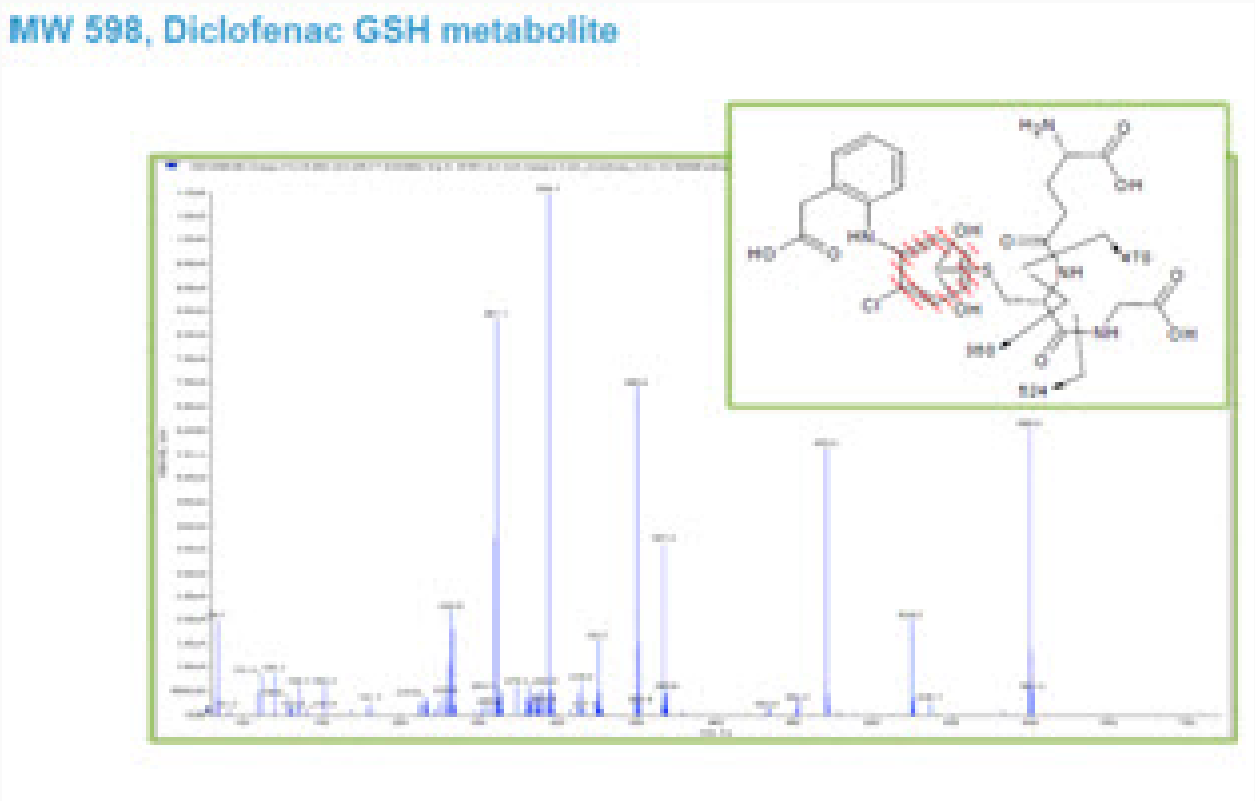
- INPUT: a raw image
- OUTPUT : a sentence (caption)



A woman holding a camera in the crowd

From mass spectrometry to metabolite

- INPUT: mass spectra
- OUTPUT : metabolite(s)



Difficulties of structured output prediction

- Need to compose several elements to make a prediction
- Interdependent outputs
- Huge set of possible outputs

Approaches to structured output prediction

- Scoring or energy-based approaches
- Regression + pre-image or output decoding

Supervised learning (reminder)

Probabilistic and statistical framework

Let X be a random vector $\mathcal{X} = \mathbb{R}^p$

X describes the properties of a message (say, features)

Let Y be a binary discrete variable $\mathcal{Y} = \{-1, 1\}$ (classification) or Y is a continuous random variable

Let P be the joint probability distribution of (X,Y), P is supposed to be fixed but unknown

Let $S_{train} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ be a i.i.d. sample from P.

Classic Supervised learning

Probabilistic and statistical framework (approach 1)

- $R_{emp}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(x_i))$
- From S_{train}, determine $h \in \mathcal{H}$ that minimizes $R_{emp}(h) + \lambda\Omega(h)$
- Classification : discriminant approach (aims at predicting the good class - but no modeling)
- ℓ: a local loss function that measures how the predicted class differs from the true class

Classic Supervised learning

Probabilistic and statistical framework (approach 2)

- Classification: From S_{train}, build an estimate of $P(Y = j|x)$ by modeling $p(x|Y = j)$
- Regression: From S_{train}, model $p(y|x)$ (regression)
- How ? for instance maximum likelihood, MAP, maximum entropy,. . .

Difficulties of structured output prediction

The function h to be learned is from $\mathcal{X} \to \mathcal{Y}$ where $\mathcal{Y}$ is more complex than $\mathbb{R}$ or $\{-1, +1\}$.

- Need to compose several elements to make a prediction
- Interdependent outputs
- Huge set of possible outputs

Approaches to structure output prediction (1)

Scoring / energy-based methods methods

Define g a score function that takes as inputs feature pairs on x and y:

$$f(x) = \arg\max_{y \in \mathcal{Y}} g(x, y) \tag{3}$$

$$g(x, y) \in \mathbb{R} \tag{4}$$

$$f(x) \in \mathcal{Y} \tag{5}$$

Learn f, e.g. learn g

Example of energy-based methods methods

Define g a score function that takes as inputs feature pairs on x and y:

$$f(x) = \arg\max_{y \in \mathcal{Y}} g(x, y) \tag{6}$$

$$f(x) \in \mathcal{Y} \tag{7}$$

Examples of g:

$$g(x, y) = w^T \phi(x, y)$$

$$g(x, y) = P(Y = y|x)$$

Scoring or energy-based methods

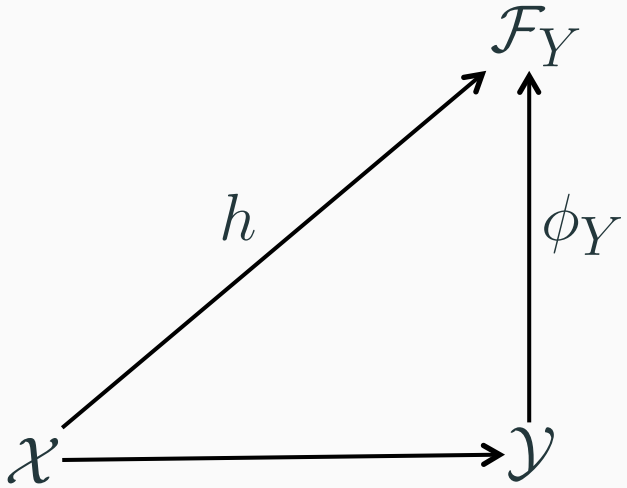
Difficulties

Prediction is expensive:

Very often, the optimization problem to solve is NP-hard

Prediction is called *inference* in the context of graphical models

Approaches to structured output prediction (2)



Output kernel regression methods

$$
\begin{aligned}
f(x) &= \phi^{-1}(h(x)) \\
h(x) &\in \mathcal{F}_y \\
f(x) &\in \mathcal{Y}
\end{aligned}
$$

Structured regression

Difficulties

Prediction requires to solve a pre-image problem except in some cases like structured multiple outputs, link prediction . . .

Predicted outputs are approximated

References (a few to begin with)

- Álvarez, M. A. and Rosasco, L. and Lawrence, N. D., Kernels for vector-valued functions: a review, Foundations and Trends in Machine Learning, 4:3,2012.
- Altun, Smola, Hofman, Exponetial families for conditional random fields, UAI 2004
- C. Brouard, F. d'Alché-Buc, M. Szafranski, Semi-supervised link prediction with penalized output kernel regression, ICML 2011
- Caponnetto, A. and Micchelli, C. A. and Pontil, M. and Ying, Y., Universal MultiTask Kernels, Journal of Machine Learning Research, 9,2008.
- Chen LC, Schwing AG, Yuille AL, Urtasun R. Learning deep structured models. InProc. ICML 2015.
- Collins, Michael, Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, 2002.
- Crammer, Koby and Singer, Yoram, On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines, J. Mach. Learn. Res., 3/1/2002

References continued...

- Daumé, Marcu, Learning as search optimization : approximate large margin structured prediction, ICML, 2005.
- Michelli and Pontil, M. On learning vector-valued functions, JMLR, 2005.
- Joder, Essid, S. and Richard, G., A conditional rendom field framework for robust audio-to-score matching, IEEE Trans. ASLP, 19(8), 2011.
- C. Sutton, McCallum, An introduction to Conditional Random Field for relational learning, `http://people.cs.umass.edu/~mccallum/papers/crf-tutorial.pdf`, Foundations and Trends in Machine Learning, to appear.
- Christoph H. Lampert, Matthew B. Blaschko, Structured Prediction by Joint Kernel Support Estimation Machine Learning, vol 77, number 2-3, pages 249-269, Springer, 2009
- Yann LeCun and Fu Jie Huang: Loss Functions for Discriminative Training of Energy-Based Models,AIStats'05, 2005
- Ben Taskar, Learning structured prediction models, a large margin approach, PhD thesis (`http://www.seas.upenn.edu/~taskar/pubs/thesis.pdf`), U. Pennsylvany, USA, 2004

Score-based methods: multiclass classification

From multi class classification to structured prediction

- Scoring functions:
  - Model of the form:$h(x) = \arg\max_{y \in \mathcal{Y}} score(x, y)$
- (1) Solve the problem for multiple classes (today's program)
- (2) Solve the problem in general for any structured prediction problem

Document classification

Example 1

- INPUT : ”. . . run a health care insurance program ...”
- OUTPUT : politics

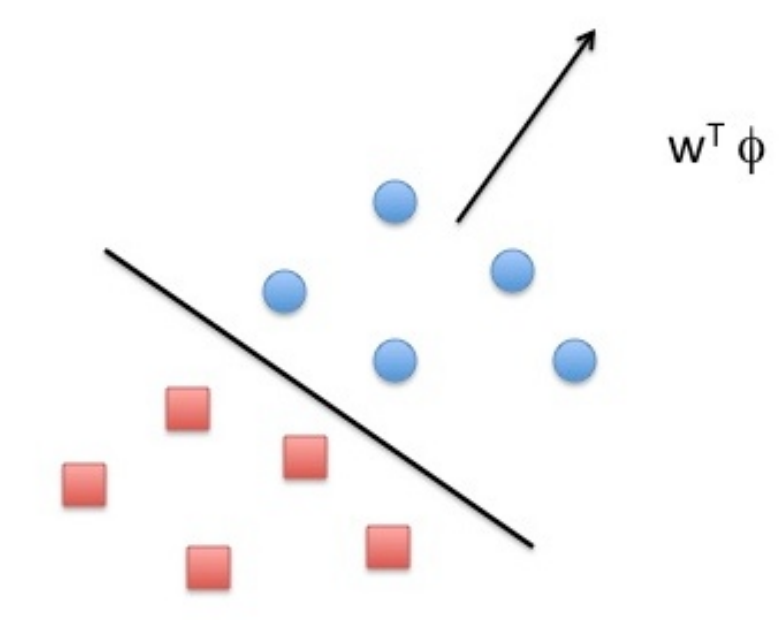
Example 2

- INPUT: ”. . . run the marathon ...”
- OUTPUT : sports

A binary classification task

Two classes: politics, sports
Using a linear model

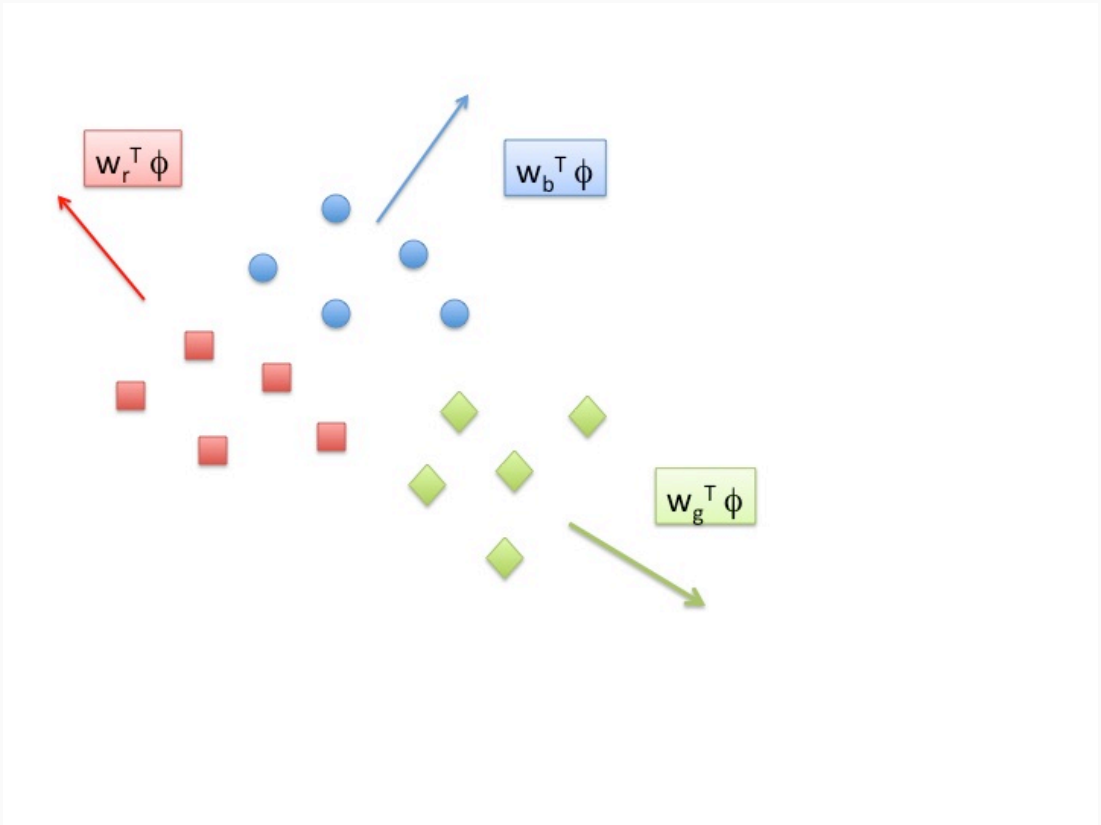
- Input features: for instance, bag of words
- Prediction: $h_w(x) = sgn(w^T \phi(x))$



Now a multiclass classification task

Multiple classes : economics, politics, sports

- Each class y defined by a linear model of the following form: $h_y(x) = w_y^T \phi(x)$



Linear models for Multiclass classification

With p classes:

$$\phi(x, y)^T = [0 \dots 0 \ \phi(x)^T \ 0 \dots 0]$$
$$w^T = [\mathbf{w}_1^T \dots, \mathbf{w}_y^T, \dots \mathbf{w}_p^T]$$

Remember : **here** y is a class label

$$score(x_i, y, w) = \mathbf{w}^T \phi(x_i, y) = \mathbf{w}_y^T \phi(x_i)$$

- Whatever y, w_y's have the same dimension, say p.
- The vector $\mathbf{w}$ is the stack of all $\mathbf{w}_y$ with $y \in$ the finite set $\mathcal{Y}$
- NB : we will note: $\phi(x_i, y) = \phi_i(y)$

Linear models for Multiclass classification

Scoring methods for multiclass classification

$$score(x_i, y, w) = \mathbf{w}^T \phi(x_i, y) = \mathbf{w}^T \phi_i(y)$$

$$prediction(x^i, w) = \arg\max_{y \in \mathcal{Y}} \mathbf{w}^T \phi_i(y)$$

Learning linear models: the perceptron rule

Simple discriminative method

$$
\begin{aligned}
y' &= \arg\max_y \mathbf{w}^T \phi_i(y) \\
\mathbf{w} &\leftarrow \mathbf{w} + \eta(\phi_i(y_i) - \phi_i(y'))
\end{aligned}
$$

Remember the idea: if there is a mistake, I add the right vector and susbtract the wrong vector.
Note that later we will use the following notation:
$\Delta_i(y') = (\phi_i(y_i) - \phi_i(y'))$

Learning linear models by minimizing a loss function

- What is a training error here ?
- $error = \sum_i step(\mathbf{w}^T \phi_i(y_i) - \max_{y \neq y_i} \mathbf{w}^T \phi_i(y))$
- with $step(z) = 1$ if $z < 0$ and 0, otherwise
  - zero-one loss : discontinuous, minimization is NP-complete
  - Turn to convexified losses

Maximum entropy (log loss)

- Posterior probabilities

$$P(y|x, \mathbf{w}) = \frac{\exp(\mathbf{w}^T \phi(x, y))}{\sum_{y'} \exp(\mathbf{w}^T \phi(x, y')}$$

- Maximize the log conditional likelihood of training data

$$\max_{\mathbf{w}} \log \prod_i P(y_i|x_i, \mathbf{w}) = \sum_i \log \left(\frac{\exp(\mathbf{w}^T \phi_i(y_i)}{\sum_y \exp(\mathbf{w}^T \phi_i(y))} \right)$$

$$\max_{\mathbf{w}} \sum_i (\mathbf{w}^T \phi_i(y_i) - \log \sum_y \exp(\mathbf{w}^T \phi_i(y)))$$

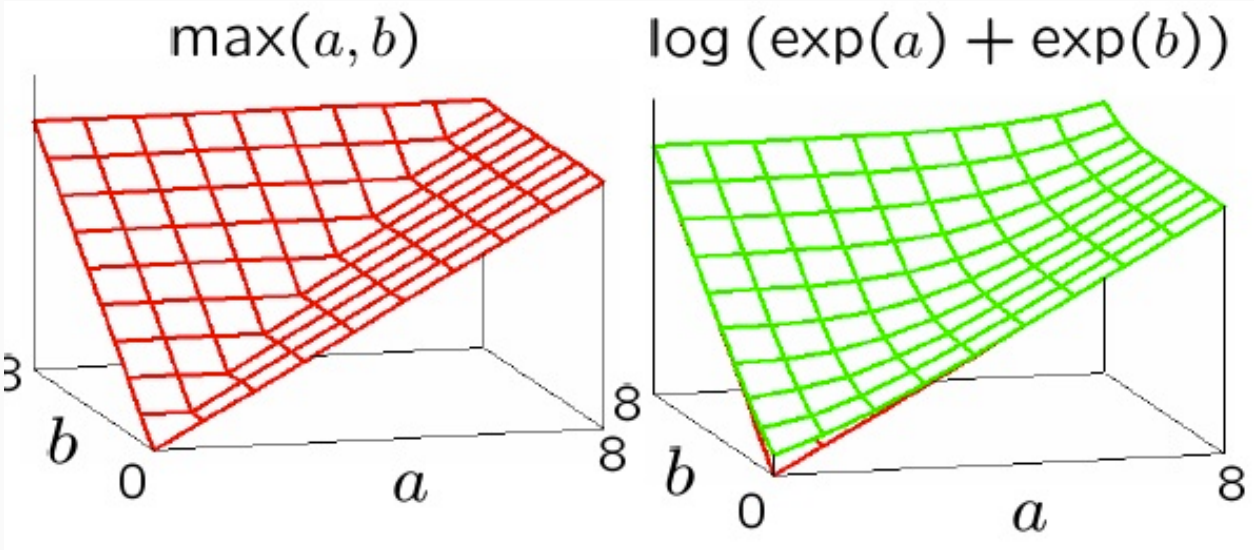
Maximum entropy with regularization

$$\max_{\mathbf{w}} \sum_i (\mathbf{w}^T \phi_i(y_i) - \log \sum_y \exp(\mathbf{w}^T \phi_i(y))) - \lambda ||\mathbf{w}||^2$$

equivalent to

$$\min_{\mathbf{w}} \lambda ||\mathbf{w}||^2 - \sum_i (\mathbf{w}^T \phi_i(y_i) - \log \sum_y \exp(\mathbf{w}^T \phi_i(y)))$$

soft-max



Now let us try to maximize a margin

If we just want to separate the data we would impose:

$$\forall i, \forall y \neq y_i, \mathbf{w}^T \phi_i(y_i) \geq \mathbf{w}^T \phi_i(y)$$

but we need to define what is a good separator ! solution: margin maximization

Now maximizing a margin

On our example:

$\mathbf{w}^T\phi(\textit{run the marathon}, \textit{sports}) \geq \mathbf{w}^T\phi(\textit{run the marathon}, \textit{politics}) + \gamma$

$\mathbf{w}^T\phi(\textit{run the marathon}, \textit{sports}) \geq \mathbf{w}^T\phi(\textit{run the marathon}, \textit{economics}) + \gamma$

$\mathbf{w}^T\phi(\textit{run the marathon}, \textit{sports}) \geq \mathbf{w}^T\phi(\textit{run the marathon}, \textit{sports})$

Margin maximization: Pb 1

$\max_{\|\mathbf{w}\| \leq 1} \gamma$

s.t. :

$\forall i, \forall y, \mathbf{w}^T\phi_i(y_i) \geq \mathbf{w}^T\phi_i(y) + \gamma\ell_i(y)$

$\mathbf{w} = \gamma\mathbf{u}$

$\gamma = \frac{1}{\|\mathbf{u}\|}$

With $\ell_i(y) = 0$ if $y = y_i$, 1 otherwise.

Minimizing the norm of "canonical hyperplane"

Pb1 is equivalent to :

$\max_{\|\gamma \mathbf{u}\| \leq 1} \frac{1}{\|\mathbf{u}\|^2}$

s.t. :

$\forall i, \forall y, \mathbf{u}^T \phi_i(y_i) \geq \mathbf{u}^T \phi_i(y) + \ell_i(y)$

which is equivalent to:

$\min_{\|\mathbf{u}\| \leq 1} \|\mathbf{u}\|^2$

s.t. :

$\forall i, \forall y, \mathbf{u}^T \phi_i(y_i) \geq \mathbf{u}^T \phi_i(y) + \ell_i(y)$

which is equivalent to :

$\min_{\mathbf{u}} \frac{1}{2} \|\mathbf{u}\|^2$

s.t. :

$\forall i, \forall y, \mathbf{u}^T \phi_i(y_i) \geq \mathbf{u}^T \phi_i(y) + \ell_i(y)$

Minimizing the norm of "canonical hyperplane"

$\min_{\mathbf{u}} \frac{1}{2} \|\mathbf{u}\|^2$
s.t. :
$\forall i, \forall y, \mathbf{u}^T \phi_i(y_i) \geq \mathbf{u}^T \phi_i(y) + \ell_i(y)$
Equivalent to:

Margin maximization: Pb 2

$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2$
s.t. :
$\forall i, \forall y, \mathbf{w}^T \phi_i(y_i) \geq \mathbf{w}^T \phi_i(y) + \ell_i(y)$

Allowing for non-separability (adding slack variables)

Margin maximization with slack variables: Pb 3

$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i$

s.t. :

$\forall i, \forall y, \mathbf{w}^T \phi_i(y_i) + \xi_i \geq \mathbf{w}^T \phi_i(y) + \ell_i(y)$

$\forall i, \xi_i \geq 0$

A min-max formulation (Tsochantaridis et al. 2005)

We solve ξ_i: $\forall i, \forall y, \xi_i \geq \mathbf{w}^T \phi_i(y) + \ell_i(y) - \mathbf{w}^T \phi_i(y_i)$

$\forall i, \xi_i = \max_y [\mathbf{w}^T \phi_i(y) + \ell_i(y)] - \mathbf{w}^T \phi_i(y_i)$

Pb 4

$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \max_y [\mathbf{w}^T \phi_i(y) + \ell_i(y)] - \mathbf{w}^T \phi_i(y_i)$

Compare max-margin and maxent (log-loss)

Maxent (in logistic regression)

$$\min_{\mathbf{w}} \lambda \|w\|^2 - \sum_i (\mathbf{w}^T \phi_i(y_i) - \log(\sum_i y \exp(\mathbf{w}^T \phi_i(y)))$$

SVM

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \max_y [\mathbf{w}^T \phi_i(y) + \ell_i(y)] - \mathbf{w}^T \phi_i(y_i)$$

Both try to make the true score better than a function of the other score

Solving the optimization pb

- Solving the problem in the dual using formulation of Pb3 is possible
- However there are many constraints (size $|\mathcal{Y}|$ n)
- Working set training (active constraints)
- Eventually : easy to kernelize (we'll see that later)

Algorithm

1. input: $(x_1, y_1), (x_n, y_n)$, C, ϵ
2. $S_i \leftarrow 0 \forall i$
3. repeat
   - for i=1,. . . , n
     - Define: H(y) = $\ell_i(y) - \mathbf{w}^T \phi_i(y_i) + \mathbf{w}^T \phi_i(y)$
     - with $\mathbf{w} := \sum_j \sum_{y' \in S_j} \alpha_{jy'} (\phi(x_j, y_j) - \phi(x_j, y))$
     - Compute $\hat{y} = \arg\max_{y \in \mathcal{Y}} H(y)$
     - Compute $\xi_i = \max(0, \max_{y \in S_i} H(y))$
     - If $H(\hat{y}) > \xi_i + \epsilon$ then
       - $S_i \leftarrow S_i \cup \{\hat{y}\}$
       - $\alpha_S \leftarrow$ optimize dual over $S = \cup S_i$
     - Endif
   - End for
4. until no S_i has changed during iteration

Tasks solved with this approach

: Tsochantaridis et al. ICML 2004.

- multi-class classification
- hierarchical/structured classification
- sequence labelling

Scoring methods for structured output prediction

Structured prediction

Let $p(x, y)$ be the unknown true distribution of the data
Let $\mathcal{S} = \{(x_i, y_i), i = 1, \ldots, n\}$ be iid sample from p.
|et $\phi : \mathcal{X} \to \mathbb{R}^p$ a feature function
Let $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a loss function

- Find a vector $\mathbf{w}^*$ that minimize the expected loss $E_{(x,y)}[\Delta(y, f(x))]$
- With $f(x) = \arg\max_{y \in \mathcal{Y}} g(x, y, \mathbf{w})$
- and $g(x, y, \mathbf{w}) =< \mathbf{w}, \phi(x, y) >$

Structured prediction

$Min_{\mathbf{w}} E_{(x,y)}[\Delta(y, \arg\max'_y g(x, y', w))]$ Two problems :

1. p is unknown as usual : need to define an empirical loss function and a penalty term
   - Solve instead: $min_{\mathbf{w}} \lambda \|\mathbf{w}\|^2 + \frac{1}{n} \sum_i \Delta(y_i, \arg\max_y g(x_i, y, \mathbf{w}))$
2. $\arg\max_y g(x_i, y, \mathbf{w})$ is discontinuous versus $\mathbf{w}$
   - Replace $\Delta(y, y')$ with a well-behaved $\ell(x, y, \mathbf{w})$
   - ℓ be chosen as an upper bound of Δ, continuous and convex

Structured output SVM

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{n} \sum_i \max_y [\mathbf{w}^T \phi_i(y) + \ell_i(y) - \mathbf{w}^T \phi_i(y_i)]$$

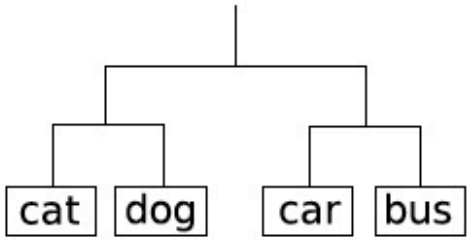
Structured output SVM for a hierarchy of classes

Hierarchical Multiclass Loss:

$$\Delta(y, y') := \frac{1}{2}(\text{distance in tree})$$

$$\Delta(\texttt{cat}, \texttt{cat}) = 0, \quad \Delta(\texttt{cat}, \texttt{dog}) = 1,$$

$$\Delta(\texttt{cat}, \texttt{bus}) = 2, \quad etc.$$



Solve: $$\min_{w,\xi} \frac{1}{2}\|w\|^2 + \frac{C}{N}\sum_{n=1}^{N} \xi^n$$

subject to, for $i = 1, \ldots, n$,

$$\langle w, \phi(x^n, y^n)\rangle - \langle w, \phi(x^n, y)\rangle \geq \Delta(y^n, y) - \xi^n \quad \text{for all } y \in \mathcal{Y}.$$

References for this lecture

- Crammer, Koby and Singer, Yoram, On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines, J. Mach. Learn. Res., 3/1/2002.
- Tsochantaridis, I. and Joachims, T. and Hofmann, T. and Altun, Y., Large margin methods for structured and interdependent output variables, JMLR, 6,2005
- Collins, Michael, Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10,2002.
- Ben Taskar, Learning structured prediction models, a large margin approach, PhD thesis (`http://www.seas.upenn.edu/~taskar/pubs/thesis.pdf`), U. Pennsylvany, USA, 2004.