Linear Algorithms for On-Line Multitask/Multiview Learning

Claudio Gentile
DICOM
Universita’ dell’Insubria, Varese, Italy

Workshop on Learning Theory
Foundations of Computational Mathematics
Hong Kong, June 2008

Joint with: G. Cavallanti, N. Cesa-Bianchi
Goal:

- Provide formalization of **task relatedness** suitable to learning **linear** functions in **on-line** fashion
- Design algorithms and provide (worst-case) competitive analysis
- Focus on simple problem: **Binary classification** via **linear-threshold** functions
Learning a single linear-threshold function/1: The (first-order) Perceptron algorithm

Keep weight vector $\mathbf{w}_t \in \mathbb{R}^d$

In trial $t$:

- Get instance $\mathbf{x}_t \in \mathbb{R}^d$
- Predict with $\hat{y}_t = \text{SGN}(\mathbf{w}_t^\top \mathbf{x}_t) \in \{-1, +1\}$
- Get label $y_t \in \{-1, +1\}$
- **If** mistake $(y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0)$ **then** update $\mathbf{w}_{t+1} := \mathbf{w}_t + y_t \mathbf{x}_t$
Learning a single linear-threshold function/2

[Bl,No,...]

Compete against (single) task \( \mathbf{u} \in \mathbb{R}^d \)

Arbitrary sequence \( S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T) \in \mathbb{R}^d \times \{-1, +1\} \)

\[ |\mathcal{M}| \leq \inf_{\mathbf{u}} \left( \frac{D(\mathbf{u}; S)}{||x||_2^2 ||u||_2^2 + ||x||_2 ||u||_2 \sqrt{D(\mathbf{u}; S)}} \right) \]

"loss" of \( \mathbf{u} \)

\( \mathcal{M} \) is set of mistaken trials \( t \),

\[ D(\mathbf{u}; S) = \sum_{t \in \mathcal{M}} \max \{0, 1 - y_t \mathbf{u}^\top \mathbf{x}_t\} \]
Learning a single linear-threshold function/3:
The \( p \)-norm Perceptron algorithm \([\text{GLS97,GL99, ...}]\)

Fix pair of (vector) dual norms \( \| \cdot \| \) and \( \| \cdot \|\ast \)
E.g. \( \| \cdot \|_q \) and \( \| \cdot \|_p \), with \( p \) dual to \( q \), \( q \geq 1 \)
Mappings: \( f = \frac{1}{2} \nabla \| \cdot \|_p^2 \), \( f^{-1} = \frac{1}{2} \nabla \| \cdot \|_q^2 \)

Keep weight vector \( \mathbf{w}_t \in \mathbb{R}^d \)
In trial \( t \):

- Get instance \( \mathbf{x}_t \in \mathbb{R}^d \)
- Predict with \( \hat{y}_t = \text{SGN}(f(\mathbf{w}_t)^\top \mathbf{x}_t) \in \{-1,+1\} \)
- Get label \( y_t \in \{-1,+1\} \)
- If mistake then update \( \mathbf{w}_{t+1} := \mathbf{w}_t + y_t \mathbf{x}_t \)
Learning a single linear-threshold function/4

[GLS97,G99,...]

Compete against (single) sparse task $u \in \mathbb{R}^d$

Arbitrary sequence $S = (x_1, y_1), \ldots, (x_T, y_T) \in \mathbb{R}^d \times \{-1, +1\}$

$$|\mathcal{M}| \leq \inf_{u} \left( \underbrace{D(u; S)}_{\text{"loss" of } u} + (p - 1)\|x\|_p^2 \|u\|_q^2 + \|x\|_p \|u\|_q \sqrt{(p - 1)D(u; S)} \right)$$

$\mathcal{M}$ is set of mistaken trials $t$,

$$D(u; S) = \sum_{t \in \mathcal{M}} \max\{0, 1 - y_t u^\top x_t\}$$
Multitask/Multiview

Want now to compete against a set of $K$ related tasks $u_1, \ldots, u_K$

Issues:

• What’s task relatedness? (what’s the motivating regularization?)

• What’s an appropriate on-line learning protocol?

• What can we hope for in worst-case scenarios?
Task relatedness/1

Vectors $\mathbf{u}_1, \ldots, \mathbf{u}_K \in \mathbb{R}^d$ are related when algorithm can bet on small task variance

$$\frac{1}{K} \sum_{i=1}^{K} \|\mathbf{u}_i - \overline{\mathbf{u}}\|^2$$

Task centroid $\overline{\mathbf{u}} = (\mathbf{u}_1 + \cdots + \mathbf{u}_K)/K$
Given $K$ (adversarially chosen) training sequences

$$S_i = (x_{i,1}, y_1), (x_{i,2}, y_2), \ldots, (x_{i,T}, y_T) \in \mathbb{R}^{d_i} \times \{-1, 1\}$$

for $i = 1, \ldots, K$,
tasks $u_i \in \mathbb{R}^{d_i}$ are related when algorithm can bet on small margin (or view) variance

$$\sum_{t=1}^{T} (u_i^T x_{i,t} - \hat{\gamma}_t)^2$$

for all tasks $i$,

$$\hat{\gamma}_t = \frac{1}{K} \sum_{i=1}^{K} u_i^T x_{i,t}$$

Note: Vectors $u_i$ can live in heterogeneous spaces
Task relatedness/3 [AZ05, TRW05, AMPY07, ...]

Vectors $\mathbf{u}_1, \ldots, \mathbf{u}_K \in \mathbb{R}^d$ are related when algorithm can bet on $d \times K$ task matrix

$$U = [\mathbf{u}_1 | \mathbf{u}_2 | \ldots | \mathbf{u}_K]$$

being low rank (and having small singular values)

Measure $U$ via unitarily invariant norms

E.g. Schatten $q$-norm

$$\|U\|_{S_q} = \|\sigma_U\|_q = (\text{tr}(U^\top U)^q)^{1/q}$$

$\sigma_U$ vector of singular values of $U$

- $\|U\|_{S_2}$ is Frobenius norm
- $\|U\|_{S_1}$ is trace norm $\approx \|U\|_{S_0} = \text{rank of } U$
Task relatedness/4: Regularization

View task vectors $\mathbf{u}_1, \ldots, \mathbf{u}_K$ as:

- long (column) vector $\mathbf{u} = [\mathbf{u}_1^\top \mid \mathbf{u}_2^\top \mid \ldots \mid \mathbf{u}_K^\top]^\top$

measure relatedness by

$$\|A\mathbf{u}\|_q^2$$

- projected onto data $\implies$ proceed in one dimension

- matrix $\mathbf{U} = [\mathbf{u}_1 \mid \mathbf{u}_2 \mid \ldots \mid \mathbf{u}_K]$

measure relatedness by

$$\|\mathbf{U}\|_2^2$$
Multitask/Multiview protocols [ABR07,TRW05]

Multitask:
- adversary chooses task number $i_t$ and instance $x_t$
- required to predict $y_t \in \{-1, +1\}$ via $\hat{y}_t \in \{-1, +1\}$
- compete against $u_{i_t}$:
  measure ”regret” by $\{\hat{y}_t \neq y_t\} - \max \{0, 1 - y_t u_{i_t}^\top x_t\}$

Multiview:
- adversary chooses task instances $x_{1,t}, x_{2,t}, \ldots, x_{K,t}$
- required to predict $y_t \in \{-1, +1\}$ via $\hat{y}_t \in \{-1, +1\}$
- compete against $u_1, \ldots, u_K$ at once:
  measure ”regret” by $\{\hat{y}_t \neq y_t\} - \frac{1}{K} \sum_{i=1}^{K} \max \{0, 1 - y_t u_i^\top x_{i,t}\}$
What we are after in worst-case scenarios

**Multitask:**
Receive one example \((x_t, y_t)\) per trial
If tasks very related, as measured by \(\|Au\|_q^2\)

\[
\text{[Cumulative "regret"]} \leq \text{[single task bound]} + \text{[smaller terms]}
\]

**Multiview:**
Like receiving \(K\) examples \((x_{1,t}, y_t), \ldots, (x_{K,t}, y_t)\) per trial
If tasks very related, as measured by \(\|U\|_{S_q}^2\)

\[
\text{[Cumulative "regret"]} \leq \frac{\text{[single task bound]}}{K} + \text{[smaller terms]}
\]
Some algorithms/1: Multitask $p$-norm Perceptron

Keep $K$ weight vectors $w_{i,t}$, one per task

In trial $t$:

- Get task $i_t$ and instance $x_t$
- Predict with $\hat{y}_t = \text{sgn}(f(w_{i_t,t})^\top x_t)$
- Get label $y_t \in \{-1,+1\}$
- **If** mistake **then** update
  - $w_{i_t,t+1} := w_{i_t,t} + y_t x_t$
  - $w_{j,t+1} := w_{j,t} + \frac{1}{2}y_t x_t \quad \forall j \neq i_t$

**Note:** When $p = q = 2$ ($f$ = ident.) equivalent to single task on ”long” vector RKHS with product $\langle u, v \rangle = u^\top A v$ [EMP05]
Some bounds/1: Multitask $p = 2$-norm Perceptron

Arbitrary sequence

$S = (x_1, y_1, i_1), \ldots, (x_T, y_T, i_1) \in \mathbb{R}^d \times \{-1, +1\} \times \{1, \ldots, K\}$

\[
|\mathcal{M}| \leq \inf_{u=[u_1|\ldots|u_K]} \left( D(u; S) + \|x\|_2^2 \|Au\|_2^2 \\
+ \|x\|_2 \|Au\|_2 \sqrt{D(u; S)} \right)
\]

\[
D(u; S) = \sum_{t \in \mathcal{M}} \max\{0, 1 - y_t u_{i_t}^\top x_t\}
\]

\[
\|Au\|_2^2 \simeq \underbrace{\|u\|_2^2}_{\text{cost for single task } \overline{u}} + \underbrace{\sum_{i=1}^{K} \|u_i - \overline{u}\|_2^2}_{\text{cost for task variance}}
\]
Some algorithms/2: Multiview $p$-norm Perceptron

Fix pair of unitarily invariant dual norms $\| \cdot \|$ and $\| \cdot \|_*$

E.g. $\| \cdot \|_S^q$ and $\| \cdot \|_S^p$ with $p$ dual to $q$, $q \geq 1$

Mappings: $F = \frac{1}{2} \nabla \| \cdot \|_S^2$, $F^{-1} = \frac{1}{2} \nabla \| \cdot \|_S^2$ \cite{Le95}

Keep $d \times K$ matrix $W_t$

In trial $t$:

- Get $d \times K$ instance matrix $X_t = [\mathbf{x}_{1,t} | \mathbf{x}_{2,t} | \ldots | \mathbf{x}_{K,t}]$
- Predict with $\hat{y}_t = \text{sgn}(\text{tr}(F(W_t)^\top X_t)) \in \{-1, +1\}$
- Get label $y_t \in \{-1, +1\}$
- If mistake then update $W_{t+1} := W_t + y_t X_t$
Some algorithms/3: Remarks on multiview $p$-norm

How does mapping $F$ look like?

$$F(W) \approx W (W^\top W)^{p/2-1}$$

- $p = 2$: $F = \text{identity}$

  $K$ perceptrons in parallel (no column $x_{i,t}$ sharing)
  "Cooperation" only via prediction

  $$\hat{y}_t = \text{SGN}(\text{TR}(W_t^\top X_t)) = \text{SGN}\left(\sum_{i=1}^{K} w_{i,t}^\top x_{i,t}\right)$$

- $p > 2$: column sharing

- $p \approx \log K$: similar to Matrix EG
  (but with neither learning rates nor matrix exponentials)
Some bounds/2: Multiview $p$-norm Perceptron

Arbitrary sequence

$$S = (X_1, y_1), ..., (X_T, y_T) \in \mathbb{R}^{d \times K} \times \{-1, +1\}$$

$$|\mathcal{M}| \leq \inf_{U} \left( D(U; S) + \frac{(p - 1)\|X\|_{S_p}^2 \|U\|_{S_q}^2}{K^2} ight.$$  

$$\left. + \frac{\|X\|_{S_p} \|U\|_{S_q}}{K} \sqrt{(p - 1)D(U; S)} \right)$$

$$D(U; S) = \sum_{t \in \mathcal{M}} \frac{1}{K} \sum_{i=1}^{K} \max\{0, 1 - y_t u_i^\top x_{i,t}\}$$
Some bounds/3: Remarks on multiview $p$-norm bound

Meaningful when competitor $U$ is low rank

Set $p \approx \log r$, $r = \min\{d, K\}$:

\[
\|X\|_{S_p}^2 \approx \|X\|_{S_\infty}^2 \approx \|X\|_{S_2}^2 / K \approx \|x\|_2^2
\]

(singletask complexity)

\[
\|U\|_{S_q}^2 \leq \|U\|_{S_1}^2 \approx \|U\|_{S_2}^2 \approx K\|u\|_2^2
\]

($K$ times singletask complexity)

\[
\frac{(p-1)\|X\|_{S_p}^2 \|U\|_{S_q}^2}{K^2} \approx \frac{(\log r)\|x\|_2^2 \|u\|_2^2}{K}
\]

(singletask 2-norm bound over $K$)
Conclusions and open questions

- Formalization of on-line multitask/multiview learning
- Ideas taken from many papers
  \[\text{[ABR07,EMP05,AMPY08,SNB05,TRW05, ...]}\]
- Mistake bound (”convergence”) analysis
- Further (non presented) results
- Several open questions:
  - Multitask with $K$ labels per trial
  - Computational issues on $p$-norm multiview
    (would like to avoid struggling with SVD ...)
  - Etc.