A General Regression Framework for Learning String-to-String Mapping

Corinna Cortes, Mehryar Mohri, and Jason Watson

Abhishek Tripathi

Department of Computer Science

University of Helsinki

Introduction

- Application in text and speech processing
  - A pronunciation model – phonemic transcriptions of a word
  - Natural language processing – part of speech tagging
- Learning string-to-string mapping $\approx$ Regression estimation of learning a real valued mapping
- Key aspects: Structures of strings can be exploited in learning
- Main techniques
  - Maximum-margin Markov Networks
  - Support vector machine learning for interdependent and structures output spaces
- A general and simple regression formulation of the problem is introduced

General Formulation

- X, Y – alphabets of the input and output strings
- (x_1,y_1)................ (x_m,y_m) in X* x Y*, training sample of size m drawn according to some distribution D
- Find a hypothesis f : X* $\longrightarrow$ Y* that predicts accurately the label y in Y* of a string x in X* drawn randomly according to D

$$
\begin{tikzcd}
X^* \arrow[r, "f"] \arrow[d, "\Phi_X"'] \arrow[dr, "g"] & Y^* \arrow[d, "\Phi_Y"'] \\
F_X \arrow[r, "W"'] & F_Y \arrow[u, bend right, "\Phi_Y^{-1}"']
\end{tikzcd}
$$

Learning as two-step approach

- Regression Problem :

  Hypothesis g : X* $\longrightarrow$ F_Y predicting $\Phi_Y(y)$ for x in X* with a label y in Y*, drawn randomly according to D.

- Pre-image problem:

  To predict the output string f(x) in Y* associated to x in X*.

  $$f(x) = \arg\min_{y \in Y^*} \left\| g(x) - \Phi_Y(y) \right\|^2,$$

  which provides an approximate pre-image when an exact pre-image does not exist. $(\Phi_Y^{-1}(g(x)) = \phi)$

Regression Problems and Algorithms

- Define $$K_X(x,x') = \Phi_X(x).\Phi(x'), \forall x, x' \in X^*$$ $$K_Y(y,y') = \Phi_Y(y).\Phi_Y(y') \forall y, y' \in Y^*$$
- K_X and K_Y are positive definite symmetric kernels mapping X* and Y* to the Hilbert spaces F_X and F_Y respectively.
- Dimension(F_X) = N_1, Dimension(F_Y) = N_2
- If W : $F_X \longrightarrow F_Y$ a linear function, $W \in \Re^{N2 \times N1}$
  - g is modeled as

$$\forall x \in X^*, g(x) = W(\Phi_X(x))$$

Kernel Ridge Regression with Vector Space Images

- The optimization problem,

$$\arg\min_{W \in \Re^{N2 \times N1}} F(W) = \sum_{i=1}^{m} \left\| W M_{x_i} - M_{y_i} \right\|^2 + \gamma \left\| W \right\|_F^{\ 2}$$

- Let M_X = [M_{x1},...,M_{xm}], and M_Y =[M_{y1},...,M_{ym}], the optimization problem can be re-written as

$$\arg\min_{W \in \Re^{N2 \times N1}} F(W) = \left\| W M_X - M_Y \right\|_F^2 + \gamma \left\| W \right\|_F^{\ 2}$$

- The unique solution of the optimization problem

$$W = M_Y M_X^T (M_X M_X^T + \gamma I)^{-1}$$ Primal Solution

$$W = M_Y (K_X + \gamma I)^{-1} M_X^T$$ Dual Solution

$\mathbf{K_X}$ is R^{mxm} Gram matrix associated to the kernel K_X : $\mathbf{K_{ij}} = K_X(x_i,x_j)$

Generalization to Regression with Constraints

- Use the string's structure to restrict the hypothesis space to achieve better result
- In part of speech tagging, a tag must match the word at the same position both in output and input sequences
- Incorporate input-output constraints via regularization on the regression matrix W
- The generalization leads to
  - a closed form solution
  - And to an efficient iterative algorithm

Pre-Image Solution for Strings

- Finding Pre-images: Common to all kernel-based structured output problems, including M^3N and SVM-ISOS
- Determine the predicted output : given $z \in F_Y$, the problem consists of finding $y \in Y^*$ such that $\Phi_Y(y) = z$
- Problem is trivial when Φ_Y corresponds to polynomial kernels of odd degree since Φ_Y is then invertible.
- Pre-image problem for n-gram Kernels for strings

N-gram Kernels

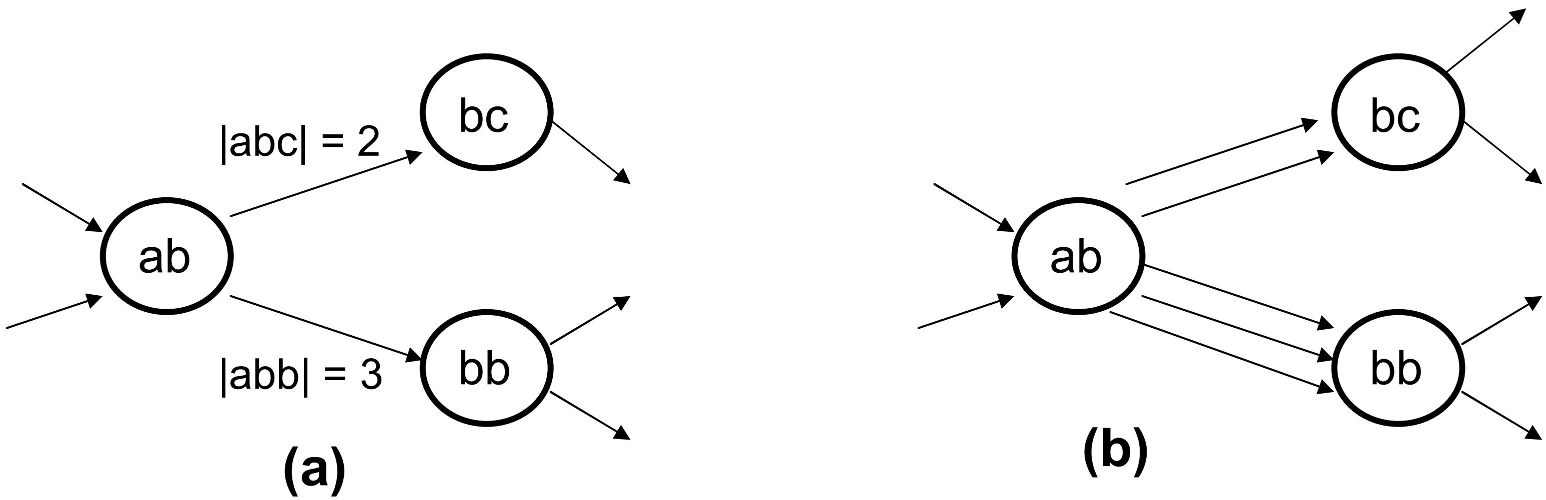
- N-gram kernels form a general family of kernels between strings
- Measure the similarity between two strings using their common n-gram sequences
- Let $|x|_u$ denote the number of occurrences of u in a string x
  - the n-gram kernel k_n between two strings y_1 and y_2 in Y*, n >= 1 is defined by
    $$k_n(y_1, y_2) = \sum_{|u|=n} |y_1|_u |y_2|_u,$$
  - Where the sum runs over all strings u of length n

Pre-Image Problem for n-gram Kernels

- Let Σ be the alphabet of strings
- $z = (z_1, \ldots\ldots\ldots, z_l)$, where $l = |\Sigma|^n$ and z_k is the count for an n-gram sequence u_k
- Find string y such that for k= 1,2,....., l, $|y|_{u_k} = z_k$
- Equivalent Graph-Theoretic Formulation of the problem
  - Can be formulated as De Bruijn graph
  - Finding a string y is then equivalent to finding an Euler circuit

Graph-Theoretic Formulation

- De Bruijn graphs
  - $G_{z,n}$ associated with n and the vector z. It is constructed in the following way:
  - Associate a vertex to each (n-1)-gram sequence
  - Add an edge from the vertex identified with $a_1a_2a_3$......... a_{n-1} to the vertex identified with a_2a_3......... a_n weighted with the count of n-gram $a_1a_2a_3$......... a_n
  - Replace each edge carrying weight c with c identical unweighted edges with the same origin and destination vertices.
  - Let $H_{z,n}$ be the resulting unweighted graph.
- Euler circuit of $H_{z,n}$ is a circuit on the graph in which each edge is traversed exactly once

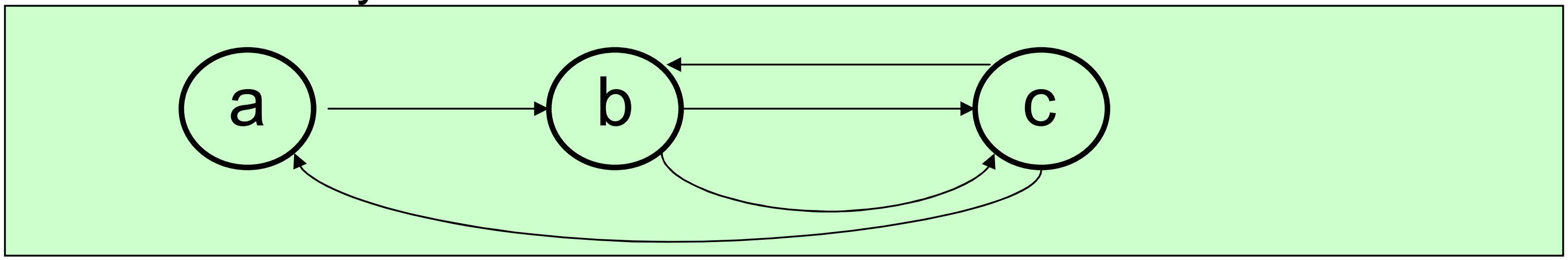


(a) $G_{z,3}$ associated with the vector z in the case of trigrams(n = 3). The weight carried by the edge from vertex ab to vertex bc is the number of occurrences of the trigram abc as specified by the vector z.

(b) The expanded graph $H_{z,3}$ associated with $G_{z,3}$. An edge in $G_{z,3}$ is repeated as many times as there were occurrences of the corresponding trigram.

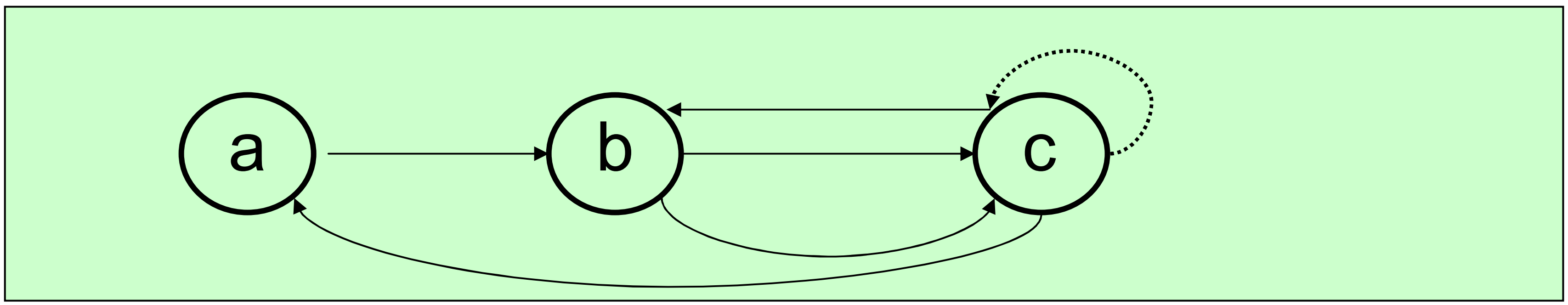
Eulerian Circuit of a Graph

- Existence of pre-image
  - In-degree(q) : number of incoming edges of vertex q
  - Out-degree(q) : Number of outgoing edges
  - *Theorem:* The vector z admits a pre-image iff for any vector q of $H_{z,n}$, in-degree(q) = out-degree(q).
- Example: z = (0,1,0,0,0,2,1,1,0) , co-ordinates indicate the counts of the bigrams aa,ab,ac,ba,bb,bc,ca,cb,cc. Graph satisfies the condition of the theorem, thus it admits an Eulerian circuit. The pre-image y = bcbca, if we start from the vertex a which is both the start and the end symbol.



Uniqueness of Pre-Images

- Linear-time algorithm for determining Eulirean circuit of a graph exists.
- In general, when it exists, the pre-image sequence is not unique
- Case of non-unique pre-images.



- Both bcbcca and bccbca are possible pre-images.

Experiments

- Description of datasets(Taskar et al, 2004b)
  - Subset of the hand-written words, MIT Spoken Language Systems Group
  - 6877 word instances with a total of 52,152 characters
  - First character of each word is removed to keep only lower case characters
  - The image of each character in 16x8 = 128 binary-pixel representation
  - Ten fold crosvalidation process, ten times one fold is used for training, and the remaining nine are used for testing

The General Handwriting Recognition Problem

- Determine a word y given the sequence of pixel-based images of its handwritten segmented characters $x = x_1, \ldots\ldots, x_k$
- Perfect Segmentation: one-to-one mapping of images to characters
  - Image segment x_i corresponds exactly to one word character, the character, y_i, of y in position i.
- General Regression(REG) and REG-constraints are used with polynomial kernel of third degree
- The best empirical value for ridge regression coefficient, $\gamma = 0.01$
- The REG-constraints, with Regularization parameter $\eta = 1$, performs significantly better than no-constraints REG.

The Comparison

Technique	Accuracy	
REG-constraint $\eta = 0$	84.1%	+/- 0.8%
REG-constraint $\eta = 1$	88.5%	+/- 0.9%
REG	79.5%	+/- 0.4%
REG-Viterbi(n = 2)	86.1%	+/- 0.7%
REG-Viterbi(n = 3)	98.2%	+/- 0.3%
SVMs(Cubic kernel)	80.9%	+/- 0.5%
M^3Ns(Cubic kernel)	87.0%	+/- 0.4%