NIPS workshop on RLCO 2014

Implicit Learning of Simpler Output Kernels for Multi-label Prediction

Hanchen Xiong, Sandor Szedmak, Justus Piater

University of Innsbruck, Austria

Montreal, Canada. 2014.12.13

Support Vector Machines

$$\min_{\mathbf{w}\in\mathbb{R}^{\mathcal{H}_\phi}} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m} \xi^{(i)}$$

$$\text{s.t.} \quad y^{(i)}\left(\mathbf{w}^\top\phi(\mathbf{x}^{(i)})\right) \geq 1-\xi^{(i)}, \xi^{(i)} \geq 0, i \in \{1,\dots,m\}$$

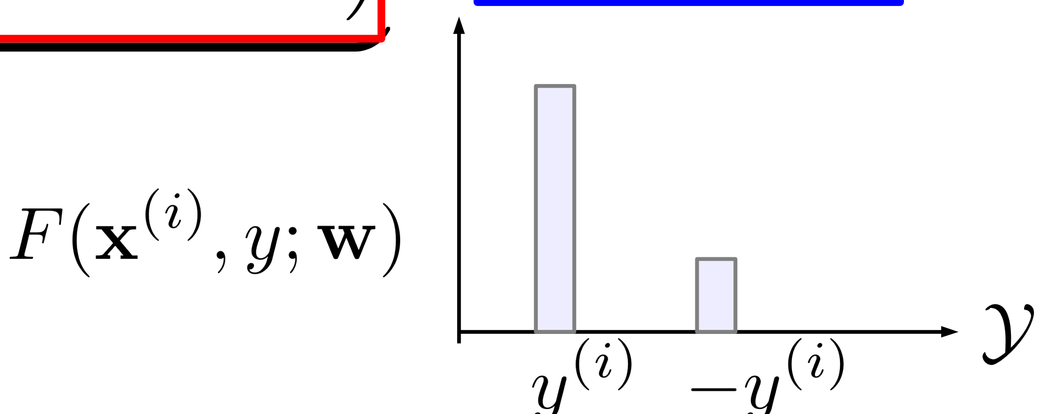
define a score function $\boxed{F(\mathbf{x}^{(i)}, y^{(i)}; \mathbf{w})} = y^{(i)}\left(\mathbf{w}^\top\phi(\mathbf{x}^{(i)})\right)$

and output distance $\boxed{d(y^{(i)}, -y^{(i)})} = |y^{(i)} - (-y^{(j)})| = 2$

$$\boxed{\min_{\mathbf{w}\in\mathbb{R}^{\mathcal{H}_\phi}} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m} \max\{0, d(y^{(i)}, -y^{(i)}) - \Delta_F(y^{(i)}, -y^{(i)})\}}$$

$$\underbrace{\boxed{F\left(\mathbf{x}^{(i)}, y^{(i)}; \mathbf{w}\right)} - \boxed{F\left(\mathbf{x}^{(i)}, -y^{(i)}; \mathbf{w}\right)}}_{\Delta_F(y^{(i)}, -y^{(i)})} \geq \boxed{d(y^{(i)}, -y^{(i)})} - \xi^{(i)}$$

$$\xi^{(i)} \geq 0$$



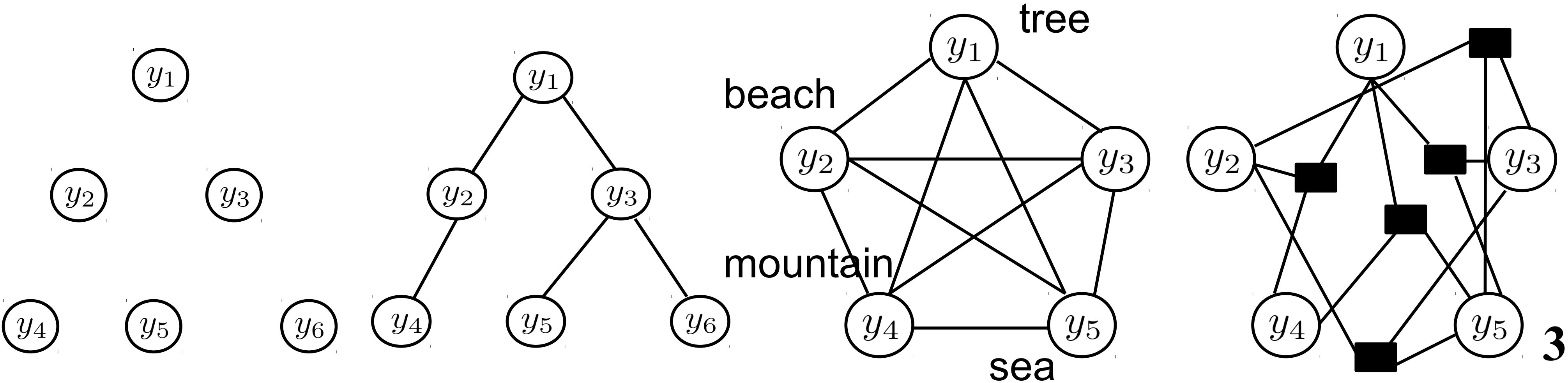
Inter-Label Dependencies

Structural Outputs: $\mathcal{Y} = \{-1, +1\}^T$

$$\mathbf{y}^{(i)} = [y_1^{(i)}, y_2^{(i)}, y_3^{(i)}, \cdots, y_d^{(i)}]^\top$$

tree	mountain	beach	sea
+1	+1	-1	-1
-1	-1	+1	+1

Labels are inter-dependent: e.g. Markov network

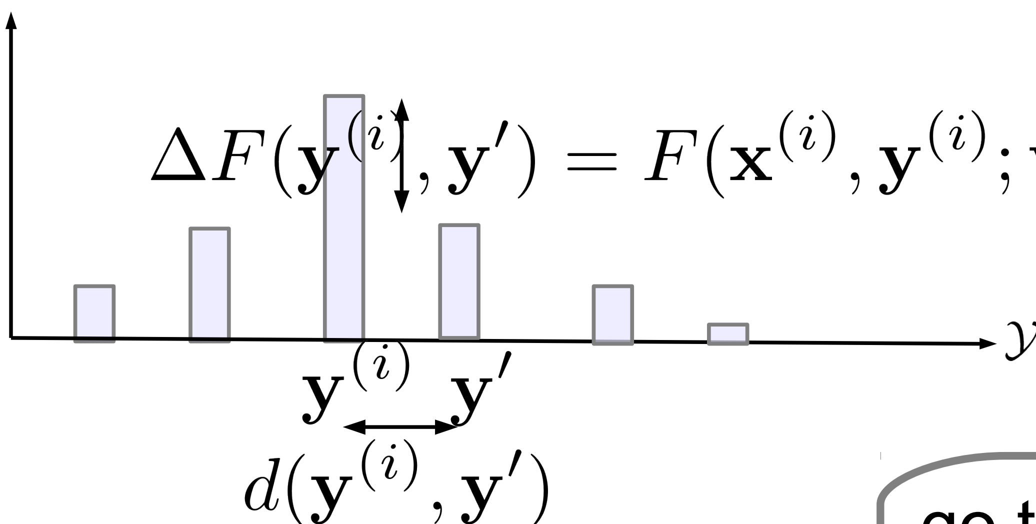


Structural Support Vector Machine

$$\underbrace{F\left(\mathbf{x}^{(i)}, y^{(i)}; \mathbf{w}\right) - F\left(\mathbf{x}^{(i)}, -y^{(i)}; \mathbf{w}\right)}_{\Delta_F(y^{(i)}, -y^{(i)})} \geq d(y^{(i)}, -y^{(i)}) - \xi^{(i)}$$

generalized to structured outputs

$F(\mathbf{x}^{(i)}, \mathbf{y}; \mathbf{w})$



$$\Delta F(\mathbf{y}^{(i)}, \mathbf{y}') = F(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}; \mathbf{w}) - F(\mathbf{x}^{(i)}, \mathbf{y}'; \mathbf{w})$$

go through all possible outputs is too expensive

Structural SVM:

$$\min_{\mathbf{W} \in \mathbb{R}^{\Psi}} \frac{1}{2}||\mathbf{W}||^2 + C \sum_{i=1}^{m} \max_{\mathbf{y}' \in \mathcal{Y}} \left\{ d(\mathbf{y}^{(i)}, \mathbf{y}') - \Delta_F(\mathbf{y}^{(i)}, \mathbf{y}') \right\}$$

SSVM for Multi-Label Prediction

Structural SVM

$$\min_{\mathbf{W}\in\mathbb{R}^{\Psi}} \frac{1}{2}||\mathbf{W}||^2 + C\sum_{i=1}^{m} \max_{\mathbf{y}'\in\mathcal{Y}} \left\{ d(\mathbf{y}^{(i)},\mathbf{y}') - \Delta_F(\mathbf{y}^{(i)},\mathbf{y}') \right\}$$

We define function as $\mathrm{F}\left(\mathbf{x}^{(i)},\mathbf{y}^{(i)};\mathbf{W}\right) = \langle \mathbf{W}, \phi(\mathbf{x}^{(i)}) \otimes \mathbf{y}^{(i)} \rangle_F$
and use *Hamming distance* on outputs:

because of linear decomposibility

$$\min_{\mathbf{W}\in\mathbb{R}^{\mathcal{H}_\phi}\times\mathbb{R}^{T}} \frac{1}{2}||\mathbf{W}||_F^2 + C\sum_{i=1}^{m}\sum_{t=1}^{T} \max_{y'_t=\{-1,+1\}} \left\{ d(y_t^{(i)},y'_t) - \Delta_F(y_t^{(i)},y'_t) \right\}$$

Linear Score Function in Structural SVM

Structural SVM:

$$\min_{\mathbf{W}\in\mathbb{R}^{\mathcal{H}_\phi}\times\mathbb{R}^T} \frac{1}{2}||\mathbf{W}||_F^2 + C\sum_{i=1}^{m}\sum_{t=1}^{T}\max_{y'_t=\{-1,+1\}}\left\{d(y_t^{(i)}, y'_t) - \Delta_F(y_t^{(i)}, y'_t)\right\}$$

$$\downarrow$$

$$\min_{\mathbf{w}_1,\cdots,\mathbf{w}_T\in\mathbb{R}^{\mathcal{H}_\phi}} \sum_{t=1}^{T}\left\{\boxed{\frac{||\mathbf{w}_t||^2}{2} + C\sum_{i=1}^{m}\max\left\{0, d(y_t^{(i)}, -y_t^{(i)}) - \Delta_F(y_t^{(i)}, -y_t^{(i)})\right\}}\right\}$$

Regular SVM:

$$\min_{\mathbf{w}\in\mathbb{R}^{\mathcal{H}_\phi}} \boxed{\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m}\max\{0, d(y^{(i)}, -y^{(i)}) - \Delta_F(y^{(i)}, -y^{(i)})\}}$$

linearly decomposable

$$\downarrow$$

Structural SVM → training *T* SVMs jointly.

Joint SVM

training *T* SVMs jointly:

$$\begin{aligned} \min \quad & \frac{1}{2} \boxed{\sum_{t=1}^{T}} ||\mathbf{w}_t||^2 + C \boxed{\sum_{t=1}^{T}} \sum_{i=1}^{m} \xi_t^{(i)} \\ \text{w.r.t.} \quad & \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_T \in \mathbf{R}^{\mathcal{H}_\phi \times 1} \\ \text{s.t.} \quad & \boxed{\sum_{t=1}^{T}} y_t^{(i)} \left(\mathbf{w}_t^\top \phi(x^{(i)}) \right) \geq T - \sum_{t=1}^{T} \xi_t^{(i)} \end{aligned}$$

$$\mathrm{y}^{(i)} = [y_1^{(1)}, \ldots, y_T^{(i)}], \text{ and } \mathbf{W} = [\frac{\mathbf{w}_1^\top}{T}; \ldots; \frac{\mathbf{w}_T^\top}{T}]^\top, \bar{\xi}^{(i)} = \frac{\sum_{t=1}^{T} \xi_t^{(i)}}{T}$$

Joint SVM:

$$\begin{aligned} \min_{\mathbf{W} \in \mathbb{R}^T \times \mathcal{H}_\phi} \quad & \frac{1}{2} ||\mathbf{W}||_F^2 + C \sum_{i=1}^{m} \bar{\xi}^{(i)} \\ \text{s.t.} \quad & \boxed{\langle \mathbf{y}^{(i)}}, \mathbf{W}\phi(x^{(i)}) \rangle \geq 1 - \bar{\xi}^{(i)}, \bar{\xi}^{(i)} \geq 0, i \in \{1, \ldots, m\} \end{aligned}$$

define linear kernels on structured outputs :

$$K_\psi(\mathbf{y}^{(i)}, \mathbf{y}^{(j)}) = \langle \psi(\mathbf{y}^{(i)}), \psi(\mathbf{y}^{(j)}) \rangle$$

$$\begin{aligned} \min_{\mathbf{W} \in \mathbb{R}^{H_\psi \times \mathcal{H}_\phi}} \quad & \frac{1}{2} ||\mathbf{W}||_F^2 + C \sum_{i=1}^{m} \bar{\xi}^{(i)} \\ \text{s.t.} \quad & \langle \boxed{\psi(\mathbf{y}^{(i)})}, \mathbf{W}\phi(x^{(i)}) \rangle \geq 1 - \bar{\xi}^{(i)}, \xi_i \geq 0, i \in \{1, \ldots, m\} \end{aligned}$$

Low Training Complexity:

Dual Joint SVM:

$$\begin{array}{rl} \max_{\alpha_1,\cdots,\alpha_m} & \sum_{i=1}^{m} \alpha_i - \sum_{i,j=1}^{m} \alpha_i \alpha_j \boxed{K_\psi(\mathbf{y}^{(i)}, \mathbf{y}^{(j)})} \boxed{K_\phi(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})} \\ \text{s.t} & \forall i, 0 \leq \alpha_i \leq C \end{array}$$

the same computational complexity as a single SVM

Dual SVM:

$$\begin{array}{rl} \max_{\boldsymbol{\alpha} \in \mathbb{R}^m} & \sum_{i=1}^{m} \alpha_i - \sum_{i,j=1}^{m} \alpha_i \alpha_j \boxed{y^{(i)} y^{(j)}} \boxed{K_\phi(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})} \\ \text{s.t.} & 0 < \alpha_i < C, i \in \{1, \ldots, m\} \end{array}$$

Joint SVM : Implicit Output Kernel Learning and Regularization

When linear output is used:

$$\min_{\mathbf{W}\in\mathbb{R}^{H_\psi\times\mathcal{H}_\phi}} \quad \frac{1}{2}||\mathbf{W}||_F^2 + C\sum_{i=1}^m \bar{\xi}^{(i)}$$
$$\text{s.t.} \quad \boxed{\left\langle \psi(\mathbf{y}^{(i)}), \mathbf{W}\phi(x^{(i)})\right\rangle} \geq 1-\bar{\xi}^{(i)}, \xi_i \geq 0, i\in\{1,\dots,m\}$$

$$\psi^{Lin}(\mathbf{y}^{(i)}) = \mathbf{P}\mathbf{y}^{(i)} \qquad K_\psi^{Lin}(\mathbf{y}^{(i)},\mathbf{y}^{(j)}) = \mathbf{y}^{(i)\top}\mathbf{P}^\top\mathbf{P}\mathbf{y}^{(j)}$$

$$\min_{\mathbf{W}\in\mathbb{R}^{H_\psi\times\mathcal{H}_\phi}} \quad \frac{1}{2}||\mathbf{W}||_F^2 + C\sum_{i=1}^m \bar{\xi}^{(i)}$$
$$\text{s.t.} \quad \boxed{\left\langle \mathbf{y}^{(i)}, \mathbf{P}^\top\mathbf{W}\phi(x^{(i)})\right\rangle} \geq 1-\bar{\xi}^{(i)}, \xi_i \geq 0, i\in\{1,\dots,m\}$$

$$\mathbf{U} = \mathbf{P}^\top\mathbf{W}$$

$$\min_{\mathbf{W}\in\mathbb{R}^{H_\psi\times\mathcal{H}_\phi}} \quad \boxed{\frac{1}{2}||\mathbf{W}||}_F^2 + C\sum_{i=1}^m \bar{\xi}^{(i)}$$
$$\text{s.t.} \quad \boxed{\left\langle \mathbf{y}^{(i)}, \mathbf{U}\phi(x^{(i)})\right\rangle} \geq 1-\bar{\xi}^{(i)}, \xi_i \geq 0, i\in\{1,\dots,m\}$$

a new regularization: $\dfrac{1}{2}\mathbf{tr}(\mathbf{W}^\top\mathbf{P}\mathbf{P}^\top\mathbf{W})$

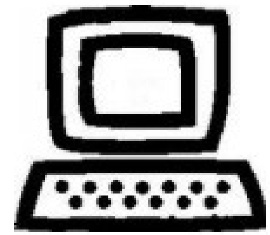
$$\min_{\mathbf{U}\in\mathbb{R}^{H_\psi\times\mathcal{H}_\phi}} \quad \boxed{\frac{1}{2}||\mathbf{U}||}_F^2 + C\sum_{i=1}^m \bar{\xi}^{(i)}$$
$$\text{s.t.} \quad \left\langle \mathbf{y}^{(i)}, \mathbf{U}\phi(x^{(i)})\right\rangle \geq 1-\bar{\xi}^{(i)}, \xi_i \geq 0, i\in\{1,\dots,m\}$$

Results and Comparisons

Benchmark Databases for Image Annotation

bug, green, insect, tree, wood

blue, cloud, ocean, sky, water



black, computer, drawing handle, screen

asian, boy, gun, man, white

anime, comic, people, red, woman

feet, flower, fur. red, shoes

Dataset	Number of labels	Number of training instances	Number of test instances	Number of average labels
Corel5k	260	4500	500	3.3965
Espgame	268	18689	2081	4.6859
Iaprtc12	291	17665	1962	5.7187

Results and Contributions : Comparison with State-of-the-arts

Comparable Results:

	Corel5K			Espgame			Iaprtc12		
Method	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
MBRM	24.0	25.0	24.0	18.0	19.0	18.0	24.0	23.0	23.0
JEC	27.0	32.0	29.0	24.0	19.0	21.0	29.0	19.0	23.0
TagProp	33.0	42.0	37.0	39.0	27.0	**32.0**	45.0	**34.0**	**39.0**
FastTag	32.0	**43.0**	37.0	**46.0**	22.0	30.0	**47.0**	26.0	34.0
JSVM	**48.5**	38.0	**42.6**	32.7	**31.6**	**32.2**	42.2	29.4	34.6
JSVM+Pol(2)	46.6	37.0	41.3	32.6	24.4	27.9	37.9	26.6	31.2
JSVM+Pol(3)	41.5	31.3	35.7	28.5	21.3	24.4	38.0	26.1	31.0

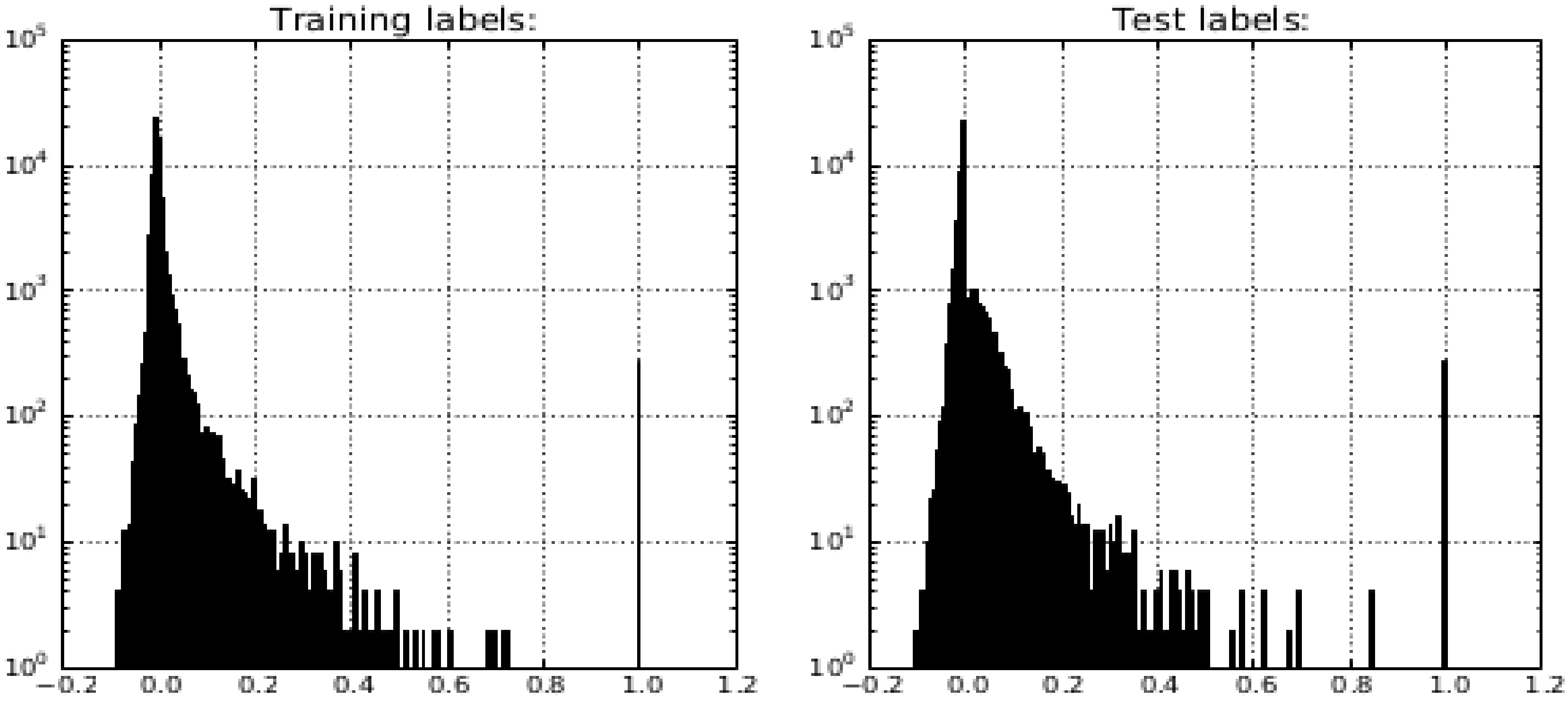
MBRM: *S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In Computer Vision and Pattern Recognition, 2004.*

JEC: *Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. Baselines for image annotation. International Journal of Computer Vision, 90:88–105, 2010.*

TagProp: *Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In International Conference on Computer Vision, 2009.*

FasTag: *Minmin Chen, Alice Zheng, and Kilian Q. Weinberger. Fast image tagging. In International Conference on Machine Learning, 2013*

Preference to Simpler Label-Dependence



Conclusions

A take-home message: A simpler output kernel is desirable to avoid overfitting in output structural dependencies.

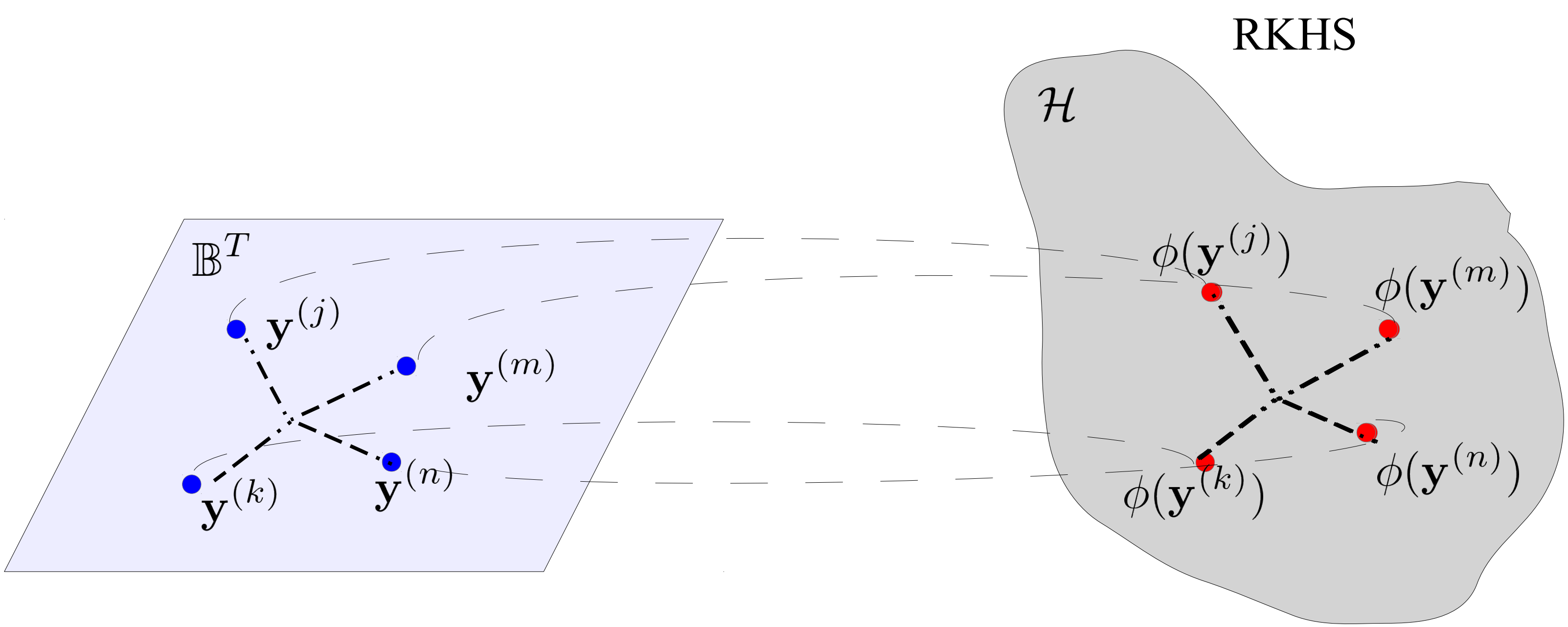
Thanks for your attention !

Questions and Answers ?

Joint SVM : Pre-Image

$$\begin{aligned}\hat{\mathbf{y}}^* &= \arg\max_{\mathbf{y}\in\{+1,-1\}^T} \langle \psi(\mathbf{y}), \mathbf{W}\phi(\hat{\mathbf{x}})\rangle \\ &= \arg\max_{\mathbf{y}\in\{+1,-1\}^T} \textstyle\sum_{i=1}^m \alpha_i \underbrace{K_\phi(\mathbf{x}^{(i)}, \hat{\mathbf{x}})}_{\beta_i} K_\psi(\mathbf{y}^{(i)}, \mathbf{y})\end{aligned}$$

$$\hat{\mathbf{y}}^* = \left(\sum_{k=1}^K \mathbf{y}^{(k)} w_k\right) \Big/ \sum_{k=1}^K w_k \qquad w_j = \sum_{i=1}^m \alpha_i \beta_i K_\psi(\mathbf{y}^{(i)}, \mathbf{y}^{(j)})$$



Understanding New Regularization

$$\frac{1}{2}\mathbf{tr}(\mathbf{W}^\top\mathbf{P}\mathbf{P}^\top\mathbf{W}) = \frac{1}{2}\mathbf{tr}(\mathbf{P}\mathbf{P}^\top\mathbf{W}\mathbf{W}^\top)$$

lemma: for positive (semi-)definite matrices A and B:

$$\mathbf{tr}(AB)^m \leq \{\mathbf{tr}(A)^{2m}\mathbf{tr}(B)^{2m}\}^{1/2}$$

where m is positive integer.

$$\frac{1}{2}\mathbf{tr}(\mathbf{P}\mathbf{P}^\top\mathbf{W}\mathbf{W}^\top) \leq \frac{1}{2}\mathbf{tr}(\mathbf{P}\mathbf{P}^\top)\mathbf{tr}(\mathbf{W}\mathbf{W}^\top) = \frac{1}{2}||\mathbf{P}||_F^2||\mathbf{W}||_F^2$$

"*On Some Matrix Trace Inequalities*", Zubeyde Ulukok and Ramazan Trkmen, Journal of Inequalities and Applications, 2010

Multiple SVMs: train T SVMs independently

- too expensive (T can be very large)
- ignore inter-label dependencies

	Training Time (sec)	Testing Time (sec)	Testing Performance Precision (%)	Recall (%)	F1 (%)
Independent SVMs (Gau)	6285.11	117.20	15.3	22.1	18.1
Independent SVMs (Pol)	4612.23	147.9	15.1	29.7	20.0
Joint SVM (Gau)	80.68	**6.92**	40.8	37.1	38.9
Joint SVM (Pol)	**76.48**	9.11	**48.5**	**38.0**	**42.6**