Modern Machine Learning: New Challenges and Connections

Maria-Florina Balcan
Machine Learning is Changing the World

“Machine learning is the hot new thing”
(John Hennessy, President, Stanford)

“A breakthrough in machine learning would be worth ten Microsofts”
(Bill Gates, Microsoft)

“Web rankings today are mostly a matter of machine learning”
(Prabhakar Raghavan, VP Engineering at Google)
The World is Changing Machine Learning

New applications

Explosion of data
Modern ML: New Learning Approaches

Modern applications: **massive amounts** of raw data.

Only a tiny fraction can be annotated by human experts.

Protein sequences  Billions of webpages  Images
Modern applications: massive amounts of raw data.

Techniques that best utilize data, minimizing need for expert/human intervention.

- Semi-supervised Learning, (Inter)active Learning.
Modern ML: New Learning Approaches

Modern applications: \textit{massive amounts} of data \textit{distributed} across multiple locations.
Modern ML: New Learning Approaches

Modern applications: massive amounts of data distributed across multiple locations.

E.g.,
- video data
- scientific data

Key new resource communication.
The World is Changing Machine Learning

New approaches. E.g.,

- Semi-supervised learning
- Interactive learning
- Distributed learning
- Multi-task/transfer learning
- Community detection
- Never ending learning

Resource constraints. E.g.,

- Computational efficiency (noise tolerant algos)
- Human labeling effort
- Statistical efficiency
- Communication
- Privacy/Incentives
The World is Changing Machine Learning

New approaches. E.g.,

- Semi-supervised learning
- Interactive learning
- Distributed learning
- Multi-task/transfer learning
- Community detection
- Never ending learning

My research: foundations, robust algorithms.

Applications: vision, comp. biology, social networks.
Machine Learning is Changing the World

My research: MLT way to model and tackle challenges other fields. E.g.,

- Game Theory
- Discrete Optimization
- Machine Learning Theory
- Mechanism Design
- Matroid Theory
- Control Theory
- Approximation Algorithms
Outline of the talk

- Modern Learning Paradigms
  - Interactive Learning

- Machine Learning Lenses in Other Areas
  - Submodularity, implications to Matroid Theory, Algorithmic Game Theory, Optimization

- Discussion, Other Exciting Directions
Supervised Learning

• E.g., which emails are spam and which are important.

Not spam

spam

• E.g., classify objects as chairs vs non chairs.
Statistical / PAC learning model

- Algo sees $(x_1,c^*(x_1)),..., (x_m,c^*(x_m))$, $x_i$ i.i.d. from $D$
  - Does optimization over $S$, finds hypothesis $h \in C$.
  - Goal: $h$ has small error, $\text{err}(h)=\Pr_{x \in D}(h(x) \neq c^*(x))$
- $c^*$ in $C$, realizable case; else agnostic
Two Main Aspects in Classic Machine Learning

Algorithm Design. How to optimize?
Automatically generate rules that do well on observed data.

E.g., Boosting, SVM, etc.

Generalization Guarantees, Sample Complexity
Confidence for rule effectiveness on future data.

\[ O\left( \frac{1}{\epsilon} \left( \text{VCdim}(C) \log \left( \frac{1}{\epsilon} \right) + \log \left( \frac{1}{\delta} \right) \right) \right) \]
Interactive Machine Learning

- Active Learning
- Learning with more general queries; connections
Active Learning

Classifier

Learning Algorithm

Unlabeled data

Expert Labeler

raw data

face

not face
Active Learning in Practice

- **Text classification: active SVM** (Tong & Koller, ICML2000).
  - e.g., request label of the example closest to current separator.

- **Video Segmentation** (Fathi-Balcan-Ren-Regh, BMVC 11).
Provable Guarantees, Active Learning

- Canonical theoretical example [CAL92, Dasgupta04]

\[
\begin{bmatrix}
+ & - & + \\
\end{bmatrix}
\]

Active Algorithm

- Sample with \(1/\varepsilon\) unlabeled examples; do binary search.

\[
\begin{bmatrix}
- & - & + \\
\end{bmatrix}
\]

Passive supervised: \(\Omega(1/\varepsilon)\) labels to find an \(\varepsilon\)-accurate threshold.

Active: only \(O(\log 1/\varepsilon)\) labels. Exponential improvement.
Provable Guarantees, Active Learning

Lots of exciting activity in recent years.

My work:

• **First noise tolerant algo** [Balcan, Beygelzimer, Langford, ICML’06]
  
  Generic (any class), adversarial label noise.
  
  Substantial follow-up on work.

  [Hanneke07, DasguptaHsuMontleoni’07, Wang’09, Fridman’09, Koltchinskii10, BHW’08, BeygelzimerHsuLangfordZhang’10, Hsu’10, Minsker’12, Ailon’12, ...]

• **First computationally efficient algos**

  [Balcan-Long COLT’13] [Awasthi-Balcan-Long STOC’14]
Margin Based Active Learning

Learning linear separators, when $D$ logconcave in $\mathbb{R}^d$.

[Balcan-Long COLT'13] [Awasthi-Balcan-Long STOC'14]

• Realizable: **exponential improvement** in label complexity over passive [only $O(d \log 1/\epsilon)$ labels to find $w$ error $\epsilon$].

• Agnostic: **poly-time AL algo** outputs $w$ with $\text{err}(w) = O(\eta)$, $\eta = \text{err}(\text{best lin. sep})$.

• **Unexpected Implications for Passive Learning!** 😊

  • **Computational complexity:** improves significantly on the best guarantees known for passive [KKMS’05], [KLS’09].

  • **Sample complexity:** resolves open question about ERM.
Margin Based Active-Learning, Realizable Case

Draw $m_1$ unlabeled examples, label them, add them to $W(1)$.

Iterate $k = 2, \ldots, s$

- find a hypothesis $w_{k-1}$ consistent with $W(k-1)$.
- $W(k) = W(k-1)$.
- sample $m_k$ unlabeled samples $x$ satisfying $|w_{k-1}^T \cdot x| \leq \gamma_{k-1}$
- label them and add them to $W(k)$.
Margin Based Active-Learning, Realizable Case

Log-concave distributions: log of density fnc concave.
- wide class: uniform distr. over any convex set, Gaussian, etc.

**Theorem**

If $\log\text{-concave in } \mathbb{R}^d$. If $\gamma_k = O\left(\frac{1}{2^k}\right)$ then $\text{err}(w_s) \leq \varepsilon$ after $s = \log\left(\frac{1}{\varepsilon}\right)$ rounds using $\tilde{O}(d)$ labels per round.

**Active learning**
- $O\left(d \log\left(\frac{1}{\varepsilon}\right)\right)$ label requests
- $\Omega\left(\frac{d}{\varepsilon}\right)$ unlabeled examples

**Passive learning**
- $\Theta\left(\frac{d}{\varepsilon}\right)$ label requests
Induction: all $w$ consistent with $W(k)$, $\text{err}(w) \leq 1/2^k$
Analysis: Aggressive Localization

Induction: all $w$ consistent with $W(k)$, $\text{err}(w) \leq 1/2^k$. 

Suboptimal
Analysis: Aggressive Localization

Induction: all $w$ consistent with $W(k)$, $\text{err}(w) \leq 1/2^k$

\[
\text{err}(w) = \Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \geq \gamma_{k-1}) + \Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \leq \gamma_{k-1}) \leq 1/2^{k+1}
\]
Analysis: Aggressive Localization

Induction: \textit{all w consistent with $W(k)$, $err(w) \leq 1/2^k$}

\[ err(w) = \Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \geq \gamma_{k-1}) + \]
\[ \Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \leq \gamma_{k-1}) \Pr(|w_{k-1} \cdot x| \leq \gamma_{k-1}) \]

Enough to ensure $\Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \leq \gamma_{k-1}) \leq C$

Need only $m_k = \tilde{O}(d)$ labels in round $k$. 
Margin Based Active-Learning, Agnostic Case

Draw $m_1$ unlabeled examples, label them, add them to $W$.

**iterate** $k=2, \ldots, s$

• find $w_{k-1}$ in $B(w_{k-1}, r_{k-1})$ of small
  $\tau_{k-1}$ hinge loss wrt $W$.
  • Clear working set.

• sample $m_k$ unlabeled samples $x$
  satisfying $|w_{k-1} \cdot x| \leq \gamma_{k-1}$;
  • label them and add them to $W$.

end iterate

• Analysis: exploit localization & variance analysis control the gap
  between hinge loss and 0/1 loss in each round.
# Improves over Passive Learning too!

<table>
<thead>
<tr>
<th>Passive Learning</th>
<th>Prior Work</th>
<th>Our Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malicious</td>
<td>$\text{err}(w) = O \left( \eta^{1/3} \log^{2/3} \left( \frac{d}{\eta} \right) \right)$</td>
<td>$\text{err}(w) = O \left( \eta \log^2 \left( \frac{1}{\eta} \right) \right)$</td>
</tr>
<tr>
<td>Agnostic</td>
<td>$\text{err}(w) = O \left( \eta^{1/3} \log^{1/3} \left( \frac{1}{\eta} \right) \right)$</td>
<td>$\text{err}(w) = O \left( \eta \log^2 \left( \frac{1}{\eta} \right) \right)$</td>
</tr>
<tr>
<td>Active Learning</td>
<td>NA</td>
<td>$\text{err}(w) = O \left( \eta \log^2 \left( \frac{1}{\eta} \right) \right)$</td>
</tr>
<tr>
<td>[agnostic/malicious]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Improves over Passive Learning too!

<table>
<thead>
<tr>
<th>Passive Learning</th>
<th>Prior Work</th>
<th>Our Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malicious</td>
<td>$\text{err}(w) = O(\eta d^{1/4})$&lt;br&gt;$\text{err}(w) = O(\sqrt{\eta \log(d/\eta)})$&lt;br&gt;[$\text{KKMS'05}$]&lt;br&gt;[$\text{KLS'09}$]</td>
<td>$\text{err}(w) = O(\eta)$&lt;br&gt;Info theoretic optimal</td>
</tr>
<tr>
<td>Agnostic</td>
<td>$\text{err}(w) = O(\eta \sqrt{\log(1/\eta)})$&lt;br&gt;[$\text{KKMS'05}$]</td>
<td>$\text{err}(w) = O(\eta)$</td>
</tr>
<tr>
<td>Active Learning</td>
<td>NA</td>
<td>$\text{err}(w) = O(\eta)$&lt;br&gt;Info theoretic optimal</td>
</tr>
<tr>
<td>[$\text{agnostic/malicious}$]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Slightly better results for the uniform distribution case.
Localization both algorithmic and analysis tool!

Useful for active and passive learning!
Important direction: richer interactions with the expert.

Better Accuracy

Fewer queries

Natural interaction
**New Types of Interaction**

**Class Conditional Query**

Classifier ➔ Learning Algorithm ➔ raw data ➔ Expert Labeler

**Mistake Query**

Classifier ➔ Learning Algorithm ➔ raw data ➔ Expert Labeler

dog, cat, penguin, wolf
Class Conditional & Mistake Queries

- Used in practice, e.g. Faces in IPhoto.
- Lack of theoretical understanding.
- Realizable (Folklore): much fewer queries than label requests.

Our Work [Balcan-Hanneke COLT’12]

Tight bounds on the number of CCQs to learn in the presence of noise (agnostic and bounded noise)
Important direction: richer interactions with the expert.

Better Accuracy  Fewer queries

Natural interaction
Interaction for Faster Algorithms

Expert an expensive computer program.

E.g., clustering protein sequences by function

Voevodski-Balcan-Roglin-Teng-Xia, [UAI 10, SIMBAD 11, JMLR’12]

• BLAST to computes similarity to all other objects in dataset (one-vs-all queries).

• Strong guarantees under natural **stability** condition: produce accurate clusterings with only $k$ one-vs-all queries.

• Better accuracy than state of the art techniques using limited info on standard datasets (pFam and SCOP datasets).
**Interaction for Faster Algorithms**

Voevodski-Balcan-Roglin-Teng-Xia [UAI 10, SIMBAD 11, JMLR'12]

Better accuracy than state of the art techniques using limited info on standard datasets (pFam and SCOP datasets).

---

**Figure 1:** Comparing the performance of $k$-means in the embedded space (blue) and *Landmark-Clustering* (red) on 10 datasets from Pfam. Datasets 1-10 are created by randomly choosing 8 families from Pfam of size $s$, $1000 \leq s \leq 10000$.

**Figure 2:** Comparing the performance of spectral clustering (blue) and *Landmark-Clustering* (red) on 10 datasets from SCOP. Datasets A and B are the two main examples from [10], the other datasets (1-8) are created by randomly choosing 8 superfamilies from SCOP of size $s$, $20 \leq s \leq 200$. 
Interactive Learning

Summary:
• First noise tolerant poly time, label efficient algos for high dim. cases. [BL’13] [ABL’14]
• Learning with more general queries. [BH’12]

Related Work:
• Active & Differentially Private [Balcan-Feldman, NIPS’13]

Cool Implications:
• Sample & computational complexity of passive learning
• Communication complexity, distributed learning.
  [Balcan-Blum-Fine-Mansour, COLT’12] Runner UP Best Paper
Learning Theory Lenses on Submodularity
Submodular functions

- Discrete fns that model laws of diminishing returns.
  - Optimization, operations research
  - **Algorithmic Game Theory** [Lehman-Lehman-Nisan’01], ....
  - **Social Networks** [Kleinberg-Kempe-Tardos’03]
  - **Machine Learning** [Bilmes’03] [Guestrin-Krause’07], ...

---

[Image 181x239 to 388x364]

[Image 404x238 to 490x364]
New Lenses on Submodularity
Learning Submodular Functions

[Balcan&Harvey, STOC 2011 & Necktar Track, ECML-PKDD 2012]

Novel structural results

Alg. Game Theory
Economics

Matroid Theory

Discrete Optimization

Applications: Economics, Social Networks, …
Submodular functions

- \( V = \{1, 2, \ldots, n\} \); set-function \( f: 2^V \rightarrow \mathbb{R} \) submodular if

For \( T \subseteq S, x \notin S \),

\[
f(T \cup \{x\}) - f(T) \geq f(S \cup \{x\}) - f(S)
\]

E.g.,

\[ + \cdot x \] Large improvement

\[ + \cdot x \] Small improvement
Submodular functions

More examples:

• **Concave Functions**  Let \( h : \mathbb{R} \rightarrow \mathbb{R} \) be concave. For each \( S \subseteq V \), let \( f(S) = h(|S|) \)

• **Vector Spaces**  Let \( V = \{v_1,\ldots,v_n\} \), each \( v_i \in \mathbb{R}^n \). For each \( S \subseteq V \), let \( f(S) = \text{rank}(V[S]) \)

• **Coverage Fns**  Let \( A_1,\ldots,A_n \) be sets. For each \( S \subseteq V \), let \( f(S) = |\bigcup_{j \in S} A_j| \)
Learning submodular functions

• Valuation Functions in Economics
  • $V$ = all the items you sell.
  • $f(S)$ = valuation on set of items $S$.

• Influence Function in Social Networks
  • $V$ = set of nodes.
  • $f(S)$ = expected influence when $S$ is the originating set of nodes in the diffusion.
PMAC model for learning real valued functions

- Algo sees \((S_1, f(S_1)), \ldots, (S_m, f(S_m))\), \(S_i\) i.i.d. from \(D\), produces \(g\).
- With probability \(\geq 1-\delta\) we have \(\text{Pr}_S[g(S) \leq f(S) \leq \alpha g(S)] \geq 1-\epsilon\)

**Probably Mostly Approximately Correct**
Learning submodular functions \cite{BH'11}

**Theorem:** (General upper bound)

Poly time alg. for PMAC-learning (w.r.t. an arbitrary distribution) with an approx. factor $\alpha=O(n^{1/2})$.

**Theorem:** (General lower bound)

No algo can PMAC learn the class of submodular fns with approx. factor $\tilde{O}(n^{1/3})$.

- Even if value queries allowed; even for rank fns of matroids.

**Corollary:** Matroid rank fns do not have a concise, approximate representation.

Surprising answer to open question in Economics of

Paul Milgrom  Noam Nisan
A General Lower Bound

Theorem

No algo can PMAC learn the class of non-neg., monotone, submodular fns with an approx. factor $\tilde{o}(n^{1/3})$.

Construct a hard family of matroid rank functions.

High=$n^{1/3}$

Low=$\log^2 n$

Vast generalization of partition matroids.
Exploit Additional Properties

• E.g., product distribution
  • BH’11, Lipschitz submodular
  • FV’13 general submodular

• Learning valuation fns from AGT and Economics.

• Learning influence function in social networks.
Learning Valuation Functions

Target dependent learnability for classes from Algorithmic Game Theory and Economics

[Balcan-Constantin-Iwata-Wang, COLT 2012]

Additive $\subseteq$ OXS $\subseteq$ GS $\subseteq$ Submodular $\subseteq$ XOS $\subseteq$ Subadditive

Theorem: (Poly number of XOR trees) $O(n^\epsilon)$ approximation in time $O(n^{1/\epsilon})$. 
Learning the Influence Function

Influence function in social networks: coverage function.

- **Maximum likelihood approach; promising performance**
  [Du, Liang, Balcan, Song 2014]

- **Memetracker Dataset, blog data cascades**: “apple and jobs”, “tsunami earthquake”, “william kate marriage”
New Lenses on Submodularity
Learning Submodular Functions

[Balcan&Harvey, STOC 2011 & Necktar Track, ECML-PKDD 2012]

Novel structural results

Algo Game Theory Economics
Matroid Theory
Discrete Optimization

Follow-up work: Algo Game Theory, Machine Learning, Discrete Optimization, Privacy.
Discussion. Other Exciting Directions
Themes of My Research

- Foundations for Modern ML and Applications

- Learning Lens on Other Fields, Connections
Foundations for Modern Paradigms

New approaches. E.g.,

- Semi-supervised learning
- Interactive learning
- Distributed learning
- Multi-task/transfer learning
- Community detection
- Never ending learning

Resource constraints. E.g.,

- Computational efficiency (noise tolerant algos)
- Human labeling effort
- Statistical efficiency
- Communication
- Privacy/Incentives
Analysis Beyond the Worst Case

• Identify ways to capture structure of real world instances, analysis beyond the worst case.

• E.g., circumvent NP-hardness barriers for objective based clustering (e.g., k-means, k-median) [BBG’09] [BB’09] [BRT’10] [BL’12]

• Perturbation resilience: small changes to input distances shouldn’t move optimal solution by much.

• Approximation stability: near-optimal solutions shouldn’t look too different from optimal solution.
Analysis Beyond the Worst Case

• Identify ways to capture structure of real world instances, analysis beyond the worst case.

Major Direction:

• Understand characteristics of practical tasks that makes them easier to learn.

• Develop learning tools that take advantage of such characteristics.
Understanding & Influencing Multi-agent Systems

What do we expect to happen?

**Traditional:** analyze *equilibria* *(static concept)*

**More realistic:** view agents as adapting, learning entities.

**Major Question:** Can we “guide” or “nudge” dynamics to *high-quality states quickly* [using minimal intervention]?
Themes of My Research

- Foundations for Modern ML and Applications

- Learning Lens on Other Fields, Connections

Approx. Algorithms  Game Theory  Discrete Optimization
Control Theory  Machine Learning Theory  Matroid Theory
Mechanism Design