Homework II

ABSTRACT

L'oggetto di studio è il dataset "Playlist", creato da noi utilizzando la piattaforma "Sort your music" (http://sortyourmusic.playlistmachinery.com/), che utilizza le API di Spotify per estrarre dati relativi alle canzoni presenti sulla piattaforma. L'obiettivo della nostra analisi è duplice:

- Metodo non supervisionato: raggruppare brani con caratteristiche simili.
  Poiché in realtà conosciamo anche i veri gruppi, avremo anche la possibilità di verificare la corrispondenza con la reale suddivisione.
- Metodo supervisionato: allenare un classificatore in grado di prevedere il cluster di appartenenza per nuove canzoni.

IL DATASET

Il dataset è composto da 555 osservazioni provenienti da tre playlist, ognuna caratterizzata da un diverso genere musicale: trap, rock e ballads. Sono presenti le seguenti 15 variabili:

ID	Numero progressivo che identifica la canzone all'interno delle diverse playlist
Title	Titolo della canzone
Artist	Nome dell'artista
Release	Data di pubblicazione della canzone
Bpm	"Beats per minute", indica l'andamento e la velocità del tempo del brano
Energy	Energia della canzone; compresa tra 1 e 100
Dance	"Danceability" della canzone; compresa tra 1 e 100
Loud	Volume della canzone in LUFS (unità di misura del volume audio)
Valence	Capacità della canzone di creare un'atmosfera positiva; compresa tra 1 e 100
Length	Durata del brano nel formato "MM:SS"
Acoustic	Acustica della canzone; compresa tra 1 e 100. Maggiore è il valore, maggiore è l'utilizzo di strumenti acustici, anziché elettrici o elettronici
Pop	Popolarità del brano in base al numero di riproduzioni; compresa tra 1 e 100.
Asep	Massimizza la separazione degli artisti nel set
Rnd	Numero casuale, viene utilizzato per poter riprodurre casualmente i brani
Playlist	Playlist originale di appartenenza – etichetta

ANALISI PRELIMINARI E ESPLORATIVE

Per prima cosa abbiamo eliminato le variabili considerate irrilevanti agli scopi delle nostre analisi ("ID", "release", "asep" e "rnd") e trasformato il formato della variabile "length" da "MM:SS" a intero indicante la lunghezza della canzone in secondi.
Osservando il barplot riportante le frequenze assolute delle modalità della variabile "playlist", abbiamo verificato che la risposta è bilanciata. Abbiamo inoltre notato che sono presenti tre coppie di brani con lo stesso titolo ma di artisti diversi: abbiamo quindi deciso di unire titolo e artista in un'unica variabile al fine di ottenere una chiave univoca.
Dopo aver escluso la presenza di missing values, abbiamo calcolato la correlazione tra le variabili quantitative e escluso la presenza di multicollinearità tra gli input. Abbiamo dunque ritenuto opportuno mantenere tutte le variabili.

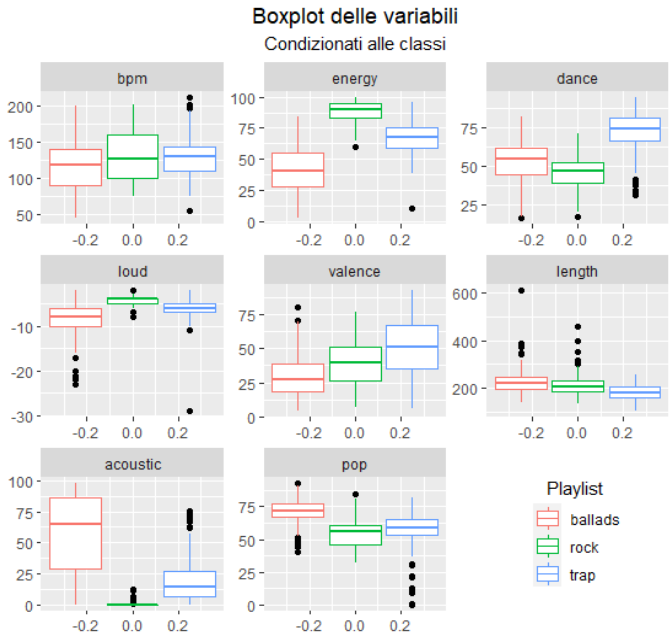

Figura 1: Boxplot di tutte le variabili condizionati alle classi di appartenenza

Osservando i boxplot condizionati alle classi di ogni variabile (Figura 1), abbiamo notato che quelle con maggiore potere discriminante rispetto alla playlist di appartenenza sembrano essere “energy”, “dance”, “valence” e “acoustic”. Dagli istogrammi e dalle curve di densità non parametrica (Figura 2) notiamo inoltre che “bpm”, “energy”, “dance” e “valence” sono quelle che si avvicinano maggiormente ad un andamento multimodale.

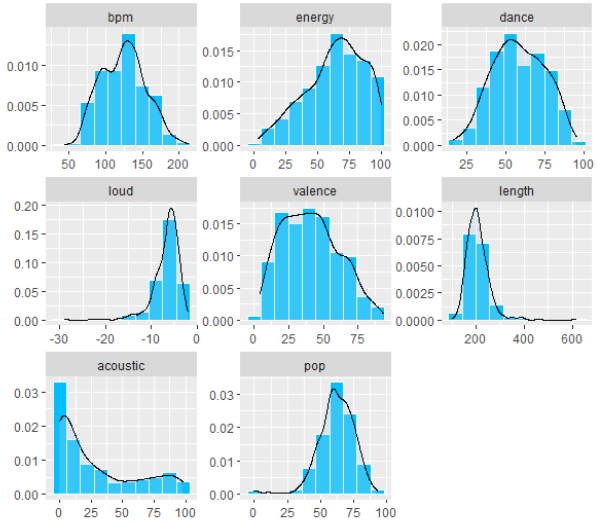

Figura 2: Istogrammi e curve di densità non parametrica delle variabili

Da un’attenta osservazione degli scatterplot tra tutte le possibili coppie di variabili e delle stime di densità non parametrica, entrambi colorati e distinti per classe di appartenenza, abbiamo inoltre notato che le classi non sembrano essere sempre ben “separate” e abbiamo avuto nuovamente conferma di quali siano le variabili con maggiore potere discriminante (ricordando comunque sempre che si tratta di una visualizzazione ristretta a due dimensioni e quindi marginale).
Abbiamo inoltre notato che le variabili “loud”, “length” e “acoustic” sono asimmetriche e per questo abbiamo optato per una trasformazione logaritmica. Dal momento che “loud” assume

valori negativi, ne abbiamo invertito il segno, ricordando di interpretarla in seguito in senso opposto.
Infine abbiamo standardizzato le variabili al fine di uniformare la variabilità senza modificare la distribuzione delle stesse. In particolare, per il dataset per il clustering abbiamo utilizzato medie e standard deviations calcolate su tutte le osservazioni, mentre nella classificazione abbiamo utilizzato le medie e le sd del training set per standardizzare sia training che test.
Tenendo presente che le variabili in esame non sembrano seguire una distribuzione normale condizionatamente alla classe di appartenenza, abbiamo comunque deciso di applicare i metodi, consapevoli dell'approssimazione.

MODEL-BASED CLUSTERING

In una prima fase abbiamo "finto" di non conoscere nulla sulla variabile "playlist" e abbiamo clusterizzato le osservazioni senza imporre restrizioni sul numero di cluster. Abbiamo quindi applicato sia la funzione *Mclust*, che considera solo la bontà di adattamento del modello ai dati tramite BIC, sia la funzione *MclustICL*, che invece tiene conto anche della bontà dei cluster tramite ICL.
Nel caso del BIC il clustering ottimale è VEE (volume variabile, stessa forma e orientamento) con 10 gruppi. Il numero di gruppi è elevato e il modello, che richiede la stima di 134 parametri, è eccessivamente complesso. Anche nel caso dell'ICL i tre modelli migliori sono VEE, con numero di gruppi pari 8, 10 e 12.
In entrambi i casi, il trade-off tra complessità del modello e numero dei cluster porta ad individuare un modello abbastanza semplice, ma che necessita di un numero molto elevato di cluster per adattarsi alla realtà, cadendo nell'overfitting (con tutti i problemi annessi).
Concentrandoci sul modello identificato secondo BIC, abbiamo rappresentato la classificazione e l'incertezza, ottenendo, come ci saremmo aspettati, grafici molto disordinati e cluster poco sensati/interpretabili (Figura 2).

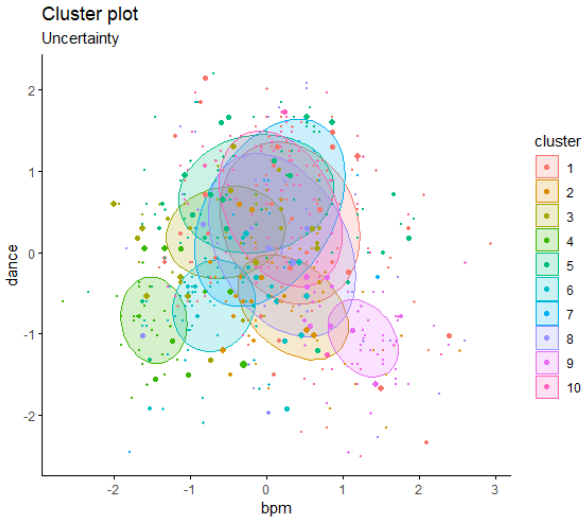

Figura 3: Uncertainty plot con le variabili "dance" e "bpm"

Abbiamo comunque proseguito valutando la bontà di questo primo tentativo di clusterizzazione, tenendo sempre conto del fatto che l'overfitting determina una distorsione dei risultati. L'entropia relativa, pari a 0.11, sembrerebbe indicare un buon raggruppamento. Il valore di R^2 è invece pari a 0.43 se calcolato con la traccia e 0.99 se calcolato con il determinante: questo può essere dovuto al fatto che nel secondo caso raccogliamo più informazione. Infine abbiamo calcolato l'incertezza (un valore per ogni brano), notando che

la probabilità di allocare l'unità ad un cluster non corretto è molto elevata per un numero significativo di osservazioni, e la divergenza KL simmetrizzata (un valore per ogni coppia di cluster), notando che la distanza tra molte coppie di cluster è abbastanza ridotta.
Abbiamo pertanto concluso che il modello VEE con 10 gruppi non produce un buon raggruppamento dei brani.
In una seconda fase, abbiamo deciso di sfruttare la nostra conoscenza sul numero di classi e di imporre quindi il numero dei cluster pari a 3, ottenendo come modello migliore, sia secondo BIC che secondo ICL, EEV (stesso volume e forma, diverso orientamento), il quale prevede la stima di 118 parametri.
Dalla Figura 3 emergono gruppi abbastanza definiti, considerando anche che stiamo visualizzando soltanto 2 variabili alla volta sulle 8 a disposizione nell'analisi. Nella figura possiamo vedere sia quali brani sono stati classificati correttamente sia l'incertezza ad essi relativa, che rimane comunque alta per alcuni brani.

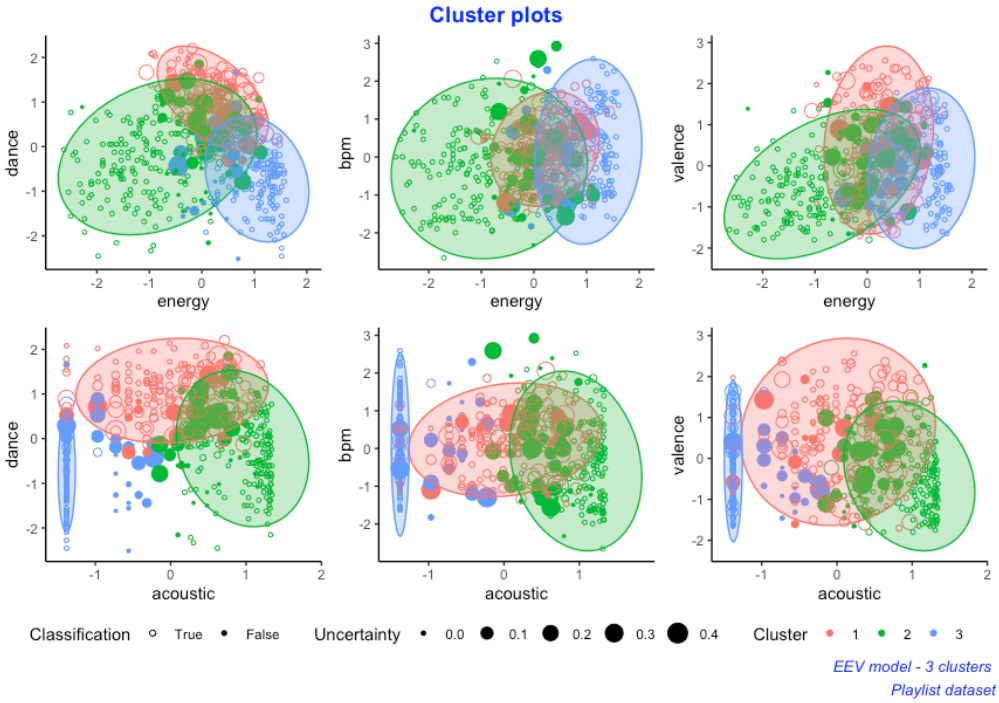

Figura 4: Classificazione, incertezza e cluster dei brani

Per cercare di aumentare la dimensionalità e dare uno sguardo più completo, abbiamo rappresentato anche degli scatterplot interattivi in 3D (Figura 4).

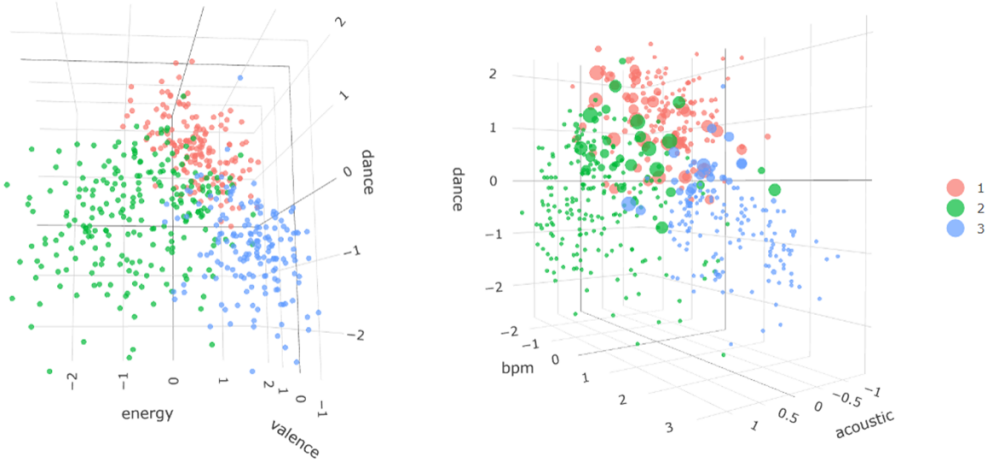

Figura 5: Scatterplot 3D della classificazione (sinistra) e dell'incertezza (destra) considerando diverse variabili

La scelta delle variabili per effettuare i grafici è stata dettata dalle osservazioni effettuate nella prima fase di analisi esplorative, che ci avevano permesso di individuare quelle a maggiore potere discriminante rispetto alle classi.
A differenza delle normali situazioni in cui si sfruttano modelli non supervisionati, in questo caso abbiamo anche a disposizione le vere labels: possiamo quindi valutare tramite CER (Classification Error Rate) e ARI (Adjusted Rand Index) la bontà della classificazione:
- le osservazioni mal classificate sono 87 su un totale di 555: la percentuale di brani erroneamente classificati è pertanto pari al 16%
- la similarità tra le due partizioni è invece pari a 0.58

Per verificare che l'accuratezza espressa dal CER fosse quella corretta, abbiamo testato manualmente, permutando l'ordine dei factor nella classificazione e testando tutte le possibili combinazioni.
Entrambi i risultati indicano che il clustering sia buono. Nonostante siano meno indicativi rispetto a CER e ARI, abbiamo calcolato anche in questo caso entropia relativa, R^2 e divergenza KL simmetrizzata, ottenendo risultati soddisfacenti. Da quest'ultimo indicatore abbiamo ricavato che i cluster delle classi *trap* e *rock* sono i più vicini, mentre quelli di *ballads* e *rock* sono i più distanti.
Osservando le stime dei parametri delle componenti della mistura, e in particolare le medie delle nostre variabili, possiamo facilmente dare un'interpretazione dei cluster ottenuti.
In linea con le nostre aspettative, dal momento che *ballads* sono canzoni tipicamente lente e rilassanti, i brani con valore medio più elevato di "bpm" e "energy" sono *trap* e *rap*. Per quanto riguarda "dance", i brani *trap* sono quelli che presentano valori mediamente maggiori, mentre *rock* quelli minori; ci saremmo aspettati un valore più alto per *ballads*, mentre quello ottenuto per le canzoni *rock* è più coerente con le nostre aspettative. La variabile "loud" è da interpretare con segno opposto, quindi i risultati ottenuti sono sensati: valori elevati per canzoni *rock* e bassi per *ballads*. Come ci aspettavamo, *ballads* presenta valori di "valence" (che indica la capacità di trasmettere un'atmosfera positiva) mediamente bassi. I valori di "length" confermano invece che le canzoni *trap* sono mediamente più brevi rispetto a quelle di altri generi. Le *ballads* usano principalmente strumenti acustici, a differenza dei brani *rock*. La variabile "pop" mostra invece le preferenze del pubblico in questo momento, orientate verso il genere delle *ballads*.
A titolo esemplificativo, abbiamo selezionato alcuni dei brani misclassificati per cercare di capirne il motivo. Ad esempio:
- "VVS – Capo Plaza ft. Gunna" è stata classificata come *rock* invece che come *trap*: una delle possibili cause marginali può essere il suo valore di "acoustic" molto più vicino alla media dei brani *rock*.
- "Oioioi – Edo Fendi" è stata classificata come *ballads* invece che come *trap*: una delle possibili cause marginali può essere il suo valore di "energy" molto più vicino alla media delle *ballads*.

MODEL-BASED CLASSIFICATION

Abbiamo proseguito il lavoro applicando tecniche di classificazione.
Abbiamo innanzitutto suddiviso le osservazioni in training (80%) e test set (20%) e applicato, come specificato in precedenza, la trasformazione logaritmica e la standardizzazione.
Abbiamo quindi scelto due diversi modelli per la classificazione.
In un primo momento abbiamo utilizzato il modello EDDA. Nello specifico, abbiamo allenato il classificatore attarverso la funzione *mixmodLearn* con pesi variabili, essendo in ambito di mixture sampling. Abbiamo svolto l'operazione sul training set ripetutamente tramite un ciclo,

dal momento che la suddivisione delle unità per la cross validation avviene in modo casuale e per uno stesso modello restituisce valori del MER (Misclassification Error Rate) sempre diversi. Il modello scelto è “Gaussian pk_L_Bk”, ovvero un modello EVI, con uguale volume, forma variabile e assi allineati agli assi cartesiani. In Figura 5 l’andamento del MER per i vari modelli fino all’ottenimento di quello ottimale.

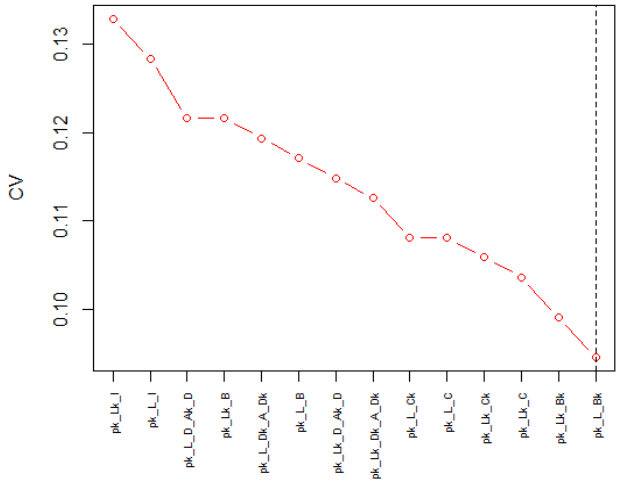

Figura 6: Andamento del MER per i vari modelli fino all’ottenimento di quello ottimale

Abbiamo infine effettuato l’assegnazione delle labels per i brani del test set. Conoscendo le vere etichette, abbiamo potuto calcolare la percentuale di unità misclassificate, pari a 0.117. Nonostante i risultati ottenuti possano essere considerati soddisfacenti, abbiamo comunque deciso di proseguire applicando anche il modello MDA, utile nel caso in cui le componenti non siano normali, ma ognuna a sua volta una mistura. Il modello ottenuto è composto da:

- Per ballads, modello VVE (volume e forma variabili, stesso orientamento) con 2 componenti;
- Per rock, modello EEI (volume e forma costanti, orientamento degli assi parallelo agli assi cartesiani) con 3 componenti;
- Per trap, modello VVE (volume e forma variabili, stesso orientamento) con 2 componenti.

Estraendo le informazioni relative alla mistura di ogni classe, notiamo che la loro suddivisione in componenti sembra essere sensata. Anche in questo caso abbiamo allenato il modello, assegnato le classi alle unità del test e calcolato il MER, che in questo caso assume valore 0.108.