Convex Two-Layer Modeling

Özlem Aslan, *Hao Cheng*, *Dale Schuurmans*
UNIVERSITY OF ALBERTA
Xinhua Zhang
NICTA

Question

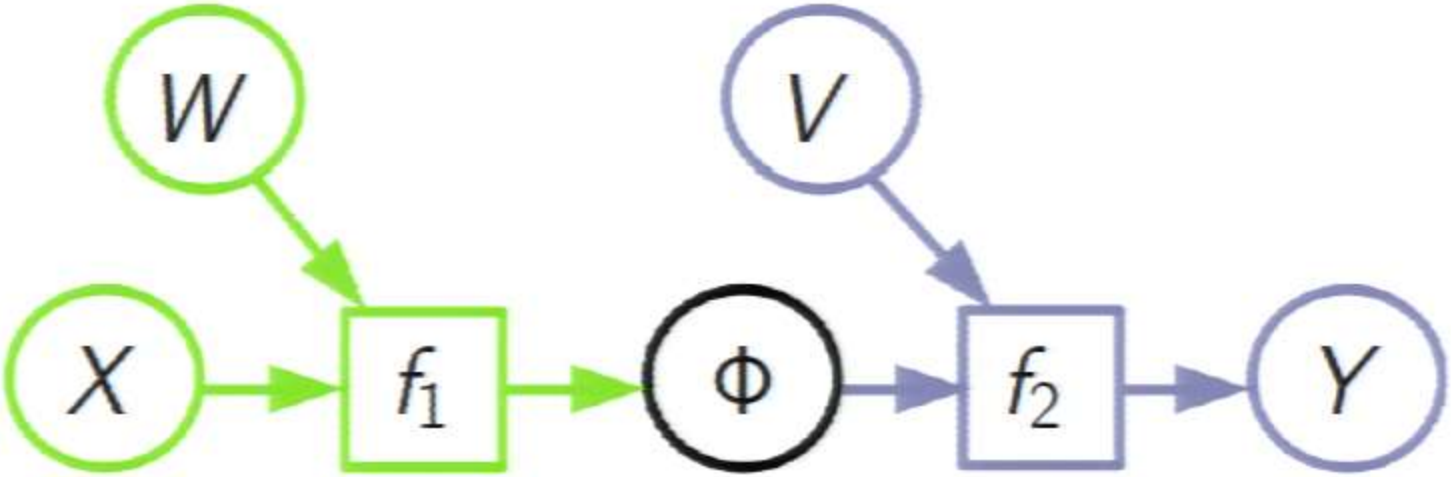
- Useful convex form of multi-layer training?

✔ Yes (two nonlinear layers)

- Scale beyond small problems?

✔ Yes

Two-Layer Conditional Model



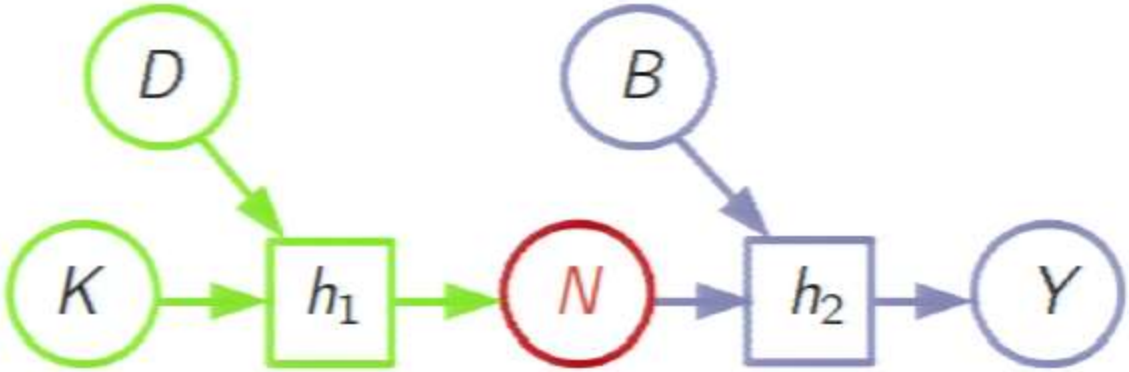
$$L_1(WX, \Phi) + \frac{\alpha}{2}\|W\|_F^2 \qquad L_2(V\Phi, Y) + \frac{\beta}{2}\|V\|_F^2$$

$$\min_{W,V} \min_{\Phi} F_1(W, \Phi) + \gamma F_2(\Phi, V)$$

Not jointly convex

Reformulate

- Latent feature kernel $N = \Phi'\Phi$

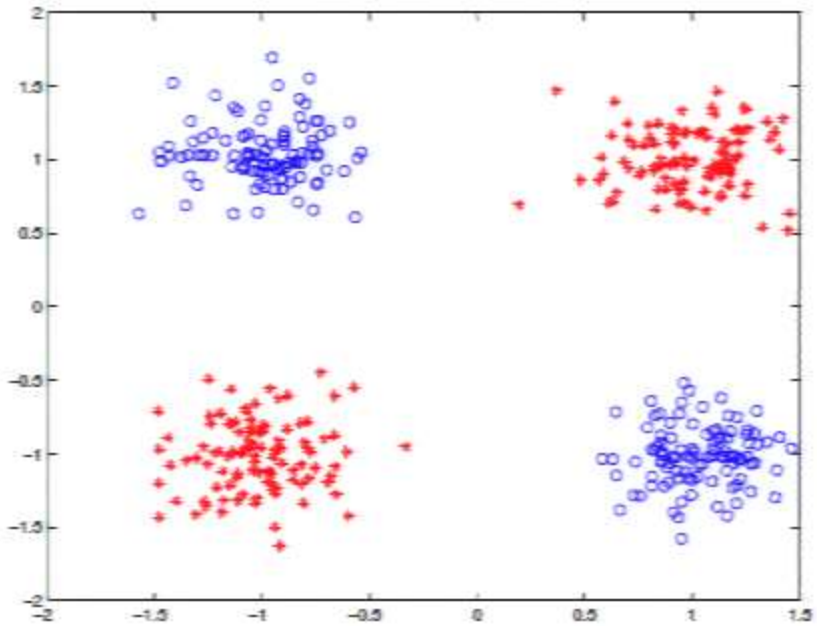


$\underbrace{\tilde{L}_1(DK, N) + \frac{\alpha}{2}\operatorname{tr}(D'N^\dagger DK)}$ $\qquad$ $\underbrace{L_2(B, Y) + \frac{\beta}{2}\operatorname{tr}(BN^\dagger B')}$

$$\min_{N \in \mathcal{N}_0} \min_{D \in \mathrm{Im}(N)} \min_{B \in \mathrm{Im}(N)} H_1(N, D) + \gamma H_2(B, N)$$

- Relax $\mathcal{N}_0 \subseteq \mathcal{N}_2$
- Low rank optimization for scalability

Sample Evaluations



Poster: Fri81

	XOR	MNIST	USPS	COIL
Relaxed Clust 2 layer	49.8 $\pm$0.7	13.7 $\pm$0.6	46.6 $\pm$1.0	45.0 $\pm$0.8
Trans SVM 1 layer	50.2 $\pm$1.2	11.4 $\pm$0.5	11.3 $\pm$0.4	14.9 $\pm$0.4
SVM 1 layer	50.3 $\pm$1.1	11.2 $\pm$0.4	10.7 $\pm$0.4	15.6 $\pm$0.5
Local 2 layer	4.2 $\pm$0.9	16.3 $\pm$0.6	9.7 $\pm$0.5	12.8 $\pm$0.6
Convex 2 layer	**0.2** $\pm$0.1	**8.8** $\pm$0.4	**6.6** $\pm$0.4	**8.2** $\pm$0.4