Fondamenti di Analisi Dati e Laboratorio

Terza Prova in Itinere: Unsupervised Learning

1. Obiettivo della prova

L'obiettivo di questo terzo assignment è **arricchire** l'analisi svolta nelle prove precedenti integrando tecniche di **Apprendimento Non Supervisionato (Unsupervised Learning)**.

Non si tratta di ricominciare da capo, ma di utilizzare questi nuovi strumenti per migliorare la comprensione dei dati, visualizzare strutture complesse o raffinare i modelli predittivi già creati. Gli studenti dovranno valutare quali tecniche (tra PCA, Clustering e Stima della Densità) siano **opportune** per il proprio dataset e applicarle laddove possano portare valore aggiunto, ad esempio per semplificare i dati, scoprire nuovi gruppi o visualizzare meglio le distribuzioni.

2. Tecniche ed esempi di uso

Gli studenti dovranno aggiungere al notebook esistente una nuova sezione dedicata a queste analisi, motivando la scelta delle tecniche proposte.

Riduzione della Dimensionalità e Visualizzazione (PCA)

La Principal Component Analysis (PCA) può essere utilizzata con due scopi principali: **ridurre la complessità** (se ci sono molte variabili correlate) o **visualizzare** i dati in uno spazio 2D/3D.

Spunti operativi (applicare se pertinente):

1. **Visualizzazione:** Se il dataset ha molte dimensioni, proiettare i dati sulle prime 2 o 3 Componenti Principali.
    - *Insight:* Colorare i punti nel grafico a dispersione utilizzando la variabile target (classi) usata nella prova precedente ove opportuno. Le classi si separano visivamente nello spazio ridotto? Questo aiuta a capire se il task di classificazione è "facile" o "difficile" geometricamente.
2. **Riduzione del Rumore:** Se nella fase di modellazione (Prova 2) avete riscontrato overfitting o tempi di calcolo lunghi, provate a selezionare un numero di componenti che spieghi una buona percentuale di varianza (es. 90%) e usate queste componenti trasformate come input per i vostri modelli precedenti. Migliorano le performance o la stabilità?
3. **Interpretazione:** Analizzare i "loadings" (i pesi delle variabili originali sulle componenti) per capire quali variabili originali "pesano" di più nella struttura latente dei dati.

Riorganizzazione dei Dati tramite Clustering

Il Clustering permette di trovare raggruppamenti "naturali" nei dati senza usare etichette predefinite. Questo è utile per confermare le vostre classi o per scoprire segmenti completamente nuovi.

Spunti operativi (applicare se pertinente):

1. **Segmentazione Esplorativa:** Applicare un algoritmo (es. K-Means, Clustering Gerarchico) per dividere i dati in gruppi omogenei.

    - *Analisi:* I gruppi trovati hanno senso logico? (Es. "Gruppo A: Clienti alto-spendenti", "Gruppo B: Clienti occasionali").
2. **Confronto con le Classi Note:** Se il vostro problema originale era di classificazione, confrontate i cluster trovati dall'algoritmo (Unsupervised) con le classi reali (Supervised). C'è sovrapposizione?
    - *Esempio:* Se state classificando "Tipi di Vino", il clustering riesce a ritrovare le varietà di uva basandosi solo sulle proprietà chimiche, o fa confusione?
3. **Preprocessing per Classificazione:** In alcuni casi, l'etichetta del cluster può diventare una *nuova feature* (variabile in ingresso) per i modelli di classificazione della Fase 2, arricchendo il dataset.

Stima della Densità (Density Estimation)

Questa tecnica è utile per visualizzare la distribuzione dei dati in modo più fluido rispetto agli istogrammi e per identificare anomalie.

Spunti operativi:

1. **Visualizzazione Avanzata:** Utilizzare la *Kernel Density Estimation (KDE)* per visualizzare la distribuzione di variabili chiave, specialmente se multimodali (con più "picchi").
2. **Anomaly Detection (Opzionale):** Osservare le aree a "bassa densità". Ci sono punti isolati (outlier) che i boxplot classici non avevano evidenziato? Potrebbero essere errori nel dataset o casi di studio interessanti.

3. Conclusioni del Progetto

Il notebook deve concludersi con una sintesi che unisca i puntini di tutte e tre le prove.

Il gruppo procederà dunque a redarre le slide per la presentazione finale come da indicazioni condivise.