Machine Learning and Artificial Intelligence

John Shawe-Taylor

Department of Computer Science
University College London

ECSS, Gothenburg, 2018

Aims

- Give a personal perspective on the recent progress and resultant buzz around artificial intelligence

Aims

- Give a personal perspective on the recent progress and resultant buzz around artificial intelligence
- Throw light on the place of machine learning in these developments

Aims

- Give a personal perspective on the recent progress and resultant buzz around artificial intelligence
- Throw light on the place of machine learning in these developments
- Highlight the leading role that Europe has played

Aims

- Give a personal perspective on the recent progress and resultant buzz around artificial intelligence
- Throw light on the place of machine learning in these developments
- Highlight the leading role that Europe has played
- Suggest promising directions for further attention

Machine Learning (ML)

- Machine Learning seeks patterns in data: based on probabilistic analysis rather than logical inference

Machine Learning (ML)

- Machine Learning seeks patterns in data: based on probabilistic analysis rather than logical inference
- Simplest problems are supervised learning: data such as images labelled with content (eg contains bicycle)

Machine Learning (ML)

- Machine Learning seeks patterns in data: based on probabilistic analysis rather than logical inference
- Simplest problems are supervised learning: data such as images labelled with content (eg contains bicycle)
- Task is to use this data to identify a function that classifies new images (ie image contains bicycle)

Machine Learning (ML)

- Machine Learning seeks patterns in data: based on probabilistic analysis rather than logical inference
- Simplest problems are supervised learning: data such as images labelled with content (eg contains bicycle)
- Task is to use this data to identify a function that classifies new images (ie image contains bicycle)
- Initial enthusiasm in 1980's was followed by disillusionment over unreliable and frequently poor results

- One criticism of this early work was that it was heuristic and ad-hoc.

Principled Machine Learning

- One criticism of this early work was that it was heuristic and ad-hoc.
- A series of EU Networks have promoted principled machine learning over a 20 year period (1993-2013):
  - NeuroCOLT, NeuroCOLT2, PASCAL, PASCAL2

Principled Machine Learning

- One criticism of this early work was that it was heuristic and ad-hoc.
- A series of EU Networks have promoted principled machine learning over a 20 year period (1993-2013):
  - NeuroCOLT, NeuroCOLT2, PASCAL, PASCAL2
  - Influential in promoting a paradigm shift both in ML, as well as uptake of ML in Computer Vision and Natural Language Processing

Principled Machine Learning

- One criticism of this early work was that it was heuristic and ad-hoc.
- A series of EU Networks have promoted principled machine learning over a 20 year period (1993-2013):
  - NeuroCOLT, NeuroCOLT2, PASCAL, PASCAL2
  - Influential in promoting a paradigm shift both in ML, as well as uptake of ML in Computer Vision and Natural Language Processing
  - In its final two years 30% of papers at top two ML conferences included an author from the PASCAL2 network.

Principled Machine Learning

- One criticism of this early work was that it was heuristic and ad-hoc.
- A series of EU Networks have promoted principled machine learning over a 20 year period (1993-2013):
  - NeuroCOLT, NeuroCOLT2, PASCAL, PASCAL2
  - Influential in promoting a paradigm shift both in ML, as well as uptake of ML in Computer Vision and Natural Language Processing
  - In its final two years 30% of papers at top two ML conferences included an author from the PASCAL2 network.
- Will give an example of results on generalisation of learning systems

Generalisation of a learner

- A learning system uses a sample of data to try to identify the pattern

Generalisation of a learner

- A learning system uses a sample of data to try to identify the pattern
- But we need the pattern to perform well on new (previously unseen) data: this is generalisation

Generalisation of a learner

- A learning system uses a sample of data to try to identify the pattern
- But we need the pattern to perform well on new (previously unseen) data: this is generalisation
- Can be analysed in a statistical framework assuming training and test data are drawn from the same distribution (PAC or probably approximately correct framework)

Generalisation of a learner

- A learning system uses a sample of data to try to identify the pattern
- But we need the pattern to perform well on new (previously unseen) data: this is generalisation
- Can be analysed in a statistical framework assuming training and test data are drawn from the same distribution (PAC or probably approximately correct framework)
- PAC-Bayes bounds are based on defining a prior distribution over the functions and then choosing a posterior distribution that compromises between loss on the training data and shift from the prior

Generalisation of a learner

- A learning system uses a sample of data to try to identify the pattern
- But we need the pattern to perform well on new (previously unseen) data: this is generalisation
- Can be analysed in a statistical framework assuming training and test data are drawn from the same distribution (PAC or probably approximately correct framework)
- PAC-Bayes bounds are based on defining a prior distribution over the functions and then choosing a posterior distribution that compromises between loss on the training data and shift from the prior
- Note that bound holds for all posterior distributions

Form of the PAC-Bayes SVM bound

- Bound involves KL divergence between prior and posterior and between empirical and true loss (as distributions over discrete set $\{0, 1\}$)

Form of the PAC-Bayes SVM bound

- Bound involves KL divergence between prior and posterior and between empirical and true loss (as distributions over discrete set $\{0,1\}$)
- If we define the inverse of the KL by

$$\mathrm{KL}^{-1}(q, A) = \max\{p : \mathrm{KL}(q\|p) \leq A\}$$

then with probability $1-\delta$ over the choice of the m sample

$$Pr\left(\langle \mathbf{w}, \phi(\mathbf{x})\rangle \neq y\right) \leq 2 \min_{\mu} \mathrm{KL}^{-1}\left(\mathbb{E}_m[\tilde{F}(\mu\gamma(\mathbf{x}, y))], \frac{\mu^2/2 + \ln\frac{m+1}{\delta}}{m}\right)$$

where $\mathbb{E}_m$ is the empirical average of the cumulative normal distribution ($\tilde{F}$) of a scaling μ of the margin $\gamma(\mathbf{x}, y)$ of example $(\mathbf{x}, y)$

Definition of the Prior

- In PAC-Bayes we are free to choose the prior as long as it doesn't depend on the training data

Definition of the Prior

- In PAC-Bayes we are free to choose the prior as long as it doesn't depend on the training data
- Bound corresponds to prior at the origin

Definition of the Prior

- In PAC-Bayes we are free to choose the prior as long as it doesn't depend on the training data
- Bound corresponds to prior at the origin
- Can use part of the data to estimate a better prior and then evaluate the bound on the remaining data

Model Selection with the new bound: setup

- Comparison with X-fold Xvalidation, PAC-Bayes Bound and the Prior PAC-Bayes Bound

Model Selection with the new bound: setup

- Comparison with X-fold Xvalidation, PAC-Bayes Bound and the Prior PAC-Bayes Bound
- UCI datasets

Model Selection with the new bound: setup

- Comparison with X-fold Xvalidation, PAC-Bayes Bound and the Prior PAC-Bayes Bound
- UCI datasets
- Select C and σ that lead to minimum Classification Error (CE)

Model Selection with the new bound: setup

- Comparison with X-fold Xvalidation, PAC-Bayes Bound and the Prior PAC-Bayes Bound
- UCI datasets
- Select C and σ that lead to minimum Classification Error (CE)
    - For X-F XV select the pair that minimize the validation error

Model Selection with the new bound: setup

- Comparison with X-fold Xvalidation, PAC-Bayes Bound and the Prior PAC-Bayes Bound
- UCI datasets
- Select C and σ that lead to minimum Classification Error (CE)
  - For X-F XV select the pair that minimize the validation error
  - For PAC-Bayes Bound and Prior PAC-Bayes Bound select the pair that minimize the bound

Results

		Classifier					
		SVM				ηPrior SVM	
Problem		2FCV	10FCV	PAC	PrPAC	PrPAC	τ-PrPAC
digits	Bound	–	–	0.175	0.107	0.050	0.047
	CE	0.007	0.007	0.007	0.014	0.010	0.009
waveform	Bound	–	–	0.203	0.185	0.178	0.176
	CE	0.090	0.086	0.084	0.088	0.087	0.086
pima	Bound	–	–	0.424	0.420	0.428	0.416
	CE	0.244	0.245	0.229	0.229	0.233	0.233
ringnorm	Bound	–	–	0.203	0.110	0.053	0.050
	CE	0.016	0.016	0.018	0.018	0.016	0.016
spam	Bound	–	–	0.254	0.198	0.186	0.178
	CE	0.066	0.063	0.067	0.077	0.070	0.072

Good Old-fashioned Artificial Intelligence (GOFAI)

- First attempts at creating Artificial Intelligence were focussed on reproducing logical reasoning in automated programs

Good Old-fashioned Artificial Intelligence (GOFAI)

- First attempts at creating Artificial Intelligence were focussed on reproducing logical reasoning in automated programs
- All too frequently these approaches were unable to avoid the combinatorial explosion of possibilities as solutions were sought in very large search spaces

Good Old-fashioned Artificial Intelligence (GOFAI)

- First attempts at creating Artificial Intelligence were focussed on reproducing logical reasoning in automated programs
- All too frequently these approaches were unable to avoid the combinatorial explosion of possibilities as solutions were sought in very large search spaces
- Effective heuristics were developed when branching factors were not too large: eg deep blue Chess playing

Good Old-fashioned Artificial Intelligence (GOFAI)

- First attempts at creating Artificial Intelligence were focussed on reproducing logical reasoning in automated programs
- All too frequently these approaches were unable to avoid the combinatorial explosion of possibilities as solutions were sought in very large search spaces
- Effective heuristics were developed when branching factors were not too large: eg deep blue Chess playing
- General purpose AI seemed as remote as ever

Composing Learning Systems

- The core of the success of Computer Science has been the principle of divide and conquer

Composing Learning Systems

- The core of the success of Computer Science has been the principle of divide and conquer
  - Decompose a problem into clearly defined subproblems

Composing Learning Systems

- The core of the success of Computer Science has been the principle of divide and conquer
  - Decompose a problem into clearly defined subproblems
  - Repeat until the solution to the subproblems can be implemented directly

Composing Learning Systems

- The core of the success of Computer Science has been the principle of divide and conquer
  - Decompose a problem into clearly defined subproblems
  - Repeat until the solution to the subproblems can be implemented directly
  - Test solutions work against the definitions at each level

Composing Learning Systems

- The core of the success of Computer Science has been the principle of divide and conquer
  - Decompose a problem into clearly defined subproblems
  - Repeat until the solution to the subproblems can be implemented directly
  - Test solutions work against the definitions at each level
  - Deliver solution of overall problem

Composing Learning Systems

- The core of the success of Computer Science has been the principle of divide and conquer
  - Decompose a problem into clearly defined subproblems
  - Repeat until the solution to the subproblems can be implemented directly
  - Test solutions work against the definitions at each level
  - Deliver solution of overall problem
- What about composing learning systems

Composing Learning Systems

- The core of the success of Computer Science has been the principle of divide and conquer
  - Decompose a problem into clearly defined subproblems
  - Repeat until the solution to the subproblems can be implemented directly
  - Test solutions work against the definitions at each level
  - Deliver solution of overall problem
- What about composing learning systems
  - More tricky as functionality shifts as learning progresses

Composing Learning Systems

- The core of the success of Computer Science has been the principle of divide and conquer
  - Decompose a problem into clearly defined subproblems
  - Repeat until the solution to the subproblems can be implemented directly
  - Test solutions work against the definitions at each level
  - Deliver solution of overall problem
- What about composing learning systems
  - More tricky as functionality shifts as learning progresses
  - But first attempts to use principled approaches to do this within the PASCAL2 network: application to Go.

Composing Learning Systems

- The core of the success of Computer Science has been the principle of divide and conquer
  - Decompose a problem into clearly defined subproblems
  - Repeat until the solution to the subproblems can be implemented directly
  - Test solutions work against the definitions at each level
  - Deliver solution of overall problem
- What about composing learning systems
  - More tricky as functionality shifts as learning progresses
  - But first attempts to use principled approaches to do this within the PASCAL2 network: application to Go.
- Became the focus of a follow-on project CompLACS:

Composing Learning for Artificial Cognitive Systems (CompLACS)

- EU collaborative project looking at the potential of this strategy led by UCL, see http://complacs.cs.ucl.ac.uk/

Composing Learning for Artificial Cognitive Systems (CompLACS)

- EU collaborative project looking at the potential of this strategy led by UCL, see http://complacs.cs.ucl.ac.uk/
- Developed principles and practice for combining learning subcomponents, with guarantees on performance of the composition

Composing Learning for Artificial Cognitive Systems (CompLACS)

- EU collaborative project looking at the potential of this strategy led by UCL, see http://complacs.cs.ucl.ac.uk/
- Developed principles and practice for combining learning subcomponents, with guarantees on performance of the composition
- Example application was the control of Robots, UAVs, and web portal

Composing Learning for Artificial Cognitive Systems (CompLACS)

- EU collaborative project looking at the potential of this strategy led by UCL, see http://complacs.cs.ucl.ac.uk/
- Developed principles and practice for combining learning subcomponents, with guarantees on performance of the composition
- Example application was the control of Robots, UAVs, and web portal
- Will give one example of a system developed for reinforcement learning (RL)

Compressed Conditional Mean Embeddings for RL

- RL requires an agent to choose actions based on an observed state in order to maximise future reward: can model robots, UAVs and playing games in this framework

Compressed Conditional Mean Embeddings for RL

- RL requires an agent to choose actions based on an observed state in order to maximise future reward: can model robots, UAVs and playing games in this framework
- Key problem can be learning stochastic environment and estimating effect of actions on future rewards

Compressed Conditional Mean Embeddings for RL

- RL requires an agent to choose actions based on an observed state in order to maximise future reward: can model robots, UAVs and playing games in this framework
- Key problem can be learning stochastic environment and estimating effect of actions on future rewards
- Kernel methods enable embedding of probability distributions in kernel defined feature spaces (mean embedding):
  - evaluating an expectation becomes a simple inner product as functions are also represented as points in the feature space.

Compressed Conditional Mean Embeddings for RL

- Computing the influence of actions requires conditional mean embeddings:
  - reduces to learning a regression function from state, action pairs to distributions in the kernel defined feature space.

Compressed Conditional Mean Embeddings for RL

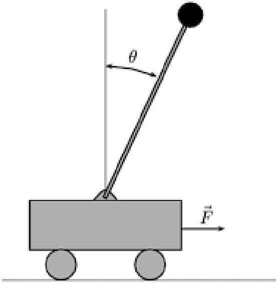
- Computing the influence of actions requires conditional mean embeddings:
  - reduces to learning a regression function from state, action pairs to distributions in the kernel defined feature space.
- Reduces continuous RL to finite state RL on training data:
  - hence can do exact planning provided not too many states

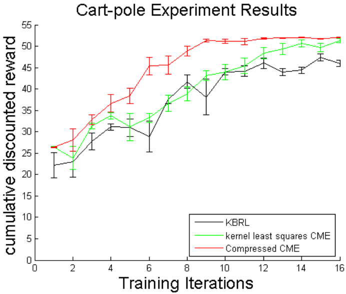
Compressed Conditional Mean Embeddings for RL

- Computing the influence of actions requires conditional mean embeddings:
  - reduces to learning a regression function from state, action pairs to distributions in the kernel defined feature space.
- Reduces continuous RL to finite state RL on training data:
  - hence can do exact planning provided not too many states
- Matching pursuit and data compression ensure computation does not explode

Simulated under-actuated cart-pole swing-up benchmark problem

- $\mathcal{S} = \mathbb{R}^2$, $s = (\theta, \dot{\theta})$, $\mathcal{A} = [-50, 50]$, horizontal force in newtons





Experiments: Quadrocopter Simulator

Simulator calibrated to model the dynamics of Pelican$^{\text{TM}}$ quadrocopter platforms

$\mathcal{S} \subset \mathbb{R}^{13}$, $s = (x, y, z, \theta\, \phi, \psi, \dot{x}, \dot{y}, \dot{z}, \dot{\theta}, \dot{\phi}, \dot{\psi}, F)$

$\mathcal{A} \subset \mathbb{R}^{3}$ represents desired velocity vectors, PID controller translates into low level commands

Tasks:

- Navigation: platform must navigate to point
- Holding pattern: platform must stay in circle and maintain minimum velocity

Experiments: Quadrocopter Results

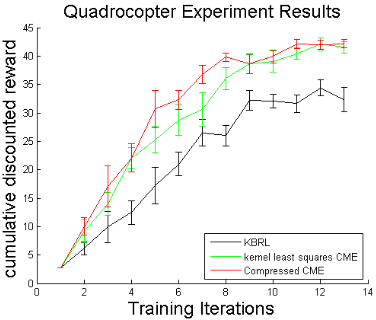


Figure: Quadrocopter tasks: navigation task

RKHS controller **better in high-dim. state-space**

Experiments: Quadrocopter Results

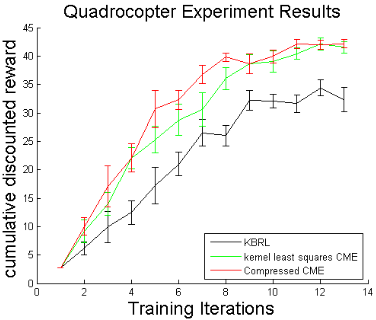


Figure: Quadrocopter tasks: navigation task

RKHS controller **better in high-dim. state-space**

- Extensions using deep learning to represent the kernel have been effective: richer representations but more data required.

Experiments: Quadrocopter Results

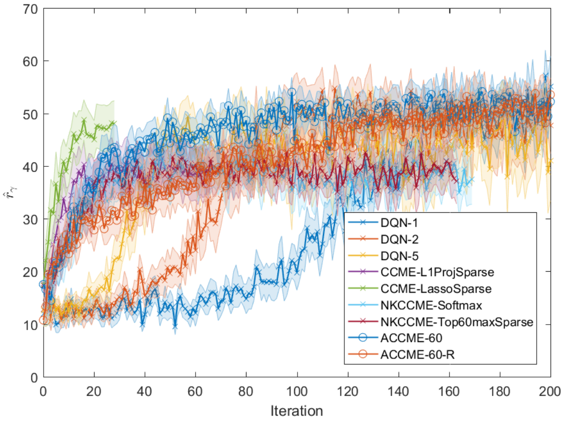


Figure: Quadrocopter tasks: holding pattern

Video

- see https://youtu.be/FNJQXRQbgX8

- Remarkable things emerge when just a few learning systems are combined

Progress on Composing Learning

- Remarkable things emerge when just a few learning systems are combined
  - IBM Watson uses information retrieval subsystems to propose potential answers combined with a machine learning method of ranking them

Progress on Composing Learning

- Remarkable things emerge when just a few learning systems are combined
  - IBM Watson uses information retrieval subsystems to propose potential answers combined with a machine learning method of ranking them
  - AlphaGo the Go playing system developed by DeepMind is based on the composition of three components: a deep learning system to evaluate board value; a prioritisation system for move planning that trades exploration and exploitation; and a deep learning system to compute the value function of a move

- Turning an AI problem into logical inference often throws the baby out with the bathwater:

- Turning an AI problem into logical inference often throws the baby out with the bathwater:
  - nuanced representations have been shown to retain semantic information, furthermore the additional information contains patterns that machine learning can for example use to prioritise the search

- Turning an AI problem into logical inference often throws the baby out with the bathwater:
  - nuanced representations have been shown to retain semantic information, furthermore the additional information contains patterns that machine learning can for example use to prioritise the search
  - Machine learning can harvest patterns in data to ensure that these clues are exploited to create effective performance

- Turning an AI problem into logical inference often throws the baby out with the bathwater:
  - nuanced representations have been shown to retain semantic information, furthermore the additional information contains patterns that machine learning can for example use to prioritise the search
  - Machine learning can harvest patterns in data to ensure that these clues are exploited to create effective performance
- But a combination of logical inference and machine learning techniques may be needed for further significant advances

Conclusion

- Principled approaches to machine learning have created reliable building blocks that when combined can generate behaviours that show progress towards general artificial intelligence

Conclusion

- Principled approaches to machine learning have created reliable building blocks that when combined can generate behaviours that show progress towards general artificial intelligence
- Caveats:
    - is intelligent behaviour the same as real intelligence?
    - still missing a general framework for creating composite learning systems.

Conclusion

- Principled approaches to machine learning have created reliable building blocks that when combined can generate behaviours that show progress towards general artificial intelligence
- Caveats:
  - is intelligent behaviour the same as real intelligence?
  - still missing a general framework for creating composite learning systems.
- But these advances do also challenge our understanding of what general artificial intelligence is:
  - Humans are expert at rationalising our actions after the event: not clear that we make them so rationally?