Semi-supervised Learning for Structured Output Variables

Ulf Brefeld Tobias Scheffer
ICML 2006

presented by Jean-Baptiste Jeannin
CS6784 February 25th, 2010

Outline

- Semi-supervised learning by co-training
- Structured output variables
- Using co-training for structured output variables

Framework and notations

- Structured input $\mathbf{x}$ and output $\mathbf{y}$ with dependencies
- Joint feature representation $\Phi(\mathbf{x}, \mathbf{y})$
- Learn $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\bar{\mathbf{y}} \in \mathcal{Y}} f(\mathbf{x}, \bar{\mathbf{y}})$$ is as desired

- Linear model

$$f(\mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \Phi(\mathbf{x}, \mathbf{y}) \rangle$$

- Labeled examples $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$
  Unlabeled examples $\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+m}$

Co-training

- Semi-supervised learning: using both labeled and unlabeled data for learning
- Idea of training: exploit two sufficiently redundant representations

$$\Phi(\mathbf{x}, \mathbf{y}) = (\Phi^0(\mathbf{x}, \mathbf{y}), \Phi^1(\mathbf{x}, \mathbf{y}))$$

  - web-page body text / hyperlinks pointing to page
  - sound of person saying hello / lip movements

Co-training

- Idea of-training: exploit two sufficiently redundant representations
  - Training example: $((\Phi^0(\mathbf{x}, \mathbf{y}), \Phi^1(\mathbf{x}, \mathbf{y})), \mathbf{y})$
  - Test example: $(\Phi^0(\mathbf{x}, \mathbf{y}), \Phi^1(\mathbf{x}, \mathbf{y}))$
  - Hypotheses

$$f^0(\mathbf{x}, \mathbf{y}) = \langle \mathbf{w}^0, \Phi^0(\mathbf{x}, \mathbf{y}) \rangle$$
$$f^1(\mathbf{x}, \mathbf{y}) = \langle \mathbf{w}^1, \Phi^1(\mathbf{x}, \mathbf{y}) \rangle$$

are compatible if and only if for all test examples

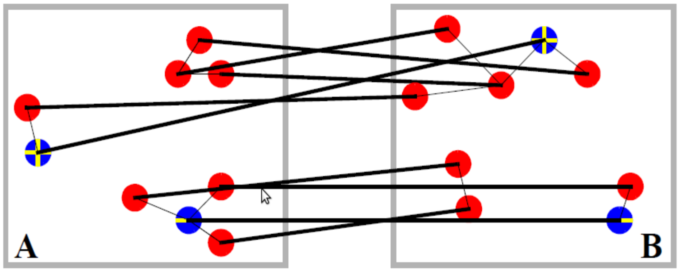
$$f^0(\mathbf{x}, \mathbf{y}) = f^1(\mathbf{x}, \mathbf{y})$$

Co-training

- Hypotheses are compatible if and only if for all examples

$$f^0(\mathbf{x}, \mathbf{y}) = f^1(\mathbf{x}, \mathbf{y})$$

- Perfect classifiers do not disagree



Co-training

- Joint decision function

$$f(\mathbf{x},\mathbf{y}) = f^0(\mathbf{x},\mathbf{y}) + f^1(\mathbf{x},\mathbf{y})$$
$$= \langle \mathbf{w}^0, \Phi^0(\mathbf{x},\mathbf{y})\rangle + \langle \mathbf{w}^1, \Phi^1(\mathbf{x},\mathbf{y})\rangle$$

Structured output variables for supervised learning [Tsochantaridis et al.]

- Support vector learning with slack variables $\xi_i \geq 0$
- Introducing a loss function $\Delta : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_0^+$
- We would like $\mathbf{y}_i = \mathrm{argmax}_{\bar{\mathbf{y}}} \langle \mathbf{w}, \Phi(\mathbf{x}_i, \bar{\mathbf{y}})\rangle$
- Minimize over all $\mathbf{w}$ and ξ_i

$$\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^n \xi_i$$

such that $\forall_{i=1}^n, \forall_{\bar{\mathbf{y}}\neq\mathbf{y}_i}$

$$\langle \mathbf{w}, \Phi(\mathbf{x}_i,\mathbf{y}_i) - \Phi(\mathbf{x}_i,\bar{\mathbf{y}})\rangle \geq 1 - \frac{\xi_i}{\Delta(\mathbf{y}_i,\bar{\mathbf{y}})}$$

Semi-supervised and co-learning

- Consensus maximizing principle:
  - Minimize the number of errors in labeled examples
  - Minimize the disagreement for unlabeled examples
- Let $\hat{\mathbf{y}}_i^1$ be the prediction of $\mathbf{x}_i$ using f^1
  $\hat{\mathbf{y}}_i^1$ is treated as correct output
- For unlabeled examples $\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+m}$

$$\hat{\mathbf{y}}_i^1 = \mathrm{argmax}_{\bar{\mathbf{y}}} \langle \mathbf{w}^0, \Phi^0(\mathbf{x}_i, \bar{\mathbf{y}})\rangle$$

$$f^0(\mathbf{x}_i, \hat{\mathbf{y}}_i^1) - \max_{\bar{\mathbf{y}}\neq\hat{\mathbf{y}}_i^1} f^0(\mathbf{x}_i,\bar{\mathbf{y}}) = \gamma_i^0 \geq 1$$

and vice-versa

Semi-supervised and co-learning

- Minimize over all $\mathbf{w}$ and ξ_i

$$\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^n \xi_i + CC_u \sum_{i=n+1}^{n+m} \min\{\gamma_i^1, 1\}\xi_i$$

such that $\forall_{i=1}^n, \forall_{\bar{\mathbf{y}}\neq\mathbf{y}_i}$

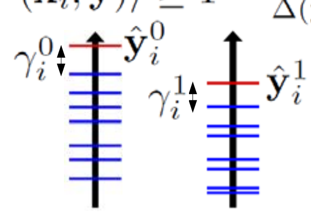
$$\langle \mathbf{w}, \Phi(\mathbf{x}_i,\mathbf{y}_i) - \Phi(\mathbf{x}_i,\bar{\mathbf{y}})\rangle \geq 1 - \frac{\xi_i}{\Delta(\mathbf{y}_i,\bar{\mathbf{y}})}$$

and $\forall_{i=n+1}^{n+m}, \forall_{\bar{\mathbf{y}}\neq\mathbf{y}_i}$

$$\langle \mathbf{w}^0, \Phi^0(\mathbf{x}_i,\hat{\mathbf{y}}_i^1) - \Phi^0(\mathbf{x}_i,\bar{\mathbf{y}})\rangle \geq 1 - \frac{\xi_i}{\Delta(\hat{\mathbf{y}}_i^1,\bar{\mathbf{y}})}$$

and vice-versa

- γ_i^1 is the margin for the prediction of $\hat{\mathbf{y}}_i^1$



Dual problem – Empirical results

- Algebra transforms this optimization problem introducing Lagrange multipliers, like in normal Support Vector Machines, for resolution
- 3 cases are studied:
  - Multi-class classification
  - Label sequence learning
  - Natural language parsing
- Co-trained SVM outperforms SVM in most tasks

Example: label sequence learning

- Mapping sequential input to sequential output
- Datasets: sentences where we discriminate gene/other or person/organization/location
- The two views are a random split of the attributes
- Results: SVM and coSVM beat HMM. SVM is outperformed by coSVM in all but one setting

Conclusion

- A semi-supervised approach for structured output variables

- Combines the ideas of:
  - Co-learning (Blum & Mitchell, 1998)
  - Structured output variables (Tsochantaridis, Joachims, Hofmann & Altun, 2005)