ALLEGATO B

UNIVERSITÀ DEGLI STUDI DI MILANO
selezione pubblica per n.1 posto di Ricercatore a tempo determinato in tenure track (RTT) per il settore concorsuale 01/B1 - Informatica, settore scientifico-disciplinare INF/01 - Informatica
presso il Dipartimento di Informatica "Giovanni degli Antoni" Codice concorso 5395

Andrea Paudice
CURRICULUM VITAE

INFORMAZIONI PERSONALI

COGNOME	**PAUDICE**
NOME	**ANDREA**
DATA DI NASCITA	24/12/1987

ATTIVITA' DI RICERCA

La mia attività di ricerca si concentra sugli aspetti fondazionali e teorici del Machine Learning. Nello specifico sono interessato alla progettazione e all'analisi di algoritmi che abbiano prestazioni matematicamente certificate e che siano utilizzabili nella risoluzione di problemi reali. Questo tipicamente si traduce: o nel progetto di algoritmi con garanzie teoriche calcolabili, in opposizione a modelli che invece richiedono all'utilizzare conoscenze di dominio spesso difficili da ottenere; o nell'analisi di problemi fondamentali sotto ipotesi deboli. Più di recente mi sto anche interessando a problemi di ottimizzazione convessa e le sue intersezioni con la teoria statistica dell'apprendimento. Sono anche interessato alle applicazioni di tali metodi a problemi di analisi di serie-temporali e adversarial machine learning.

Ricerca Attuale

Active Learning. Data una sequenza di *n* punti etichettati *S=((x1,y1),…,(xn,yn))*,un algoritmo di classificazione identifica una funzione che può etichettare dati futuri. Le prestazioni di tali algoritmi dipendono direttamente dalla dimensione *n* del dataset di input. Oggigiorno, collezionare grosse moli di punti *x* è semplice, mentre ottenere delle etichette accurate è difficoltoso. L'apprendimento attivo (*active learning*) è un'area del moderno machine learning che si interessa dello sviluppo e dell'analisi di algoritmi che, partendo da una sequenza non etichettata di punti *Sx=(x1,…,xn)*, ed interagendo interattivamente con un etichettatore, etichettano l'intera sequenza *Sx* richiedendo solamente *O(log n)* etichette; un fenomeno noto come *exponential savings*. Il risultato è un risparmio esponenziale nel costo di etichettazione. Tuttavia, non sempre è possibile riuscire ad ottenere un risparmio esponenziale nel costo di etichettatura. La mia attività di ricerca verte appunto sull'identificazione di condizioni sufficiente all'esistenza di algoritmi con vantaggi esponenziali.

I miei contributi in questo campo interessano: i) l'identificazione di condizioni geometriche generali per gli exponential savings nel caso di punti in spazi Euclidei (NeurIPS 2022, NeurIPS 2021, NeurIPS 2020), queste condizioni sono poi sfruttate per la progettazione di algoritmi efficienti; ii) l'identificazione di condizioni astratte per gli exponential savings in spazi astratti, inclusi quelli pseudo-metrici, e la progettazione di algoritmi che le sfruttano (JMLR 2023, NeurIPS 2021, COLT 2021); iii) l'approssimazione e la ricostruzione di clusters in spazi metrici semi-avversariali (NeurIPS 2019).

Robust Machine Learning. Molti degli attuali algoritmi di statistical learning, e le rispettive analisi, assumono che i dati siano identicamente distribuiti; talvolta richiedendo anche stringenti condizioni sul rumore (dette *light tails*). Nelle moderne applicazioni di machine learning tuttavia, si hanno grosse quantità di dati disponibili per il training, ma di qualità bassa. Il risultato è che molti degli algoritmi tradizionali ottengono prestazioni scadenti. Un esempio classico è l'effetto degli *outlier*; anche pochi outlier possono cambiare radicalmente il modello appreso dagli algoritmi. Lo studio di algoritmi di machine learning robusti, mira al design di metodi che offrano le circa stesse prestazioni dei modelli classici, ma sotto assunzioni di rumore più deboli.

I miei contributi in questo campo spaziano su due fronti. Nel contesto dei metodi non supervisionati (clustering, riduzione della dimensionali, codifica sparsa, representation learning), essi interessano il noto *Huber contamination model* secondo il quale i dati di training sono contaminati da una componente di rumore stocastico arbitrario. In tale setting, è stata identificata una tecnica generale (i.e. che si applica uniformemente a tutti i problemi citati) che consente di ottenere buone prestazioni sulla compente pulita del modello (ICML 2021).

Nel contesto dell'ottimizzazione stocastica, uno dei fondamenti del machine learning moderno, i contributi forniti interessano l'estensione delle proprietà di convergenza del celebre algoritmo *Stochastic Gradient Descent* (SGD) al caso di dati corrotti da rumore a code pesanti (*heavy tails*). Questo setting, è significativamente più difficile di quello light tails, e garanzie forti per problemi *non-smooth* come la regressione statistica con errore assoluto o la classificazione con hinge-loss erano mancanti. In questo contesto, i risultati ottenuti (in revisione a SJMODS 2022 e AAAI 2023) interessano la convergenza di un'ampia classe di schemi di tipo SGD, estendendo di fatto la concreta applicabilità di tali algoritmi.

Multi-task Learning. Talvolta ci si trova a dover risolvere N task simultaneamente. Una domanda centrale nel machine learning moderno è se sia possibile sfruttare eventuali similarità tra i task esistenti per ottenere dei vantaggi significativi rispetto alla risoluzione individuale degli stessi (*multi-task learning*). L'idea è che è possibile migliorare le prestazioni su un singolo task, utilizzando le informazioni degli altri task. Sebbene questo problema sia stato estremante studiato nel setting di apprendimento statistico; nella variante *online*, in cui i dati arrivano uno alla volta, molte domande sono ancora aperte.

I contributi apportati su questo problema sono: i) una famiglia di algoritmi che generalizza il noto *Online Mirror Descent* (*OMD*) al setting multi-task e che ottiene prestazioni ottime; ii) una estensione al caso in cui l'utilizzatore non ha informazioni pregresse sulle relazioni esistenti tra tasks (TMLR 2022).

Ricerca Passata

Adversarial Machine Learning. Data la grande diffusione delle tecnologie di Machine Learning, anche in applicazioni suscettibili a *cyber attacks*, è un problema di grande interesse determinare se e come un hacker può influenzare le prestazioni di tali tecnologie. Lo studio della sensibilità di tali algoritmi ai cyber attacks è parte di quello che è oggi noto come Adversarial Machine Learning. Un problema che sta ricevendo particolare attenzione di recente è quello dei cosiddetti *poisoning attacks* in cui un hacker introducendo pochi dati fabbricati ad-hoc nel training set, può influenzare pesantemente il modello appreso dall'algoritmo di learning, con conseguenze potenzialmente catastrofiche.

I miei contributi in questo campo interessano: i) la dimostrazione che di strategie di attacco (i.e. algoritmi per la generazione dei punti malevoli) semplici ed efficienti compromette seriamente le prestazioni anche di modelli di *Deep Learning* allo stato dell'arte (AISec 2017, TOSN 2018); ii) utilizzando opportune strategie di smoothing è possibile limitare gli effetti di certi poisoning attacks (ECML 2018).

Security Alerts Analysis. Molti dei moderni sistemi di monitoraggio della sicurezza delle infrastrutture informatiche (*SIEM*) generano dati ad un tasso che non è sostenibile dagli analisti umani il cui ruolo è quello di esaminare lo stato attuale del sistema sulla base di queste informazioni. Il risultato è che tali enormi moli di dati sono utilizzate solo per analisi *post-mortem* per risalire alle cause di un incidente. Ciò nonostante, identificare i pochi segni rilevanti di un incidente tra questi allarmi è un problema complesso. Inoltre tali allarmi, spesso testuali e poco strutturati; oltre che senza etichette che li distringono tra malevoli e benevoli. Il risultato è che i) i tradizionali approcci di apprendimento supervisionato non si prestano alla risoluzione del problema; ii) le metriche classiche (principalmente geometriche) dei metodi non supervisionati danno luogo a risultati scarsamente interpretabili in questo contesto.

In questo contesto, i contributi forniti sono i seguenti: i) la definizione di un metodo non supervisionato per l'analisi di dati non strutturati e di un framework di validazione interpretabile (RSDA 2014, TDSC 2019, FCGS 2016); ii) lo sviluppo di algoritmo di inferenza casuale per la determinazione dei percorsi di evoluzione dei cyber attacks (TOPS 2017).

TITOLI

TITOLI DI STUDIO

- PhD in Computer Science, *cum laude*
  - University of Milan "La Statale", Gennaio 2022.
  - Titolo: Algorithms for Clustering and Unsupervised Learning Problems.

- Supervisori: Prof. Nicolò Cesa-Bianchi, Dr. Massimiliano Pontil.

- Master of Research in Advanced Computing with *Distinction*
  - Imperial College London, Ottobre 2016.
  - Titolo: Shedding Light on Stability-based Feature Selection.
  - Supervisori: Prof. Emil Lupu, Dr. Andras Gyorgy.

- Master Degree in Computer Engineering *summa cum Laude*
  - University of Naples "Federico II", Settembre 2014.
  - Titolo: Filtering Alerts for the Analysis of a Production SaaS Cloud: a Benchmarking Study.
  - Supervisori: Prof. Domenico Cotroneo, Dr. Antonio Pecchia.

- Bachelor Degree in Computer Engineering
  - University of Naples "Federico II", Settembre 2011.
  - Titolo: Analisi Empirica di Reti IP Satellitari.
  - Supervisori: Prof. Antonio Pescapè, Dr. Alessio Botta.

CONTRATTI DI RICERCA, ASSEGNI DI RICERCA O EQUIVALENTI

- Ricercatore PostDoc
  - Università di Milano "La Statale", 01/04/2022 - 01/04/2024.
  - Temi: Progettazione ed analisi di algoritmi di active learning e ottimizzazione stocastica.

- Assegno di ricerca
  - Istituto Italiano di Tecnologia, 01/10/2021 - 01/09/2022.
  - Temi: Progettazione ed analisi di algoritmi di active learning e ottimizzazione stocastica.

- Borsa di Dottorato
  - Università degli Studi di Milano "La Statale", 01/10/2018 - 01/10/2022.
  - Temi: Progettazione ed analisi di algoritmi di active learning e ottimizzazione stocastica.

- Assegno di Ricerca
  - Imperial College London, 01/10/2015 - 01/04/2018.
  - Temi: Progettazione ed analisi di algoritmi di adversarial machine learning, relative applicazioni.

- Assegno di Ricerca
  - Università degli Studi di Napoli "Federico II", 01/01/2015 - 01/10/2015.
  - Temi: Progettazione di algoritmi di anomaly detection per dati di natura testuale.

ALTRE ATTIVITA' PROFESSIONALI

- Machine Learning Scientist
  - Intecs s.p.a., 01/04/2018 - 01/10/2018.
  - Temi: Progettazione ed analisi di algoritmi di time-series classification.

ATTIVITÀ DIDATTICA A LIVELLO UNIVERSITARIO IN ITALIA O ALL'ESTERO

Didattica

- Tutoraggio per il corso di "Statistical Methods for Machine Learning", dipartimento di Informatica, Università degli Studi di Milano "La Statale", 45 ore, 01/03/2022 - 01/09/2022.

- Tutoraggio per il corso di "Statistical Methods for Machine Learning", dipartimento di Informatica, Università degli Studi di Milano "La Statale", 30 ore, 01/03/2021 - 01/09/2021.

- Tutoraggio per il corso di "Statistical Methods for Machine Learning", dipartimento di Informatica, Università degli Studi di Milano "La Statale", 30 ore, 01/03/2020 - 01/09/2020.

- Teaching Assistant per il corso di "Machine Learning", dipartimento di ingegneria Elettronica, Imperial College London, 90 ore, 01/01/2016 - 01/06/2016.

Supervisione

- Co-supervisione di 4 tesi magistrali di studenti magistrali in Informatica, Imperial College London, 01/10/2015-01/04/2018.

DOCUMENTATA ATTIVITÀ DI FORMAZIONE O DI RICERCA PRESSO QUALIFICATI ISTITUTI ITALIANI O STRANIERI;

- Machine Learning School "RegML", Università degli studi di Genoa, 01/07/2016-31/07/2016.
  - Temi: Regolarizzazione nel Machine Learning, Complessità Computazionale nell'Ottimizzazione Convessa.

ORGANIZZAZIONE, DIREZIONE E COORDINAMENTO DI GRUPPI DI RICERCA NAZIONALI E INTERNAZIONALI, O PARTECIPAZIONE AGLI STESSI

- *European Laboratory for Learning and Intelligent Systems* (ELLIS) Member 01/01/2022 - Presente
  - ELLIS è l'iniziativa Europea per mantenere competitiva la ricerca nel Machine Learning. La membership viene concessa solo a chi ha dato contributi rilevanti al campo.

ATTIVITÀ DI RELATORE A CONGRESSI E CONVEGNI NAZIONALI E INTERNAZIONALI

- Intervento breve e presentazione poster su "*Exact Recovery of Clusters in Finite Metric Spaces Using Oracle Queries*". 34th Conference on Learning Theory (COLT), 2021.

- Intervento breve e presentazione poster su "*Robust Unsupervised Learning via L-statistic Minimization*". 38th International Conference on Machine Learning (ICML), 2021.

- Intervento breve e presentazione poster su "*On Margin-Based Cluster Recovery with Oracle Queries*". Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems (NeurIPS), 2021.

- Intervento breve e presentazione poster su "*Exact Recovery of Mangled Clusters with Same-Cluster Queries*". Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS), 2020.

- Presentazione poster su "*Correlation Clustering with Adaptive Similarity Queries*". Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS), 2019.
  - Intervento esteso a Google Zurigo, Ottobre 2019.
  - Intervento esteso a University College London, Londra, Giugno 2019.

- Intervento su "*An Experiment with Conceptual Clustering for the Analysis of Security Alerts*", 25th IEEE International Symposium on Software Reliability Engineering Workshops, 2014.

ATTIVITA' DI PEER REVIEWING E COMMISSIONE CONFERENZE

Attività di Peer Reviewing per Conferenze

- Conference on Learning Theory (COLT): 2018, 2019, 2020, 2021, 2022, 2023.

- Conference on Algorithmic Learning Theory (ALT): 2018, 2019, 2020.

- International Conference on Machine Learning (ICML): 2023.

- Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems (NeurIPS); 2020.

- AAAI Conference on Artificial Intelligence (AAAI): 2023.

- European Conference on Machine Learning (ECML): 2022, 2023.

- The Web Conference (WWW): 2020.

- IEEE International Conference on Software Reliability Engineering (ISRE): 2017.

- International Conference on Distributed Computing and Networking (ICDCN): 2017.

Attività di Peer Reviewing per Riviste

- Journal of Machine Learning Research (JMLR): 2018, 2019, 2020, 2021, 2022.

- Pattern Recognition (PR): 2018, 2019.

- IEEE Transaction of Pattern Analysis and Machine Intelligence (TPAMI): 2018, 2019, 2020, 2021, 2022.

- IEEE Transaction on Neural Networks and Learning Systems (TNNLS): 2017.

- Future Generation Computer Systems (FGCS): 2017, 2018.

- IEEE Transaction on Dependable and Secure Computing (TDSC): 2017.

Commissione conferenze

- Conference on Learning Theory (COLT) Program Committee Member: 2020, 2021, 2022, 2023.
- Association for Advancements on Artificial Intelligence (AAAI): 2024.

CONSEGUIMENTO DI PREMI E RICONOSCIMENTI NAZIONALI E INTERNAZIONALI PER ATTIVITÀ DI RICERCA

- Riconoscimento da parte di *European Laboratory for Learning and Intelligent Systems* (ELLIS) come membro.

- Vincitore di assegno di ricerca per PostDoc in Informatica all'Università degli Studi di Milano, qualificandomi primo.

- Vincitore di concorso di dottorato in Informatica all'Università degli Studi di Milano, qualificandomi primo.

- La pubblicazione "*Exact Recovery of Mangled Clusters with Same-Cluster Queries*" è stata selezionata per una full oral presentation a Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS) 2020. Un riconoscimento riservato al top 1% delle pubblicazione accettate.

- Vincitore di concorso di dottorato in Informatica all'Università di Manchester (declinato).

- Vincitore di concorso di dottorato in Informatica all'Università di Lille (declinato).

PRODUZIONE SCIENTIFICA

PUBBLICAZIONI SCIENTIFICHE

La comunità scientifica del Machine Learning è fondamentalmente conference-driven. Pertanto, i risultati rilevanti appaiono sempre prima ad una conferenza, e solo dopo (e non sempre) in versione estesa su rivista. Per questa ragione, la maggior parte delle mie pubblicazioni nell'area del machine learning teorico, sono a conferenza. La conferenza di machine learning teorico più importante è COLT; mentre le principali conferenze sono ICML e NeurIPS. La rivista principale del settore è JMLR, seguita da riviste tematiche di Ottimizzazione e Pattern Analysis. Nei lavori fondazionali è convenzione utilizzare l'ordine alfabetico per gli autori, ma il mio contributo è stato sempre di primo piano.

Nel contesto della sicurezza e dell'affibilità dei sistemi informatici, le principali aree applicative su cui ho lavorato, TDSC, FGCS, TSN, TOPS sono riviste di riferimento.

Conferenze

- Marco Bressan, , Nicolò Cesa-Bianchi, Silvio Lattanzi, Andrea Paudice, Maximilian Thiessen. "*Active Learning of Classifiers with Label and Seed Queries*". Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems (NeurIPS), 2022. CORE Classification A*. ISBN: Ancora non disponibile.

- Marco Bressan, , Nicolò Cesa-Bianchi, Silvio Lattanzi, Andrea Paudice. "*On Margin-Based Multi-Class Active Learning and Cluster Recovery*". Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems (NeurIPS), 2021. CORE Classification A*. ISBN: I9781713845393.

- Marco Bressan, , Nicolò Cesa-Bianchi, Silvio Lattanzi, Andrea Paudice. "*Exact Recovery of Clusters in Finite Metric Spaces Using Oracle Queries*". Proceedings of 34th Conference on Learning Theory (COLT), 2021. CORE Classification A*. ISBN: Ancora non disponibile.

- Andreas Maurer, Daniela Parletta, Andrea Paudice, Massimiliano Pontil. "*Robust Unsupervised Learning via L-statistic Minimization*". Proceedings of the 38th International Conference on Machine Learning (ICML). CORE Classification A*. ISBN: Ancora non disponibile.

- Marco Bressan, , Nicolò Cesa-Bianchi, Silvio Lattanzi, Andrea Paudice. "*Exact Recovery of Mangled Clusters with Same-Cluster Queries*". Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS), 2020. CORE Classification A*. ISBN: 9781713829546.

- Marco Bressan, , Nicolò Cesa-Bianchi, Andrea Paudice, Fabio Vitale. "*Correlation Clustering with Adaptive Similarity Queries*". Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS), 2019. CORE Classification A*. ISBN: 9781713807933.

- Andrea Paudice, Luis Munoz-Gonzalez, Emil. C. Lupu. "*Label Sanitization Against Label Flipping Poisoning Attacks*". Proceedings of the European Conference on Machine Learning (ECML) 2018, 2018. Workshops. ISBN: 978-3-030-13452-5.

- Luis Munoz-Gonzalez, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, Fabio Roli. "*Towards Poisoning of Deep Learning Algorithms with Back-gradient Optimization*". Proceedings of the 10th {ACM} Workshop on Artificial Intelligence and Security (AISec), 2017. DOI: 10.1145/3128572.3140451.

- Andrea Paudice, Santonu Sarkar, Domenico Cotroneo. "*An Experiment with Conceptual Clustering for the Analysis of Security Alerts*". Proceedings of the 25th IEEE International Symposium on Software Reliability Engineering Workshops, 2014. DOI: 10.1109/ISSREW.2014.82.

Riviste

- Nicolò Cesa-Bianchi, Pierre Laforgue, Andrea Paudice, Massimiliano Pontil. "*Online Multi-Task Mirror Descent*". Transaction on Machine Learning Research, 2022. ISBN: Ancora non disponibile.

- Domenico Controneo, Andrea Paudice, Antonio Pecchia. "*Empirical Analysis and Validation of Security Alerts Filtering Techniques*". IEEE Transaction on Dependable and Secure Computing, 2019, volume 16, pp. 856-870. DOI: 10.1109/TDSC.2017.2714164.

- Vittorio P. Illiano, Andrea Paudice, Luis Munoz-Gonzalez, Emil. C. Lupu. "*Determining Resilience Gains From Anomaly Detection for Event Integrity in Wireless Sensor Networks*". ACM Transaction on Sensor Networks, 2018, volume 14, number 1, pp. 5-35. DOI: 10.1145/3176621.

- Luis Munoz-Gonzalez, Daniele Sgandurra, Andrea Paudice, Emil. C. Lupu. "*Efficient Attack Graph Analysis through Approximate Inference.*". ACM Transaction on Private and Security, 2017, volume 20, number 3, pp. 10-30. DOI: 10.1145/3105760.

- Domenico Cotroneo, Andrea Paudice, Antonio Pecchia. "*Automated root cause identification of security alerts: Evaluation in a SaaS Cloud*". Future Generation and Computer Systems, 2016, volume 56, pp. 375-387. DOI: 10.1016/j.future.2015.09.009.

In revisione

- Roberto Colomboni, Emmanuel Esposito, Andrea Paudice. An Improved Uniform Convergence Bound with Fat-Shattering Dimension. (IPL), 2023.

- Francois Bachoc, Tommaso Cesari, Roberto Colombo, Andrea Paudice. "*A Near-Optimal Algorithm for Univariate Zeroth-Order Budget Convex Optimization*". 40th International Conference on Machine Learning (AAAI), 2023.

- Daniela Parletta, Andrea Paudice, Massimiliano Pontil, Saverio Salzo. "*High Probability Bounds for Subgradient Methods under Heavy Tails Noise*". SIAM Journal on Mathematics Of Data Science (SIMODS), 2023.

- Marco Bressan, Nicolò Cesa-Bianchi, Andrea Paudice, Silvio Lattanzi. "*Margin-Based Active Learning of Multiclass Classifiers*". Journal of Machine Learning Research (JMLR), Accettato con revisione minore. 2023.

COLLABORAZIONI DI RICERCA

- Nicolò Cesa-Bianchi, Università degli Studi di Milano "La Statale"
- Marco Bressan, Università degli Studi di Milano "La Statale"
- Roberto Colomboni, Università degli Studi di Milano "La Statale"
- Khaled Endowa, Università degli Studi di Milano "La Statale"
- Emmanuel Esposito, Università degli Studi di Milano "La Statale"
- Silvio Lattanzi, Google
- Andras Gyorgy, Google DeepMind
- Maximilian Thiessen, Technische Universität Wien
- Andrea Maurer, Istituto Italiano di Tecnologia
- Daniela A. Parletta, Università degli Studi di Genoa, Istituto Italiano di Tecnologia
- Massimiliano Pontil, Istituto Italiano di Tecnologia
- Saverio Salzo, Università degli Studi di Roma "La Sapienza", Istituto Italiano di Tecnologia
- Emil C. Lupu, Imperial College London
- Luis Munoz-Gonzalez, Imperial College London
- Tommaso Cesari, University of Ottawa
- Francois Bachoc, University Paul Sabatier
- Antonio Pecchia, Università degli Studi del Sannio
- Domenico Cotroneo, Università degli Studi di Napoli "Federico II"
- Battistia Biggio, Università degli Studi di Cagliari

Data	25/10/2023	Luogo	Napoli