Predicting Structured Data: introduction

Juho Rousu

17.1.2008

Motivation

- Many real-world prediction tasks involve structured objects (sequences, images, networks)
- Yet, the mainstream of machine learning has been concerned with
  - vectorial/tabular input data

    X = [temp= 3.2 °C, moisture= 92%, wind= 5.2 m/s, snow=no, sun=no]
  - Single variable as output

    Y = Helsinki

Structured objects/data?

- In this seminar by a structured object we mean that there are statistical dependency assumptions between variables:
  - Adjacent characters in a string, adjacent words in a sentence
  - Parent-child relationship in hierarchies
  - Neighborhood in a lattice, e.g. nearby pixels in an image
- We try to utilize these relationships in learning the models
- Note about conflicting terminology:
  - In some communities this kind of data is called *semi-structured* or even *unstructured* while tabular data is called structured (as data items are neatly aligned in columns)
  - For us, tabular data has *no structure,* as there are no predeclared statistical dependencies which could be utilized in prediction

Example: Statistical Machine Translation

- Input: Paired data of sentences

$(x_1,y_1),\ldots,(x_n,y_n)$

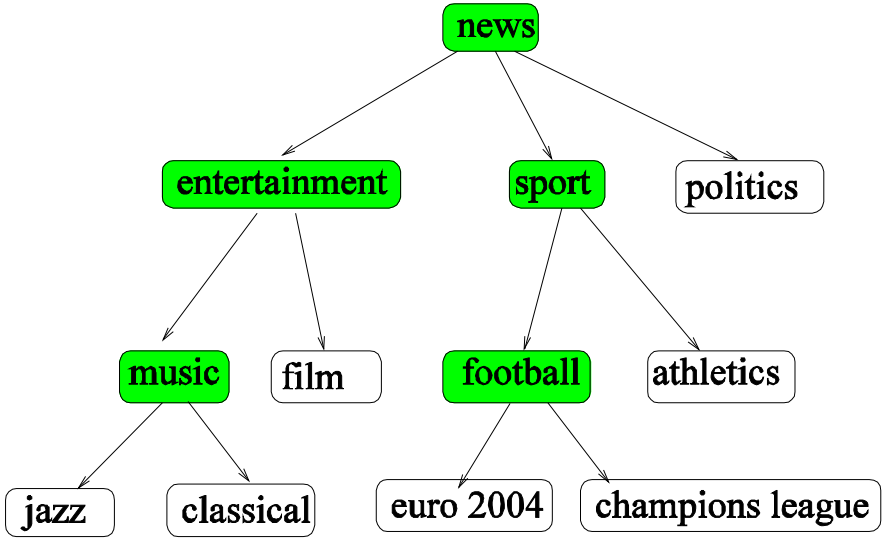
  x_i = ”Euroopan parlamentti hyväksyi kulutusluottodirektiivin”

  y_i = ”European parliament adopted consumer credit directive”,

- Output: model f(x)
  - that predicts target sentence $\hat{y}$= f(x) given a new source sentence x

Example: Hierarchical Classification

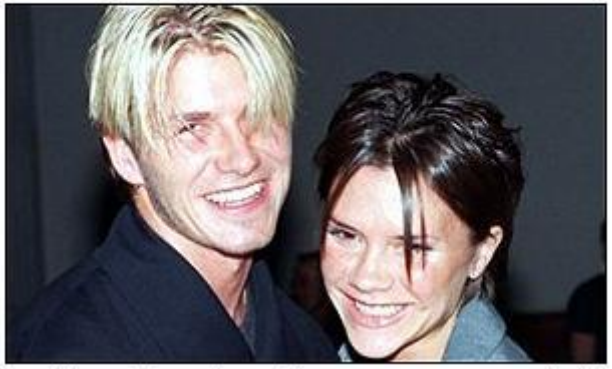
- Given a text document, predict membership in the hierarchy



You are in: **Entertainment**

Saturday, 8 January, 2000, 15:02 GMT

Football pundit accuses Posh



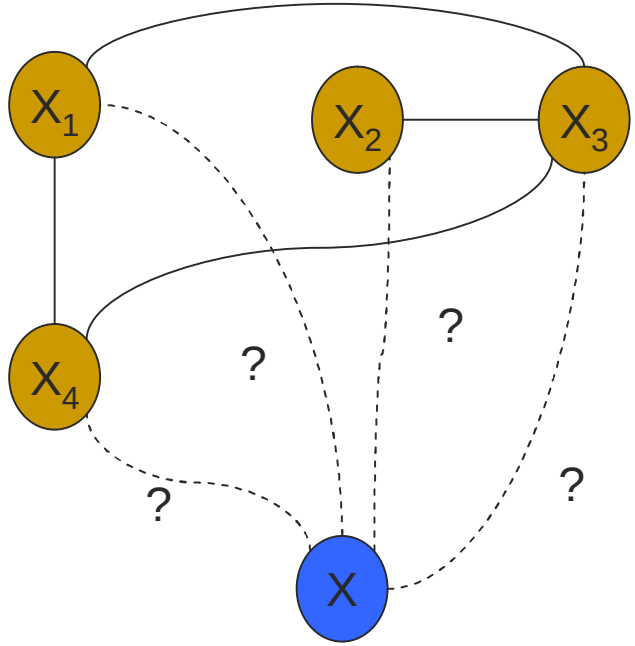
David and Victoria Beckham are permanently in the public eye

BBC football pundit Mark Lawrenson has accused David Beckham and his pop star wife Victoria of "courting publicity".

Lawrenson, an analyst on BBC1's Football Focus, spoke out during a discussion about Beckham's sending off in Thursday's World Club Championship match.

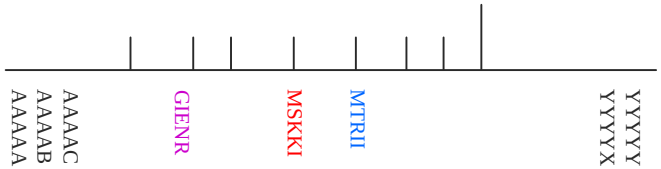
Example: Supervised network inference

- Input:
  - undirected graph G = (V, E)
  - subgraph $G_n = (V_n, E_n)$
  - feature descriptions $\phi(X)$ for the nodes X
- Output:
  - function e(X, X'): $V \times V \rightarrow \{0,1\}$, predicting if an edge should occur between an arbitrary pair of nodes



A Structured Prediction scheme

2. Use learned model: $W\,\phi(s) = \psi^*$ to obtain a output feature vector ψ^*



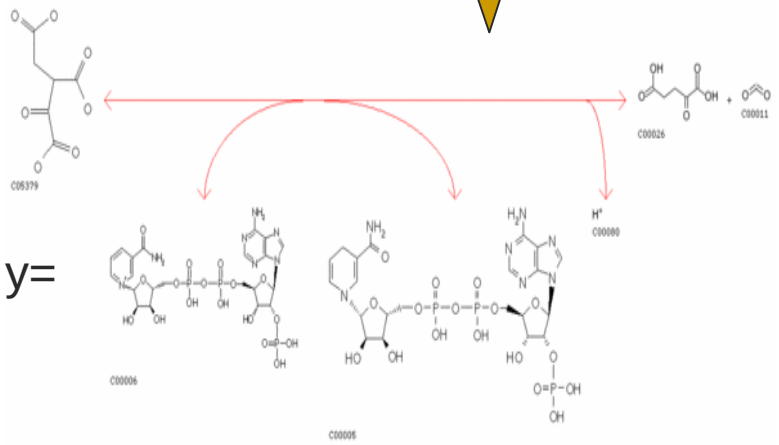
1. Map the structured object to a feature vector $\phi(x)$

3. Compute the preimage of the output feature vector $y^* = \psi^{-1}(\psi^*)$

x=

MSKKISGGSVVEMQGDEMTRIIWELIKEKLIFPYVELDLHSYDLGIENRDATNDQVTKDA
AEAIKKHNVGVKCATITPDEKRVEEFKLKQMWKSPNGTIRNILGGTVFREAIICKNIPRL
VSGWVKPIIIGRHAYGDQYRATDFVVPGPGKVEITYTPSDGTQKVTYLVHNFEEGGGVAM
GMYNQDKSIEDFAHSSFQMALSKGWPLYLSTKNTILKKYDGRFKDIFQEIYDKQYKSQFE
AQKIWYEHRLIDDMVAQAMKSEGGFIWACKNYDGDVQSDSVAQGYGSLGMMTSVLVCPDG
KTVEAEAAHGTVTRHYRMYQKGQETSTNPIASIFAWTRGLAHRAKLDNNKELAFFANALE
EVSIETIEAGFMTKDLAACIKGLPNVQRSDYLNTFEFMDKLGENLKIKLAQAKL

y=



Learning with kernels

- Many learning algorithms have natural dual formulations
  - Learning in terms of kernel functions $k(x,x') = \langle \phi(x), \phi(x') \rangle$, inner products of feature vectors
  - Efficient learning algorithms via convex optimization
  - Support vector machines for classification is the prime example
- Benefits:
  - For many types of structured objects $k(x,x')$ can be more efficiently computed than the explicit feature vectors $\phi(x)$
  - Training becomes independent of the input dimension $|\phi(x)|$
  - Same learning algorithm can be used in different purposes by the change of the kernel
- Tradeoff:
  - $O(n|\phi(x)|)$ space with feature vectors, $O(n^2)$ space with kernels