Titolo: La statistica nell'era dei big data

Autori: Eva Riccomagno, Daniele Venturoli

Testo:
Giorno dopo giorno, gli strumenti più diversi generano **una mole di dati**, i famosi 'big data'. Le centraline per misurare l'inquinamento, le carte di credito, le attività svolte da banche ed assicurazioni, e anche i social network quali facebook e twitter, sono esempi di generatori di dati. Per mettere ordine in questo continuo flusso d'informazioni e trasformarle in **conoscenza utilizzabile** occorre saperle gestire e interpretare in modo consapevole, competente e soprattutto corretto, e occorre costruire modelli matematici e statistici per **quantificare il rischio**, **fare previsioni** e simulare scenari alternativi. Il 'data analyst', o come si diceva un tempo lo statistico, si occupa proprio di questo.

La quantità di dati e la loro accessibilità, unite al fatto che gli strumenti informatici disponibili per gestirli ed analizzarli sono molti e molto raffinati, non devono però far dimenticare le **lezioni della statistica,** soprattutto quelle sulla qualità del dato e della sua rappresentatività del fenomeno di cui è istanza. Un esempio per tutti è dato dal progetto ***Google Flu Trends (GFT)***, un'applicazione ideata per prevedere i picchi influenzali basandosi sul numero di ricerche online d'informazioni sui sintomi dell'influenza. Inizialmente il progetto sembrava fornire previsioni più precise di quelle stimate dai CDC (*Centers for Disease Control and Prevention* di Atlanta, l'analogo statunitense della rete dei medici-sentinella in Italia), ma poi il progetto ha cominciato a sovrastimare il numero di persone che si sarebbero ammalate. In parte perché le ricerche sui sintomi venivano fatte anche da chi ammalato non era, ma soprattutto perché la qualità del dato non era sicura, né era direttamente dipendente dall'oggetto della previsione, a differenza dei dati forniti dai CDC.

Le lezioni derivate dall'insuccesso del progetto GFT (che come strumento complementare per le previsioni dei picchi influenzali è invece utilissimo) e analoghe esperienze possono riassumersi nella frase '**big data, old problems**' al centro di fervide discussioni in corso in diversi ambienti scientifici. È auspicabile e prevedibile che presto emergerà un'evoluzione della **Statistica Matematica** in grado di includere i disparati metodi di analisi attualmente disponibili per i big data (bootstrapping, MCMC, lasso eccetera), così come successe il secolo scorso quando le fondamenta della **Statistica Matematica** furono saldamente stabilite sulla Probabilità e sugli assiomi di Kolmogorov da scienziati quali R. A.Fisher e J. Neyman.

Un approccio consapevole all'analisi dei dati è parte integrante delle **attività di didattica e di ricerca che si conducono al Dipartimento di Matematica di Genova**. Si è rivelato vincente per creare un algoritmo per la prevenzione di frodi nell'online banking e (in collaborazione con altri dipartimenti dell'Ateneo) per studiare come valutare e combinare le diverse opinioni di esperti per lo sviluppo di un'unità navale. I temi qui accennati sono curati nel corso di laurea in Matematica e approfonditi nel corso di laurea triennale in SMID (Statistica Matematica e Trattamento Informatico dei Dati). I nostri laureati si inseriscono negli ambiti lavorativi in cui analisi dati e previsioni in presenza di incertezza sono necessarie. La presenza di SMID nella classe di laurea in Matematica, unica triennale in Italia, è motivata dalla consapevolezza che statistica, matematica e informatica sono i tre strumenti necessari al **Data Analyst** del mondo contemporaneo affinché possa offrire una chiave per interpretare i risultati di esperimenti e osservazioni, e in definitiva, una vera e propria chiave per capire meglio il mondo che ci circonda.