



Aalto University
School of Science
and Technology

Max-Margin Learning with A Random Sample of Spanning Trees

Hongyu Su

Helsinki Institute for Information Technology HIIT
Department of Computer Science
Aalto University

February 13, 2015

Multilabel classification

- ▶ Multilabel classification is an important research field in machine learning.
 - ▶ For example, a document can be classified as “science”, “genomics”, and “drug discovery”.
 - ▶ Each input variable $\mathbf{x} \in \mathcal{X}$ is associated with multiple output variables $\mathbf{y} \in \mathcal{Y}$, $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_l$, $\mathcal{Y}_i = \{+1, -1\}$.
 - ▶ The goal is to find a mapping function that predicts the best values of an output given an input $f \in \mathcal{H} : \mathcal{X} \rightarrow \mathcal{Y}$.
- ▶ The central problems of multilabel classification:
 - ▶ The size of the output space \mathcal{Y} is exponential in the number of microlabels.
 - ▶ The dependency of microlabels needs to be exploited to improve the prediction performance.

Structured output learning

- ▶ There is an *output graph* connecting multiple labels.
 - ▶ A set of nodes represents multiple labels.
 - ▶ A set of edges represents the correlation between labels.
- ▶ Hierarchical classification:
 - ▶ The output graph is a rooted tree or a directed graph defining different levels of granularities.
 - ▶ For example, SSVM, ...
- ▶ Graph labeling:
 - ▶ The output graph often takes a general form (e.g., a tree, a chain).
 - ▶ For example, M^3N , CRF, MMCrf, ...
- ▶ The output graph is assumed to be known *a priori*.

Research question

- ▶ The output graph is hidden in many applications.
 - ▶ For example, a surveillance photo can be tagged with “building”, “road”, “pedestrian”, and “vehicle”.
- ▶ We study the problem in structured output learning when the output graph is not observed.
- ▶ In particular:
 - ▶ Assume the dependency can be expressed by a complete set of pairwise correlations.
 - ▶ Build a structured output learning model with a complete graph as the output graph.
 - ▶ Solve the optimization problem and the inference problem (\mathcal{NP} -hard).

Today

- ▶ A structured prediction model which performs max-margin learning on a random sample of spanning tree.
- ▶ Two ways to combine the set of random spanning trees
 - ▶ conical combination in NIPS paper.
 - ▶ convex combination as future work.
- ▶ Derivations and the corresponding optimization problems.

Model

- ▶ Training examples comes in pair $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m \in \mathcal{X} \times \mathcal{Y}$.
- ▶ A complete graph $G = (E, V)$ is used as the output graph.
- ▶ $\varphi(\mathbf{x})$ is the input feature map, e.g., a feature vector of d dimension.
- ▶ $\Gamma_G(\mathbf{y})$ is the output feature map of \mathbf{y} on G of $4 \times |E|$ dimension

$$\begin{aligned}\Gamma_G(\mathbf{y}) &= \{\Gamma_e(\mathbf{y}_e)\}_{e \in G}, \\ \Gamma_e(\mathbf{y}_e) &= [\mathbf{1}_{\mathbf{y}_e=00}, \mathbf{1}_{\mathbf{y}_e=01}, \mathbf{1}_{\mathbf{y}_e=10}, \mathbf{1}_{\mathbf{y}_e=11}]\end{aligned}$$

- ▶ A joint feature map of $(\mathbf{x}_i, \mathbf{y}_i)$

$$\phi_G(\mathbf{x}_i, \mathbf{y}_i) = \varphi(\mathbf{x}_i) \otimes \Gamma_G(\mathbf{y}_i) = \{\phi_e(x_i, \mathbf{y}_{i,e})\}_{e \in G}.$$

- ▶ A compatibility score is defined as

$$F(\mathbf{x}, \mathbf{y}; \mathbf{w}_G) = \langle \mathbf{w}_G, \phi_G(\mathbf{x}, \mathbf{y}) \rangle = \sum_{e \in G} \langle \mathbf{w}_{G,e}, \phi_e(\mathbf{x}, \mathbf{y}_e) \rangle$$

Model (cont.)

- ▶ \mathbf{w} ensures an input \mathbf{x}_i with a correct multilabel \mathbf{y}_i achieves a higher score than with any incorrect multilabel $\mathbf{y} \in \mathcal{Y}$.
- ▶ The predicted output $\mathbf{y}(\mathbf{x})$ for a given input \mathbf{x} is computed by

$$\mathbf{y}(\mathbf{x}) = \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}} F(\mathbf{x}, \mathbf{y}; \mathbf{w}_G) = \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}} \sum_{e \in G} \langle \mathbf{w}_{G,e}, \phi_{G,e}(\mathbf{x}, \mathbf{y}_e) \rangle,$$

which is called *inference problem*.

- ▶ The inference problem is \mathcal{NP} -hard for most joint feature maps on the complete graph.

How to learn w on a complete graph?

- ▶ The *margin* of an example \mathbf{x}_i is

$$\gamma_G(\mathbf{x}_i; \mathbf{w}_G) = F(\mathbf{x}_i, \mathbf{y}_i; \mathbf{w}_G) - \max_{\mathbf{y} \in \mathcal{Y}/\mathbf{y}_i} F(\mathbf{x}_i, \mathbf{y}; \mathbf{w}_G).$$

- ▶ \mathbf{w} is solved by *max-margin principle* which aims to maximize $\gamma(\mathbf{x}_i; \mathbf{w}_G)$ over all training example $\mathbf{x}_i, i \in \{1, \dots, m\}$.
- ▶ The inference problem on a complete graph is \mathcal{NP} -hardness.
- ▶ The parameter space is quadratic in the number of microlabels k .
- ▶ We aim to use a joint feature map that allows the inference problem be solved in polynomial time.

Superposition of random trees

- ▶ $S(G)$ is a complete set of spanning tree generate from G .
- ▶ $\mathbf{w}_T = \{\mathbf{w}_{G,e}\}_{e \in T}$ is the projection of \mathbf{w}_G on T .
- ▶ $\phi_T(\mathbf{x}, \mathbf{y}) = \{\phi_e(\mathbf{x}, \mathbf{y})\}_{e \in T}$ is the projection of $\phi_G(\mathbf{x}, \mathbf{y})$ on T .
- ▶ Rewrite

$$\begin{aligned} F(\mathbf{x}, \mathbf{y}, \mathbf{w}_G) &= \sum_{e \in G} \langle \mathbf{w}_{G,e}, \phi_{G,e}(\mathbf{x}, \mathbf{y}_e) \rangle \\ &= \frac{1}{\ell^{\ell-2}} \sum_{T \in S(G)} \sqrt{\frac{\ell}{2}} \langle \mathbf{w}_T, \phi_T(\mathbf{x}, \mathbf{y}_e) \rangle \\ &= \frac{1}{n} \sum_{i=1}^n a_{T_i} \langle \hat{\mathbf{w}}_{T_i}, \hat{\phi}_{T_i}(\mathbf{x}, \mathbf{y}_e) \rangle, \\ \frac{1}{n} \sum_{i=1}^n a_{T_i}^2 &= 1, \quad \frac{1}{n} \sum_{i=1}^n a_{T_i} \leq 1, \quad a_{T_i} \geq 0, \quad n = \ell^{\ell-2}. \end{aligned}$$

How many trees?

- ▶ If there is a predictor \mathbf{w}_G on complete graph achieves a margin on some training data, with high probability we need n spanning tree predictors $\{\mathbf{w}_{T_i}\}_{i=1}^n$ to achieve a close margin. n is quadratic in terms of ℓ .
- ▶ Recall

$$F(\mathbf{x}, \mathbf{y}, \mathbf{w}_T) = \frac{1}{n} \sum_{i=1}^n a_{T_i} \langle \hat{\mathbf{w}}_{T_i}, \hat{\phi}_{T_i}(\mathbf{x}, \mathbf{y}_e) \rangle,$$

$$\frac{1}{n} \sum_{i=1}^n a_{T_i}^2 = 1, \quad \frac{1}{n} \sum_{i=1}^n a_{T_i} \leq 1, \quad a_{T_i} \geq 0, \quad n = \ell^2.$$

Conical combination

- ▶ A sample \mathcal{T} of n spanning trees drawn from G .
- ▶ Normalized feature weights $\hat{\mathbf{w}}_{T_i} = \frac{\mathbf{w}_{T_i}}{\|\mathbf{w}_{T_i}\|}$, $T_i \in \mathcal{T}$.
- ▶ Normalized feature vectors $\hat{\phi}_{T_i}(\mathbf{x}, \mathbf{y}) = \frac{\phi_{T_i}(\mathbf{x}, \mathbf{y})}{\|\phi_{T_i}(\mathbf{x}, \mathbf{y})\|}$, $T_i \in \mathcal{T}$.
- ▶ Conical combination of spanning trees

$$F(\mathbf{x}, \mathbf{y}, \mathbf{w}_{\mathcal{T}}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n q_i \langle \hat{\mathbf{w}}_{T_i}, \hat{\phi}_{T_i}(\mathbf{x}, \mathbf{y}) \rangle$$
$$\sum_{i=1}^n q_i^2 = 1, q_i \geq 0, \forall i \in \{1, \dots, n\}.$$

Conical combination (cont.)

- To solve $\{\mathbf{w}_{T_i}\}_{T_i \in \mathcal{T}}$, we need to work on the optimization problem

$$\begin{aligned} \min_{\xi, \gamma, \mathbf{q}, \mathcal{W}} \quad & \frac{1}{2\gamma^2} + \frac{C}{\gamma} \sum_{k=1}^m \xi_k \\ \text{s.t.} \quad & \frac{1}{\sqrt{n}} \sum_{i=1}^n q_i \langle \hat{\mathbf{w}}_{T_i}, \hat{\phi}_{T_i}(\mathbf{x}_k, \mathbf{y}_k) \rangle - \max_{\mathbf{y} \in \mathcal{Y}} \frac{1}{\sqrt{n}} \sum_{i=1}^n q_i \langle \hat{\mathbf{w}}_{T_i}, \hat{\phi}_{T_i}(\mathbf{x}_k, \mathbf{y}) \rangle \\ & \geq \gamma - \xi_k, \xi_k \geq 0, \forall k \in \{1, \dots, m\}, \sum_{i=1}^n q_i^2 = 1, q_i \geq 0, \forall i \in \{1, \dots, n\}. \end{aligned}$$

- This is equivalent to

$$\begin{aligned} \min_{\mathbf{w}_{T_i}, \xi_i} \quad & \frac{1}{2} \sum_{i=1}^n \|\mathbf{w}_{T_i}\|^2 + C \sum_{k=1}^m \xi_k \\ \text{s.t.} \quad & \frac{1}{\sqrt{n}} \sum_{i=1}^n \langle \mathbf{w}_{T_i}, \phi_{T_i}(\mathbf{x}_k, \mathbf{y}_k) \rangle - \max_{\mathbf{y} \neq \mathbf{y}_k} \frac{1}{\sqrt{n}} \sum_{i=1}^n \langle \mathbf{w}_{T_i}, \phi_{T_i}(\mathbf{x}_k, \mathbf{y}) \rangle \geq 1 - \xi_k, \\ & \xi_k \geq 0, \forall k \in \{1, \dots, m\}. \end{aligned}$$

Convex combination

- ▶ A sample \mathcal{T} of n spanning trees drawn from G .
- ▶ Normalized feature weights $\hat{\mathbf{w}}_{T_i} = \frac{\mathbf{w}_{T_i}}{\|\mathbf{w}_{T_i}\|}$, $T_i \in \mathcal{T}$.
- ▶ Normalized feature vectors $\hat{\phi}_{T_i}(\mathbf{x}, \mathbf{y}) = \frac{\phi_{T_i}(\mathbf{x}, \mathbf{y})}{\|\phi_{T_i}(\mathbf{x}, \mathbf{y})\|}$, $T_i \in \mathcal{T}$.
- ▶ Conical combination of spanning trees

$$F(\mathbf{x}, \mathbf{y}, \mathbf{w}_{\mathcal{T}}) = \frac{1}{n} \sum_{i=1}^n q_i \langle \hat{\mathbf{w}}_{T_i}, \hat{\phi}_{T_i}(\mathbf{x}, \mathbf{y}) \rangle$$

$$\sum_{i=1}^n q_i = 1, q_i \geq 0, \forall i \in \{1, \dots, n\}.$$

Convex combination (cont.)

- To solve $\{\mathbf{w}_{T_i}\}_{T_i \in \mathcal{T}}$, we need to work on the optimization problem

$$\begin{aligned} \min_{\xi, \gamma, \mathbf{q}, \mathcal{W}} \quad & \frac{1}{2\gamma^2} + \frac{C}{\gamma} \sum_{k=1}^m \xi_k \\ \text{s.t.} \quad & \frac{1}{n} \sum_{i=1}^n q_i \langle \hat{\mathbf{w}}_{T_i}, \hat{\phi}_{T_i}(\mathbf{x}_k, \mathbf{y}_k) \rangle - \max_{\mathbf{y} \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^n q_i \langle \hat{\mathbf{w}}_{T_i}, \hat{\phi}_{T_i}(\mathbf{x}_k, \mathbf{y}) \rangle \\ & \geq \gamma - \xi_k, \xi_k \geq 0, \forall k \in \{1, \dots, m\}, \sum_{i=1}^n q_i = 1, q_i \geq 0, \forall i \in \{1, \dots, n\}. \end{aligned}$$

- This is equivalent to

$$\begin{aligned} \min_{\mathbf{w}_{T_i}, \xi_i} \quad & \frac{1}{2} \left(\sum_{i=1}^n \|\mathbf{w}_{T_i}\| \right)^2 + C \sum_{k=1}^m \xi_k \\ \text{s.t.} \quad & \frac{1}{n} \sum_{i=1}^n \langle \mathbf{w}_{T_i}, \phi_{T_i}(\mathbf{x}_k, \mathbf{y}_k) \rangle - \max_{\mathbf{y} \neq \mathbf{y}_k} \frac{1}{n} \sum_{i=1}^n \langle \mathbf{w}_{T_i}, \phi_{T_i}(\mathbf{x}_k, \mathbf{y}) \rangle \geq 1 - \xi_k, \\ & \xi_k \geq 0, \forall k \in \{1, \dots, m\}. \end{aligned}$$

Convex combination (cont.)

- This can be expressed equivalently as

$$\begin{aligned} \min_{\mathbf{w}_{T_i}, \xi_i, \lambda_i} \quad & \frac{1}{2} \sum_{i=1}^n \frac{1}{\lambda_i} \|\mathbf{w}_{T_i}\|^2 + C \sum_{k=1}^m \xi_k \\ \text{s.t.} \quad & \frac{1}{n} \sum_{i=1}^n \langle \mathbf{w}_{T_i}, \phi_{T_i}(\mathbf{x}_k, \mathbf{y}_k) \rangle - \max_{\mathbf{y} \neq \mathbf{y}_k} \frac{1}{n} \sum_{i=1}^n \langle \mathbf{w}_{T_i}, \phi_{T_i}(\mathbf{x}_k, \mathbf{y}) \rangle \geq 1 - \xi_k, \\ & \xi_k \geq 0, \forall k \in \{1, \dots, m\}, \sum_{i=1}^n \lambda_i = 1, \lambda_i \geq 0, \forall i \in \{1, \dots, n\}. \end{aligned}$$

Conclusions

- ▶ Theoretical study shows if a large margin structured output learner exists, then the combination of a random sample of spanning trees will achieve a similar margin with a high probability.
- ▶ The K -best inference algorithm is tractable which is proved theoretically and empirically.
- ▶ RTA is not constrained by the availability of the output graph, it can therefore be applied to a wider range of multilabel classification problem where the output graph is believed to play an important role during learning.