Dati sintetici e Modelli GAN

Teoria e Fondamenti

Francesca Liguori, Manuel Caccone

11 luglio 2025

Section 1

Introduzione ai dati sintetici

Definizione e caratteristiche

- I dati sintetici sono **dataset artificiali** progettati per replicare le **proprietà statistiche**, le **correlazioni** e le **distribuzioni** dei dati reali
- Preservano la struttura e le relazioni tra variabili senza contenere riferimenti a individui o a eventi specifici
- Costituiscono una soluzione efficace al trade-off tra utilità analitica e protezione della privacy
- Permettono lo sviluppo e la validazione di modelli attuariali in contesti sicuri

Tipologie principali

- **Completamente sintetici**: generati interamente attraverso modelli statistici o algoritmi
- **Parzialmente sintetici**: solo le variabili sensibili vengono sostituite con valori artificiali
- **Ibridi**: combinazione strategica di elementi reali e sintetici per obiettivi specifici

Section 2

Vantaggi e sfide nell'adozione dei dati sintetici

Opportunità strategiche

I dati sintetici eliminano i colli di bottiglia legati alla disponibilità e alla qualità dei dati reali, offrendo i seguenti vantaggi:

- **Scalabilità**: possibilità di generare volumi illimitati per addestrare modelli complessi
- **Compliance normativa**: assenza di informazioni sensibili, in piena conformità con GDPR e normative sulla protezione dei dati
- **Disponibilità di scenari**: generazione di casistiche rare ma rilevanti per la modellazione
- **Riduzione dei bias**: capacità di correggere distorsioni storiche presenti nei dataset reali, migliorando l'equità dei modelli predittivi
- **Costi ridotti**: risparmio notevole rispetto alla raccolta e annotazione manuale

Limiti e rischi emergenti

- **Degradazione della qualità**: modelli generativi imperfetti possono introdurre artefatti o relazioni spurie, compromettendo l'affidabilità
- **Ambiguità della compliance normativa**: assenza di standard uniformi per la validazione legale, particolarmente rilevante in ambiti regolamentati come le assicurazioni
- **Complessità tecniche**: necessità di competenze specializzate in AI e statistica, spesso scarse nelle organizzazioni tradizionali

Casi d'uso specifici

- **Pricing assicurativo**: combinazione di dati sintetici con informazioni tradizionali (età, parametri geografici, . . .) per migliorare la personalizzazione del premio
- **Analisi di scenari catastrofali**: simulazione di eventi climatici estremi per migliorare la valutazione dei rischi
- **Rilevamento frodi**: creazione di pattern comportamentali anomali per training di modelli, senza esporre dati sensibili

Section 3

Ratemaking con dati sintetici

Fondamenti del ratemaking sintetico

- **Ratemaking tradizionale**: determinazione dei premi assicurativi basata su analisi statistica dei dati storici di sinistri, esposizioni e variabili di rischio
- **Innovazione sintetica**: utilizzo di dati artificiali per superare i limiti dei dataset storici e migliorare la precisione attuariale

Vantaggi specifici per il ratemaking

- **Ampliamento del dataset**: generazione di osservazioni per segmenti con pochi dati storici
- **Bias correction**: correzione di distorsioni storiche nei portafogli esistenti
- **Validazione robusta**: test di stress dei modelli su casistiche diversificate

Micro-segmentazione RC Auto: Caso Pratico

Scenario: Compagnia assicurativa vuole personalizzare il pricing per giovani conducenti (18-25 anni) con dati comportamentali limitati

Variabili reali osservate (tutto il portafoglio)

- Età: 18-25 anni
- Provincia: Milano/Roma/Napoli
- Tipo veicolo: Utilitaria/Berlina/SUV

Variabile sintetica da generare

- **Stile di guida**: Aggressivo/Normale/Prudente
- **Disponibile solo per 500 clienti su 10.000** (5% con black box)

Step logici della generazione sintetica (1/2)

Step 1: Analisi dei dati reali limitati

Dai 500 clienti con black box osservo:

- 175 Aggressivi (35%), 250 Normali (50%), 75 Prudenti (15%)
- **Correlazioni scoperte:**
  - Età 18-20 + SUV $\rightarrow$ 60% Aggressivo
  - Età 23-25 + Utilitaria $\rightarrow$ 70% Prudente
  - Roma + Berlina $\rightarrow$ 55% Normale

Step 2: Training della GAN

La GAN impara le regole:

- INPUT: Età, Provincia, Tipo veicolo
- OUTPUT: Probabilità per ogni stile di guida
- **Vincolo**: Mantenere distribuzione 35%/50%/15%

Step logici della generazione sintetica (2/2)

Step 3: Generazione per i restanti 9.500

Per ogni cliente senza black box:

- Cliente: Età=19, Provincia=Roma, Veicolo=SUV
- **GAN predice**: 65% Aggressivo, 30% Normale, 5% Prudente
- **Assegnazione**: Aggressivo

Risultati della micro-segmentazione

Segmento tradizionale: - Premio base giovani: €1.200 - Coefficiente unico: 2.5x rispetto al base

Micro-segmenti con dati sintetici:

- Giovane prudente, utilitaria: €950 (-21%)
- Giovane aggressivo, SUV: €1.650 (+38%)
- Giovane normale, berlina: €1.180 (-2%)

Benefici ottenuti

- **Precisione**: Riduzione della varianza non spiegata
- **Equità**: Pricing più giusto per conducenti virtuosi
- **Competitività**: Acquisizione di segmenti redditizi

Section 4

Generazione dei dati sintetici

Processo di generazione

1. **Preparazione e analisi dei dati reali**: pulizia e codifica dei dati, seguite dalle analisi statistiche necessarie a comprenderne le caratteristiche fondamentali
2. **Scelta e addestramento del modello generativo**: selezione del modello e addestramento sui dati reali, per apprendere le distribuzioni e le relazioni tra le variabili
3. **Generazione e post-processing**: generazione dei dati sintetici, poi trasformati per renderli coerenti con il formato e la struttura dei dati reali di partenza
4. **Validazione e utilizzo**: verifica della coerenza tra i dati sintetici e i dati reali rispetto alle distribuzioni e alle relazioni tra le variabili e verifica dell'anonimizzazione del dataset sintetico prima dell'utilizzo effettivo

Approcci tradizionali

- **Modelli parametrici**: utilizzo di distribuzioni note (Pareto, Gamma, Weibull) per la simulazione
- **Tecniche di ricampionamento**: bootstrap e varianti per la generazione di nuovi dataset
- **Simulazione Monte Carlo**: generazione di scenari basata su modelli probabilistici definiti

Approcci basati su Intelligenza Artificiale

- **Generative Adversarial Networks (GAN)**: competizione tra reti neurali per la generazione ottimale
- **Variational Autoencoders (VAE)**: codifica e decodifica attraverso spazi latenti
- **Transformer e Large Language Models**: generazione di dati strutturati tramite modelli linguistici

Section 5

Modelli *Generative Adversarial Networks* (GAN)

Fondamenti teorici delle GAN

- **Introdotte da Ian Goodfellow nel 2014**
- **Framework basato sulla teoria dei giochi a somma zero**
- Due reti neurali che competono in un processo antagonistico:
  - **Discriminator (D)**: impara a distinguere dati reali dai dati sintetici generati
  - **Generator (G)**: impara a creare dati sintetici sempre più realistici per ingannare il discriminatore
- **Minmax Game**: G cerca di minimizzare la capacità di D di distinguere i dati generati da quelli reali

Principio di funzionamento

- **Architettura competitiva**: due reti neurali (Generator e Discriminator) in competizione
- **Processo iterativo**: miglioramento progressivo della qualità dei dati generati
- **Equilibrio di Nash**: convergenza verso dati sintetici indistinguibili da quelli reali
- **Preservazione delle correlazioni**: mantenimento di relazioni complesse tra variabili

Formulazione matematica delle GAN

- **Funzione obiettivo**:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{\xi\sim p_\xi(\xi)}[\log(1 - D(G(\xi)))]$$

Dove

- $p_{data}(x)$: distribuzione dei dati reali
- $p_\xi(\xi)$: distribuzione del rumore ξ in input al generatore (spesso normale multivariata)
- $D(x)$: probabilità che x provenga dai dati reali secondo il discriminatore
- $G(\xi)$: output del generatore dato un input casuale ξ

Funzionamento del training GAN (1/3)

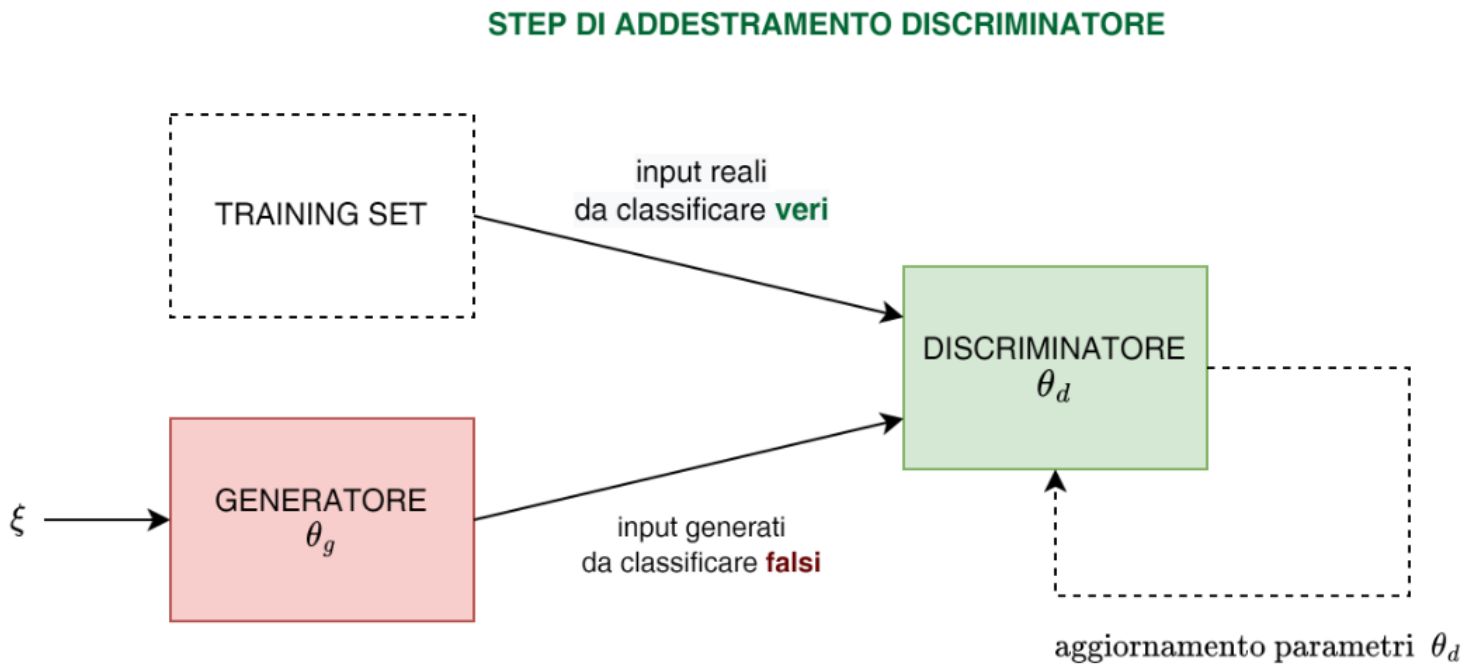

Figure 1: Step 1

Funzionamento del training GAN (2/3)

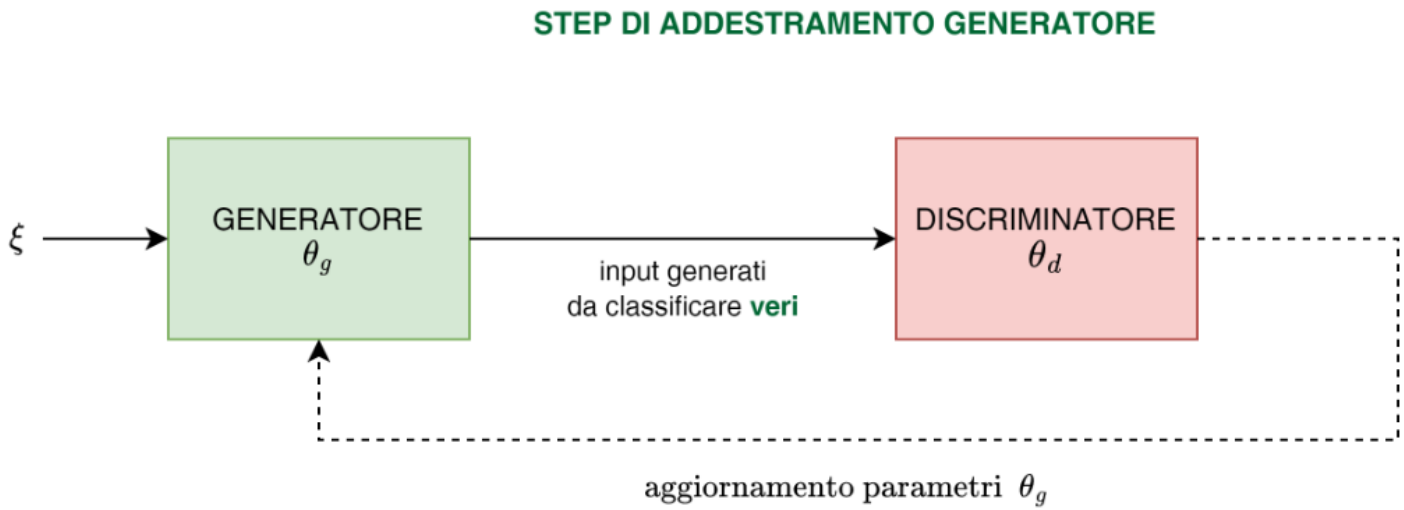

Figure 2: Step 2

Funzionamento del training GAN (3/3)

Il training è un processo iterativo che si interrompe quando si raggiunge l'**equilibrio di Nash**, ossia quando $D(x) = \frac{1}{2}$ per tutti gli x

Vantaggi delle GAN in ambito attuariale

- **Alta qualità del dato sintetico**:
  - Dati sintetici molto simili a quelli reali grazie al lavoro collaborativo tra le due reti neurali
- **Preservazione delle correlazioni complesse**:
  - Catturano relazioni non lineari tra variabili
  - Mantengono la struttura dei dati multivariati
- **Flessibilità**:
  - Capacità di generare scenari rari ma plausibili, utile per analisi di scenario
- **Scalabilità**:
  - Applicabilità a dataset di grandi dimensioni

Sfide nell'implementazione GAN

- **Mode collapse**: il generatore produce solo pochi esempi, con una ridotta capacità di generalizzazione
- **Training instabile**:
  - Difficoltà nel bilanciare l'apprendimento di G e D
  - Vanishing gradients
- **Valutazione**: difficoltà nel misurare oggettivamente la qualità dei dati generati
  - Metriche: Inception Score, Fréchet Inception Distance, Maximum Mean Discrepancy che misurano la *fidelity* (qualità) e la *diversity* (varietà) dei dati generati

Section 6

Validazione e Controllo Qualità

Metriche di validazione

- **Congruenza univariata**: test di Kolmogorov-Smirnov per la distribuzione delle singole variabili
- **Preservazione delle correlazioni**: analisi della matrice di correlazione tra dati reali e sintetici
- **Stabilità statistica**: confronto di momenti, quantili e statistiche descrittive

Section 7

Aspetti Normativi e Privacy

Conformità GDPR

- **Privacy by design**: integrazione della protezione privacy nel processo di generazione
- **Esenzione dal consenso**: non richiedono consenso esplicito per il trattamento
- **Status di anonimato**: i dati sintetici correttamente generati sono considerati anonimi
- **Verifica dell'anonimato**: necessità di test per escludere rischi di re-identificazione

Valutazione del rischio re-identificazione

- **k-anonymity**: verifica che ogni combinazione di attributi quasi-identificativi sia condivisa da almeno k record diversi
- **l-diversity**: controllo della diversità dei valori sensibili all'interno dei gruppi
- **Differential privacy**: aggiunta di rumore calibrato per garantire la privacy matematica

Esempio pratico: k-anonymity e l-diversity

Dataset RC Auto originale (10.000 record):

Età	Sesso	Provincia	Classe_Merito	Importo_Sinistro
22	M	Milano	15	€3.200
23	M	Milano	16	€2.800
22	F	Milano	14	€1.500

Verifica k-anonymity (k=3)

- Gruppo {22, M, Milano}: 850 record ***OK***
- Gruppo {45, F, Roma}: 2 record ***KO*** (< 3)

Verifica l-diversity (l=2)

- Gruppo {22, M, Milano}: Sinistri da €500 a €15.000 ***OK***
- Gruppo {35, F, Napoli}: Tutti sinistri ~€2.000 ***KO***

Miglioramento con dati sintetici

Dati sintetici generati (50.000 record):

- Tutti i gruppi hanno k $\geq$ 5
- Importi sinistri diversificati per ogni gruppo
- Preservate le correlazioni età-sinistralità
- **Risultato**: Privacy garantita mantenendo l'utilità analitica

Section 8

Conclusioni

Section 9

Implicazioni per la professione attuariale

Per l'Attuario

- **Espansione delle possibilità analitiche**: accesso a scenari prima inaccessibili
- **Accelerazione dell'innovazione**: sviluppo più rapido di prodotti e servizi
- **Standardizzazione**: evoluzione verso best practice condivise nel settore

Raccomandazioni per l'implementazione

- **Formazione continua**: investimento in competenze specifiche per massimizzare i benefici
- **Sperimentazione controllata**: approccio incrementale per acquisire esperienza pratica
- **Collaborazione professionale**: condivisione di esperienze e best practice
- **Monitoraggio normativo**: attenzione all'evoluzione del quadro regolamentare

Section 10

Bibliografia

Riferimenti teorici e metodologici

- Goodfellow, I., et al. (2014). "Generative Adversarial Nets". *Advances in Neural Information Processing Systems*, 27.
- Arjovsky, M., Chintala, S., & Bottou, L. (2017). "Wasserstein Generative Adversarial Networks". *International Conference on Machine Learning*.
- Jordon, J., Yoon, J., & van der Schaar, M. (2019). "PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees". *International Conference on Learning Representations*.

Applicazioni in ambito assicurativo

- Kuo, K., & Lupton, D. (2020). "Towards Explainability of Machine Learning Models in Insurance Pricing". *ASTIN Bulletin*, 50(1), 267-296.
- Lindholm, M., et al. (2022). "Machine Learning in Life Insurance: A Review of Methods and Applications". *European Actuarial Journal*, 12(1), 333-394.
- Denuit, M., & Hainaut, D. (2021). "Machine Learning Models for Motor Insurance Ratemaking". *European Actuarial Journal*, 11(1), 199-217.

Privacy e aspetti normativi

- El Emam, K., et al. (2020). "A systematic review of re-identification attacks on health data". *PLoS ONE*, 15(12), e0243633.
- Rubin, D. B. (1993). "Statistical Disclosure Limitation". *Journal of Official Statistics*, 9(2), 461-468.
- European Commission (2016). "General Data Protection Regulation (GDPR)". *Official Journal of the European Union*.

Section 11

Grazie per l'attenzione!

Prossimi passi e argomenti

- **Demo interattiva**: Webapp Shiny con dati sintetici
- **Vibe coding for actuaries**, pacchetti R essenziali