Università degli Studi di Padova

Machine Learning
2024/2025

AMCO
Artificial Intelligence, Machine Learning and Control Research Group

Lecture #14
Recap + Theory Exam Example

Gian Antonio Susto

Exam – theoretic/numeric exercise part

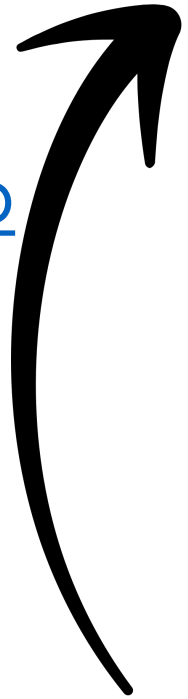
A 45-60 minutes exam, multiple choices (main reference: slides)

Exam – theoretic/numeric exercise part

A 45-60 minutes exam, multiple choices (main reference: slides)

➔ Let’s start with a simulation!

https://forms.gle/Bjgar3TRfy8ydbgA7



Domanda #01: mean, median, variance, mode

Qual è la mediana del vettore?

✅ Risposta corretta: A. 5
(V ordinato: [1, 3, 5, 5, 5, 7, 9] ⇒ valore centrale = 5)

[Domanda 02] Quali attività fanno parte di un preprocessing robusto? 1 points

- (●) Normalizzazione/standardizzazione, gestione valori mancanti, encoding delle variabili categoriche, gestione outlier ✓
- ( ) Costruzione del modello, test di accuratezza, cross-validazione
- ( ) Addestramento, ottimizzazione iperparametri, fine-tuning
- ( ) Plotting, valutazione bias/variance, calcolo della varianza

[Domanda 03] In quale ordine logico andrebbero effettuate le operazioni seguenti?

1. Pulizia dati

2. Split in train/test

3. Feature engineering

4. Analisi esplorativa

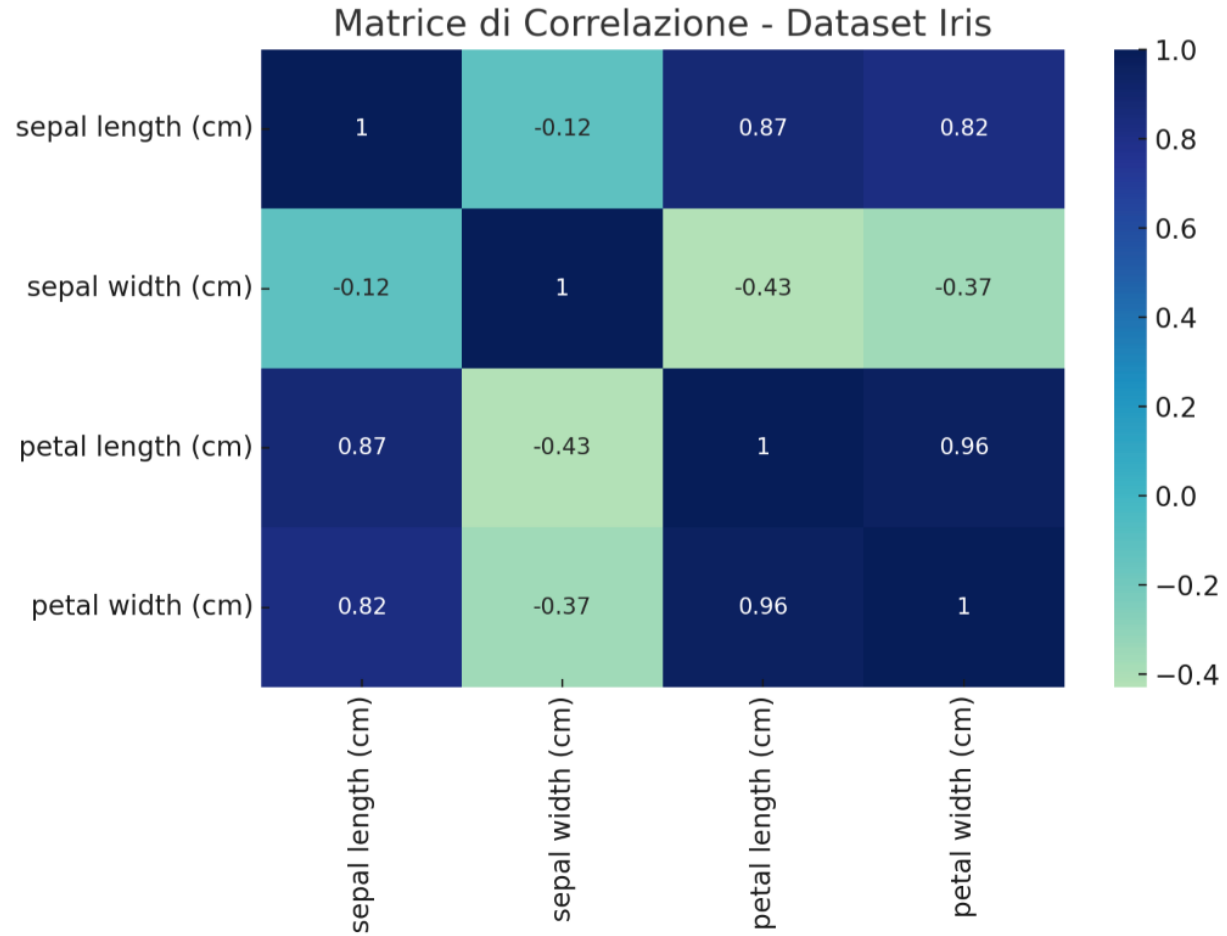
5. Riduzione dimensionalità

1 points

- ○ 1 → 4 → 3 → 2 → 5
- ◉ 4 → 1 → 3 → 5 → 2 ✓
- ○ 1 → 2 → 4 → 3 → 5
- ○ 4 → 1 → 2 → 5 → 3

[Domanda 04] Si consideri questa matrice di correlazione. Che impatto potrebbe avere sugli step successivi di sviluppo?

1 points

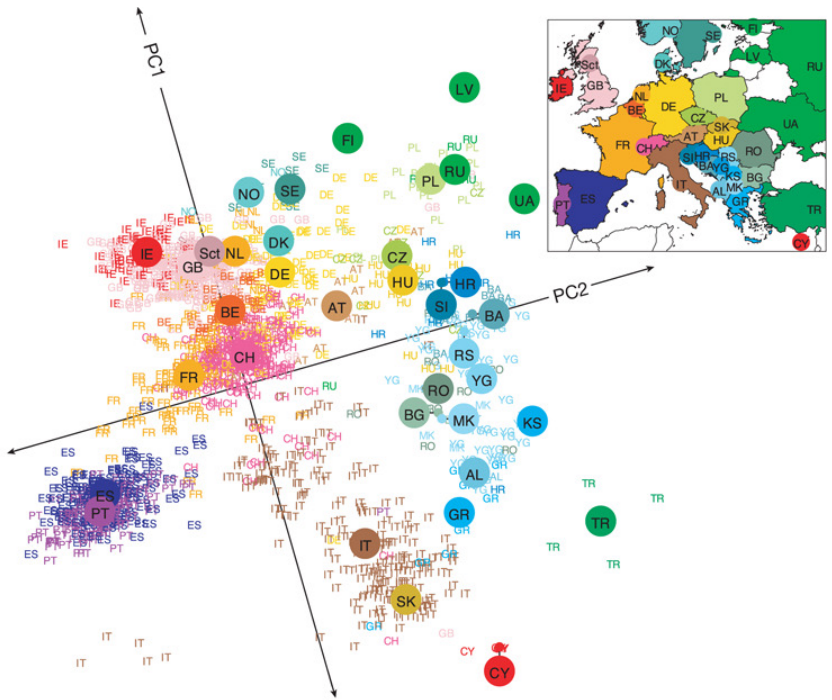


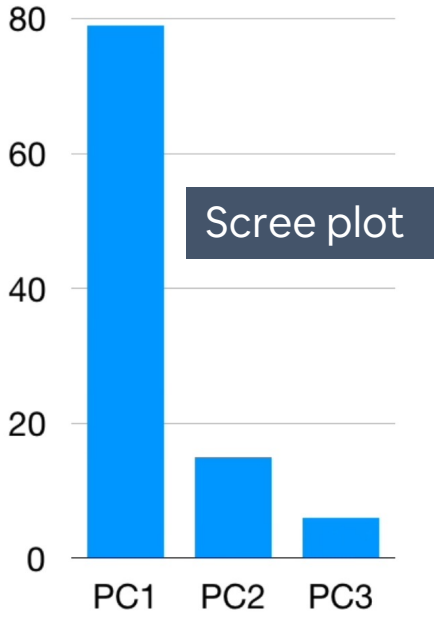
- (●) Impatto sulla feature selection ✓
- ( ) Considerazioni sullo sbilanciamento del dataset
- ( ) Rimozione outlier

Choose correct answers:

[Domanda 05] Per cosa può rivelarsi utile la PCA 1 points

- [x] Ridurre la dimensionalità dei dati ✓
- [x] Decorrelare le variabili ✓
- [x] Capire la struttura dei dati ✓
- [ ] Fare - con immediatezza - feature selection





$X_c = X - \mu$ Dati 'centrati'

$\Sigma = \frac{1}{n} X_c^\top X_c \quad \in \mathbb{R}^{p \times p}$ Matrice di covarianza

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$

$v_1, v_2, \ldots, v_p$ Autovalori e corrispondenti autovettori

$W_k = [v_1, v_2, \ldots, v_k] \in \mathbb{R}^{p \times k}$

Prime k componenti principali

$Z = X_c W_k \quad \in \mathbb{R}^{n \times k}$

Proiezione dati originali

Domanda #06: PCA

Input: matrice X (n x d), numero componenti k

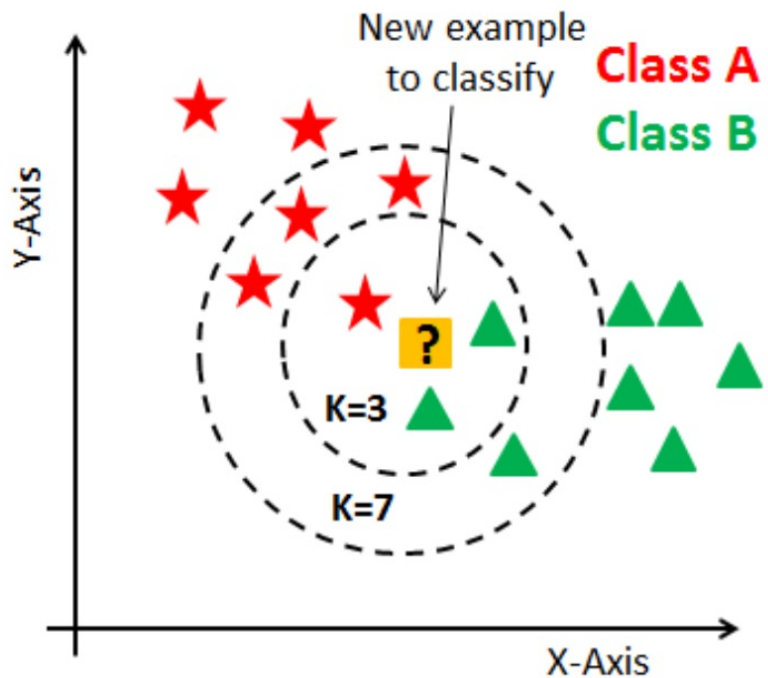
1. Centrare X sottraendo la media da ogni colonna
2. Calcolare la matrice di covarianza: $C = (1/n) * X^t \cdot X$
3. Calcolare autovalori e autovettori di C
4. Ordinare gli autovettori in base agli autovalori decrescenti
5. Selezionare i primi k autovettori → matrice W (d x k)
6. Proiettare i dati nello spazio ridotto: $X\_proj = X \cdot W$

Output: X_proj

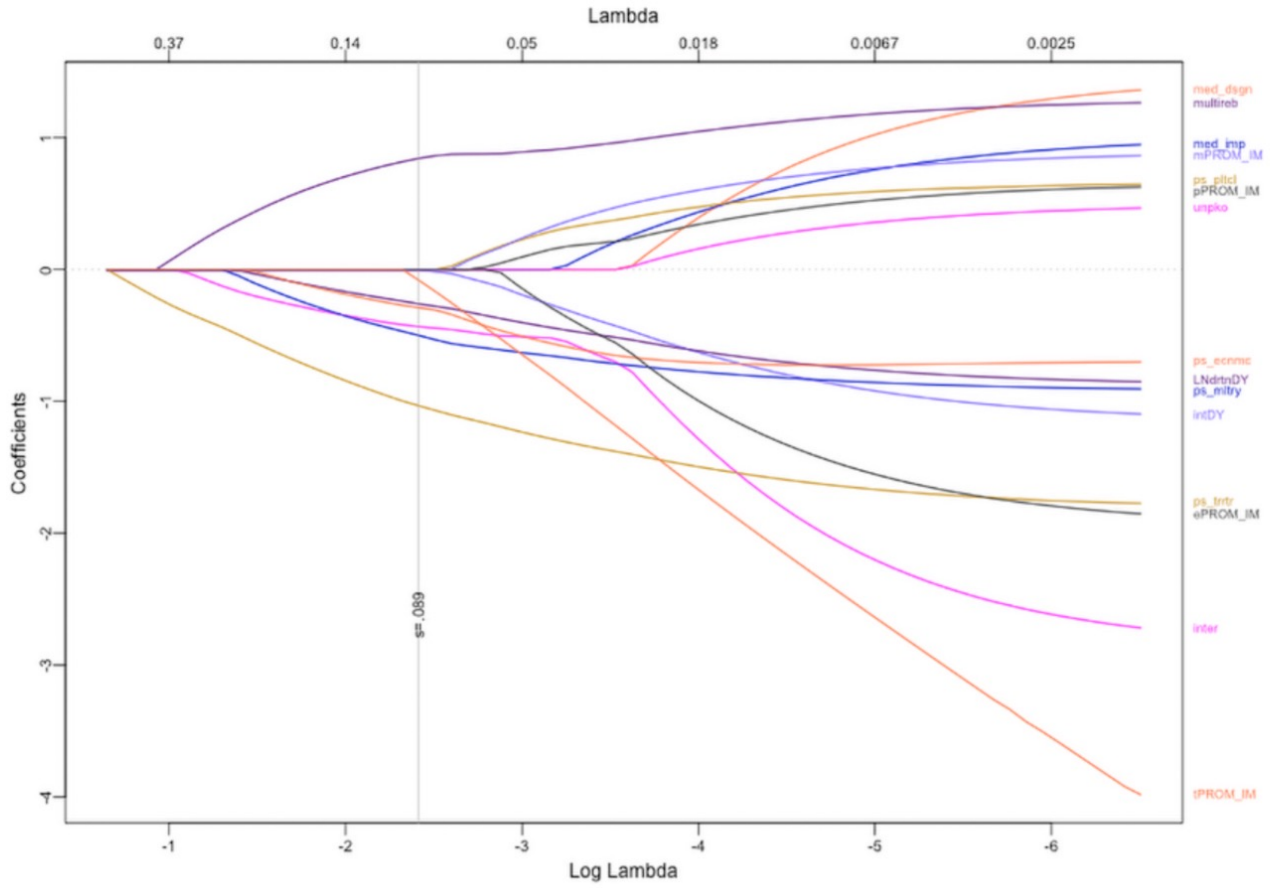
[Domanda #07] Quali sono gli iper-parametri/scelte di design in k-nn classifier 1 points

- [ ] Il parametro di regolarizzazione che gestisce il trade-off fra bias e varianza
- [x] La misura di distanza ✓
- [ ] La dimensione del dataset
- [x] La cardinalità dei vicini nel decretare la moda ✓

Add answer feedback



[Domanda #08] Questo traceplot a che tipo di modello potrebbe essere associato? 1 points



- [ ] OLS
- [x] LASSO ✓
- [ ] Rigde Regression
- [x] Elastic Net ✓

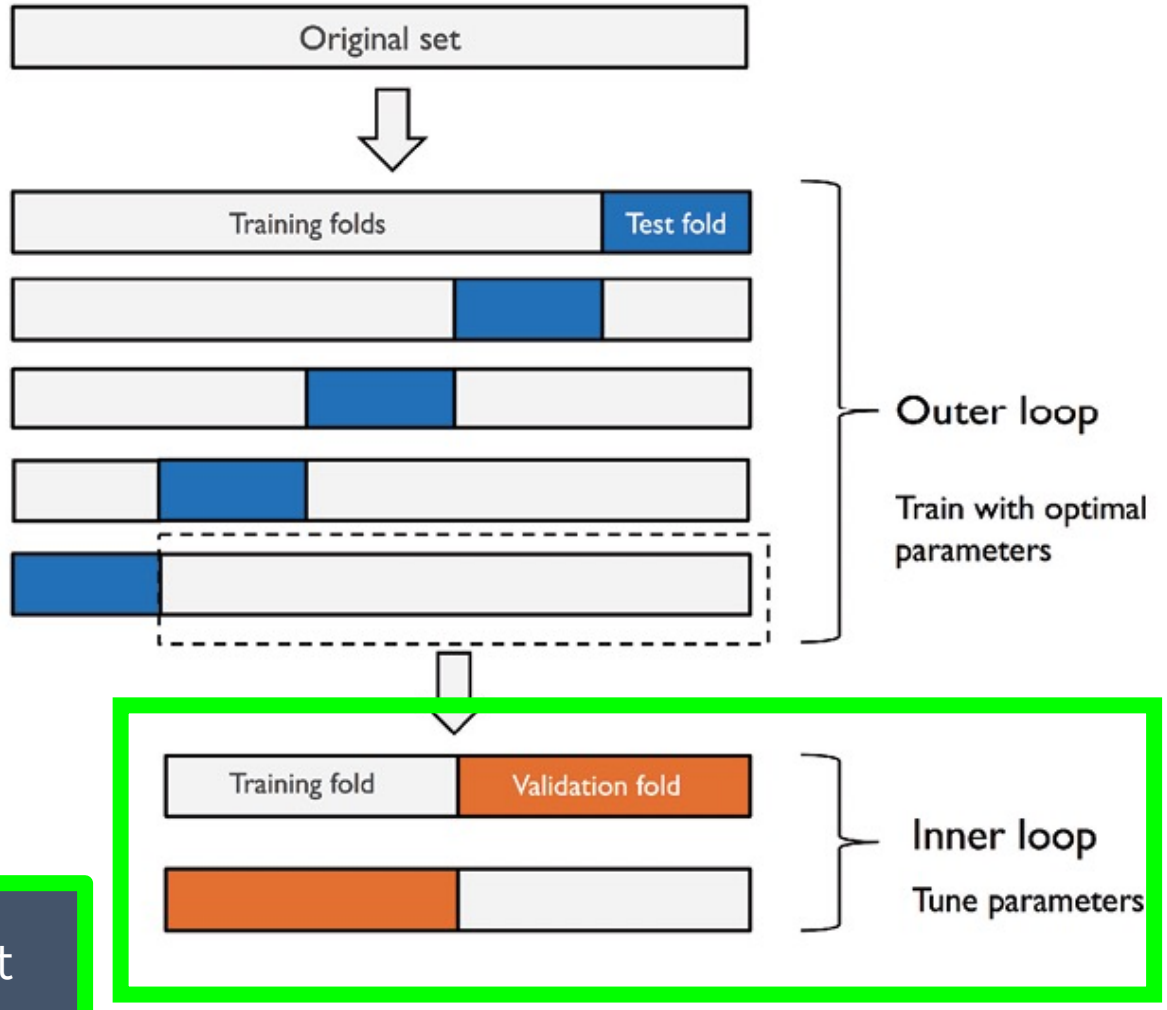
[Domanda #09] Benefici Ridge Regression 1 points

- [x] Gestisce bene la multicollinearità ✓
- [x] Stabilizza la regressione OLS ✓
- [x] Soluzione in forma chiusa ✓
- [ ] Selezione delle variabili

Domanda #10: Cross-validazione

Nested cycle of CV

1. Inner
   - Training data for model construction
   - Validation data for choosing the hyperparameter(s)
2. Outer
   - Training data (training+validation) for model building
   - Test data for performance evaluation



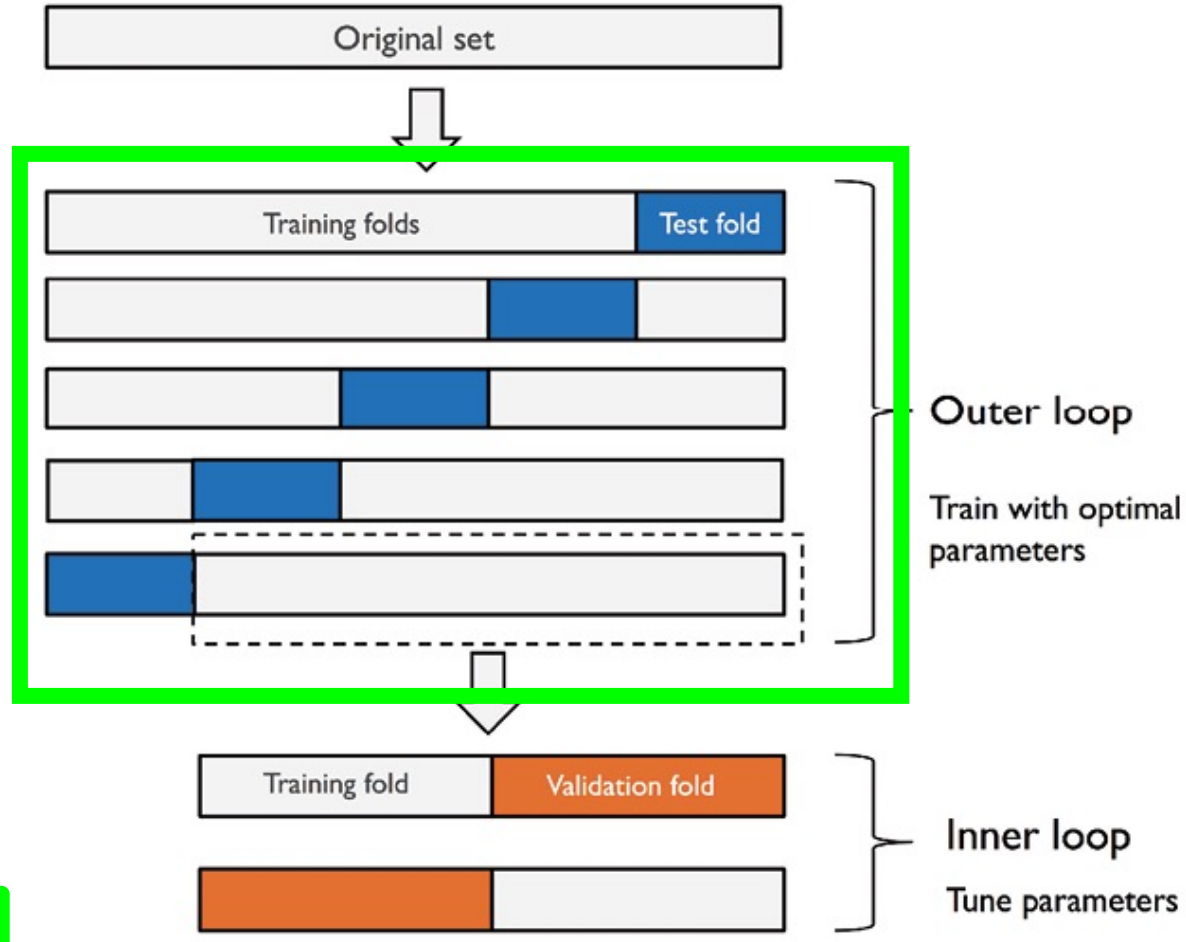
We are answering the question: what is the best hyperparameter for this approach?

Domanda #10: Cross-validazione

Nested cycle of CV

1. Inner
   - Training data for model construction
   - Validation data for choosing the hyperparameter(s)
2. Outer
   - Training data (training+validation) for model building
   - Test data for performance evaluation

We are answering the question: what the performance will be?

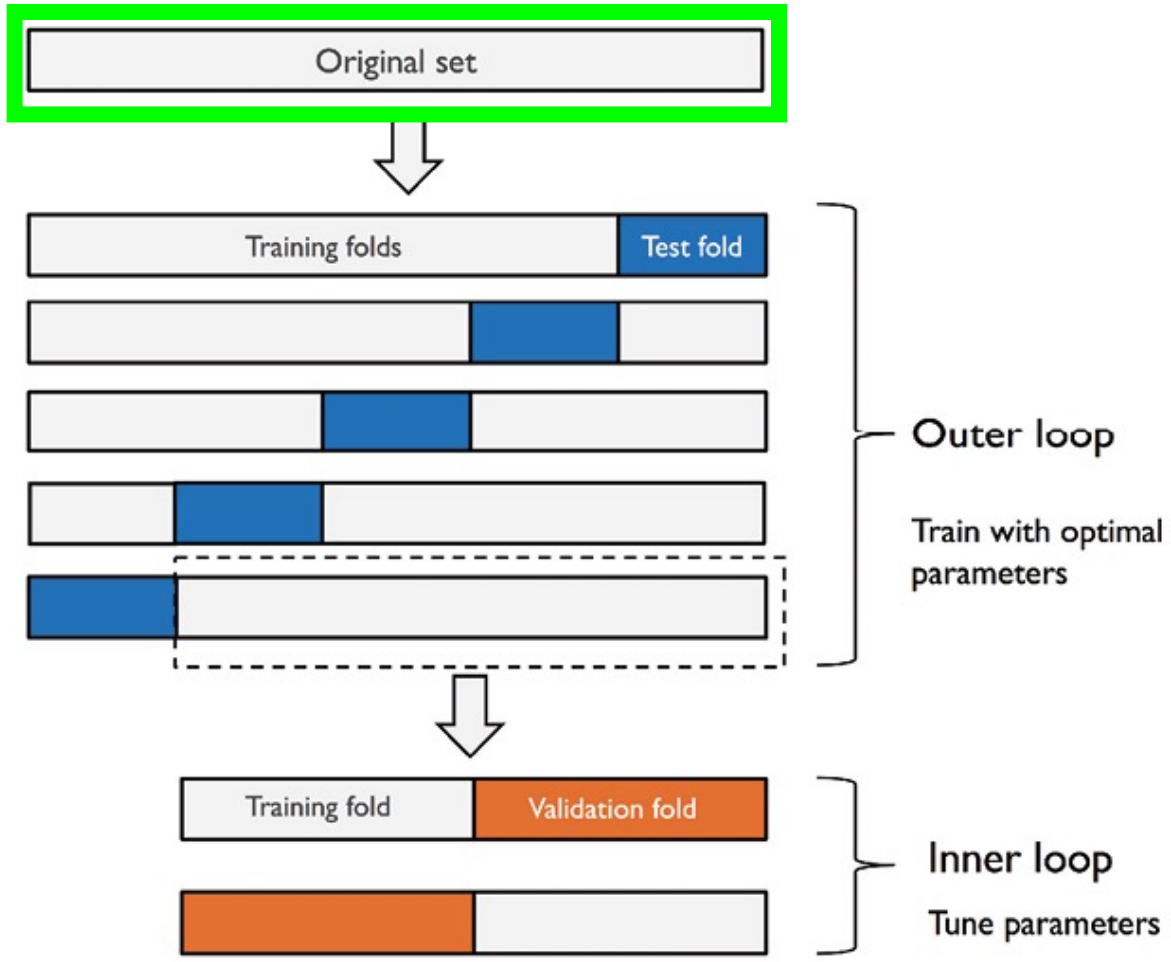


Domanda #10: Cross-validazione

We are answering the question: what will be the 'final' model?

Nested cycle of C

1. Inner
   - Training data for model construction
   - Validation data for choosing the hyperparameter(s)
2. Outer
   - Training data (training+validation) for model building
   - Test data for performance evaluation



Domanda #10: Cross-validazione

[Domanda #10] Cross-validazione: segnare le risposte vere 1 points

- [x] Alla fine della procedura di validazione, tutto il dato viene considerato di training e utilizzato per trovare i parametri del modello finale ✓
- [x] Durante il ciclo interno di cross-validazione, faccio delle valutazioni che mi consentono di capire i migliori iperparametri ✓
- [x] Durante il ciclo esterno di cross-validazione, posso valutare le prestazioni di diversi modelli e diverse scelte di feature engineering ✓
- [x] Durante il ciclo esterno di cross-validazione, posso utilizzare sia k-fold cross-validation che Monte Carlo CV ✓
- [x] Nel ciclo esterno il dato è diviso in train e test, in quello interno in train e validazione ✓

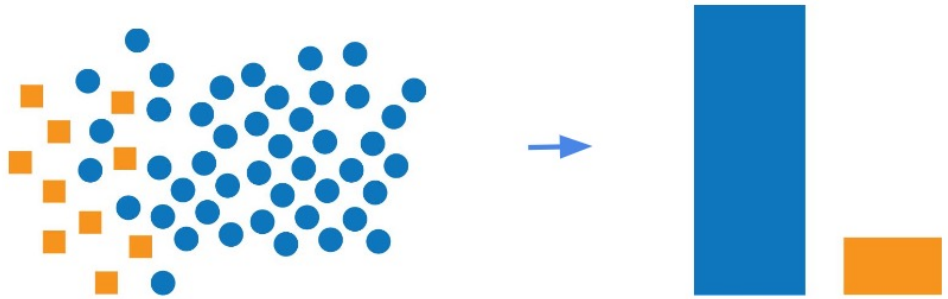
Unbalanced Data

1. The model may ignore the minority class

If 90% of the data belongs to class A and only 10% to class B, a model can get 90% accuracy by always predicting A. This leads to misleading metrics — high accuracy but poor performance on the minority class.

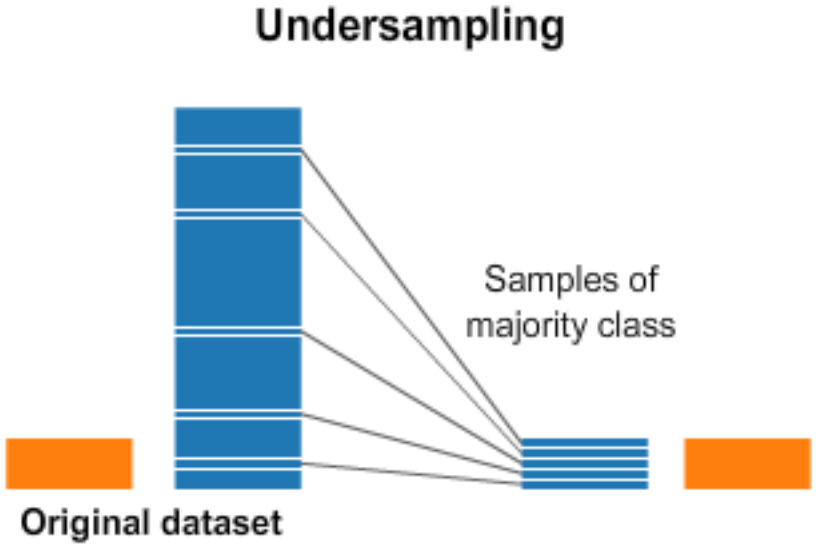
2. Training becomes biased

Some algorithms (e.g., logistic regression, SVM, neural networks) can focus on optimizing the majority class, leading to: (i) Poor generalization on the minority class; (ii) Skewed decision boundaries.



Handling unbalanced data: undersampling

Undersampling is a technique used to balance imbalanced datasets by (randomly) reducing the number of samples in the majority class.



✅ Pros:

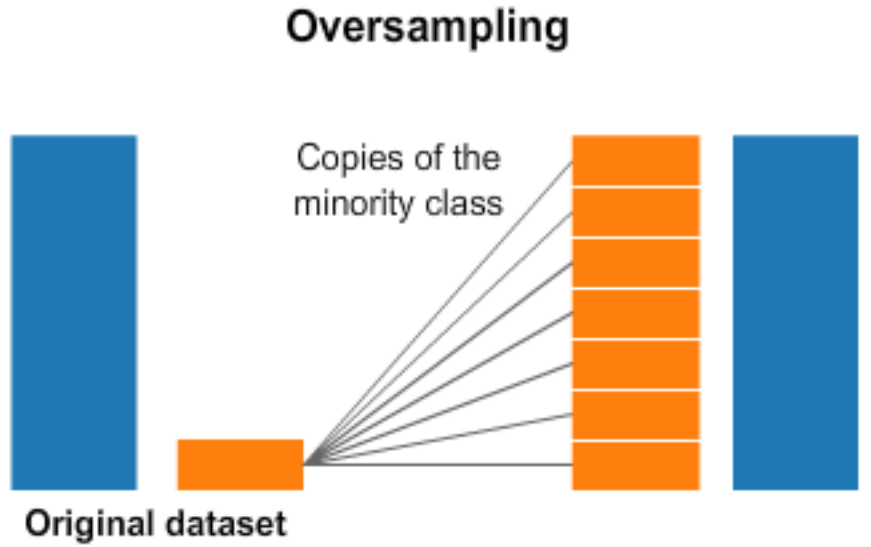
- Simple and fast
- Reduces training time
- Can improve performance on the minority class

❌ Cons:

- Risk of losing useful information (especially if I don't have 'lots' of data)
- May cause underfitting if too much data is discarded

Handling unbalanced data: oversampling

Oversampling is a technique used to balance imbalanced datasets by increasing the number of samples in the minority class. Simply duplicates existing samples!



✅ Pros:

- Balances the dataset without losing information
- Helps the model learn patterns in the minority class better

❌ Cons:

- Can lead to overfitting
- Synthetic data might not always represent the true distribution

Handling unbalanced data: oversampling

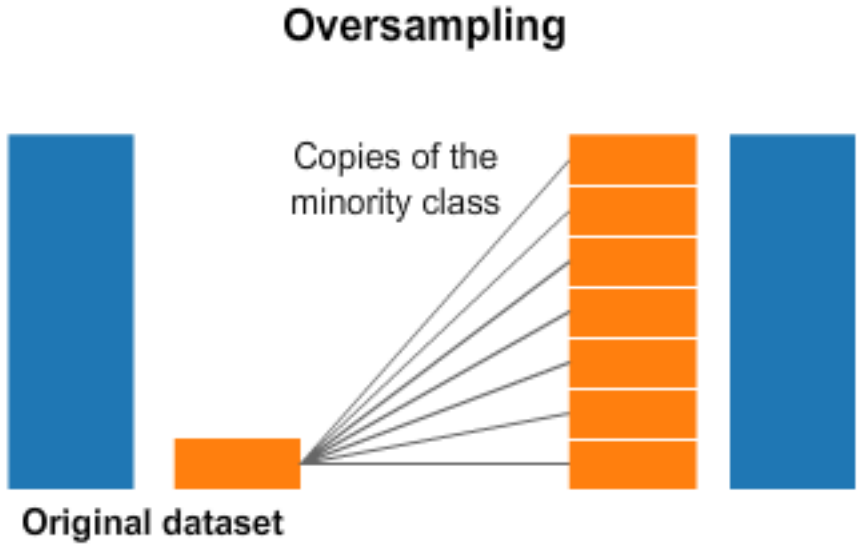
Oversampling is a technique used to balance imbalanced datasets by increasing the number of samples in the minority class. Simply duplicates existing samples!

✅ Pros:

- Balances the dataset without losing information
- Helps the model learn patterns in the minority class better

❌ Cons:

- Can lead to overfitting
- Synthetic data might not always represent the true distribution



In the literature you'll find a procedure called SMOTE: don't use it!!!!!!

Even if popular there is no scientific evidence that it actually works!

Domanda #11: Dati sbilanciati

[Domanda #11] Procedure che possono essere utili nel gestire dati-sbilanciati 1 points

- [ ] Rimozione outlier
- [x] Manipolazione del dataset ('aggiungere' e/o toglierne osservazioni) ✓
- [x] Cross-validazione stratificata ✓

Add answer feedback

Thank you!

Gian Antonio Susto