Collaborative filtering in Hilbert spaces with spectral regularization

Jacob Abernethy[1] Francis Bach[2]
Theodoros Evgeniou[3] Jean-Philippe Vert[4]

[1]UC Berkeley

[2]INRIA / Ecole normale superieure de Paris

[3]INSEAD

[4]Mines ParisTech / Institut Curie / INSERM

Göttingen University, Centre for Statistics seminar, Nov 27, 2009.

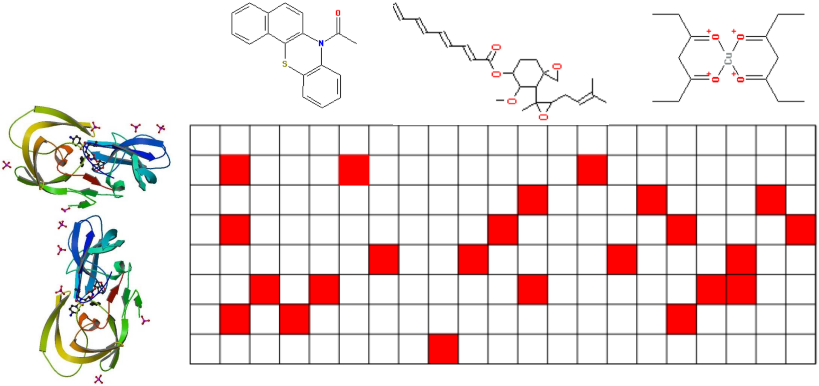
The NETFLIX challenge

18,000 movies

480,000 users

x	1	1	x	...	x
x	x	x	5	...	x
x	x	3	x	...	x
x	4	3	x	...	2
...	x	x	x	...	x
x	5	x	1	...	x
x	x	3	3	...	x
x	1	x	x	...	2

In silico chemogenomics



Formalization

The problem

- $\mathcal{X}$ and $\mathcal{Y}$ two sets ("customers" and "movies").
- Training data: $(\mathbf{x}_i, \mathbf{y}_i, t_i)_{i=1,\ldots,n} \in (\mathcal{X}, \mathcal{Y}, \mathbb{R})^n$ some ratings t_i by customer $\mathbf{x}_i$ for movie $\mathbf{y}_i$
- $n_{\mathcal{X}} \leq n$ (resp. $n_{\mathcal{Y}} \leq n$) the number of different customers (resp. movies) in the training data.
- Goal: learn the "rating function" $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Existing strategies

1. Collaborative filtering
2. Regression over pairs

Formalization

The problem

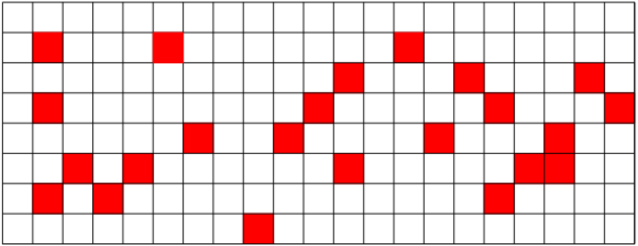
- $\mathcal{X}$ and $\mathcal{Y}$ two sets ("customers" and "movies").
- Training data: $(\mathbf{x}_i, \mathbf{y}_i, t_i)_{i=1,\ldots,n} \in (\mathcal{X}, \mathcal{Y}, \mathbb{R})^n$ some ratings t_i by customer $\mathbf{x}_i$ for movie $\mathbf{y}_i$
- $n_{\mathcal{X}} \leq n$ (resp. $n_{\mathcal{Y}} \leq n$) the number of different customers (resp. movies) in the training data.
- Goal: learn the "rating function" $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$.

Existing strategies

1. Collaborative filtering
2. Regression over pairs

Strategy 1: Collaborative Filtering (CF)

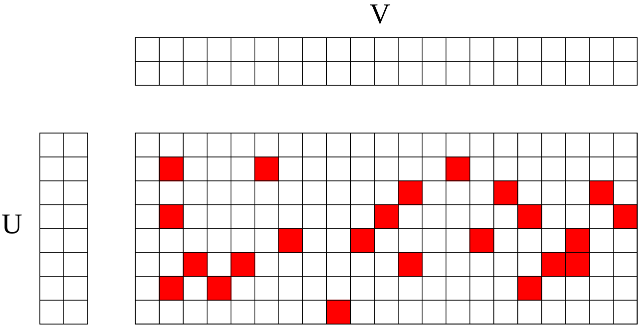
- Ignore any information about movies and customers
- $\mathcal{X} = \{\mathbf{x}^1, \ldots, \mathbf{x}^{n_\mathcal{X}}\}$ and $\mathcal{Y} = \{\mathbf{y}^1, \ldots, \mathbf{y}^{n_\mathcal{Y}}\}$ are finite
- Training data: large $n_\mathcal{X} \times n_\mathcal{Y}$ incomplete matrix F that describes the known ratings of some customers for some movies
- Goal: complete the matrix.



	🟥				🟥								🟥							
											🟥				🟥				🟥	
	🟥									🟥						🟥				🟥
						🟥			🟥					🟥				🟥		
		🟥		🟥							🟥						🟥	🟥		
	🟥		🟥													🟥				
								🟥												

CF by low-rank matrix approximation

- A common strategy for CF
- F has rank less than k ⇔ $\boxed{F = UV^\top}$ $U \in \mathbb{R}^{n_\mathcal{X} \times k}$, $V \in \mathbb{R}^{n_\mathcal{Y} \times k}$
- Examples: PLSA (Hoffmann, 2001), MMMF (Srebro et al, 2004)
- Numerical and statistical efficiency



CF by low-rank matrix approximation example

Fitting low-rank models (Srebro et al, 2004)

- Relax the (non-convex) rank of F into the (convex) trace norm of F: if $\sigma_i(F)$ are the singular values of F,

$$\mathrm{rank} F = \sum_i 1_{\sigma_i(F)>0} \qquad \|F\|_* = \sum_i \sigma_i(F)\,.$$

- i-th observation t_i corresponding to $\mathbf{x}_i = \mathbf{x}^{u(i)}$ and $\mathbf{y}_i = \mathbf{y}^{v(i)}$:

$$\min_{F\in\mathbb{R}^{n_\mathcal{X}\times n_\mathcal{Y}}} \sum_{i=1}^{n} \ell(t_i, F_{u(i),v(i)}) + \lambda\|F\|_*\,,$$

where $\ell(z,z')$ is a convex loss function.
- This is an SDP if ℓ is SDP-representable

CF by low-rank matrix approximation example

Fitting low-rank models (Srebro et al, 2004)

- Relax the (non-convex) rank of F into the (convex) trace norm of F: if $\sigma_i(F)$ are the singular values of F,

$$\mathrm{rank} F = \sum_i \mathbf{1}_{\sigma_i(F)>0} \qquad \|F\|_* = \sum_i \sigma_i(F) .$$

- i-th observation t_i corresponding to $\mathbf{x}_i = \mathbf{x}^{u(i)}$ and $\mathbf{y}_i = \mathbf{y}^{v(i)}$:

$$\min_{F \in \mathbb{R}^{n_{\mathcal{X}} \times n_{\mathcal{Y}}}} \sum_{i=1}^{n} \ell(t_i, F_{u(i),v(i)}) + \lambda \|F\|_* ,$$

  where $\ell(z, z')$ is a convex loss function.
- This is an SDP if ℓ is SDP-representable

Strategy 2: Regression over pairs

- $\mathcal{X}$ and $\mathcal{Y}$ represent the attributes of each customer/movie
- This is a classical regression problem over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- For example, take $\mathbf{z} = \mathbf{x} \otimes \mathbf{y}$ and find

$$f(\mathbf{x}, \mathbf{y}) = \mathbf{w}^\top \mathbf{z} = \mathbf{w}^\top (\mathbf{x} \otimes \mathbf{y})$$

by solving

$$\min_{\mathbf{w} \in \mathcal{X} \otimes \mathcal{Y}} \sum_{i=1}^{n} \ell(t_i, \mathbf{w}^\top (\mathbf{x}_i \otimes \mathbf{y}_i)) + \lambda \|\mathbf{w}\|^2 .$$

Regression over pairs with kernels

- Kernel methods (SVM...) are efficient methods to solve problems of the form

$$\min_{\mathbf{w}\in\mathcal{Z}} \sum_{i=1}^{n} \ell(t_i, \mathbf{w}^\top \mathbf{z}_i) + \lambda \|\mathbf{w}\|^2 \, .$$

- They require the definition of the kernel:

$$\begin{aligned} K_Z(\mathbf{z}, \mathbf{z}') &= \mathbf{z}^\top \mathbf{z}' \\ &= (\mathbf{x} \otimes \mathbf{y})^\top (\mathbf{x} \otimes \mathbf{y}) \\ &= (\mathbf{x}^\top \mathbf{x}') \times (\mathbf{y}^\top \mathbf{y}') \\ &= K_X(\mathbf{x}, \mathbf{x}') K_Y(\mathbf{y}, \mathbf{y}') \, . \end{aligned} \tag{1}$$

Comparison of both strategies

Collaborative filtering

$$\min_{F \in \mathbb{R}^{n_{\mathcal{X}} \times n_{\mathcal{Y}}}} \sum_{i=1}^{n} \ell(t_i, F_{u(i),v(i)}) + \lambda \|F\|_* \,.$$

- Use various spectral penalties of the matrix (rank, trace norm)
- No use of attribute, no prediction outside the training set

Regression over pairs

$$\min_{\mathbf{w} \in \mathcal{X} \otimes \mathcal{Y}} \sum_{i=1}^{n} \ell(t_i, \mathbf{w}^\top (\mathbf{x}_i \otimes \mathbf{y}_i)) + \lambda \|\mathbf{w}\|^2 \,.$$

- Flexible use of attributes with kernels
- No special treatment of repetitions in the training set

Our contribution

Goal

- Make a link between collaborative filtering and regression over pairs
- Develop methods that combine the advantages of both strategies

Contributions

- A general framework for CF with or without attributes, using kernels to describe attributes (“kernel-CF”)
- A family of algorithms in this setting

Our contribution

Goal

- Make a link between collaborative filtering and regression over pairs
- Develop methods that combine the advantages of both strategies

Contributions

- A general framework for CF with or without attributes, using kernels to describe attributes (“kernel-CF”)
- A family of algorithms in this setting

From CF to regression over pairs

- Represent the i-th customer $\mathbf{x}^i \in \mathcal{X}$ (resp. j-th movie $\mathbf{y}^j \in \mathcal{Y}$) by the i-th basis vector $e_i \in \mathbb{R}^{n_\mathcal{X}}$ (resp. $f_j \in \mathbb{R}^{n_\mathcal{Y}}$):

$$\phi_X(\mathbf{x}^i) = e_i\,, \quad \phi_Y(\mathbf{y}^j) = f_j\,.$$

- The rating $F_{i,j}$ of $\mathbf{x}^i$ for $\mathbf{y}^j$ is given by

$$F_{i,j} = e_i^\top F y_j = Tr\left(F^\top(\phi_X(\mathbf{x}^i) \otimes \phi_Y(\mathbf{y}^j))\right)\,.$$

- We can thus rewrite CF as

$$\min_{F \in \mathbb{R}^{n_\mathcal{X} \times n_\mathcal{Y}}} \sum_{i=1}^{n} \ell(t_i, Tr\left(F^\top(\phi_X(\mathbf{x}_i) \otimes \phi_Y(\mathbf{y}_j))\right)) + \lambda \|F\|_*\,.$$

The idea

$$\min_{\mathbf{w}\in\mathcal{X}\otimes\mathcal{Y}} \sum_{i=1}^{n} \ell(t_i, \mathbf{w}^\top(\mathbf{x}_i \otimes \mathbf{y}_i)) + \lambda\|\mathbf{w}\|^2 \,.$$

$$\min_{F\in\mathbb{R}^{n_\mathcal{X}\times n_\mathcal{Y}}} \sum_{i=1}^{n} \ell(t_i, Tr\left(F^\top(\phi_X(\mathbf{x}_i)\otimes\phi_Y(\mathbf{y}_j))\right)) + \lambda\|F\|_* \,.$$

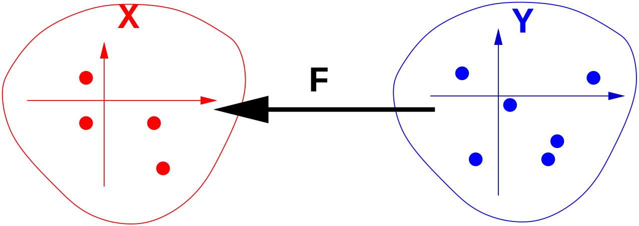
- Put the attribute informations in $\phi_X(\mathbf{x})$ and $\phi_Y(\mathbf{y})$, like in regression
- Investigate penalties beyond the ℓ_2 norm, like in CF
- For this we need to work with "infinite-dimensional matrices", i.e., compact operators

Setting

- Movies: points in a Hilbert space $\mathcal{X}$
- Customers: points in a Hilbert space $\mathcal{Y}$
- We model the preference of customer $\mathbf{y}$ for a movie $\mathbf{x}$ by a bilinear form:

$$f(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, F\mathbf{y} \rangle_{\mathcal{X}} \ ,$$

where $F \in \mathcal{B}_0\left(\mathcal{Y}, \mathcal{X}\right)$ is a compact linear operator (i.e., a "matrix").



Spectra of compact operators

Classical results

- For $(\mathbf{x},\mathbf{y})$ in $\mathcal{X}\times\mathcal{Y}$ the tensor product $\mathbf{x}\otimes\mathbf{y}$ is the operator

$$\forall \mathbf{h}\in\mathcal{Y}\,,\quad (\mathbf{x}\otimes\mathbf{y})\,\mathbf{h} = \langle \mathbf{y},\mathbf{h}\rangle_{\mathcal{Y}}\,\mathbf{x}\,.$$

- Any compact operator $F:\mathcal{Y}\to\mathcal{X}$ admits a spectral decomposition:

$$F = \sum_{i=1}^{\infty}\sigma_i \mathbf{u}_i\otimes\mathbf{v}_i\,.$$

  where the $\sigma_i \geq 0$ are the singular values and $(\mathbf{u}_i)_{i\in\mathbb{N}}$ and $(\mathbf{v}_i)_{i\in\mathbb{N}}$ are orthonormal families in $\mathcal{X}$ and $\mathcal{Y}$.

- The spectrum of F is the set of singular values sorted in decreasing order: $\sigma_1(F)\geq\sigma_2(F)\geq\ldots\geq 0$.
- This is the natural generalization of singular values for matrices.

Spectra of compact operators

Classical results

- For $(\mathbf{x}, \mathbf{y})$ in $\mathcal{X} \times \mathcal{Y}$ the tensor product $\mathbf{x} \otimes \mathbf{y}$ is the operator

$$\forall \mathbf{h} \in \mathcal{Y}, \quad (\mathbf{x} \otimes \mathbf{y})\, \mathbf{h} = \langle \mathbf{y}, \mathbf{h} \rangle_{\mathcal{Y}}\, \mathbf{x} \,.$$

- Any compact operator $F : \mathcal{Y} \to \mathcal{X}$ admits a spectral decomposition:

$$F = \sum_{i=1}^{\infty} \sigma_i \mathbf{u}_i \otimes \mathbf{v}_i \,.$$

  where the $\sigma_i \geq 0$ are the singular values and $(\mathbf{u}_i)_{i \in \mathbb{N}}$ and $(\mathbf{v}_i)_{i \in \mathbb{N}}$ are orthonormal families in $\mathcal{X}$ and $\mathcal{Y}$.

- The spectrum of F is the set of singular values sorted in decreasing order: $\sigma_1(F) \geq \sigma_2(F) \geq \ldots \geq 0$.
- This is the natural generalization of singular values for matrices.

Spectra of compact operators

Classical results

- For $(\mathbf{x}, \mathbf{y})$ in $\mathcal{X} \times \mathcal{Y}$ the tensor product $\mathbf{x} \otimes \mathbf{y}$ is the operator

$$\forall \mathbf{h} \in \mathcal{Y}, \quad (\mathbf{x} \otimes \mathbf{y})\, \mathbf{h} = \langle \mathbf{y}, \mathbf{h} \rangle_{\mathcal{Y}}\, \mathbf{x}\,.$$

- Any compact operator $F : \mathcal{Y} \rightarrow \mathcal{X}$ admits a spectral decomposition:

$$F = \sum_{i=1}^{\infty} \sigma_i \mathbf{u}_i \otimes \mathbf{v}_i\,.$$

  where the $\sigma_i \geq 0$ are the singular values and $(\mathbf{u}_i)_{i \in \mathbb{N}}$ and $(\mathbf{v}_i)_{i \in \mathbb{N}}$ are orthonormal families in $\mathcal{X}$ and $\mathcal{Y}$.

- The spectrum of F is the set of singular values sorted in decreasing order: $\sigma_1(F) \geq \sigma_2(F) \geq \ldots \geq 0$.
- This is the natural generalization of singular values for matrices.

Useful classes for operators

Operators of finite rank

- The rank of an operator is the number of strictly positive singular values.
- Hence operators of rank smaller or equal to k are characterized by:

$$\sigma_{k+1}(F) = 0 \, .$$

Trace-class operators

The trace-class operators are the compact operators F that satisfy:

$$\| F \|_* := \sum_{i=1}^{\infty} \sigma_i(F) < \infty \, .$$

$\| F \|_*$ is a norm over the trace-class operators, called the trace norm.

Useful classes for operators

Operators of finite rank

- The rank of an operator is the number of strictly positive singular values.
- Hence operators of rank smaller or equal to k are characterized by:

$$\sigma_{k+1}(F) = 0 \, .$$

Trace-class operators

The trace-class operators are the compact operators F that satisfy:

$$\| F \|_* := \sum_{i=1}^{\infty} \sigma_i(F) < \infty \, .$$

$\| F \|_*$ is a norm over the trace-class operators, called the trace norm.

Useful classes for operators (cont.)

Hilbert-Schmidt operators

- The Hilbert-Schmidt operators are compact operators F that satisfy:

$$\| F \|_{Fro}^2 := \sum_{i=1}^{\infty} \sigma_i(F)^2 < \infty \, .$$

- They form a Hilbert space with inner product:

$$\left\langle \mathbf{x} \otimes \mathbf{y}, \mathbf{x}' \otimes \mathbf{y}' \right\rangle_{\mathcal{X} \otimes \mathcal{Y}} = \left\langle \mathbf{x}, \mathbf{x}' \right\rangle_{\mathcal{X}} \left\langle \mathbf{y}, \mathbf{y}' \right\rangle_{\mathcal{Y}} \, .$$

- It is isomorphic to the reproducing kernel Hilbert space used in regression over pairs

Spectral penalty function

Definition

A function $\Omega : \mathcal{B}_0(\mathcal{Y}, \mathcal{X}) \mapsto \mathbb{R} \cup \{+\infty\}$ is called a spectral penalty function if it can be written as:

$$\Omega(F) = \sum_{i=1}^{\infty} s_i(\sigma_i(F)) \,,$$

where for any $i \geq 1$, $s_i : \mathbb{R}^+ \mapsto \mathbb{R}^+ \cup \{+\infty\}$ is a non-decreasing penalty function satisfying $s_i(0) = 0$.

Spectral penalty function

Examples

- Rank constraint: take $s_{k+1}(0) = 0$ and $s_{k+1}(u) = +\infty$ for $u > 0$, and $s_i = 0$ for $i \geq k$. Then

$$\Omega(F) = \begin{cases} 0 & \text{if } rank(F) \leq k\,, \\ +\infty & \text{if } rank(F) > k\,. \end{cases}$$

- Trace norm: take $s_i(u) = u$ for all i, then:

$$\Omega(F) = \| F \|_* \,.$$

- Hilbert-Schmidt norm: take $s_i(u) = u^2$ for all i, then

$$\Omega(F) = \| F \|_{Fro}^2 \,.$$

Spectral penalty function

Examples

- Rank constraint: take $s_{k+1}(0) = 0$ and $s_{k+1}(u) = +\infty$ for $u > 0$, and $s_i = 0$ for $i \geq k$. Then

$$\Omega(F) = \begin{cases} 0 & \text{if } rank(F) \leq k\,, \\ +\infty & \text{if } rank(F) > k\,. \end{cases}$$

- Trace norm: take $s_i(u) = u$ for all i, then:

$$\Omega(F) = \| F \|_*\,.$$

- Hilbert-Schmidt norm: take $s_i(u) = u^2$ for all i, then

$$\Omega(F) = \| F \|_{Fro}^2\,.$$

Spectral penalty function

Examples

- Rank constraint: take $s_{k+1}(0) = 0$ and $s_{k+1}(u) = +\infty$ for $u > 0$, and $s_i = 0$ for $i \geq k$. Then

$$\Omega(F) = \begin{cases} 0 & \text{if } rank(F) \leq k\,, \\ +\infty & \text{if } rank(F) > k\,. \end{cases}$$

- Trace norm: take $s_i(u) = u$ for all i, then:

$$\Omega(F) = \parallel F \parallel_* \,.$$

- Hilbert-Schmidt norm: take $s_i(u) = u^2$ for all i, then

$$\Omega(F) = \parallel F \parallel_{Fro}^2 \,.$$

Learning operator with spectral regularization

Setting

- Training set: $(\mathbf{x}_i, \mathbf{y}_i, t_i)_{i=1,\ldots,N}$ a set of (movie,customer,preference).
- Loss function $l(t, t')$: cost of predicting preference t instead of t'.
- Empirical risk of an operator F:

$$R_N(F) = \frac{1}{N}\sum_{i=1}^{N} l\left(\langle \mathbf{x}_i, F\mathbf{y}_i\rangle_{\mathcal{X}}, t_i\right) .$$

Learning an operator

$$\min_{F\in\mathcal{B}_0(\mathcal{Y},\mathcal{X}),\ \Omega(F)<\infty} \left\{R_N(F) + \lambda\Omega(F)\right\} .$$

Learning operator with spectral regularization

Setting

- Training set: $(\mathbf{x}_i, \mathbf{y}_i, t_i)_{i=1,\ldots,N}$ a set of (movie,customer,preference).
- Loss function $l(t, t')$: cost of predicting preference t instead of t'.
- Empirical risk of an operator F:

$$R_N(F) = \frac{1}{N} \sum_{i=1}^{N} l\left(\langle \mathbf{x}_i, F\mathbf{y}_i \rangle_{\mathcal{X}}, t_i\right) .$$

Learning an operator

$$\min_{F \in \mathcal{B}_0(\mathcal{Y},\mathcal{X}),\ \Omega(F) < \infty} \left\{ R_N(F) + \lambda \Omega(F) \right\} .$$

Particular cases

$$\min_{F\in\mathcal{B}_0(\mathcal{Y},\mathcal{X}),\ \Omega(F)<\infty} \left\{R_N(F)+\lambda\Omega(F)\right\} .$$

CF

- $K_X(\mathbf{x},\mathbf{x}') = \delta(\mathbf{x},\mathbf{x}')$, $K_Y(\mathbf{y},\mathbf{y}') = \delta(\mathbf{y},\mathbf{y}')$
- $\Omega(F) = \|F\|_*$ or $rank(F)$

Pairwise regression

- $K_X(\mathbf{x},\mathbf{x}')$ and $K_Y(\mathbf{y},\mathbf{y}')$ defined by attributes
- $\Omega(F) = \|F\|^2_{Fro}$

Many variants, e.g., multitask learning

- $K_X(\mathbf{x},\mathbf{x}') = \delta(\mathbf{x},\mathbf{x}')$ and $K_Y(\mathbf{y},\mathbf{y}')$ defined by attributes
- $\Omega(F) = \|F\|_*$

Questions

Theory

Is it a "good" algorithm in theory?

- To be investigated...
- See Srebro et al. (2004), Bach (2007) for preliminary results with the trace norm

Practice

Can we implement it? Does it work on real data?

- Optimization problem in the space of compact operators... but we show later that it boils down to a finite-dimensional optimization problem
- Promising results on real data

Questions

Theory

Is it a "good" algorithm in theory?

- To be investigated...
- See Srebro et al. (2004), Bach (2007) for preliminary results with the trace norm

Practice

Can we implement it? Does it work on real data?

- Optimization problem in the space of compact operators... but we show later that it boils down to a finite-dimensional optimization problem
- Promising results on real data

Questions

Theory

Is it a "good" algorithm in theory?

- To be investigated...
- See Srebro et al. (2004), Bach (2007) for preliminary results with the trace norm

Practice

Can we implement it? Does it work on real data?

- Optimization problem in the space of compact operators... but we show later that it boils down to a finite-dimensional optimization problem
- Promising results on real data

Questions

Theory

Is it a "good" algorithm in theory?

- To be investigated...
- See Srebro et al. (2004), Bach (2007) for preliminary results with the trace norm

Practice

Can we implement it? Does it work on real data?

- Optimization problem in the space of compact operators... but we show later that it boils down to a finite-dimensional optimization problem
- Promising results on real data

A classical representer theorem

Theorem

If $\hat{F}$ is a solution the problem:

$$\min_{F \in \mathcal{B}_2(\mathcal{Y},\mathcal{X})} \left\{ R_N(F) + \lambda \sum_{i=1}^{\infty} \sigma_i(F)^2 \right\} ,$$

then it is necessarily in the linear span of $\{\mathbf{x}_i \otimes \mathbf{y}_i \, : \, i = 1, \ldots, N\}$, i.e., it can be written as:

$$\hat{F} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i \otimes \mathbf{y}_i \, ,$$

for some $\alpha \in \mathbb{R}^N$.

Proof sketch

- $\mathcal{B}_2(\mathcal{Y},\mathcal{X})$ is isomorphic to the RKHS of the tensor product kernel:

$$k_{\otimes}\left((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}')\right) = \left\langle \mathbf{x},\mathbf{x}' \right\rangle_{\mathcal{X}} \left\langle \mathbf{y},\mathbf{y}' \right\rangle_{\mathcal{Y}} \ ,$$

  by $f(\mathbf{x},\mathbf{y}) = \langle \mathbf{x}, F\mathbf{y}\rangle_{\mathcal{X}}$. In particular,

$$\| f \|^2_{\mathcal{H}_{\otimes}} = \| F \|^2 = \Omega(F) \, .$$

- The problem is therefore a classical kernel method:

$$\min_{f\in\mathcal{H}_{\otimes}} \left\{ R_N(f) + \lambda \| f \|^2_{\otimes} \right\} \ ,$$

  so the classical representer theorem can be used. □

A generalized representer theorem

Theorem

For any spectral penalty function $\Omega : \mathcal{B}_0\left(\mathcal{Y},\mathcal{X}\right) \mapsto \mathbb{R}$, let the optimization problem:

$$\min_{F \in \mathcal{B}_0(\mathcal{Y},\mathcal{X}), \Omega(F) < \infty} \left\{ R_N(F) + \lambda \Omega(F) \right\} .$$

If the set of solutions is not empty, then there is a solution F in $\mathcal{X}_N \otimes \mathcal{Y}_N$, i.e., there exists $\alpha \in \mathbb{R}^{m_\mathcal{X} \times m_\mathcal{Y}}$ such that:

$$F = \sum_{i=1}^{m_\mathcal{X}} \sum_{j=1}^{m_\mathcal{Y}} \alpha_{ij} \mathbf{u}_i \otimes \mathbf{v}_j ,$$

where $(\mathbf{u}_1, \ldots, \mathbf{u}_{m_\mathcal{X}})$ and $\left(\mathbf{v}_1, \ldots, \mathbf{v}_{m_\mathcal{Y}}\right)$ form orthonormal bases of $\mathcal{X}_N$ and $\mathcal{Y}_N$, respectively.

Proof sketch

- For any operator $F \in \mathcal{B}_0(\mathcal{Y}, \mathcal{X})$, let

$$G = \Pi_{\mathcal{X}_N} F \Pi_{\mathcal{Y}_N},$$

  where Π_U is the orthogonal projection onto U.
- Lemma: we can show that for all $i \geq 0$:

$$\sigma_i(G) \leq \sigma_i(F).$$

- Therefore $\Omega(G) \leq \Omega(F)$.
- On the other hand $R_N(G) = R_N(F)$.
- Consequently for any solution F we have another solution $G \in \mathcal{X}_N \otimes \mathcal{Y}_N$. $\square$

Practical consequence

Theorem (cont.)

The coefficients α that define the solution by

$$F = \sum_{i=1}^{m_{\mathcal{X}}} \sum_{j=1}^{m_{\mathcal{Y}}} \alpha_{ij} \mathbf{u}_i \otimes \mathbf{v}_j \,,$$

can be found by solving the following finite-dimensional optimization problem:

$$\min_{\alpha \in \mathbb{R}^{m_{\mathcal{X}} \times m_{\mathcal{Y}}}, \Omega(\alpha) < \infty} R_N \left(diag \left(X \alpha Y^\top \right) \right) + \lambda \Omega(\alpha) \,,$$

where $\Omega(\alpha)$ refers to the spectral penalty function applied to the matrix α seen as an operator from $\mathbb{R}^{m_{\mathcal{Y}}}$ to $\mathbb{R}^{m_{\mathcal{X}}}$, and X and Y denote any matrices that satisfy $K = XX^\top$ and $G = YY^\top$ for the two Gram matrices K and G of $\mathcal{X}_N$ and $\mathcal{Y}_N$.

Summary

We obtain various algorithms by choosing:

1. A loss function (depends on the application)
2. A spectral regularization (that is amenable to optimization)
3. Two Gram matrices (aka kernel matrices)

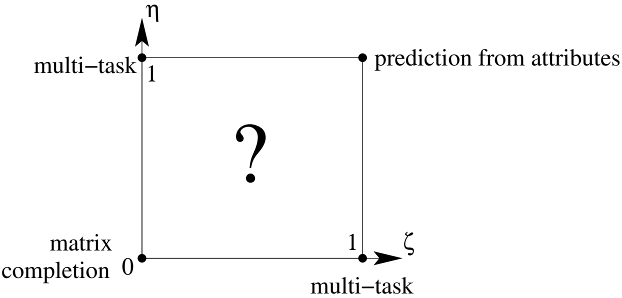
Both kernels and spectral regularization can be used to constrain the solution

A family of kernels

Taken $K_{\otimes} = K \times G$ with

$$\begin{cases} K = \eta K^{x}_{\mathit{Attribute}} + (1-\eta) K^{x}_{\mathit{Dirac}}, \\ G = \zeta K^{y}_{\mathit{Attribute}} + (1-\zeta) K^{y}_{\mathit{Dirac}}, \end{cases}$$

for $0 \leq \eta \leq 1$ and $0 \leq \zeta \leq 1$



Simulated data

Experiment

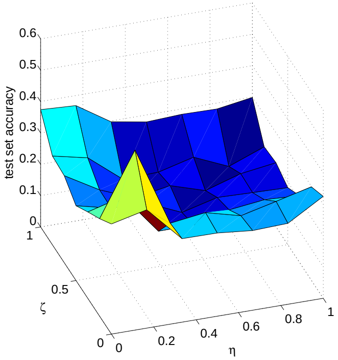
- Generate data $(\mathbf{x}, \mathbf{y}, z) \in \mathbb{R}^{f_X} \times \mathbb{R}^{f_Y} \times \mathbb{R}$ according to

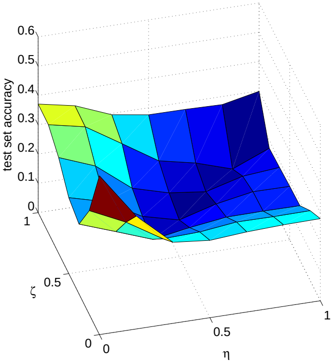
$$z = \mathbf{x}^\top B \mathbf{y} + \varepsilon$$

- Observe only $n_X < f_X$ and $n_Y < f_Y$ features
  - Low-rank assumption will find the missing features
  - Observed attributes will help the low-rank formulation to concentrate mostly on the unknown features
- Comparison of
  - Low-rank constraint without tracenorm (note that it requires regularization)
  - Trace-norm formulation (regularization is implicit)

Simulated data: results

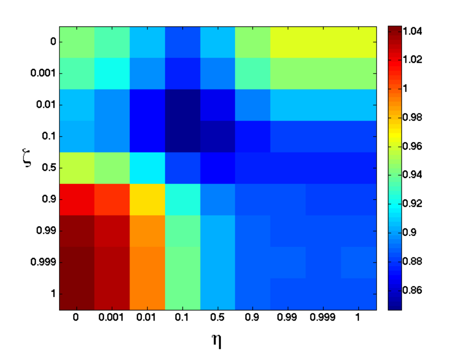
- Compare MSE
- Left: rank constraint (best: 0.1540), right: trace norm (best: 0.1522)





Movies

- MovieLens 100k database, ratings with attributes
- Experiments with 943 movies and 1,642 customers, 100,000 rankings in $\{1, \dots, 5\}$
- Train on a subset of the ratings, test on the rest
- error measured with MSE (best constant prediction: 1.26)



Conclusion

What we saw

- A general framework for CF with or without attributes
- A generalized representation theorem valid for any spectral penalty function
- A family of new methods

Future work

- The bottleneck is often practical optimization. Online version possible.
- Automatic choice of the kernel

Reference

J. Abernethy, F. Bach, T. Evgeniou and J.-P. Vert, “A new approach to collaborative filtering: operator estimation with spectral regularization”, *Journal of Machine Learning Research*, 10:803-826, 2009.