Alternative Ways to Solve Optimization Problem in Support Vector Decomposition Machine

Group Presentation

Yongxin (Taylor) Xi

May 22th, 2007

OVERVIEW

- **Part I**

  Introduction and problem formulation

- **Part II**

  Relaxation of bilinear terms into convexity

- **Part III**

  Sequential optimization: an efficient way to achieve suboptimality

- **Part IV**

  Reduction of problem complexity via linear algebra theory

- **Part V**

  Problem reformulation and improvements

Part I Problem Introduction

Task

- Machine learning:

  Use part of the samples with labels to train a classifier. After the learning phase, given new sample without labels, let the machine tell which class does the new sample belongs to.

Data

n samples: (n=392: 7 people, 7 stimuli, 8 rounds)

A certain person looking at a certain stimulus

m features: (m=2048)

IT cortex brain response: fMRI signal of m voxels

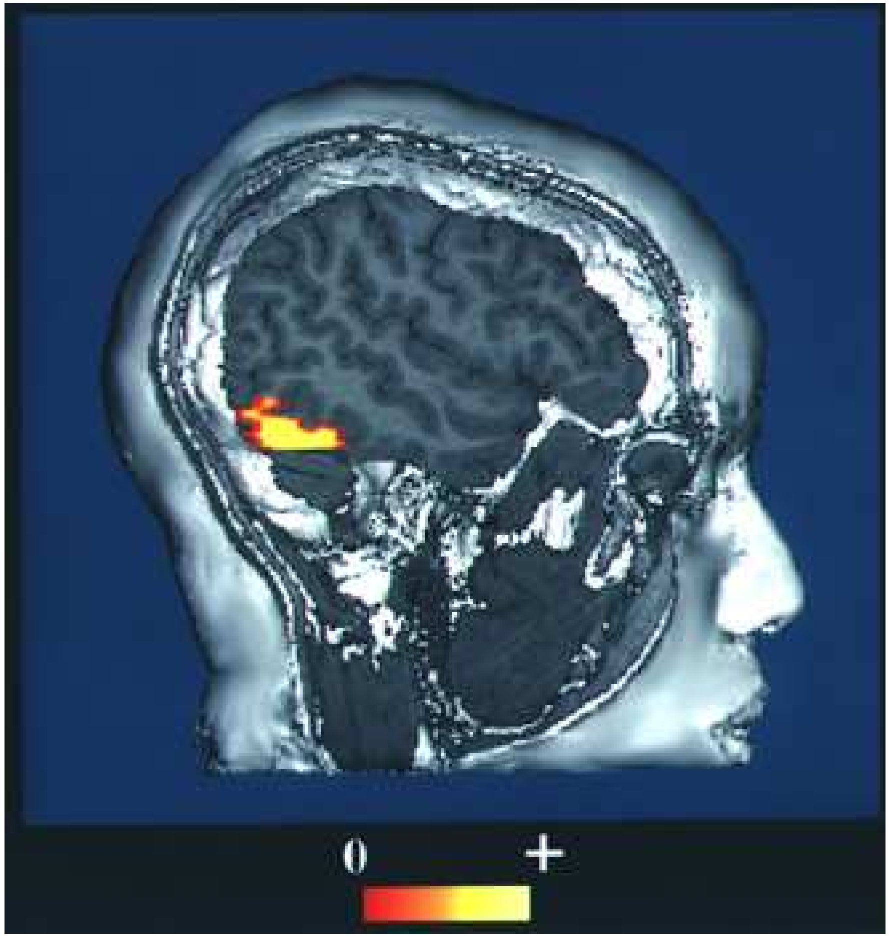
K classes: (K=7)

- female-face
- male-face
- monkey-face
- dog-face
- house
- chair
- shoe

$$X_{nxm} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix} = \begin{pmatrix} x_1 \\ \ldots \\ x_n \end{pmatrix}$$

$$Y_{nxK} = \begin{pmatrix} Y_{11} & \cdots & Y_{1K} \\ \vdots & \ddots & \vdots \\ Y_{n1} & \cdots & Y_{nK} \end{pmatrix}$$

Imaging



- This is how our brain normally responses when viewing a face

MRI machine is huge and expensive…

SVD

$$X_{nxm} \approx Z_{nxl} W_{lxm}$$

$$X_{nxm} = U_{nxn} \Sigma_{nxn} V^{T}{}_{nxm}$$

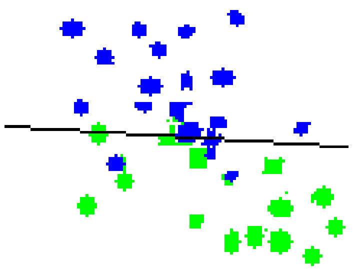
$$U^T U = I \quad V^T V = I$$

$$X_{nxm} \approx U(:,1:l)\Sigma(1:l,1:l)V^T(1:l,:)$$

$$Z_{nxl} = U(:,1:l)\Sigma(1:l,1:l)$$

$$W_{lxm} = V^T(1:l,:)$$

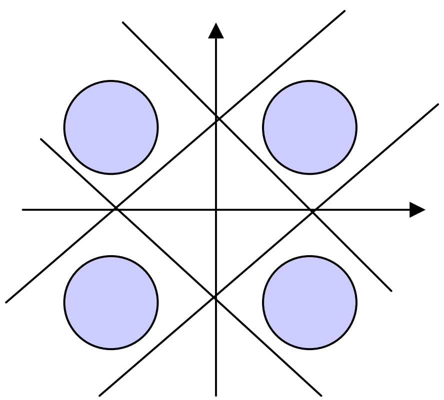
SVM



$$\min_{w,b,\xi_i} \quad \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$\text{Subject to}: \; y_i(x_i^T w + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0, \forall i$$



$$\text{Min} \quad \frac{1}{2}\sum_{j=1}^{K}\|w_j\|^2 + C\sum_{j=1}^{K}\sum_{i=1}^{N}\xi_i^j$$

$$\text{Subject to}: \; Y_{ij}(w_j^T x_i + b_j) \geq 1 - \xi_i^j, i = 1,2,\ldots,N; \; j = 1,2,\ldots,K$$

$$\xi_i^j \geq 0, \forall i, j$$

$$f(x) = \arg\max_j((w_j^T x) + b_j), \; j = 1,2,\ldots,K$$

SVDM Problem formulation (Francisco 06)

$$\underset{Z,W,Q}{Min} \| X - ZW \|_F^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{K} \max(0, \mu - Y_{ij}[ZQ]_{ij})$$

$$Subject\ to: \quad Z_{i,1} = 1$$

$$Z_{i,2:end} \le 1 \ \ , i = 1, 2, ..., n$$

$$\| Q_{:,j} \|^2 \le 1, \ i = 1, 2, ..., l; \ \ \text{j=1,2,....,k}$$

$$X_{nxm} \in R^{nxm}, Y_{nxK} \in \{1, -1\}^{nxK} \qquad \lambda > 0, u > 0, l \in Z^{++}$$

$$Z \in R^{nxl}, W \in R^{lxm}, Q \in R^{lxK}$$

My Constraints:

$$Subject\ to: \quad | W(i,:) |_2 = 1 \ \ , i = 1, 2, ..., l$$

$$| Q_{i,j} | \le 1, \ i = 1, 2, ..., l; \ \ \text{j=1,2,....,k}$$

What's changed

- $\| Q \|_F^2$ omitted from the objective

  Instead of minimizing to achieve the largest margin of fuzzy separation, the constraint $\|Q_{:,j}\|^2 \leq 1$ regulates the margin in each SVM to be greater than 2.
- intercept b removed
- threshold changed from 1 to parameter u

Part II
Relaxation of bilinear terms into convexity

Non-convexity

$$\underset{Z,W,Q}{Min} \| X - ZW \|_F^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{K} \max(0, \mu - Y_{ij}[ZQ]_{ij})$$

$$Subject\ to: \quad |W(i,:)|_2 = 1 \ , i = 1, 2, ..., l$$

$$|Q_{i,j}| \leq 1, \ i = 1, 2, ..., l; \quad \text{j=1,2,....,k}$$

- All variables appear in bilinear forms: overwhelming non-convexity!

Alternative Relaxation Methods

- GP
- Convex enclosure relaxation

  Reformulation-Linearization Technique

  (RLT, SA92, She02)
- Taylor Expansion Approximation

Convex enclosure relaxation (RLT)

- Simple z=xy

$$(x-x^L)(y-y^L)\geq 0 \qquad z\geq x^L y + xy^L - x^L y^L$$

$$(x-x^L)(y^U-y)\geq 0 \qquad z\leq x^L y + xy^U - x^L y^U$$

$$(x^U-x)(y-y^L)\geq 0 \qquad z\leq x^U y + xy^L - x^U y^L$$

$$(x^U-x)(y^U-y)\geq 0 \qquad z\geq x^U y + xy^U - x^U y^U$$

- Problem:

  W, Z and Q are not nicely bounded

$$x^L=(1-\alpha)x, y^L=(1-\alpha)y, x^U=(1+\alpha)x, y^U=(1+\alpha)y$$

Taylor Expansion Approximation

$$z = f(x, y) = xy$$

$$f(x + \Delta x, y + \Delta y) \approx f(x, y) + \frac{\partial}{\partial x} f(x, y) \cdot \Delta x + \frac{\partial}{\partial y} f(x, y) \cdot \Delta y$$

$$\Delta z \approx \frac{\partial}{\partial x} f(x, y) \cdot \Delta x + \frac{\partial}{\partial y} f(x, y) \cdot \Delta y = y \cdot \Delta x + x \cdot \Delta y$$

- Δx, Δy cannot be too big

$$|\Delta x| \leq \alpha |x|, \quad |\Delta y| \leq \alpha |y|$$

Result of relaxation to the problem

$$\underset{Z,W,Q}{Min} \| X - ZW \|_F^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{K} \max(0, \mu - Y_{ij}[ZQ]_{ij})$$

- Although we pick a nice point to start, both methods do NO good to the original problem
- Reasons:

  1) error accumulation

  2) tradeoff of α selection (variation range vs. approximation accuracy)

Part III

Sequential optimization:

--an efficient way to achieve suboptimality

Sequential Optimizing Scheme

- Fixing two of the variable matrices while optimizing the other

$$\min_{Z,W,Q} \| X - ZW \|_F^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{K} \max(0, \mu - Y_{ij}[ZQ]_{ij})$$

- Given Z, Q, solve for W—linear regression
- Given Z, W, solve for Q—LP
- Given W, Q, solve for Z—QP problem
- Converge to local optimum

Given Z, W, solve for Q—LP

$$\min_{Z,W,Q} \| X - ZW \|_F^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{K} \max(0, \mu - Y_{ij}[ZQ]_{ij})$$

$$\min_{Q} \sum_{i=1}^{n} \sum_{j=1}^{K} \max(0, \lambda(\mu - Y_{ij}[ZQ]_{ij})) \qquad (5)$$

$$Subject\ to: \quad |Q_{i,j}| \le 1,\ i = 1, 2, \ldots, l; \quad \text{j=1,2,....,k} \qquad (6)$$

$$\min_{Q(:,j),hi} \quad \sum_{i=1}^{n} hi \qquad (7)$$

$$Subject\ to: \ |Q_{i,j}| \le 1,\ i = 1, 2, \ldots, l \qquad (8)$$

$$hi \ge 0,\ i = 1, 2, \ldots, n \qquad (9)$$

$$hi \ge \lambda(\mu - Y_{ij}[ZQ]_{ij}) \qquad (10)$$

Given W, Q, solve for Z—QP problem

$$\underset{Z}{Min} \| X - ZW \|_F^2 + \sum_{i=1}^{n} \sum_{j=1}^{K} \max(0, \lambda(\mu - Y_{ij}[ZQ]_{ij})) \qquad (11)$$

$$\underset{Z(i,:)}{Min} \| X(i,:) - Z(i,:)W \|_F^2 + \sum_{j=1}^{K} hj \qquad (12)$$

$$Subject\ to: \ hj \geq 0,\ j = 1, 2, \ldots, K \qquad (13)$$

$$hj \geq \lambda(\mu - Y_{ij}[Z(i,:)Q]_j),\ j = 1, 2, \ldots, K \qquad (14)$$

Given Z, Q, solve for W
—linear regression

$$\min_{Z,W,Q} \| X - ZW \|_F^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{K} \max(0, \mu - Y_{ij} [ZQ]_{ij})$$

$$\min_{W} \| X - ZW \|_F^2 \qquad (15)$$

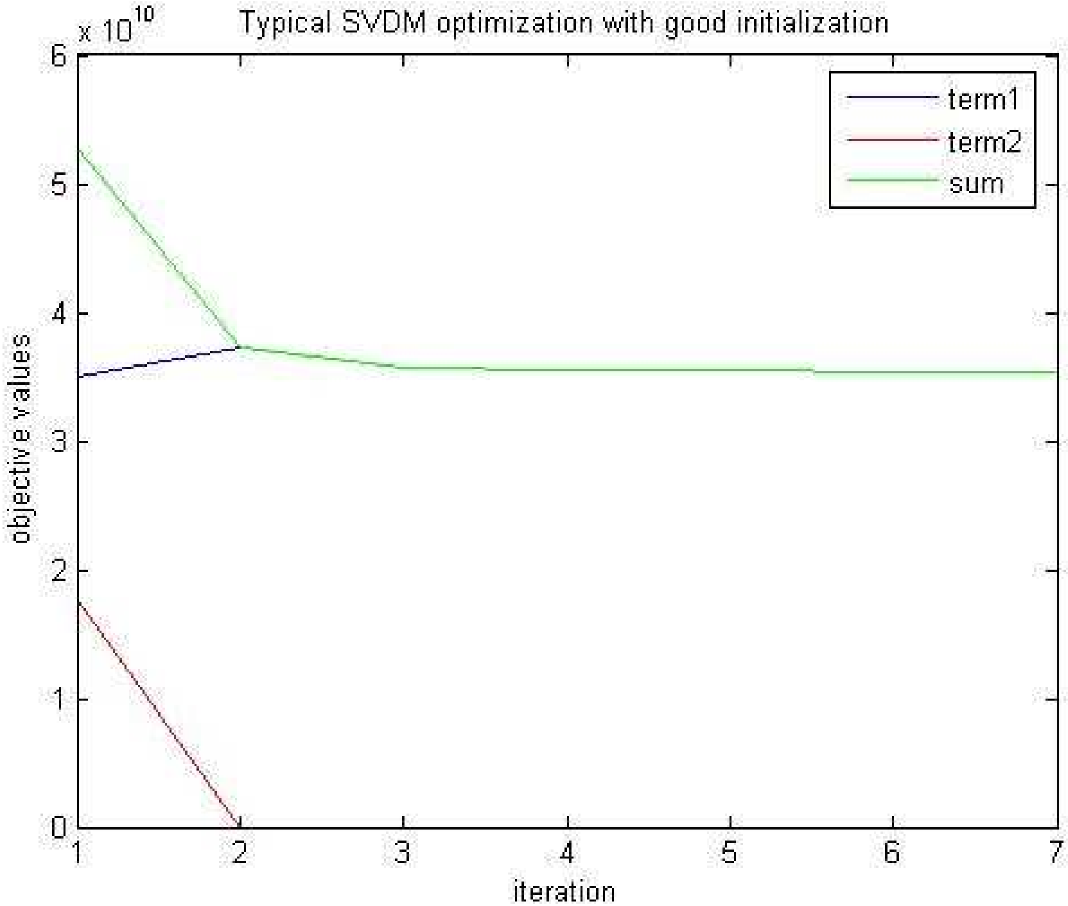
$$Subject\ to: \ |W(i,:)|_2 = 1 \ , \ i = 1, 2, ..., l \qquad (16)$$

- Small trick to make |W(i,:)|=1, i=1,2,…,l
- Or:

$$|W(i,:)|_2 \leq 1 \ , \ i = 1, 2, ..., l \qquad (17)$$

Convergence

- During each iteratio[n] three matrices Z, W and Q get updated alternatively



SVDM Testing Phase

- Given a new data, first find its reduced dimensional representation, then time the classification matrix Q to form the prediction vector

- The class which has highest prediction score wins:

$$f(x) = \arg\max_j ((w_j^T x) + b_j),\ j = 1, 2, \ldots, K$$

Testing methods

- **Cross Validation (leave one out)**
  - a. 7 category test
  - b. 2 category test (face versus object)
- **Independent Testing**

  Independent among subjects

  - I. Testing on new subjects (train on first 70% subjects, test on 30%)
    - a. 7 category test
    - b. 2 category test (face versus object)
  - II. Testing on new runs (train on first 75% collectively from each subject)
    - a. 7 category test
    - b. 2 category test (face versus object)

  Independent within subject (train on first 75% of one subject only)

  - a. 7 category test
  - b. 2 category test (face versus object)

Testing results

Test of SVDM on new data		7 Categories (K=7)	2 Categories (K=2)
Random guessing		14.28%	50%
Independent among subjects	Generalization on new subject	30%	69%
	Generalization on new sample	48%	93.9%
Independent within subjects		53%	92%
Cross-Validation		50%	81%

An interesting experiment

Independent among subjects

accuracy	F face	M face	Monkey	Dog	House	Chair	Shoe
F face	0	0.46429	0.60714	0.64286	0.85714	0.89286	0.67857
M face	0.46429	0	0.64286	0.60714	0.89286	0.96429	0.82143
Monkey	0.60714	0.64286	0	0.67857	0.75	0.75	0.78571
Dog	0.64286	0.60714	0.67857	0	0.96429	0.96429	0.92857
House	0.85714	0.89286	0.75	0.96429	0	0.75	0.75
Chair	0.89286	0.96429	0.75	0.96429	0.75	0	0.64286
Shoe	0.67857	0.82143	0.78571	0.92857	0.75	0.64286	0

Part IV
Reduction of problem complexity via linear algebra theory

QR Factorization

- QR decomposition of a real square matrix *A* is a decomposition of *A* as

$$A = QR$$

$$A \in R^{mxn}, m > n;\ Q \in R^{mxn}, Q^T Q = I;\ \ R \in R^{nxn}$$

- Factorize our data matrix X $\quad X = LQ^T = [L\ 0]\begin{bmatrix} Q^T \\ \tilde{Q}^T \end{bmatrix}$
- Denote $\quad B = \begin{bmatrix} Q & \tilde{Q} \end{bmatrix}$
- Then when solving for W:

$$\begin{aligned} \| X - ZW \|_F^2 &= \| \begin{bmatrix} L & 0 \end{bmatrix} B^T - ZW \|_F^2 \\ &= \| (\begin{bmatrix} L & 0 \end{bmatrix} - ZWB) B^T \|_F^2 \\ &= \| (\begin{bmatrix} L & 0 \end{bmatrix} - ZWB) \|_F^2 \\ &= \| (\begin{bmatrix} L & 0 \end{bmatrix} - Z[\tilde{W}_1\ \ \tilde{W}_2]) \|_F^2 \\ &= \| L - Z\tilde{W}_1 \|_F^2 + \| 0 - Z\tilde{W}_2 \|_F^2 \end{aligned}$$

Results

- Complexity of solving W reduces from n by m (392 by 2048) to n by n (392 by 392).
- Approximately 4 times faster in overall convergence

Part V
Problem reformulation and improvements

Problem Reformulation

- Original problem *(Francisco Pereira 06)*

$$\underset{Z,W,Q}{Min} \| X - ZW \|_F^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{K} \max(0, \mu - Y_{ij}[ZQ]_{ij})$$

- New formulation

$$\underset{Z,W,Q}{Min} \| X - ZW \|_F^2 + \lambda (\frac{1}{2} \| Q \|_F^2 + C \sum_{i=1}^{n} \sum_{j=1}^{K} \max(0, 1 - Y_{ij}([ZQ]_{ij} + b_j)))$$

$$Subject\ to: \quad |W(i,:)|_2 = 1 \ \ , i = 1, 2, ..., l$$

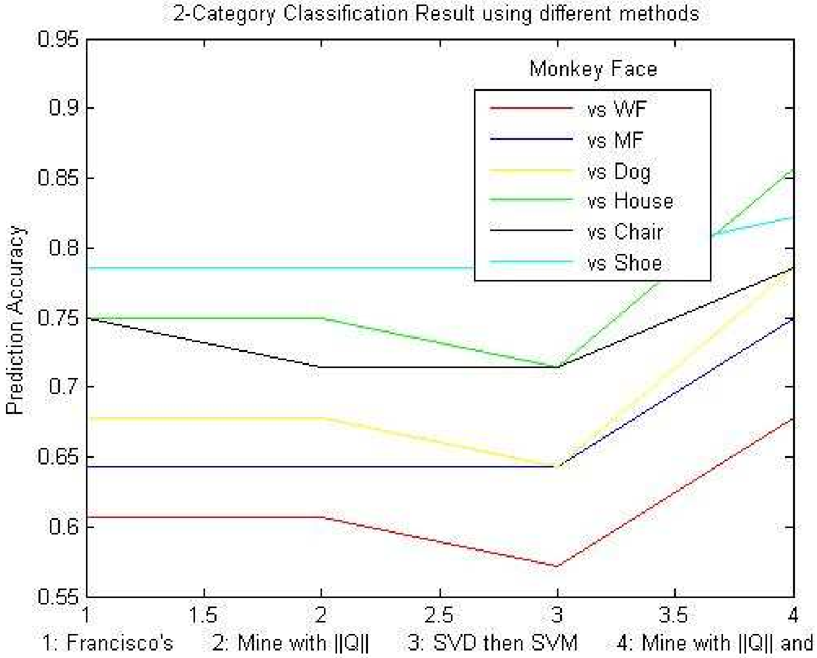
1. Standard SVM (to maximize the margin of classes)
2. In this way u can be set to 1, rather than tedious searching for the best parameter.
3. No constraints on Q   Only change: LP→QP (when solve for Q)
4. With bj, j=1,2,…K added

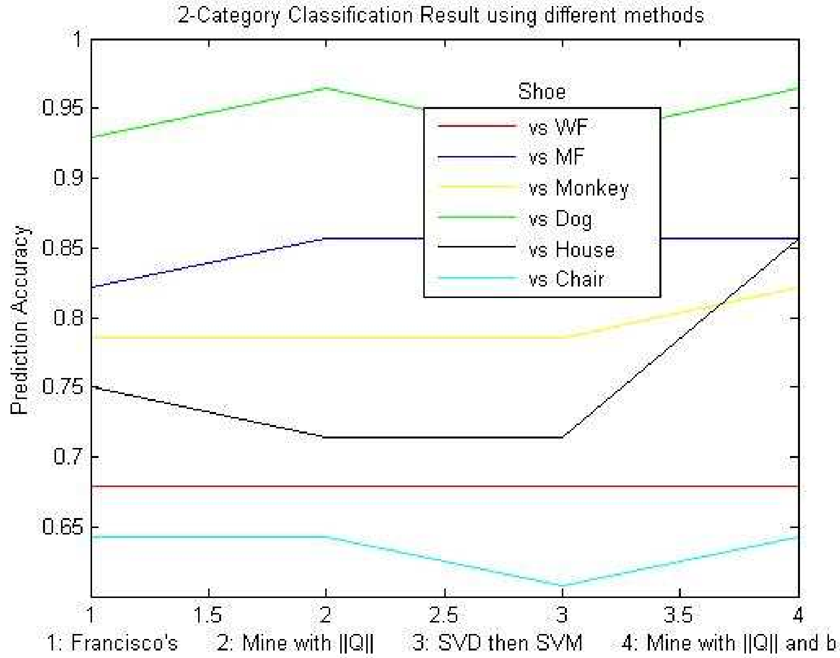
Independent among subjects

accuracy	F face	M face	Monkey	Dog	House	Chair	Shoe
F face	0	0.46429	0.60714	0.64286	0.85714	0.89286	0.67857
M face	0.46429	0	0.64286	0.60714	0.89286	0.96429	0.82143
Monkey	0.60714	0.64286	0	0.67857	0.75	0.75	0.78571
Dog	0.64286	0.60714	0.67857	0	0.96429	0.96429	0.92857
House	0.85714	0.89286	0.75	0.96429	0	0.75	0.75
Chair	0.89286	0.96429	0.75	0.96429	0.75	0	0.64286
Shoe	0.67857	0.82143	0.78571	0.92857	0.75	0.64286	0

accuracy	F face	M face	Monkey	Dog	House	Chair	Shoe
F face	0	0.57143	0.67857	0.57143	0.82143	0.92857	0.67857
M face	0.57143	0	0.75	0.53571	0.96429	0.96429	0.85714
Monkey	0.67857	0.75	0	0.78571	0.85714	0.78571	0.82143
Dog	0.57143	0.53571	0.78571	0	0.96429	0.96429	0.96429
House	0.82143	0.96429	0.85714	0.96429	0	0.75	0.85714
Chair	0.92857	0.96429	0.78571	0.96429	0.75	0	0.64286
Shoe	0.67857	0.85714	0.82143	0.96429	0.85714	0.64286	0

Fixing training set and testing set



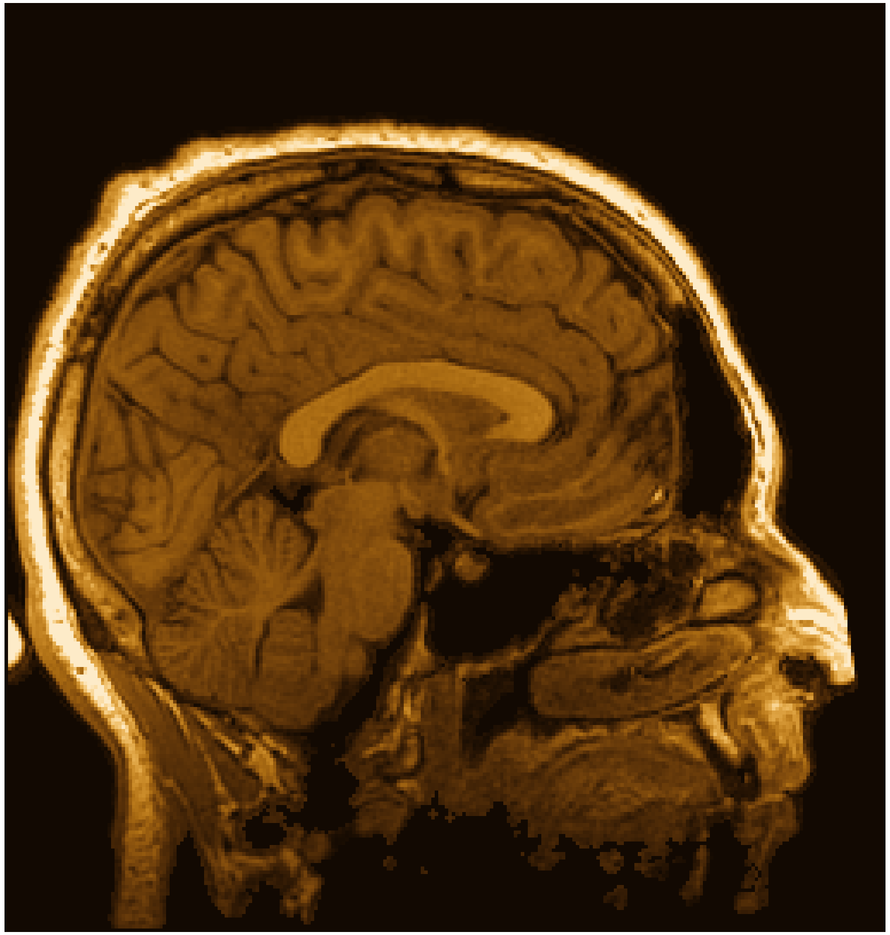


- 24/42 classifiers increase in performance;
- 12/42 classifiers remain the same;
- 6/42 classifiers decrease
- Woman vs. dog; man vs. dog; dog vs. woman;
- Dog vs. man; House vs. woman; woman vs. House

Wrap-up

- Introduction and problem formulation
- Relaxation of bilinear terms into convexity
- Sequential optimization: an efficient way to achieve suboptimality
- Reduction of problem complexity via linear algebra theory
- Problem reformulation and improvements (future work)

Thanks for your participation!



- Thinking of any questions?
  - –Ask them ☺