Analisi dei Dati

Marco Romito
marco.romito@unipi.it
http://people.dm.unipi.it/romito

Informazioni generali

Pagina del corso:
`http://people.dm.unipi.it/romito/Teaching/2024/data`

Prerequisiti: probabilità e statistica elementari, algebra lineare, elementi di ottimizzazione, contenuti del corso di *Statistica Superiore*, familiarità con `R` o `python`.

Contenuti:

- La teoria dell'apprendimento
- Valutazione dei modelli di inferenza e previsione.
- "Supervised learning" (classificazione mediante regressione, svm, random forests, reti neurali, etc.)
- "Unsupervised learning" (clustering, etc.)

Orari: Il corso si svolge nel secondo semestre

Modalità d'esame

Modalità standard: svolta negli appelli in calendario.

- Risoluzione del 100% degli esercizi assegnati durante il corso, da svolgere singolarmente, da consegnare entro la data dell'esame.
- Prova orale sui contenuti del corso e sulla soluzione di uno o più degli esercizi proposti.

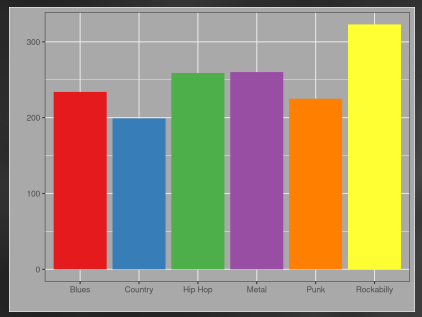
Modalità progetto: solo per studentesse e studenti frequentanti.

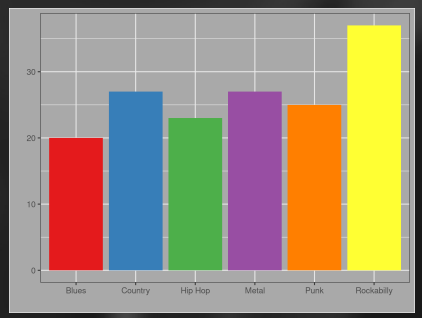
- Risoluzione del 30% degli esercizi assegnati durante il corso, lavorando eventualmente anche in gruppo.
- Relazione sull'analisi di un problema, sulla proposta di una soluzione e sulla sua implementazione.

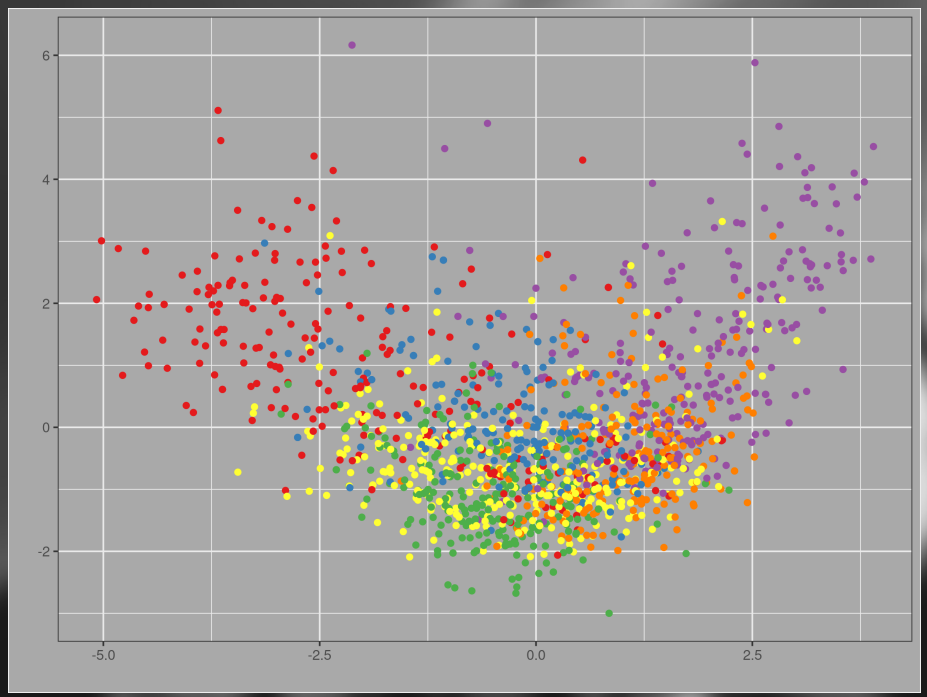
Prova d'esame – 2019/20

Sviluppo di un modello che permetta la classificazione di una traccia musicale in un genere musicale. Sono assegnate 1500 tracce dei generi

`Punk` `Rockabilly` `Country` `Metal` `Hip Hop` `Blues`







Spotify API: `duration_ms`, `key`, `mode`, `time_signature`, `acousticness`, `danceability`, `energy`, `instrumentalness`, `liveness`, `loudness`, `speechiness`, `valence`, `tempo`

Prova d'esame - 2020/21

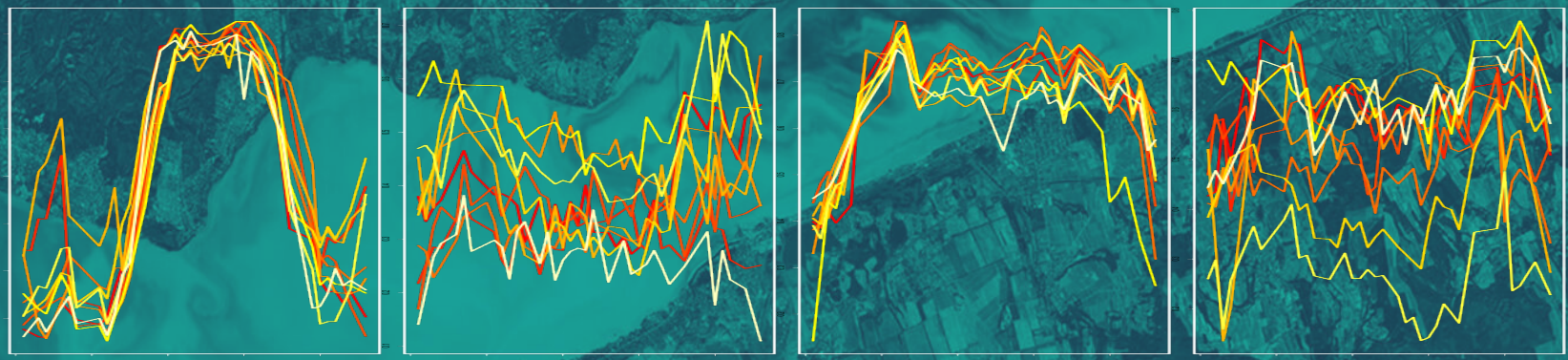
Sviluppo di un modello che permetta di riconoscere il codice MSC di un articolo scientifico basandosi sul titolo e sull'abstract. Sono state assegnate due tabelle, di *training* contenente titolo e abstract di 2837 articoli scientifici di ambito probabilità, e di *validation* contenente 1419 articoli scientifici. Ad esempio

Solving the KPZ equation
Martin Hairer
We introduce a new concept of solution to the KPZ equation which is shown to extend the classical Cole-Hopf solution. This notion provides a factorisation of the Cole-Hopf solution map into a "universal" measurable map from the probability space into an explicitly described auxiliary metric space, composed with a new solution map that has very good continuity properties. The advantage of such a formulation is that it essentially provides a pathwise notion of a solution, together with a very detailed approximation theory. In particular, our construction completely bypasses the Cole-Hopf transform, thus laying the groundwork for proving that the KPZ equation describes the fluctuations of systems in the KPZ universality class.
Comments: 102 pages, many figures
Subjects: Probability (math.PR); Mathematical Physics (math-ph); Analysis of PDEs (math.AP)

MSC classes: 60H15, 35Q82, 60K35

Prova d'esame - 2021/22

Ideazione e implementazione di un metodo di classificazione che permetta di predire, sulla base di una serie temporale corta di un'area geografica, quale sia l'impiego del suolo dell'area.
Le serie storiche sono state acquisite dai dati aperti della missione *Sentinel-2* dell'ESA, progettata per monitorare le aree verdi del pianeta e fornire supporto in caso di disastri naturali.



https://sentinel.esa.int/web/sentinel/missions/sentinel-2

Prova d'esame - 2022/23

L'obiettivo preliminare è quello di analizzare e sviluppare in modo estensivo metodi di classificazione che permettano di fornire una previsione accurata dello stadio istologico della cirrosi biliare di una serie di pazienti, sulla base di dati tabellari (numerici) o eventualmente immagini, riguardanti il loro stato di salute.

Una volta sviluppati metodi soddisfacenti di classificazione, il compito principale è di esplorare possibili metodi per produrre misure o stime sull'incertezza nella classificazione.

In effetti una sfida importante per l'implementazione di sistemi automatici in applicazioni critiche (come la diagnostica medica, i veicoli a guida autonoma, ecc.) è in che misura il sistema può fallire una previsione. Dopo tutto, la valutazione dell'incertezza è incorporata nel nostro comportamento. Ad esempio, un guidatore umano rallenta, o un medico richiede approfondimenti nel caso di incertezze significative.

`https://www.kaggle.com/datasets/fedesoriano/cirrhosis-prediction-dataset`
`https://www.kaggle.com/datasets/pranavraikokte/covid19-image-dataset`