Elaborazioni di immagini biomediche

Tesina corso di Data Mining

SAPIENZA
UNIVERSITÀ DI ROMA

Simmaco Di Lillo
Tommaso Tenna

Università degli Studi di Roma "La Sapienza"

Roma, 19/07/2022

Introduzione

Lo studio è stato svolto in collaborazione con il Dipartimento di Matematica dell'Università di Genova, sulla base di dati su pazienti con meningioma forniti dall'IRCCS Ospedale Policlinico San Martino di Genova.

I pazienti sono suddivisi secondo due *labels* :

- La **stadiazione** (o grado del tumore) descrive l'estensione rispetto alla sede originale di sviluppo;
- **PD-L1** (*programmed death-ligand 1*) è una proteina che inibisce i linfociti-T, cellule del sistema immunitario deputate alla difesa.

Struttura del progetto

1. Trattamento di immagini biomediche
2. PCA
3. Metodi di clusterizzazione
4. Alcune definizioni e strumenti
5. Analisi dei dati sperimentali
6. Bibliografia

Trattamento di immagini biomediche

Tecniche di filtraggio

Per il trattamento delle immagini MRI all'interno di questo progetto, utilizzeremo tecniche di *mask processing*.
Il filtro è definito da una cosiddetta "maschera", ovvero una matrice di dimensioni fissate che premoltiplica la matrice dell'immagine.
Una tipica maschera 3×3 è della forma

ω (-1,-1)	ω (-1,0)	ω (-1,1)
ω (0,-1)	ω (0,0)	ω (0,1)
ω (1,-1)	ω (1,0)	ω (1,1)

$$\tilde{I}(x,y) = \sum_{s=-1}^{1} \sum_{t=-1}^{1} \omega(s,t)\, I(x+s, y+t).$$

Filtri lineare

I filtri lineari diminuiscono o eliminano le componenti più intense di un'immagine, come ad esempio i bordi o punti di alta intensità derivanti da errori imputabili al macchinario di acquisizione.

$\frac{1}{9}\times$	1	1	1
	1	1	1
	1	1	1

$\frac{1}{16}\times$	1	2	1
	2	4	2
	1	2	1

$\frac{1}{4}\times$	0	1	0
	1	0	1
	0	1	0

Table: Maschere per il filtraggio lineare.

Filtri mediani

L'azione del filtro mediano su un ogni pixel dell'immagine può essere schematizzata nel modo seguente:

1. Tutti i primi vicini del pixel di riferimento vengono inseriti all'interno di un vettore;

Filtri mediani

L'azione del filtro mediano su un ogni pixel dell'immagine può essere schematizzata nel modo seguente:

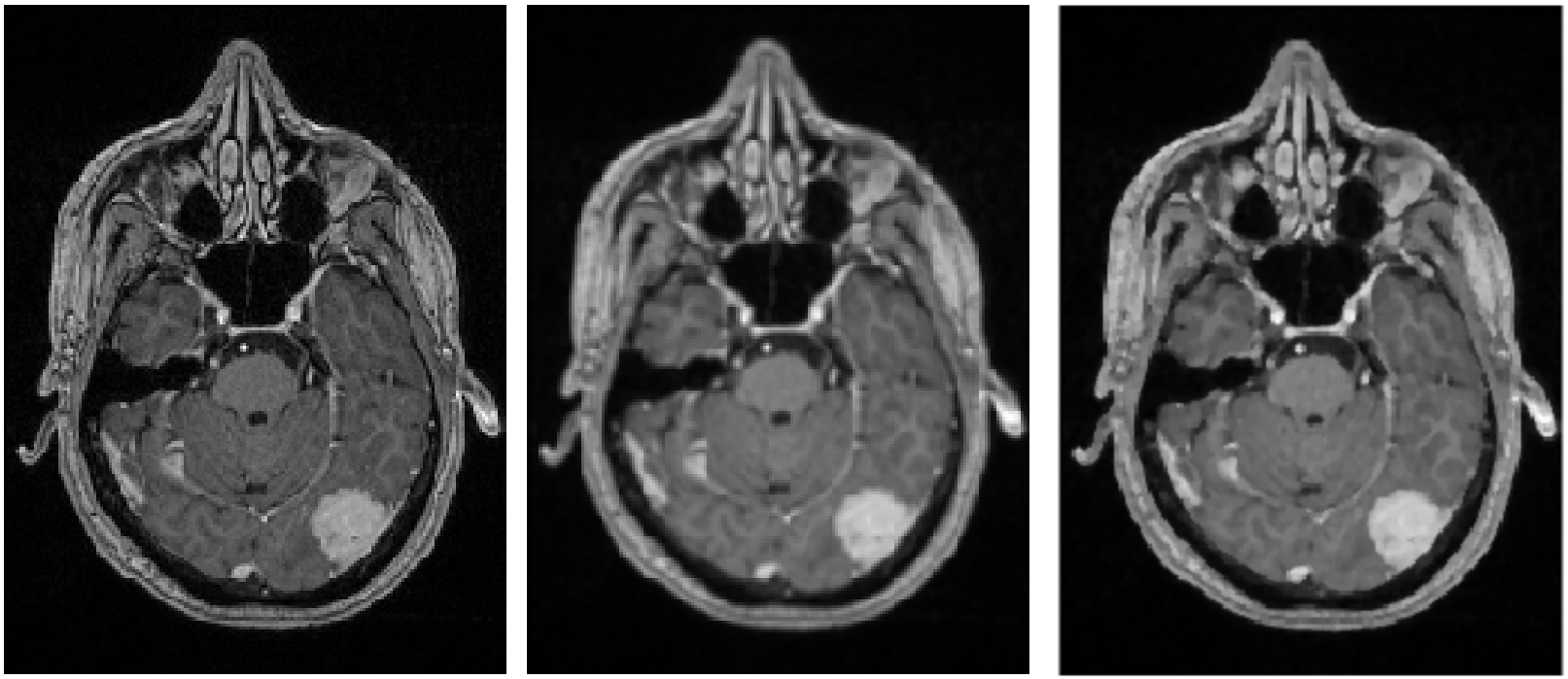
1. Tutti i primi vicini del pixel di riferimento vengono inseriti all'interno di un vettore;
2. Il vettore viene ordinato in ordine crescente (o decrescente);

Filtri mediani

L'azione del filtro mediano su un ogni pixel dell'immagine può essere schematizzata nel modo seguente:

1. Tutti i primi vicini del pixel di riferimento vengono inseriti all'interno di un vettore;
2. Il vettore viene ordinato in ordine crescente (o decrescente);
3. Il valore mediano del vettore ordinato è il nuovo valore assegnato al pixel di riferimento nell'immagine trattata.

Applicazione dei filtri alle immagini biomediche



(a) (b) (c)

Figure: Confronto tra l'immagine originale (a), l'immagine filtrata con un filtro Lineare (b) e l'immagine filtrata con un filtro Mediano (c).

PCA

Analisi della componente principale

Obiettivo: Ridurre la dimensionalità dell'insieme dei dati eliminando la ridondanza di informazioni.

- Si sostituiscono alle variabili originali nuove variabili.
- Le nuove variabili saranno non correlate e ordinate rispetto alla percentuale di variabilità presente nei dati originali.

Definizione

Detta X la matrice camponiaria, diremo che $\{y_1, \ldots, y_p\}$ sono componenti principali se y_i è tra le combinazione lineare del vettore X scorrelate da y_j con $j < i$ quella di varianza massima.

Operativamente

- Si determina y_1 risolvendo

$$\max_{a_1 \in \mathbb{R}^p, ||a_1||=1} Var(Xa_1).$$

Operativamente

- Si determina y_1 risolvendo

$$\max_{a_1 \in \mathbb{R}^p, ||a_1||=1} Var(Xa_1).$$

- Si determina y_2 risolvendo

$$\max_{a_2 \in \mathbb{R}^p, ||a_2||=1} Var(Xa_2) \quad Cov(Xa_1, Xa_2) = 0.$$

Operativamente

- Si determina y_1 risolvendo

$$\max_{a_1 \in \mathbb{R}^p, ||a_1||=1} Var(Xa_1).$$

- Si determina y_2 risolvendo

$$\max_{a_2 \in \mathbb{R}^p, ||a_2||=1} Var(Xa_2) \quad Cov(Xa_1, Xa_2) = 0.$$

- In generale, si determina y_j risolvendo

$$\max_{a_j \in \mathbb{R}^p, ||a_j||=1} Var(Xa_j) \quad Cov(Xa_i, Xa_j) = 0 \quad \forall i < j.$$

Operativamente

Poichè

$$Var(Xa_j) = a_j^T Var(X) a_j.$$

Se $\lambda_1 \geq \cdots \geq \lambda_n$ sono gli autovalori di $Var(X)$ allora

$$y_j = v_j \qquad Var(X)v_j = \lambda_j v_j$$

Quante componenti principali?

- Valutazione grafica

Quante componenti principali?

- Valutazione grafica
- Se i primi k autovalori descrivono almeno l' $80 - 90\%$ della varianza, scelgo k componenti principali

Quante componenti principali?

- Valutazione grafica
- Se i primi k autovalori descrivono almeno l' $80 - 90\%$ della varianza, scelgo k componenti principali
- Considero gli autovalori superiori ad una data soglia

Metodi di clusterizzazione

Metodi di classificazione

Obiettivo: individuare all'interno di un insieme di dati alcuni sottoinsiemi che hanno caratteristiche comuni.

Definizione

Sia $\mathcal{D} = \{x_1, x_2, \ldots x_n\}$ contenente n vettori. Se consideriamo K sottoinsiemi disgiunti dell'insieme $\mathcal{D}$, indicati con $C_1, \ldots C_K$ allora $\mathcal{C} = \{C_1, \ldots C_K\}$ è detto clustering. Ad esso associamo il costo k-means definito da

$$cost(\mathcal{C}) = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

dove μ_i è il centroide di C_i

Algoritmo di Lloyd

Dato $\mathcal{D} = \{x_1, x_2, \ldots x_n\}$ e k

1. Si scelgono i centroidi $z_1, \ldots, z_k \in \mathcal{D}$ casualmente. Si assegna x_i al centroide più vicino determinando una partizione iniziale $\mathcal{C}^{(0)}$.

Algoritmo di Lloyd

Dato $\mathcal{D} = \{x_1, x_2, \ldots x_n\}$ e k

1. Si scelgono i centroidi $z_1, \ldots, z_k \in \mathcal{D}$ casualmente. Si assegna x_i al centroide più vicino determinando una partizione iniziale $\mathcal{C}^{(0)}$.
2. Per ogni i, si assegna x_i al centroide più vicino. Determinando $\mathcal{C}^{(i)}$;

Algoritmo di Lloyd

Dato $\mathcal{D} = \{x_1, x_2, \ldots x_n\}$ e k

1. Si scelgono i centroidi $z_1, \ldots, z_k \in \mathcal{D}$ casualmente. Si assegna x_i al centroide più vicino determinando una partizione iniziale $\mathcal{C}^{(0)}$.
2. Per ogni i, si assegna x_i al centroide più vicino. Determinando $\mathcal{C}^{(i)}$;
3. Si calcolano i centroidi della partizione $\mathcal{C}^{(i)}$.

Algoritmo di Lloyd

Dato $\mathcal{D} = \{x_1, x_2, \ldots x_n\}$ e k

1. Si scelgono i centroidi $z_1, \ldots, z_k \in \mathcal{D}$ casualmente. Si assegna x_i al centroide più vicino determinando una partizione iniziale $\mathcal{C}^{(0)}$.
2. Per ogni i, si assegna x_i al centroide più vicino. Determinando $\mathcal{C}^{(i)}$;
3. Si calcolano i centroidi della partizione $\mathcal{C}^{(i)}$.
4. Si ripetono i passi 2-3 finche nessun punto cambia cluster.

Algoritmo di Lloyd

Svantaggi

- Le clusterizzazioni dipendono dall'inizializzazione
- k deve essere noto a priori
- Funzione male con outlier
- Produce cluster con dimensioni uniformi

Vantaggi.

- Facile implementazione.
- Velocità di convergenza.
- Utilizzabile con ampi set di dati.

Alcune definizioni e strumenti

Specificità e Sensibilità

- La **sensibilità** si indica la capacità intrinseca di un test di screening di individuare nella popolazione di riferimento i soggetti positivi

$$\text{sensibilità} = \frac{\text{veri positivi}}{\text{veri positivi} + \text{falsi positivi}}$$

Specificità e Sensibilità

- La **sensibilità** si indica la capacità intrinseca di un test di screening di individuare nella popolazione di riferimento i soggetti positivi

$$\text{sensibilità} = \frac{\text{veri positivi}}{\text{veri positivi} + \text{falsi positivi}}$$

- La **specificità**, invece, rappresenta la capacità del test di individuare come veri negativi i soggetti “sani".

$$\text{specificità} = \frac{\text{veri negativi}}{\text{veri negativi} + \text{falsi positivi}}$$

Matrici di confusione

Le matrici di confusione contengono informazioni riguardanti il confronto tra la reale classificazione e la classificazione effettuata mediante un metodo di clustering.

		Predicted	
		Negative	Positive
Actual	Negative	**veri negativi**	**falsi positivi**
	Positive	**falsi negativi**	**veri positivi**

Curve ROC ("*Receiver Operating Characteristic Curves*")

L'algoritmo per la determinazione delle curve ROC si struttura nel modo seguente:

1. si considera la feature j e si determina il valore massimo M e il valore minimo m che la *feature* assume;

Curve ROC ("*Receiver Operating Characteristic Curves*")

L'algoritmo per la determinazione delle curve ROC si struttura nel modo seguente:

1. si considera la feature j e si determina il valore massimo M e il valore minimo m che la *feature* assume;
2. si partiziona l'intervallo di valori $[-m, M]$ (ad esempio mediante una partizione uniforme);

Curve ROC ("*Receiver Operating Characteristic Curves*")

L'algoritmo per la determinazione delle curve ROC si struttura nel modo seguente:

1. si considera la feature j e si determina il valore massimo M e il valore minimo m che la *feature* assume;
2. si partiziona l'intervallo di valori $[-m, M]$ (ad esempio mediante una partizione uniforme);
3. si esegue un ciclo sui valori della partizione:

Curve ROC ("*Receiver Operating Characteristic Curves*")

L'algoritmo per la determinazione delle curve ROC si struttura nel modo seguente:

1. si considera la feature j e si determina il valore massimo M e il valore minimo m che la *feature* assume;
2. si partiziona l'intervallo di valori $[-m, M]$ (ad esempio mediante una partizione uniforme);
3. si esegue un ciclo sui valori della partizione:
   - per ogni paziente i, si valuta se la feature j ha un valore maggiore o minore della soglia fissata;

Curve ROC ("*Receiver Operating Characteristic Curves*")

L'algoritmo per la determinazione delle curve ROC si struttura nel modo seguente:

1. si considera la feature j e si determina il valore massimo M e il valore minimo m che la *feature* assume;
2. si partiziona l'intervallo di valori $[-m, M]$ (ad esempio mediante una partizione uniforme);
3. si esegue un ciclo sui valori della partizione:
   - per ogni paziente i, si valuta se la feature j ha un valore maggiore o minore della soglia fissata;
   - si individua il cluster di appartenenza del paziente j.

Curve ROC ("*Receiver Operating Characteristic Curves*")

L'algoritmo per la determinazione delle curve ROC si struttura nel modo seguente:

1. si considera la feature j e si determina il valore massimo M e il valore minimo m che la *feature* assume;
2. si partiziona l'intervallo di valori $[-m, M]$ (ad esempio mediante una partizione uniforme);
3. si esegue un ciclo sui valori della partizione:
   - per ogni paziente i, si valuta se la feature j ha un valore maggiore o minore della soglia fissata;
   - si individua il cluster di appartenenza del paziente j.
4. Si calcolano **veri positivi** e **falsi positivi** sulla base della classificazione reale;

Curve ROC ("*Receiver Operating Characteristic Curves*")

L'algoritmo per la determinazione delle curve ROC si struttura nel modo seguente:

1. si considera la feature j e si determina il valore massimo M e il valore minimo m che la *feature* assume;
2. si partiziona l'intervallo di valori $[-m, M]$ (ad esempio mediante una partizione uniforme);
3. si esegue un ciclo sui valori della partizione:
   - per ogni paziente i, si valuta se la feature j ha un valore maggiore o minore della soglia fissata;
   - si individua il cluster di appartenenza del paziente j.
4. Si calcolano **veri positivi** e **falsi positivi** sulla base della classificazione reale;
5. si rappresenta la curva ROC.

Rappresentazione di una curva ROC

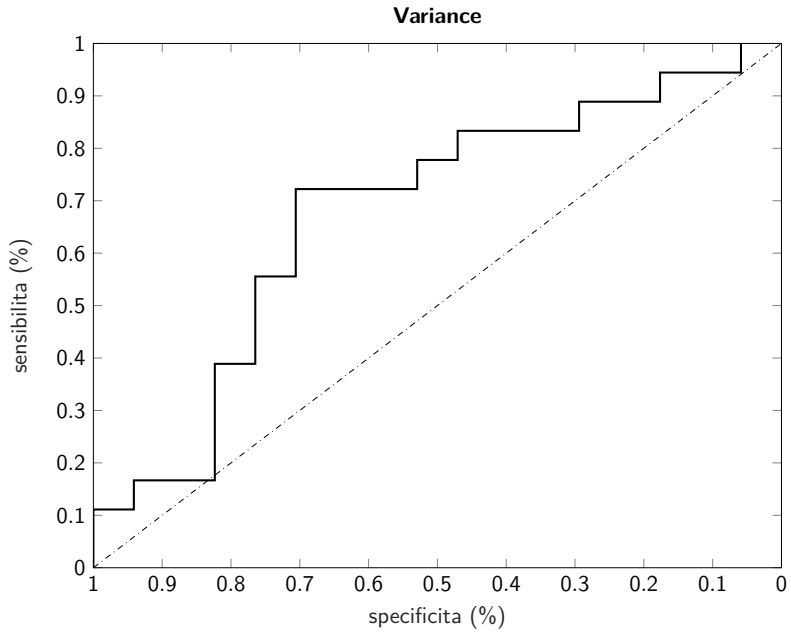


Figure: Rappresentazione di una curva ROC.

Analisi dei dati sperimentali

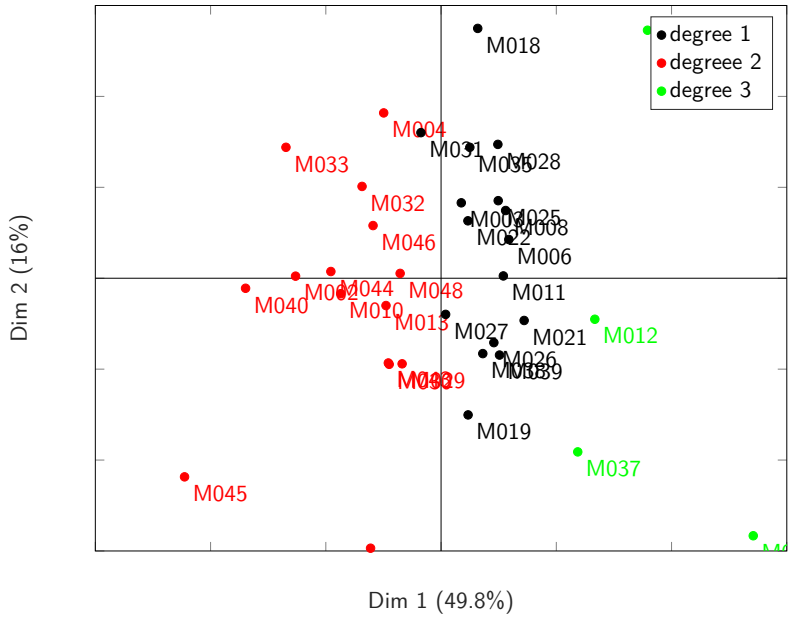
Diversi algoritmi di quantizzazione

Un algoritmo di quantizzazione riscala l'intero range di livelli di grigio della regione del tumore in un numero minore di livelli di grigio N_g (nel nostro caso $N_g = 256$). Il pacchetto di radiomica utilizzato sfrutta tre diversi algoritmi di quantizzazione:

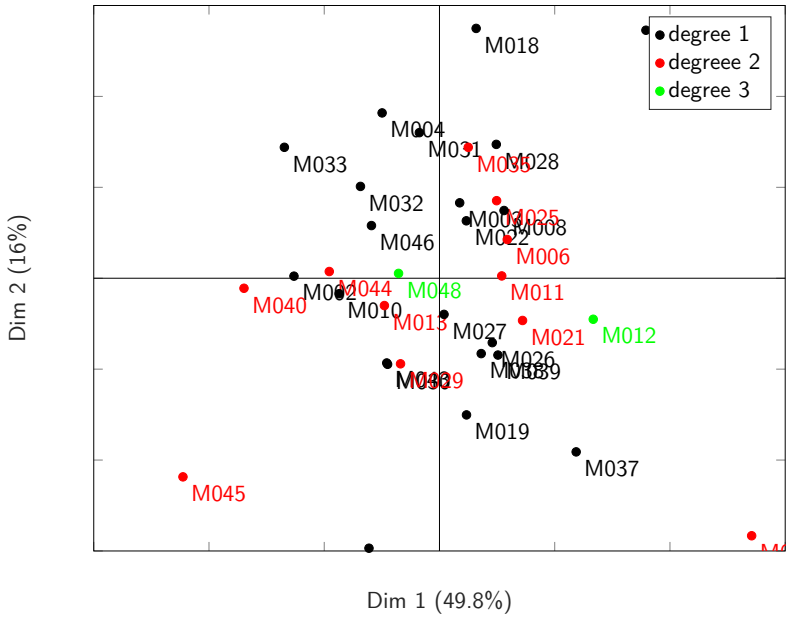
- **Equal-probability**
- **Lloyd-Max**
- **Uniform-probability**

. Per ciascun algoritmo, si ottengono *features* diverse, perciò abbiamo effettuato un confronto dei risultati ottenuti nei tre diversi casi e nel caso in cui non venga utilizzato nessun algoritmo di quantizzazione.

PCA e 3-means su Equal per il grado del tumore

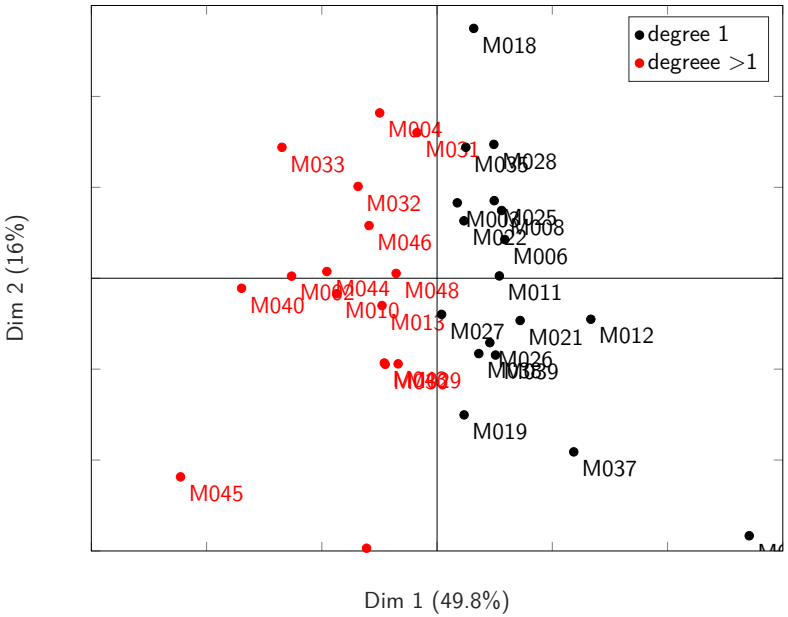


(a) Divisione con 3−means

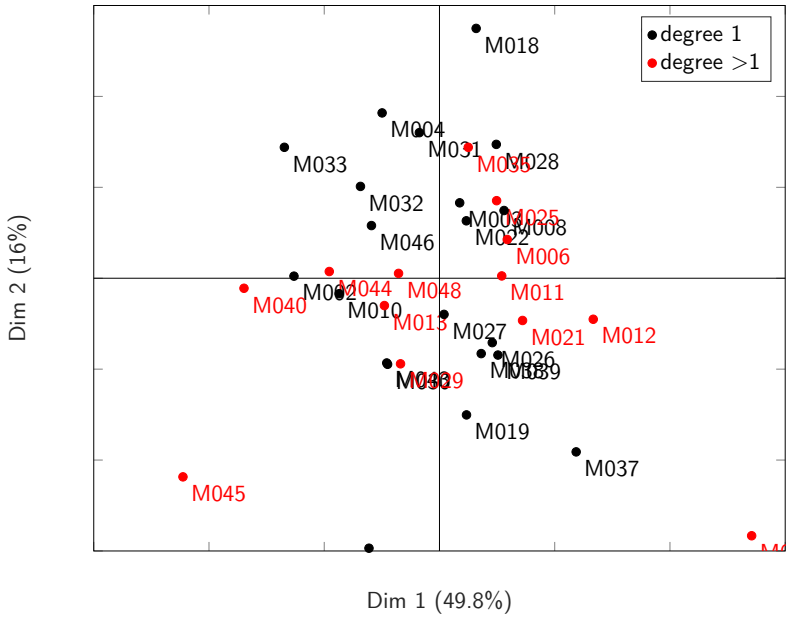


(b) Divisione reale dei pazienti

PCA e 2-means su Equal per il grado del tumore

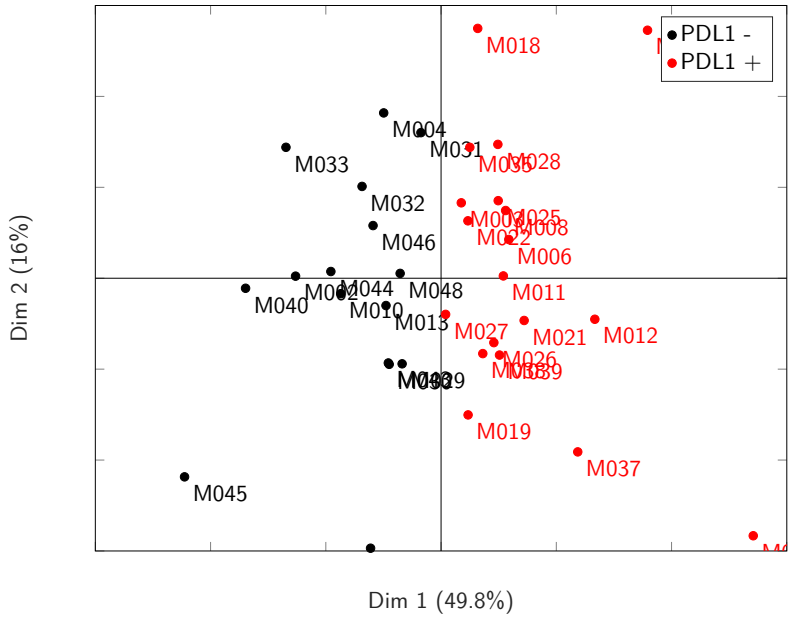


(a) Divisione con 2−means

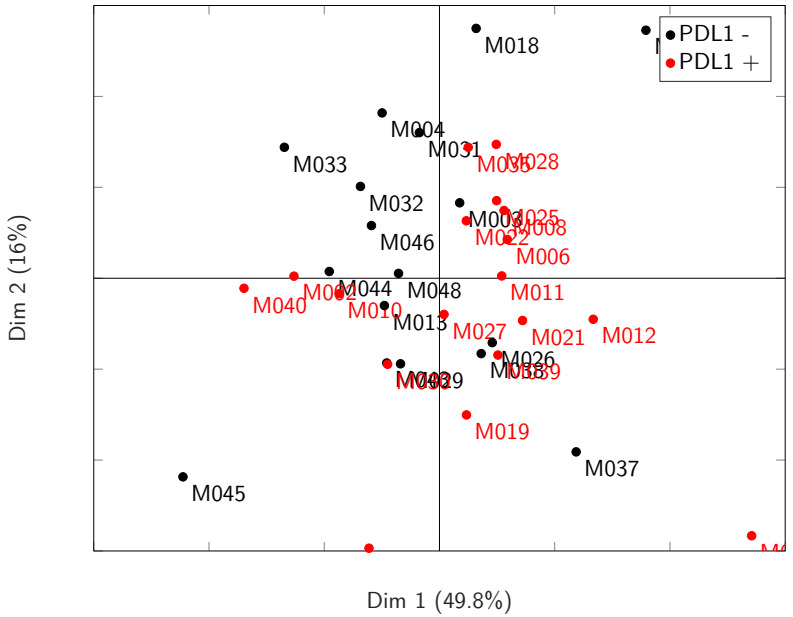


(b) Divisione reale dei pazienti

PCA e 2-means su Equal per PD-L1

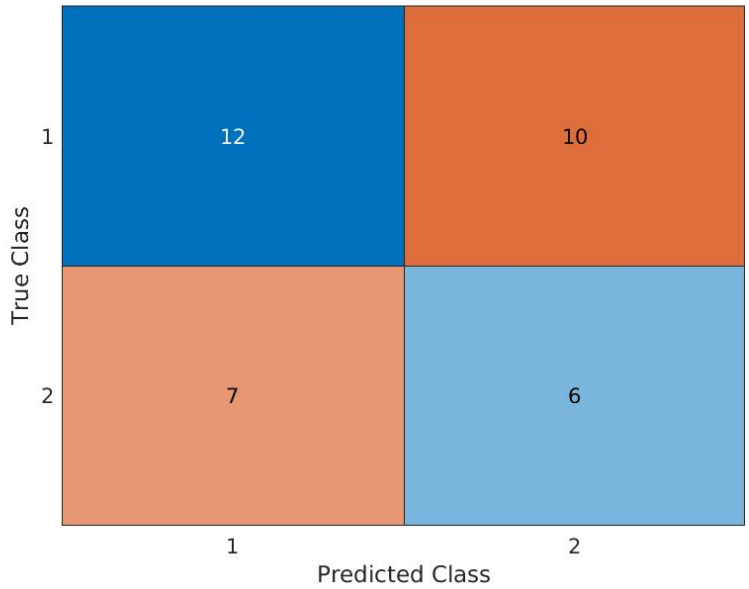


(a) Divisione con 2−means

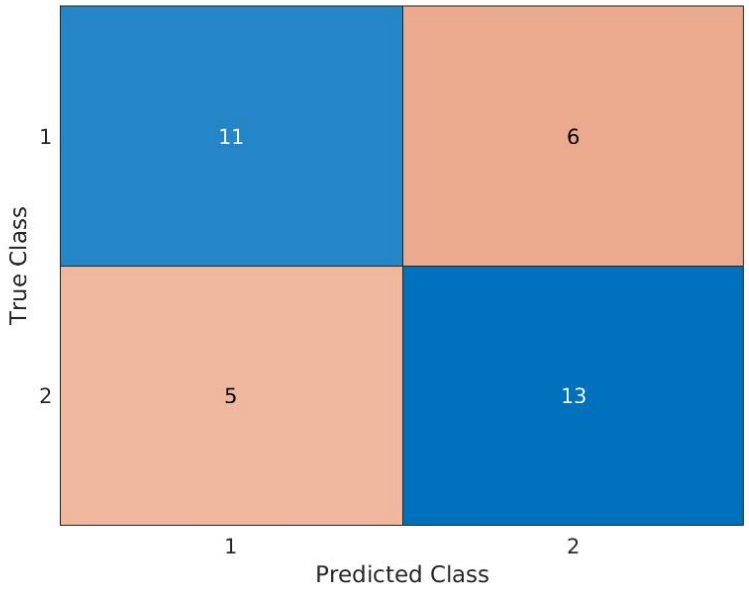


(b) Divisione reale dei pazienti

Matrici di confusione su Equal



(a) Matrice di confusione per il grado di tumore



(b) Matrice di confusione per PD-L1

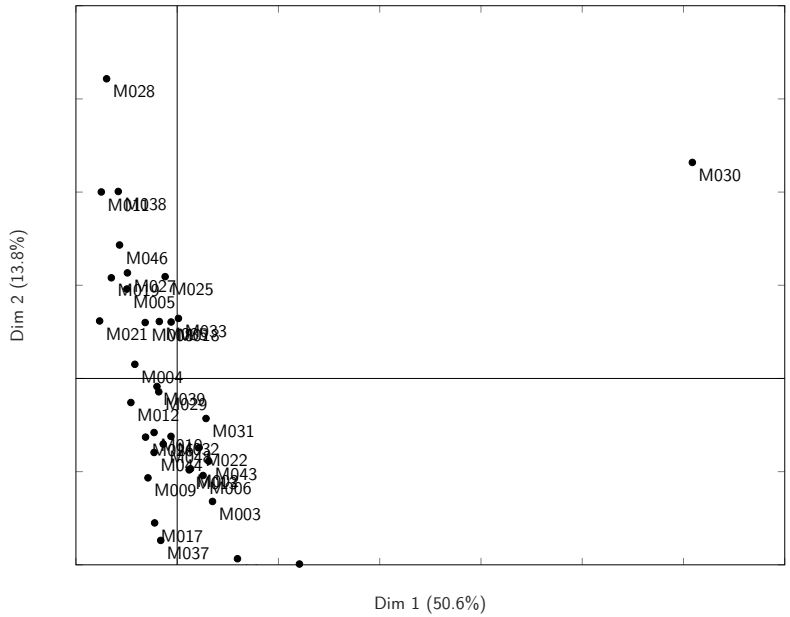
PCA e k-means per le altre quantizzazioni

Le altre quantizzazioni forniscono risultati di classificazione diversi.

	Accuratezza	Specificità	Sensibilità
Equal	68	64	65
Lloyd	65	58	65
Noquant	51	100	100
Uniform	65	58	65

Table: Confronto tra i vari algoritmi di quantizzazione nella classificazione a seconda della positività alla mutazione PD-L1.

Outlier in Noquant



(a) Disposizione dei dati Noquant nello spazio delle prime due componenti principali

PCA e k-means per le altre quantizzazioni

	Accuratezza	Specificità	Sensibilità
Equal	68	64	65
Lloyd	65	58	65
Noquant	58	41	56
Uniform	65	58	65

Table: Confronto tra i vari algoritmi di quantizzazione nella classificazione a seconda della positività alla mutazione PD-L1. Per Noquant il paziente M030 non è considerato

Grazie per l'attenzione.

Tutto il materiale è reperibile sul repository: `https://github.com/simmaco99/Analisi_dati_MRI`

Bibliografia

Bibliografia (1)

[1] Gobert N Lee and Hiroshi Fujita. "K-means clustering for classifying unlabelled MRI data". In: *9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications (DICTA 2007)*. IEEE. 2007, pp. 92–98.

[2] Lars Eldén. *Matrix methods in data mining and pattern recognition*. Vol. 15. Fundamentals of Algorithms. Second edition of [MR2314399]. SIAM, Philadelphia, PA, 2019.

[3] Richard Arnold Johnson, Dean W Wichern, et al. *Applied multivariate statistical analysis*. Vol. 6. Pearson London, UK: 2014.

Bibliografia (2)

[4] Kayvan Najarian and Robert Splinter. *Biomedical signal and image processing*. Taylor & Francis, 2012.

[5] Martin Vallières et al. "A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities". In: *Physics in Medicine & Biology* 60.14 (2015), p. 5471.

[6] Peter A Flach. "ROC analysis". In: *Encyclopedia of machine learning and data mining*. Springer, 2016, pp. 1–8.

[7] Amelia Swift, Roberta Heale, and Alison Twycross. "What are sensitivity and specificity?" In: 23.1 (2020), pp. 2–4. DOI: 10.1136/ebnurs-2019-103225.

Bibliografia (3)

[8] Mahlon D Johnson. "PD-L1 expression in meningiomas". In: *Journal of Clinical Neuroscience* 57 (2018), pp. 149–151.

[9] Andrew Ng. *The k−means clustering algorithm*. University Lecture Notes. 2021.

[10] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8 (2006), pp. 861–874.