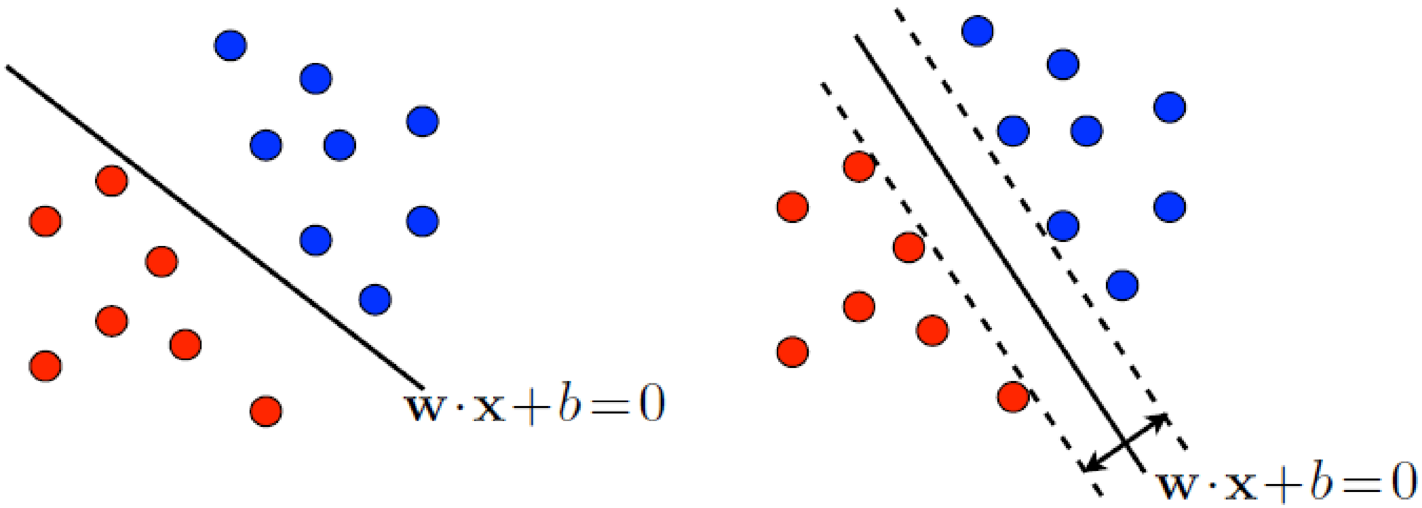
# Maximum Margin Clustering
# NIPS 2004

# SVM Classification




- Maximum Margin Classification

# SVM Classification

$$\gamma^{*-2} = \min_{\mathbf{w},b,\epsilon} \|\mathbf{w}\|^2 + C\epsilon^\top \mathbf{e} \quad \text{subject to} \quad y^i(\mathbf{w}^\top \phi(\mathbf{x}^i) + b) \geq 1 - \epsilon_i, \forall_{i=1}^N, \; \epsilon \geq 0$$

$$= \max_{\boldsymbol{\lambda}} 2\boldsymbol{\lambda}^\top \mathbf{e} - \langle K \circ \boldsymbol{\lambda}\boldsymbol{\lambda}^\top, \mathbf{y}\mathbf{y}^\top \rangle \quad \text{subject to} \quad 0 \leq \boldsymbol{\lambda} \leq C, \; \boldsymbol{\lambda}^\top y = 0 \qquad (3)$$

$$\langle a, b \rangle = \sum_{ij} a_{ij} b_{ij}$$

$$\boldsymbol{\lambda}^\top (K \circ \mathbf{y}\mathbf{y}^\top)\boldsymbol{\lambda} = \langle K \circ \mathbf{y}\mathbf{y}^\top, \boldsymbol{\lambda}\boldsymbol{\lambda}^\top \rangle = \langle K \circ \boldsymbol{\lambda}\boldsymbol{\lambda}^\top, \mathbf{y}\mathbf{y}^\top \rangle$$

- Supervised maximum margin training
- This is an standard QP problem:
  - active sets methods
  - interior point method

# Problem Definition

- Given data x1,x2,...,xn, we wish to assign the data points to two classes {-1,+1} such that separation between the two classes is as wide as possible

$$\min_{y \in \{-1,+1\}^N} \quad \max_{\lambda} 2\lambda^\top e - \langle K \circ \lambda\lambda^\top, yy^\top \rangle \quad \text{subject to} \quad 0 \le \lambda \le C \quad \lambda^\top y = 0$$

- Integer programming
- May leads to trivial solution, highly unbalanced clusters
- Not a convex function of y!

# Problems & solutions

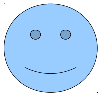
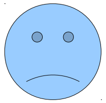
$$\min_{\mathbf{y}\in\{-1,+1\}^N} \quad \max_{\lambda} 2\lambda^\top e - \langle K \circ \lambda\lambda^\top, yy^\top \rangle \quad \text{subject to} \quad 0 \le \lambda \le C \quad \lambda^\top y = 0$$

- Unbalanced cluster problem
  - Impose a constrain on $-\ell \le e^\top y \le \ell$
- Integer programming
  - Soft clustering
- Non-convexity
  - Set b=0 to drop the constraint $\lambda^\top y = 0$
  - Centering the data at the origin and more...

# Re-express optimization problem

$$\gamma^{*-2}(\mathbf{y}) \;=\; \max_{\boldsymbol{\lambda}} \; 2\boldsymbol{\lambda}^\top \mathbf{e} - \langle K \circ \boldsymbol{\lambda}\boldsymbol{\lambda}^\top, \mathbf{y}\mathbf{y}^\top \rangle \quad \text{subject to} \quad 0 \le \lambda \le C$$

$$\gamma^{*-2}(M) \;=\; \max_{\boldsymbol{\lambda}} \; 2\boldsymbol{\lambda}^\top \mathbf{e} - \langle K \circ \boldsymbol{\lambda}\boldsymbol{\lambda}^\top, M \rangle \quad \text{subject to} \quad 0 \le \lambda \le C \qquad M = \mathbf{y}\mathbf{y}^\top$$

- Objective function is linear 
- M=yy' and M is $[-1,+1]^n$ is not convex 

# Indirectly enforce $M=yy^T$

- M encodes equivalence relation
  - transitive,reflexive and symmetric
- M has two equivalence classes

$\mathcal{L}_1$: $m_{ii} = 1;\ m_{ij} = m_{ji};\ m_{ik} \geq m_{ij} + m_{jk} - 1;\ \forall_{ijk}$

$\mathcal{L}_2$: $m_{jk} \geq -m_{ij} - m_{ik} - 1;\ \forall_{ijk}$

$\mathcal{L}_3$: $\sum_i m_{ij} \leq N - 2;\ \forall_j$

$\mathcal{L}_4$: $-\ell \leq \sum_i m_{ij} \leq \ell;\ \forall_j$

$$\min_{M \in [-1,+1]^{N \times N}} \max_{\boldsymbol{\lambda}} 2\boldsymbol{\lambda}^\top \mathbf{e} - \langle K \circ \boldsymbol{\lambda}\boldsymbol{\lambda}^\top, M \rangle$$

subject to $0 \leq \boldsymbol{\lambda} \leq C, \mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_4, M \succeq 0$

$$\mathbf{y} = \sqrt{\lambda_1}\mathbf{v}_1$$

# Semi-supervised learning

- Combine both unlabeled data and labeled data to produce a more accurate classification result

$\mathcal{S}_1$: $m_{ij} = y_i y_j$ for *labeled* examples $i, j \in \{1, ..., n\}$

$\mathcal{S}_2$: $\sum_{i=1}^{n} m_{ij} \geq 2 - n$ for *unlabeled* examples $j \in \{n+1, ..., N\}$

# Experiment Results

| | Gaussians | Circles | A I | Joined Circles | Digits | Faces |
|---|---|---|---|---|---|---|
| Maximum Margin | 1.25 | 0 | 0 | 1 | 3 | 0 |
| Spectral Clustering | 1.25 | 0 | 0 | 24 | 6 | 16.7 |
| K-means | 5 | 50 | 38.5 | 50 | 7 | 24.4 |

Table 1: Percentage misclassification errors of the various clustering algorithms on the various data sets.

| | HWD 1-7 | HWD 2-3 | UCI Austra. | UCI Flare | UCI Vote | UCI Diabet. |
|---|---|---|---|---|---|---|
| Max Marg | 3.3 | 4.7 | 32 | 34 | 14 | 35.55 |
| Spec Clust | 4.2 | 6.4 | 48.7 | 40.7 | 13.8 | 44.67 |
| TSVM | 4.6 | 5.4 | 38.7 | 33.3 | 17.5 | 35.89 |
| SVM | 4.5 | 10.9 | 37.5 | 37 | 20.4 | 39.44 |

Table 2: Percentage misclassification errors of the various semisupervised learning algorithms on the various data sets. SVM uses no unlabeled data. TSVM is due to [8].

# My Comments

- Class balance problem
- For pure clustering problem, soft margin is not necessary. It leads to C value can be arbitrary, which is consistent with experiment report by the author.