

Fondamenti di Analisi dei Dati

Appunti e dispense



Università degli Studi di Catania
Dipartimento di Matematica e Informatica
Corso di Laurea Triennale in Informatica (L-31)

Autore

Emanuele Galiano

Anno accademico

Anno Accademico 2025/2026

Prefazione

Queste dispense sono nate come appunti personali per il corso di *Fondamenti di Analisi Dati* tenuto presso il Dipartimento di Matematica e Informatica dell’Università di Catania tenuto dal Prof. Antonino Furnari nell’anno accademico 2025/2026. L’obiettivo principale di questo materiale è stato fornire una risorsa personale di studio sulla parte teorica del corso, raggruppando concetti chiave, definizioni, teoremi, immagini ed esempi in unico documento.

Questo file non deve essere visto come un testo ufficiale o completo sull’argomento, in quanto potrebbero esserci errori, omissioni o imprecisioni. Invito pertanto chi legge questo documento a consultare le dispense ufficiali del corso proposte dal docente ed eventuali testi di riferimento consigliati. Inoltre, invito chiunque noti errori o abbia suggerimenti a contattarmi via email:

- **Email personale:** galianoo.emanuele@gmail.com,
- **Email universitaria:** emanuele.galiano@studium.unict.it;

oppure se ancora presente online, aprire una **issue** o una **pull request** nel repository GitHub associato a queste dispense:

<https://github.com/emanuelegaliano/Fundamentals-of-Data-Analysis>

In particolare, all’interno del repository GitHub sono presenti il codice open-source dei sorgenti \LaTeX utilizzati per generare queste dispense, oltre a tutti i file di supporto come immagini e codici per la generazione di alcuni grafici presenti nel testo.

Licenza

Questo documento, intitolato *Fondamenti di Analisi dei Dati – Appunti e dispense*, è distribuito sotto licenza **Creative Commons Attribution–ShareAlike 4.0 International (CC BY-SA 4.0)**.

La presente licenza è stata scelta con l'obiettivo di favorire la diffusione, la condivisione e il riutilizzo del materiale didattico, garantendo al contempo il riconoscimento dell'autore originale e la preservazione della natura open delle opere derivate.

Diritti concessi

In conformità con i termini della licenza Creative Commons BY-SA 4.0, è consentito a chiunque di:

- copiare e ridistribuire il materiale in qualsiasi mezzo o formato;
- adattare, modificare e trasformare il materiale;
- utilizzare il materiale anche per scopi commerciali.

Tali diritti sono concessi a titolo gratuito e non possono essere revocati, purché siano rispettate le condizioni indicate nella sezione seguente.

Condizioni

L'utilizzo del materiale è subordinato al rispetto delle seguenti condizioni:

- **Attribuzione (BY):** deve essere fornita un'adeguata attribuzione dell'opera, citando l'autore originale, il titolo del documento e la fonte. L'attribuzione deve essere effettuata in modo ragionevole e non tale da suggerire che l'autore originale approvi l'uso o le modifiche apportate.
- **Condividi allo stesso modo (SA):** nel caso in cui il materiale venga modificato, trasformato o utilizzato per creare opere derivate, tali opere devono essere distribuite sotto la *stessa licenza* Creative Commons Attribution–ShareAlike 4.0 International.

Non è consentito applicare termini legali o misure tecnologiche che limitino giuridicamente altri utenti dall'esercitare i diritti concessi dalla licenza.

Assenza di garanzia

Il materiale è fornito “*così com'è*”, senza garanzie di alcun tipo, esplicite o implicite. In particolare, l'autore non garantisce l'accuratezza, la completezza o l'assenza di errori nel contenuto del documento e declina ogni responsabilità per eventuali danni derivanti dall'uso del materiale.

Testo completo della licenza

Il testo legale completo della licenza Creative Commons Attribution–ShareAlike 4.0 International è disponibile al seguente indirizzo:

<https://creativecommons.org/licenses/by-sa/4.0/>

Copyright © 2026 Emanuele Galiano

Indice

Licenza	v
Diritti concessi	v
Condizioni	v
Assenza di garanzia	vi
Testo completo della licenza	vi
1 Concetti chiave dell'analisi dei dati	1
1.1 Concetti fondamentali	1
1.1.1 Data	1
1.1.2 Popolazione	1
1.1.3 Osservazioni	1
1.1.4 Campione	2
1.1.5 Variabile	2
1.1.6 Scale di misurazione	2
1.2 Organizzazione dei dati	3
1.2.1 Datasets	3
1.2.2 Design matrix	3
1.3 Collezione dei dati	3
1.3.1 Surveys	4
1.3.2 Esperimenti	4
1.3.3 Dati osservabili	4
1.4 Data Wrangling	5
1.4.1 Gestire i valori nulli o mancanti	5
1.4.2 Conversione dei dati	5
1.4.3 Rinominazione e riformattazione dei dati	5
1.4.4 Creazione di nuove feature	5
1.4.5 Filtraggio e selezione dei dati	5
1.5 Formato dei dati	5
1.5.1 Formato largo	5
1.5.2 Formato lungo	6
1.5.3 Altre tecniche di Wrangling	6
1.6 Il flusso di lavoro dell'Analisi dei dati	6
1.6.1 Passaggi fondamentali	6
2 Visualizzazione e descrizione dei dati	9

2.1	Frequenze assolute	9
2.2	Frequenze relative	9
2.2.1	Grafici a barre	10
2.2.2	Grafici a torta	10
2.3	ECDF - Empirical Cumulative Distribution Function	10
2.4	Istogrammi	11
2.5	Stima di densità	11
2.6	Statistiche di sommario	11
2.6.1	Dimensione	11
2.6.2	Misure di tendenza centrale	12
2.7	Misure di dispersione	13
2.7.1	Minimo, Massimo	13
2.7.2	Intervallo	13
2.7.3	Intervallo interquantile (IQR)	14
2.7.4	Varianza	14
2.7.5	Deviazione standard	14
2.8	Normalizzazione	14
2.8.1	Min-Max Scaling	15
2.8.2	Normalizzazione tra -1 e 1	15
2.8.3	Standardizzazione (Z-score)	15
2.9	Indicatori di forma	15
2.9.1	Asimmetria	16
2.9.2	Curtosi	16
2.10	Statistiche descrittive	16
3	Probabilità nell'analisi dei dati	17
3.1	Esperimenti casuali	17
3.1.1	Spazio degli eventi	17
3.2	Variabili Aleatorie	18
3.3	Definizione dei dati	18
3.4	Probabilità	18
3.4.1	Assiomi	18
3.4.2	Proprietà	18
3.4.3	Probabilità di Laplace	19
3.5	Stima di probabilità dalle osservazioni	19
3.5.1	Approccio frequentista	19
3.5.2	Approccio bayesiano	19
3.6	Probabilità congiunta	20
3.6.1	Regola della somma	20
3.7	Probabilità condizionata	21
3.7.1	Regola del prodotto	21
3.8	Regola della catena per le probabilità condizionate	22
3.9	Indipendenza	22
3.9.1	Indipendenza condizionata	23

4 Associazioni di variabili	25
4.1 Misure di associazioni tra variabili discrete	25
4.1.1 Indipendenza	25
4.1.2 Statistica di Pearson	26
4.1.3 Statistica di Cramér	26
4.1.4 Rischio relativo	27
4.1.5 Odds Ratio	27
4.2 Misure di associazioni tra variabili continue	28
4.2.1 Visualizzazione grafica dell'associazione	28
4.2.2 Covarianza	29
4.2.3 Coefficiente di correlazione di Pearson	30
4.2.4 Coefficiente di correlazione di Spearman	30
4.2.5 Coefficiente di correlazione di Kendall	31
4.3 Consigli utili su come scegliere la misura di associazione	32
5 Distribuzione dei dati	33
5.1 Distribuzione di probabilità	33
5.2 Distribuzioni discrete	33
5.2.1 Funzione di massa di probabilità (PMF)	33
5.2.2 Funzione di distribuzione cumulativa (CDF)	34
5.3 Distribuzioni continue	34
5.3.1 Funzione di densità di probabilità (PDF)	34
5.3.2 Funzione di distribuzione cumulativa (CDF)	36
5.4 Distribuzioni di probabilità comuni	37
5.4.1 Distribuzione uniforme discreta	37
5.4.2 Distribuzione di Bernoulli	37
5.4.3 Distribuzione binomiale	38
5.4.4 Distribuzione categorica	39
5.4.5 Distribuzione multinomiale	40
5.4.6 Distribuzione Gaussiana (Normale)	42
5.4.7 Teorema del limite centrale	43
5.4.8 Distribuzione Gaussiana Multivariata	44
5.5 Descrivere una distribuzione di probabilità	47
5.5.1 Aspettativa (media)	47
5.5.2 Varianza e deviazione standard	47
5.5.3 Covarianza	48
5.5.4 Entropia	48
5.5.5 Standardizzazione	50
6 Inferenza Statistica	53
6.1 Campionamento	53
6.1.1 Campionamento casuale semplice	53
6.1.2 Campionamento stratificato	55
6.2 Campionare la distribuzione della media	56
6.2.1 Errore standard	57

6.2.2	Distribuzione t-Student	57
6.2.3	Intervallo di confidenza	58
6.3	Bootstrapping	60
6.4	Stimatori	61
6.4.1	Stimatore del bias	61
6.4.2	Stimatore della varianza	61
6.4.3	Varianza di uno stimatore	62
6.4.4	Bias-Varianza Tradeoff	62
6.5	Test statistici	62
6.5.1	Test di ipotesi	62
6.5.2	T-test a un campione	66
6.5.3	T-test a due campioni	66
6.5.4	Test χ^2 per indipendenza	66
6.5.5	Test χ^2 di bontà di adattamento	66
6.5.6	Test di correlazione di Pearson	66
6.5.7	Test di correlazione di Spearman	66
6.6	Valutare quando un campione è distribuito normalmente	67
6.6.1	Grafici Q-Q	67
6.6.2	Test di normalità di Shapiro-Wilk	68
6.6.3	Test K^2 di D'Agostino	68
7	Analisi predittiva	69
7.1	Modello	69
7.1.1	Modelli predittivi	69
7.2	Predizione vs Spiegazione	70
7.2.1	Predizione	70
7.2.2	Spiegazione	71
7.2.3	Compromesso tra Predizione e Spiegazione	71
7.3	Statistica vs Machine Learning	71
7.3.1	Approccio statistico	72
7.3.2	Approccio di Machine Learning	72
7.3.3	Trade-Off di Complessità-Interpretabilità	72
7.4	Tipologie di problema	72
7.4.1	Regressione	73
7.4.2	Classificazione	73
7.4.3	Clustering	75
7.5	Modelli parametrici vs Modelli non parametrici	75
7.5.1	Modelli parametrici	75
7.5.2	Modelli non parametrici	76
7.6	Learning	76
7.6.1	Definizione formale	76
7.6.2	Il processo di Learning	77
7.6.3	ERM: Empirical Risk Minimization	77
7.7	Capacità del modello	78
7.7.1	Misurare la capacità del modello	78

7.7.2	Bias e Varianza	79
7.7.3	Parametri vs Iperparametri	80
7.8	Selezione del modello	80
7.8.1	Approccio 1: selezione statistica	80
7.8.2	Approccio 2: selezione predittiva	81
7.8.3	Validazione Holdout	81
7.8.4	K-Fold Cross-Validation	82
7.8.5	Leave-One-Out Cross-Validation (LOOCV)	82
7.8.6	Ottimizzazione degli iperparametri	82
8	Regressione lineare	87
8.1	Formalizzazione della regressione	87
8.2	Regressione lineare semplice	87
8.2.1	Analogia geometrica con una retta	88
8.2.2	OLS: Ordinary Least Squares	88
8.2.3	Intervalli di confidenza per i coefficienti	90
8.2.4	Test statisticci per la significatività dei coefficienti	91
8.3	Valutazione del modello di regressione	93
8.3.1	Metriche per la bontà del modello	93
8.3.2	Grafici di diagnostica	94
8.4	Regressione lineare multivariata	95
8.4.1	Interpretazione geometrica	96
8.4.2	Interpretazione statistica	96
8.4.3	Stima dei coefficienti di regressione	98
8.4.4	F-Test	99
8.4.5	Eliminazione backward delle variabili	99
8.4.6	Colinearità e instabilità di OLS	100
8.4.7	Adjusted R^2	100
8.5	Predittori qualitativi	101
8.5.1	Variabili dummy	101
9	Oltre la regressione lineare	103
9.1	Interazione tra variabili	103
9.1.1	Interpretazione dei coefficienti	103
9.2	Regressione polinomiale	104
9.2.1	Regressione quadratica	104
9.2.2	Polinomi di grado superiore	105
9.2.3	Problema dell'interpretabilità	105
9.2.4	Metriche di predizione	106
9.2.5	MSE: Mean Squared Error	106
9.2.6	RMSE: Root Mean Squared Error	107
9.2.7	MAE: Mean Absolute Error	107
9.3	Overfitting	108
9.3.1	Regolarizzazione	108
9.3.2	Bias-Varianza trade-off con la regolarizzazione	110

10 Classificazione	113
10.1 Definizione formale	113
10.2 Misure di valutazione	114
10.2.1 Accuratezza	114
10.2.2 Tipi di errori	114
10.2.3 Matrice di confusione	115
10.2.4 Precision e recall	116
10.2.5 F_1 -score	116
10.2.6 Matrice di confusione multiclass	117
10.2.7 ROC e AUC	117
10.3 K-Nearest Neighbors (KNN)	121
10.3.1 1-NN	121
10.3.2 K-NN	122
10.4 Curse of dimensionality	122
10.4.1 Spazio vuoto	123
10.4.2 Impatto su K-NN	123
10.5 Classificatori discriminativi e generativi	124
10.5.1 Classificatori discriminativi	124
10.5.2 Classificatori generativi	124
10.5.3 Macro, Micro e Weighted Averaging	125
10.5.4 Decision Boundary	125
10.6 K-NN per regressione	127
10.6.1 Bias-Varianza trade-off	127
11 Regressione logistica	129
11.1 Modello di regressione logistica	130
11.1.1 Funzione logistica	130
11.1.2 Modello di regressione logistica	130
11.1.3 Odds	131
11.1.4 Log-odds	131
11.2 Stima dei parametri	131
11.2.1 Interpretazione probabilistica del modello	132
11.2.2 Cross-entropy loss	132
11.2.3 Visualizzazione della cross-entropy loss	133
11.3 Interpretazione statistica dei coefficienti	133
11.3.1 Interpretazione dell'intercetta	134
11.3.2 Interpretazione dei coefficienti delle variabili indipendenti	134
11.4 Valutazione della regressione logistica	134
11.4.1 Pseudo R^2	135
11.5 Regressione logistica multiclass	135
11.5.1 Modello di regressione logistica multiclass	135
11.5.2 Interpretazione geometrica dei coefficienti	135
11.6 Regressione softmax	136
11.6.1 Funzione softmax	136
11.6.2 Interpretazione geometrica dei coefficienti	137

11.7	Altri modi di regressione multiclass	138
11.7.1	One-vs-Rest (OvR)	138
11.7.2	One-vs-One (OvO)	138
12	Classificatori generativi	139
12.1	MAP: Maximum A Posteriori	139
12.1.1	Probabilità a priori	140
12.1.2	Likelihood	140
12.2	Il problema della likelihood	140
12.2.1	Il modello ideale nei dati discreti	140
12.2.2	Il modello ideale nei dati continui	141
12.3	Naive Bayes	142
12.3.1	Assunzione di indipendenza condizionata	142
12.4	Naive Bayes Gaussiano	143
12.4.1	Implicazioni delle assunzioni di Naive Bayes gaussiano	144
12.4.2	Confronto dei decision boundaries	145
12.5	Naive Bayes Multinomiale	146
12.5.1	Stima dei parametri	147
12.5.2	Problemi di stima e smoothing	147
13	Rappresentazione dei Dati	149
13.1	Spazio di feature	149
13.1.1	Feature extraction	149
13.1.2	Proprietà dello spazio di feature	150
13.2	Metriche	151
13.2.1	Metriche euclidee	152
13.2.2	Distanza del coseno	153
13.3	Feature e funzioni di rappresentazione	154
13.3.1	Feature extraction "Black Box"	154
13.3.2	DPI: Data Processing Inequality	155
14	Clustering	157
14.1	Definizione del problema	157
14.2	K-means Clustering	157
14.2.1	Ottimizzazione	158
14.2.2	Scegliere il giusto K	159
15	Stima della densità	161
15.1	Densità di probabilità	161
15.2	Metodi non parametrici	161
15.2.1	Iistogrammi	161
15.2.2	Kernel Density Estimation (KDE)	162
15.2.3	Epanechnikov Kernel	164
15.2.4	Tradeoff bias-varianza: scelta della bandwith	165
15.3	Metodi parametrici	168
15.3.1	Distribuzione Gaussiana Multivariata	168

15.3.2 Gaussian Mixture Model	170
16 Riduzione della dimensionalità	175
16.1 Feature Selection vs Feature Extraction	175
16.1.1 Feature Selection	175
16.1.2 Feature Extraction	175
16.2 PCA: Principal Component Analysis	176
16.2.1 Interpretazione geometrica	176
16.2.2 Massimizzazione della varianza	176
16.2.3 Matrice di covarianza	176
16.2.4 Varianza della proiezione	178
16.2.5 Autovalori e autovettori	178
16.2.6 Costruzione e proiezione delle componenti principali	178
16.2.7 Data Whitening: decorrelazione e normalizzazione	179
16.2.8 Scelta del numero di componenti principali	179
16.2.9 Interpretazione delle prime componenti principali	179
17 Grafici	181
17.1 Istogramma	181
17.2 Grafico a barre	182
17.3 Boxplot	183
17.4 Grafici di dispersione	185
17.4.1 Matrice di scatter plot	186
17.4.2 Scatter plot con intervalli di confidenza	186
17.5 Hexbin plot	188
17.6 Grafici di densità e di contorno	189
17.7 Heatmaps	190
17.8 Load plots	190

Capitolo 1

Concetti chiave dell'analisi dei dati

1.1 Concetti fondamentali

1.1.1 Data

Un dato è un insieme di valori raccolti riguardanti un fenomeno, un evento o un'entità specifica. I dati possono essere numerici, testuali, visivi o di altro tipo e sono fondamentali per l'analisi statistica e il machine learning.

Definizione 1.1

1.1.2 Popolazione

La popolazione è l'insieme completo di tutte le osservazioni o unità di interesse in uno studio statistico. Può essere costituita da persone, oggetti, eventi o qualsiasi altra entità che si desidera analizzare.

Definizione 1.2

La popolazione può essere denotata dal simbolo Ω .

1.1.3 Osservazioni

Un'osservazione è un'istanza specifica di dati raccolti su un'entità o un evento. In un dataset, ogni riga rappresenta un'osservazione, che include tutti i valori delle variabili misurate per quell'istanza.

Definizione 1.3

Dato un insieme popolazione Ω , le osservazioni si indicheranno con:

$$\omega \in \Omega$$

1.1.4 Campione

Un campione è un sottoinsieme rappresentativo della popolazione, selezionato per l'analisi statistica. I campioni sono utilizzati per fare inferenze sulla popolazione più ampia, poiché spesso è impraticabile o impossibile raccogliere dati su ogni membro della popolazione.

Definizione 1.4

Si parla di un campione come la tupla:

$$\{\omega^{(1)}, \omega^{(2)}, \dots, \omega^{(n)}\} \subseteq \Omega$$

Dove ogni $\omega^{(i)}$ è un'osservazione appartenente alla popolazione Ω e n è la dimensione del campione.

1.1.5 Variabile

Una variabile, o feature, è una caratteristica misurabile o osservabile di un'osservazione. Quindi è una funzione che associa ad ogni osservazione un valore in un certo dominio.

Definizione 1.5

Definiamo la variabile X come una funzione:

$$X : \Omega \rightarrow S$$

Dove S è l'insieme dei possibili valori che la variabile può assumere. Quindi possiamo ridefinire l'osservazione come:

$$\omega \rightarrow x$$

Ovvero, l'osservazione ω viene mappata al valore x tramite la variabile X . Esistono diverse tipologie di variabili:

- **Quantitative:** assumono valori numerici e permettono operazioni aritmetiche.
- **Qualitative:** assumono valori categorici e non permettono operazioni aritmetiche.
- **Discrete:** assumono un numero finito o numerabile di valori.
- **Continue:** assumono un numero infinito di valori all'interno di un intervallo.
- **Scalare:** assumono un singolo valore per ogni osservazione.
- **Multi-dimensionali:** assumono più valori per ogni osservazione.

1.1.6 Scale di misurazione

Le scale di misurazione sono sistemi utilizzati per classificare e quantificare le variabili in base alle loro caratteristiche.

Definizione 1.6

Esistono diverse scale di misurazione:

- **Nominale:** classifica le osservazioni in categorie senza un ordine specifico.
- **Ordinale:** classifica le osservazioni in categorie con un ordine specifico, ma senza una distanza definita tra le categorie.
- **Intervallo:** misura le differenze tra le osservazioni, ma non ha un punto zero assoluto.
- **Ratio:** misura le differenze tra le osservazioni e ha un punto zero assoluto.

1.2 Organizzazione dei dati

1.2.1 Datasets

Un dataset è una raccolta strutturata di dati organizzati in righe e colonne, dove ogni riga rappresenta un'osservazione e ogni colonna rappresenta una variabile.

Definizione 1.7

1.2.2 Design matrix

Una design matrix è una rappresentazione tabellare dei dati in cui le righe corrispondono alle osservazioni e le colonne corrispondono alle variabili (o feature).

Definizione 1.8

Da questo possiamo notare che:

- Le Osservazioni sono rappresentate dalle righe del dataset.
- Le Variabili (o feature) sono rappresentate dalle colonne del dataset, catturate grazie alle osservazioni.
- I campioni e la popolazione, nel dataset, possono essere visti come una sottotabella del completo.

Valori nulli. In alcuni casi, i dati raccolti possono essere incompleti, con alcune osservazioni che mancano di valori per determinate variabili. Questi valori mancanti sono spesso indicati come "NA" (Not Available) o "null". Potrebbe essere dato da un errore durante la raccolta dati, ma in generale la gestione dei dati mancanti è importante nell'analisi dei dati, poiché può influenzare i risultati delle analisi statistiche e/o dei modelli di machine learning.

Outlier. In un dataset, un outlier è un'osservazione che si discosta significativamente dalle altre osservazioni, un valore "fuori scala". Gli outlier possono essere il risultato di errori di misurazione, errori di inserimento dati o possono rappresentare fenomeni rari ma validi.

1.3 Collezione dei dati

I dati possono essere ottenuti attraverso diverse fonti e metodi. Si parla del primo step per procedere all'analisi dei dati, in quanto la qualità e la rilevanza dei dati raccolti influenzano direttamente i risultati dell'analisi.

1.3.1 Surveys

I surveys sono strumenti di raccolta dati che consistono in questionari o interviste progettati per ottenere informazioni specifiche da un gruppo di persone.

Definizione 1.9

I surveys possono essere somministrati in vari modi, tra cui questionari cartacei, interviste telefoniche, sondaggi online o interviste faccia a faccia. La progettazione di un survey efficace richiede attenzione alla formulazione delle domande, alla selezione del campione e alla modalità di somministrazione per garantire la raccolta di dati accurati e rappresentativi.

Il problema dei survey è che spesso le risposte possono essere influenzate da bias (come il bias di desiderabilità sociale, dove i partecipanti rispondono in modo da apparire più favorevoli agli occhi degli altri, piuttosto che fornire risposte oneste), ma in generale non si parla di dati *altamente affidabili*.

1.3.2 Esperimenti

Gli esperimenti sono studi controllati in cui i ricercatori manipolano una o più variabili indipendenti per osservare l'effetto su una o più variabili dipendenti.

Definizione 1.10

Gli esperimenti possono essere condotti in laboratorio o sul campo e richiedono un'attenta progettazione per garantire che i risultati siano validi e affidabili. Gli esperimenti spesso includono gruppi di controllo e randomizzazione per minimizzare i bias e isolare gli effetti delle variabili manipolate.

Uno dei tipi di esperimenti più comuni è la tipologia RCT (Randomized Controlled Trial), in cui i partecipanti sono assegnati casualmente a gruppi sperimentali o di controllo.

Gli esperimenti hanno un difetto: sono costosi e richiedono tempo per essere condotti, ma in generale si parla di dati *molto affidabili*.

1.3.3 Dati osservabili

I dati osservabili sono dati raccolti attraverso l'osservazione diretta di fenomeni, eventi o comportamenti senza manipolazione o intervento da parte del ricercatore.

Definizione 1.11

Questi eventi sono utili quando gli esperimenti sono impraticabili, non etici oppure vanno fuori budget.

I dati osservabili hanno un difetto: possono essere influenzati da fattori esterni non controllati, ma in generale si parla di dati *affidabili*.

Fonti online

Nella data science moderna, una fonte sempre più comune di dati è rappresentata dalle fonti online. Questi dati possono essere raccolti da siti web, social media, database pubblici e altre piattaforme digitali.

1.4 Data Wrangling

Il data wrangling, o data munging, è il processo di pulizia, trasformazione e organizzazione dei dati grezzi in un formato utilizzabile per l'analisi.

Definizione 1.12

Come dicevamo nella sezione 1.2.2, i dati grezzi non sono mai perfettamente puliti: spesso contengono errori, valori mancanti, outlier e formati incoerenti che devono essere affrontati prima di procedere con l'analisi.

1.4.1 Gestire i valori nulli o mancanti

I valori nulli o mancanti possono essere gestiti in diversi modi, tra cui:

- Rimozione delle osservazioni con valori mancanti.
- Imputazione dei valori mancanti utilizzando la media, la mediana o la moda delle altre osservazioni.
- Utilizzo di modelli predittivi per stimare i valori mancanti basati su altre variabili.

1.4.2 Conversione dei dati

Alcune volte i dati vengono raccolti male, allora è necessario convertirli in formati più utili per l'analisi.

1.4.3 Rinominazione e riformattazione dei dati

Rinominare e formattare i dati in modo coerente può facilitare l'analisi e la comprensione del dataset.

1.4.4 Creazione di nuove feature

A volte è utile creare nuove feature basate su quelle esistenti per migliorare l'analisi.

1.4.5 Filtraggio e selezione dei dati

Filtrare e selezionare i dati rilevanti per l'analisi può migliorare l'efficienza e la precisione dei risultati.

1.5 Formato dei dati

I dati possono essere organizzati in formati differenti in base alle esigenze dell'analisi.

1.5.1 Formato largo

Quando si parla di formato largo (wide format), i dati sono organizzati in modo tale che:

- Ogni variabile è rappresentata da una colonna separata.
- Ogni osservazione è rappresentata da una riga separata.

I vantaggi sono la facile lettura e la struttura intuitiva, ma può essere inefficiente per dataset con molte variabili o osservazioni. Si usa infatti, nell'analisi esplorativa dei dati e nella visualizzazione.

1.5.2 Formato lungo

Nel caso del formato lungo (long format), i dati sono organizzati in modo tale che:

- Le variabili sono rappresentate in una colonna separata.
- Le osservazioni sono rappresentate in più righe, con ogni riga che rappresenta una combinazione di osservazione e variabile.

I vantaggi sono l'efficienza nella memorizzazione e la facilità di manipolazione dei dati, ma può essere più difficile da leggere e interpretare. Si usa infatti, spesso nei modelli statistici e nelle analisi di serie temporali.

1.5.3 Altre tecniche di Wrangling

Esistono altre tecniche di data wrangling che possono essere utilizzate per preparare i dati per l'analisi, tra cui:

- Conversione dell'unità di misura
- Normalizzazione dei valori
- Aggregazione dei dati
- Riduzione della dimensionalità
- Rimuovere duplicati
- Estrarre informazione e parsing di stringhe
- Ricampionamento dei dati tra formati lunghi e larghi.

1.6 Il flusso di lavoro dell'Analisi dei dati

La definizione che viene data, all'analisi dei dati, è:

L'analisi dei dati è il processo di ispezione, pulizia, trasformazione e modellazione dei dati con l'obiettivo di scoprire informazioni utili, trarre conclusioni e supportare il processo decisionale.

Definizione 1.13

1.6.1 Passaggi fondamentali

Il flusso di lavoro tipico per l'analisi dei dati include i seguenti passaggi fondamentali:

Ispezione o Data exploration: Esplorare i dati per comprendere la loro struttura, qualità e caratteristiche principali.

Pulizia dei dati: Rimuovere o correggere errori, valori mancanti e outlier nei dati.

Trasformazione dei dati: Modificare i dati per renderli più adatti all'analisi, ad esempio creando nuove feature o convertendo i dati in formati diversi.

Modellazione dei dati: Applicare tecniche statistiche o di machine learning per analizzare i dati e trarre conclusioni.

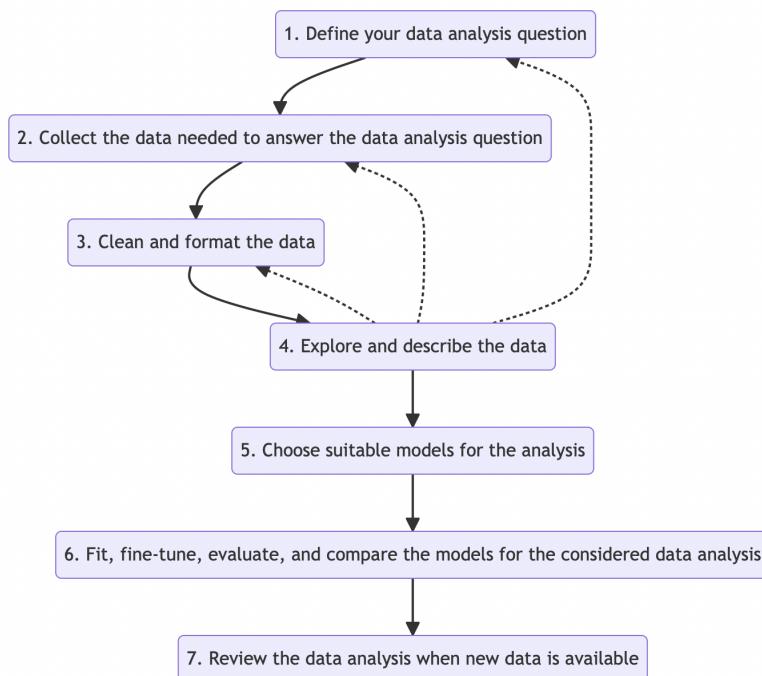


Figura 1.1: Flusso di lavoro dell'analisi dei dati.

Generalmente esiste un workflow iterativo e, ormai, standardizzato per l'analisi dei dati, che può essere riassunto nei seguenti passaggi:

1. Definire l'obiettivo dell'analisi.
2. Raccogliere i dati rilevanti.
3. Pulire e preparare i dati. Gestire i valori mancanti, rimuovere outlier, creare nuove feature, ecc.
4. Esplorare i dati. Utilizzare tecniche di visualizzazione e statistiche descrittive per comprendere la distribuzione dei dati, identificare pattern interessanti e relazioni tra le variabili.
5. Model lifecycle. Applicare un modello statistico o di machine learning per analizzare i dati.
6. Revisionare e aggiornare l'analisi. Basandosi sui risultati ottenuti, si potrebbe rivedere l'analisi, apportare modifiche ai dati o al modello e ripetere il processo se necessario.

Sebbene questo sembri un modello molto lineare e sequenziale, spesso non lo è. Infatti, si può rappresentare il flusso di lavoro dell'analisi dei dati come un ciclo iterativo, in cui si torna indietro a fasi precedenti in base ai risultati ottenuti e alle nuove informazioni scoperte durante l'analisi (come si vede in figura 1.1).

Capitolo 2

Visualizzazione e descrizione dei dati

Il problema principale dei grandi dataset di dati è che spesso sono difficili da interpretare direttamente. Per questo motivo, è necessario andare più a fondo per scoprire le strutture e i pattern che sono nascosti e non visibili semplicemente da una rapida stampa dei dati grezzi. La **statistica** aiuta in questo, in quanto fornisce uno strumento per riassumere e descrivere i dati in modo significativo.

2.1 Frequenze assolute

Un primo modo di descrivere i dati è quello di calcolare le **frequenze assolute**.

Siano a_1, \dots, a_n i valori distinti di una variabile che stiamo considerando. La frequenza assoluta n_i allora si può definire come il numero di volte che il valore a_i appare nel dataset.

Definizione 2.1

Si noti che la somma delle frequenze assolute è uguale al numero N totale di osservazioni nel dataset:

$$\sum_i n_i = N$$

Nel nostro esempio del negozio, se consideriamo la variabile `Gender`, possiamo calcolare le frequenze assolute dei valori distinti:

- Maschio: 40
- Femmina: 60

Per rappresentare le frequenze assolute, possiamo utilizzare un **grafico a barre**. Un grafico a barre mostra le categorie della variabile sull'asse delle ascisse e le frequenze assolute sull'asse delle ordinate. Ogni barra rappresenta una categoria distinta, e l'altezza della barra corrisponde alla frequenza assoluta di quella categoria.

2.2 Frequenze relative

Il problema delle frequenze assolute è che non tengono conto della dimensione totale del dataset. Per questo motivo, spesso è più utile calcolare le **frequenze relative**.

La frequenza relativa f_i di un valore a_j si definisce come il rapporto tra la frequenza assoluta n_i e il numero totale di osservazioni N :

$$f_i = \frac{n_i}{N}, \quad \forall j \in 1, \dots, n$$

Definizione 2.2

Ovviamente, da questo seguono due cose:

1. Se $n_j \leq n \Rightarrow f_j \leq 1 \forall j$
2. La somma delle frequenze relative è uguale a 1:

$$\sum_j f_j = \sum_j \frac{n_j}{N} = \frac{1}{N} \sum_j n_j = \frac{N}{N} = 1$$

Anche qui è molto utile confrontare tutti gli elementi con un grafico a barre, l'unica cosa che cambia è che le frequenze hanno una scala diversa.

2.2.1 Grafici a barre

Un altro modo per contare il numero di elementi di una determinata classe è utilizzare un **grafico a barre**. In questo tipo di grafico, ogni barra rappresenta una categoria distinta della variabile e l'altezza della barra rappresenta la frequenza assoluta o relativa di quella categoria. I grafici a barre sono particolarmente utili per confrontare le frequenze tra diverse categorie.

Esiste una variante, il **grafico a barre impilato** in cui le barre sono suddivise in segmenti che rappresentano sottocategorie. Questo tipo di grafico è utile per visualizzare la composizione delle categorie principali.

2.2.2 Grafici a torta

Un altro tipo di visualizzazione molto comune per le frequenze relative è il **grafico a torta** (o *pie chart*). In questo tipo di grafico, un cerchio è suddiviso in spicchi, ciascuno dei quali rappresenta una categoria distinta della variabile. L'area di ogni spicchio è proporzionale alla frequenza relativa della categoria corrispondente. I grafici a torta sono utili per mostrare la proporzione delle categorie rispetto al totale.

Questi grafici funzionano bene quando ci sono poche classi, ma molto male negli altri casi. Non sono infatti adatti a mostrare differenze tra categorie simili.

2.3 ECDF - Empirical Cumulative Distribution Function

Un altro modo per visualizzare la distribuzione di una variabile è utilizzare la **funzione di distribuzione cumulativa empirica** (ECDF).

La ECDF di una variabile X è una funzione che, per ogni valore x , restituisce la proporzione di osservazioni nel dataset che sono minori o uguali a x .

Definizione 2.3

Matematicamente, la ECDF si può definire come:

$$F(x) = \frac{1}{N} \sum_{i=1}^N I(X_i \leq x)$$

dove I è la funzione indicatrice che vale 1 se la condizione è vera e 0 altrimenti.

La ECDF è utile perché fornisce una rappresentazione completa della distribuzione dei dati e consente di confrontare facilmente diverse distribuzioni.

2.4 Istogrammi

Un altro strumento utile per la visualizzazione dei dati continui è l'**istogramma**. Un istogramma è un grafico che rappresenta la distribuzione di una variabile continua suddividendo l'intervallo dei valori in **classi** (o **bin**) e contando il numero di osservazioni che cadono in ciascuna classe.

Per ogni istogramma però, si deve scegliere il numero di classi e la loro ampiezza. Questo numero può essere ottenuto in modo arbitrario, oppure utilizzando delle euristiche come la **regola di Sturges**:

$$k = \lceil \log_2(N) + 1 \rceil$$

dove k è il numero di classi e N è il numero totale di osservazioni nel dataset.

2.5 Stima di densità

Un problema degli istogrammi è che la loro forma dipende fortemente dalla scelta del numero di classi e dalla loro ampiezza. Si può utilizzare la **stima di densità** per ottenere una rappresentazione più fluida della distribuzione dei dati continui.

La stima di densità è una tecnica che cerca di risolvere questo problema ottenendo un'approssimazione continua della distribuzione dei dati.

2.6 Statistiche di sommario

Per descrivere i dati in modo più sintetico, si possono calcolare alcune **statistiche di sommario** che riassumono le caratteristiche principali della distribuzione dei dati.

2.6.1 Dimensione

La dimensione di un campione univariato $\{x_i\}_i^D$ è il numero di valori che contiene:

$$|\{x_i\}_i^D| = D$$

Definizione 2.4

Le varie dimensioni possono variare, in quanto ogni colonna può contenere valori nulli.

2.6.2 Misure di tendenza centrale

Le misure di tendenza centrale sono statistiche che descrivono il centro della distribuzione dei dati.

Media

La media di un campione univariato $\{x_i\}_i^D$ è definita come:

$$\bar{x} = \frac{1}{D} \sum_{i=1}^D x_i$$

Definizione 2.5

Da notare che la media è un valore facile da interpretare, ma non sempre affidabile in quanto è molto sensibile ai valori anomali (outlier).

Quantili

Per definire la mediana, è necessario prima definire altri valori statistici chiamati **quantili**.

Un quantile di ordine α di un campione univariato $\{x_i\}_i^D$ è il valore q_α tale che una frazione α delle osservazioni è minore o uguale a q_α :

$$q_\alpha = \inf\{x | F(x) \geq \alpha\}$$

dove $F(x)$ è la funzione di distribuzione cumulativa empirica (ECDF) del campione.

Definizione 2.6

Percentili. Dalla misura dei quantili, possiamo derivare i **percentili**, che sono quantili di ordine α espressi come percentuali. Ad esempio, il 25° percentile (o primo quartile) è il valore sotto il quale si trova il 25% delle osservazioni.

Quartili. I **quartili** sono quantili che dividono il dataset in quattro parti uguali. Il primo quartile (Q1) è il 25° percentile, il secondo quartile (Q2) è la mediana (50° percentile), e il terzo quartile (Q3) è il 75° percentile.

Mediana

Una volta definiti i quantili, è facile definire la mediana.

La mediana di un campione univariato $\{x_i\}_i^D$ è il quantile di ordine 0.5:

$$\text{mediana} = q_{0.5}$$

Definizione 2.7

Un'altro modo di definire la mediana, non utilizzando i quantili, per un certo campione ordinato $x_1 \leq x_2 \leq \dots \leq x_n$ è:

$$\hat{x}_{0.5} = \begin{cases} x_{n+1}/2 & \text{se } n \text{ è dispari.} \\ \frac{x_{n/2} + x_{n/2+1}}{2} & \text{se } n \text{ è pari.} \end{cases}$$

Si nota che la mediana è meno sensibile ai valori anomali rispetto alla media, rendendola una misura più robusta della tendenza centrale in presenza di outlier.

Moda

La moda \bar{x}_M di un campione univariato $\{x_i\}_i^D$ è il valore che appare con la massima frequenza nel campione.

Definizione 2.8

Formalizzando la definizione, si può scrivere che:

$$\bar{x}_M = a_j \Leftrightarrow n_j = \max\{n_1, \dots, n_k\}$$

dove a_j sono i valori distinti della variabile e n_j le loro frequenze assolute.

2.7 Misure di dispersione

Per misure di dispersione si intendono delle statistiche che descrivono quanto i dati sono sparsi o concentrati attorno a una misura di tendenza centrale.

2.7.1 Minimo, Massimo

Il minimo x_{min} e il massimo x_{max} di un campione univariato $\{x_i\}_i^D$ sono definiti come:

$$x_{min} = \min\{x_i | i = 1, \dots, D\}$$

$$x_{max} = \max\{x_i | i = 1, \dots, D\}$$

Definizione 2.9

2.7.2 Intervallo

Dalla definizione di minimo e massimo, si può definire una misura di dispersione chiamata **intervallo**.

L'intervallo R di un campione univariato $\{x_i\}_i^D$ è definito come la differenza tra il massimo e il minimo:

$$R = x_{max} - x_{min}$$

Definizione 2.10

L'intervallo ha un problema, non è una misura molto robusta di dispersione in quanto dipende solo dai valori estremi del dataset, che possono essere influenzati da outlier.

2.7.3 Intervallo interquantile (IQR)

L'intervallo interquantile IQR di un campione univariato $\{x_i\}_i^D$ è definito come la differenza tra il terzo quartile Q3 e il primo quartile Q1:

$$IQR = Q3 - Q1$$

Definizione 2.11

Quindi, possiamo dedurre che la mediana si trova al centro dell'IQR. L'IQR è una misura più robusta di dispersione rispetto all'intervallo, in quanto non è influenzata dai valori estremi del dataset.

2.7.4 Varianza

La varianza di un campione univariato $\{x_i\}_i^D$ è definita come:

$$s^2 = \frac{1}{D-1} \sum_{i=1}^D (x_i - \bar{x})^2$$

dove \bar{x} è la media del campione.

Definizione 2.12

La varianza misura la dispersione dei dati attorno alla media. Un valore di varianza più alto indica che i dati sono più sparsi, mentre un valore più basso indica che i dati sono più concentrati attorno alla media.

Il problema di questa misura di dispersione, è che è espressa nelle unità al quadrato della variabile originale, rendendo difficile l'interpretazione diretta. Inoltre è molto sensibile agli outlier.

2.7.5 Deviazione standard

La deviazione standard di un campione univariato $\{x_i\}_i^D$ è definita come la radice quadrata della varianza:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{D-1} \sum_{i=1}^D (x_i - \bar{x})^2}$$

Definizione 2.13

Si preferisce alla varianza, la deviazione standard perché risolve il problema delle unità al quadrato, essendo espressa nelle stesse unità della variabile originale. Tuttavia, rimane sensibile agli outlier.

2.8 Normalizzazione

Il problema di utilizzare queste statistiche di sommario è che non sono comparabili tra variabili con scale diverse. Per risolvere questo problema, si può utilizzare la **normalizzazione** dei dati. Questo problema si risolve con la normalizzazione e ne esistono diverse tipologie.

2.8.1 Min-Max Scaling

La normalizzazione Min-Max di un campione univariato $\{x_i\}_i^D$ è definita come:

$$x_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

dove x_{min} e x_{max} sono il minimo e il massimo del campione.

Definizione 2.14

Questa tecnica trasforma i dati in un intervallo compreso tra 0 e 1, risolvendo il problema della scala dei dati. Tuttavia è sempre sensibile agli outliers.

2.8.2 Normalizzazione tra -1 e 1

La normalizzazione tra -1 e 1 di un campione univariato $\{x_i\}_i^D$ è definita come:

$$x_{norm} = \frac{x_{max} + x_{min}}{2} + \frac{x_i - \frac{x_{max} + x_{min}}{2}}{\frac{x_{max} - x_{min}}{2}}$$

Definizione 2.15

Questa tecnica trasforma i dati in un intervallo compreso tra -1 e 1, risolvendo il problema della scala dei dati. Tuttavia, anche questa tecnica è sensibile agli outliers.

2.8.3 Standardizzazione (Z-score)

Molto spesso è utile normalizzare i dati in modo che abbiano media 0 e deviazione standard 1. Questo processo è chiamato **standardizzazione o Z-score normalization**.

La standardizzazione (Z-score) di un campione univariato $\{x_i\}_i^D$ è definita come:

$$z_i = \frac{x_i - \bar{x}}{s}$$

dove \bar{x} è la media del campione e s è la deviazione standard del campione.

Definizione 2.16

Questa tecnica trasforma i dati in modo che abbiano media 0 e deviazione standard 1, rendendo le variabili comparabili tra loro. Inoltre, la standardizzazione è meno sensibile agli outliers rispetto alle altre tecniche di normalizzazione.

2.9 Indicatori di forma

Oltre alle misure di tendenza centrale e di dispersione, esistono anche degli indicatori di forma che descrivono la forma della distribuzione dei dati.

2.9.1 Asimmetria

L'asimmetria (skewness) di un campione univariato $\{x_i\}_i^D$ è definita come:

$$\frac{1}{D} \sum_{i=1}^D \left(\frac{x_i - \bar{x}}{s} \right)^3$$

dove \bar{x} è la media del campione e s è la deviazione standard del campione.

Definizione 2.17

L'asimmetria misura la simmetria della distribuzione dei dati attorno alla media. Un valore di asimmetria positivo indica una distribuzione asimmetrica a destra, mentre un valore negativo indica una distribuzione asimmetrica a sinistra. Quanto un valore è più vicino allo zero, tanto più la distribuzione è simmetrica.

2.9.2 Curtosi

La curtosi (kurtosis) di un campione univariato $\{x_i\}_i^D$ è definita come:

$$\frac{1}{D} \sum_{i=1}^D \left(\frac{x_i - \bar{x}}{s} \right)^4 - 3$$

dove \bar{x} è la media del campione e s è la deviazione standard del campione.

Definizione 2.18

La curtosi misura la "punta" della distribuzione dei dati. Un valore di curtosi positivo indica una distribuzione "leptocurtica" (più appuntita), mentre un valore negativo indica una distribuzione "platicurtica" (più piatta). Quanto il valore è più vicino allo zero, tanto più la distribuzione è simile a una distribuzione normale.

2.10 Statistiche descrittive

Tutte queste misure di tendenza centrale, dispersione e forma possono essere riassunte in una tabella chiamata **statistica descrittiva**. Questa tabella fornisce una panoramica completa delle caratteristiche principali della distribuzione dei dati e consente di confrontare facilmente diverse variabili.

Un grafico molto utile per rappresentare la statistica descrittiva è il **box plot** (o *diagramma a scatola*). In questo tipo di grafico, una scatola rappresenta l'intervallo interquartile (IQR) della variabile, con una linea all'interno della scatola che rappresenta la mediana. Le "whiskers" (linee esterne) si estendono fino ai valori minimo e massimo, escludendo gli outlier che sono rappresentati come punti separati. Il box plot è utile per visualizzare la distribuzione dei dati, identificare outlier e confrontare diverse variabili in modo rapido ed efficace.

Capitolo 3

Probabilità nell'analisi dei dati

Durante l'analisi dei dati, spesso ci si imbatte in situazioni in cui è necessario comprendere e quantificare l'incertezza associata ai dati raccolti. La probabilità fornisce un quadro teorico per affrontare queste situazioni, permettendo di fare inferenze basate su campioni di dati.

3.1 Esperimenti casuali

3.1.1 Spazio degli eventi

*Si definisce **spazio degli eventi** l'insieme di tutti i possibili risultati di un esperimento casuale:*

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$$

Definizione 3.1

Dallo spazio degli eventi¹, possiamo dare una seconda definizione di **evento semplice**.

*Un **evento semplice** è un sottoinsieme dello spazio degli eventi che contiene un solo elemento:*

$$A = \{\omega_i\}$$

Definizione 3.2

Un esempio di un evento semplice è l'estrazione di un asso da un mazzo di carte standard.

Molto spesso però, il risultato atteso è più grande di un singolo evento semplice. In questo caso, definiamo un **evento**.

*Un **evento** è un sottoinsieme dello spazio degli eventi che contiene più di un elemento:*

$$A = \{\omega_i, \omega_j, \dots\} \subseteq \Omega$$

Definizione 3.3

Spesso si denota $\bar{A} = \Omega \setminus A$ per indicare l'evento complementare di A . Da questa definizione possiamo trarre due semplici conclusioni:

¹Notare che lo spazio degli eventi è denominato dallo Ω come la popolazione (sotto-sezione 1.1.2), non è un errore ma una convenzione (non sono la stessa cosa)

- Lo spazio degli eventi Ω è un evento, in quanto contiene tutti gli eventi semplici. In particolare viene chiamato **evento certo**, in quanto sempre vero.
- L'insieme vuoto \emptyset è un evento che non può mai avvenire, chiamato **evento impossibile**.

3.2 Variabili Aleatorie

Una variabile aleatoria è una funzione che associa ad ogni evento semplice un numero reale:

$$X : \Omega \rightarrow E$$

Dove E è uno spazio misurabile.

Definizione 3.4

3.3 Definizione dei dati

Nel contesto dell'analisi dei dati, i dati sono considerati come realizzazioni di variabili aleatorie. Ogni osservazione raccolta durante un esperimento casuale può essere vista come un'istanza di una variabile aleatoria.

Definizione 3.5

3.4 Probabilità

Poiché i risultati di una variabile aleatoria sono legati a eventi stocastici e non deterministici, è necessario definire una misura che quantifichi la probabilità di accadimento di tali eventi.

Una misura di probabilità è una funzione P che assegna a ogni evento $A \subseteq \Omega$ un numero reale compreso tra 0 e 1.

Definizione 3.6

3.4.1 Assiomi

La misura di probabilità, generalmente soddisfa alcuni assiomi:

- $P(\Omega) = 1$: La probabilità dell'evento certo è 1.
- $P(\emptyset) = 0$: La probabilità dell'evento impossibile è 0.
- Se A_1, A_2, \dots, A_n sono eventi mutuamente esclusivi, allora:

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$$

3.4.2 Proprietà

Dalle definizioni e dagli assiomi sopra riportati, possiamo derivare alcune proprietà utili della misura di probabilità:

- **Complementarità:** $P(\bar{A}) = 1 - P(A)$
- **Monotonicità:** Se $A \subseteq B$, allora $P(A) \leq P(B)$
- **Additività:** Per due eventi qualsiasi A e B :

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

3.4.3 Probabilità di Laplace

In alcuni casi particolari, come negli esperimenti con esiti equiprobabili, possiamo utilizzare la definizione classica di probabilità proposta da Laplace.

La probabilità di Laplace di un evento A è definita come il rapporto tra il numero di esiti favorevoli all'evento A e il numero totale di esiti possibili:

$$P(A) = \frac{\text{numero di esiti favorevoli ad } A}{\text{numero totale di esiti possibili}}$$

Definizione 3.7

3.5 Stima di probabilità dalle osservazioni

Nell'analisi dei dati, spesso non conosciamo la distribuzione di probabilità sottostante. In questi casi, possiamo stimare la probabilità di un evento basandoci sulle osservazioni raccolte.

3.5.1 Approccio frequentista

L'approccio frequentista stima la probabilità di un evento A come il rapporto tra il numero di volte in cui si verifica l'evento e il numero totale di osservazioni effettuate:

$$P(A) \approx \frac{\text{numero di volte che si verifica } A}{\text{numero totale di osservazioni}}$$

Il problema di questo approccio è che la stima può essere imprecisa se il numero di osservazioni è limitato o se l'evento è raro.

3.5.2 Approccio bayesiano

L'approccio bayesiano utilizza il teorema di Bayes per aggiornare la stima della probabilità di un evento A .

Teorema di Bayes. Siano A e B due eventi con $P(B) > 0$. Allora la probabilità condizionata di A dato B è data da:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

In questo contesto, $P(A)$ rappresenta la probabilità a priori dell'evento A , mentre $P(A|B)$ è la probabilità a posteriori di A dopo aver osservato l'evento B .

Questo approccio incorpora la probabilità a priori e aggiorna la stima in base alle nuove evidenze, risultando particolarmente utile in situazioni con dati limitati o in presenza di eventi

rari. In questi contesti, infatti, vediamo la probabilità come una stima di *incertezza* piuttosto che una frequenza oggettiva.

Un esempio può essere la diagnosi medica, dove la probabilità di una malattia può essere aggiornata in base ai sintomi osservati e ai risultati dei test diagnostici.

3.6 Probabilità congiunta

In molti casi, è utile considerare la probabilità di due o più eventi che si verificano simultaneamente. Questa è nota come **probabilità congiunta**.

La probabilità congiunta di due eventi A e B è definita come la probabilità che entrambi gli eventi si verifichino.

Definizione 3.8

Un esempio comune è la probabilità che un paziente abbia sia la febbre che un'infezione batterica.

N.B. La probabilità congiunta non è una probabilità condizionata.

3.6.1 Regola della somma

Un modo di calcolare la probabilità congiunta è attraverso le tabelle di contingenza, che mostrano la frequenza con cui si verificano combinazioni di eventi.

Un altro modo è sfruttare un approccio frequentista, contando le occorrenze congiunte degli eventi nei dati osservati.

Si fa questo considerando una tabella di contingenza, ovvero una tabella che riporta le frequenze con cui si verificano le combinazioni di due o più variabili categoriali. Ad esempio, supponiamo di avere due eventi A e B con due possibili esiti ciascuno (vero/falso). La tabella di contingenza potrebbe essere strutturata come segue:

oprule	$Y = y_1$	$Y = y_2$...	$Y = y_l$	Total
$X = x_1$	n_{11}	n_{12}	...	n_{1l}	n_{1+}
$X = x_2$	n_{21}	n_{22}	...	n_{2l}	n_{2+}
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
$X = x_k$	n_{k1}	n_{k2}	...	n_{kl}	n_{k+}
Total	n_{+1}	n_{+2}	...	n_{+l}	n

Tabella 3.1: Esempio generale di tabella di contingenza

Dove:

- n_{ij} rappresenta il numero di osservazioni in cui l'evento $X = x_i$ e l'evento $Y = y_j$ si verificano contemporaneamente.
- n_{i+} è il totale delle osservazioni per l'evento $X = x_i$.
- n_{+j} è il totale delle osservazioni per l'evento $Y = y_j$.
- n è il numero totale di osservazioni.

Da questo, possiamo calcolare la probabilità congiunta $P(X = x_i, Y = y_j)$ come:

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{n}$$

Dalla tabella 3.1, la probabilità marginale è definita come:

$$P(X = x_i) = \frac{n_{i+}}{n}$$

$$P(Y = y_j) = \frac{n_{+j}}{n}$$

Definizione 3.9

Da qui, possiamo definire la probabilità marginale di un evento X come la somma delle probabilità congiunte su tutti i possibili esiti dell'altro evento Y :

$$P(X = x_i) = \frac{n_{i+}}{n} = \frac{\sum_j n_{ij}}{n} = \sum_j \frac{n_{ij}}{n} = \sum_j P(X = x_i, Y = y_j)$$

Questo risultato è conosciuto come **regola della somma**, che ci permette di calcolare la probabilità marginale di un evento sommando le probabilità congiunte con tutti gli esiti dell'altro evento (marginalizzazione).

3.7 Probabilità condizionata

In molte situazioni, è utile conoscere la probabilità di un evento dato che un altro evento si è verificato. Questa è nota come **probabilità condizionata**.

*La **probabilità condizionata** di un evento A dato che un evento B si è verificato è definita come:*

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

Definizione 3.10

Possiamo utilizzare di nuovo la tabella di contingenza per calcolare la probabilità condizionata. Ad esempio, la probabilità condizionata di $X = x_i$ dato $Y = y_j$ è:

$$P(X = x_i | Y = y_j) = \frac{\#\text{casi dove } X = x_i \text{ e } Y = y_j}{\#\text{casi dove } Y = y_j} = \frac{n_{ij}}{n} \cdot \frac{n}{n_{+j}} = \frac{n_{ij}}{n_{+j}} = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}$$

N.B. La probabilità condizionata è definita solo se $P(Y) > 0$, anche perché non si può definire la probabilità di un evento dato che un evento impossibile si è verificato: sarebbe un controsenso.

3.7.1 Regola del prodotto

Un modo alternativo di calcolare la probabilità congiunta è attraverso la regola del prodotto, che sfrutta la probabilità condizionata.

Possiamo dire, che la probabilità condizionata è:

$$P(X = x, Y = y) = \frac{P(X = x|Y = y)}{P(Y = y)}$$

Da qui segue:

$$P(X = x, Y = y) = P(X = x|Y = y) \cdot P(Y = y)$$

La probabilità del prodotto ci permette di calcolare la probabilità congiunta di due eventi moltiplicando la probabilità condizionata di uno degli eventi per la probabilità dell'altro evento.

3.8 Regola della catena per le probabilità condizionate

Quando si lavora con più variabili, la regola del prodotto può essere estesa per calcolare la probabilità condizionata di una variabile dato un insieme di altre variabili. Questa estensione è nota come **regola della catena**.

Sapendo che:

$$P(X, Y, Z) = P(X|Y, Z) \cdot P(Y, Z)$$

Possiamo espandere ulteriormente $P(Y, Z)$ usando la regola del prodotto:

$$P(Y, Z) = P(Y|Z) \cdot P(Z)$$

Sostituendo questa espressione nella precedente, otteniamo:

$$P(X, Y, Z) = P(X|Y, Z) \cdot P(Y|Z) \cdot P(Z)$$

Questa è la regola della catena per tre variabili. In generale, per n variabili X_1, X_2, \dots, X_n , la regola della catena si estende come segue:

$$P(X_1, X_2, \dots, X_n) = P(X_1) \prod_{i=2}^n P(X_i|X_1, X_2, \dots, X_{i-1})$$

3.9 Indipendenza

Due variabili aleatorie X e Y sono dette **indipendenti** se la conoscenza del valore di una non fornisce alcuna informazione sul valore dell'altra. In termini di probabilità, questo si traduce nella seguente condizione:

$$P(X|Y) = P(X) \cdot P(Y)$$

L'indipendenza si denota come:

$$X \perp Y$$

Se due variabili sono indipendenti, allora la probabilità condizionata di una variabile dato l'altra è semplicemente la probabilità marginale della prima variabile:

$$P(X|Y) = \frac{P(X, Y)}{P(Y)} = \frac{P(X) \cdot P(Y)}{P(Y)} = P(X)$$

3.9.1 Indipendenza condizionata

Due variabili aleatorie X e Y sono dette indipendenti condizionalmente rispetto a una terza variabile Z se, dato il valore di Z , la conoscenza del valore di X non fornisce alcuna informazione sul valore di Y , e viceversa. In termini di probabilità, questo si traduce nella seguente condizione:

$$P(X, Y|Z) = P(X|Z) \cdot P(Y|Z)$$

E quindi, si può denotare come:

$$X \perp Y|Z$$

Capitolo 4

Associazioni di variabili

Spesso, visualizzare statistiche su un campione di dati univariato non è sufficiente. Ad esempio, potremmo voler esaminare la relazione tra due variabili quantitative, come l'altezza e il peso di un gruppo di persone, si parla quindi di come due variabili sono associate tra loro.

Si parla di associazione tra due variabili se, osservando i valori assunti da una variabile, è possibile fare delle previsioni sui valori assunti dall'altra variabile.

Definizione 4.1

4.1 Misure di associazioni tra variabili discrete

4.1.1 Indipendenza

Prima di parlare di misure di associazione tra variabili discrete, è importante definire quando due variabili discrete sono **indipendenti**. Si parla di indipendenza, quando la conoscenza di una variabile non fornisce alcuna informazione sull'altra variabile.

La tabella di contingenza (tabella 3.1) può spiegare bene il concetto di indipendenza tra due variabili discrete: tutti quei valori, che sono marginali (cioè che riguardano una sola variabile, indicati con n_{i+} oppure n_{+j}) non forniscono alcuna informazione sui valori congiunti (cioè che riguardano entrambe le variabili, indicati con n_{ij}). Ipotizziamo quindi, di non conoscere i valori delle frequenze congiunte f_{ij} , ma solo le frequenze marginali $f_{i\cdot}$ e $f_{\cdot j}$:

oprule	$Y = y_1$	$Y = y_2$...	$Y = y_l$	Total
$X = x_1$	—	—	...	—	n_{1+}
$X = x_2$	—	—	...	—	n_{2+}
⋮	⋮	⋮	⋮	⋮	⋮
$X = x_k$	—	—	...	—	n_{k+}
Total	n_{+1}	n_{+2}	...	n_{+l}	n

Tabella 4.1: Tabella di contingenza (solo marginali)

Se dovessimo ricostruire i valori mancanti e le variabili sono indipendenti, potremmo dire:

$$P(X = x_i, Y = y_i) = P(X = x_i)P(Y = y_i)$$

Ricordando che:

$$P(X = x_i) = \frac{n_{i+}}{n} \quad , \quad P(Y = y_i) = \frac{n_{+i}}{n} \quad , \quad P(X = x_i, Y = y_i) = \frac{n_{ij}}{n}$$

Possiamo ricostruire le frequenze congiunte mancanti come:

$$\hat{n}_{ij} = n \cdot P(X = x_i, Y = y_i) = n \cdot P(X = x_i)P(Y = y_i) = n \cdot \frac{n_{i+}}{n} \cdot \frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}$$

Da cui possiamo scrivere:

$$\hat{n}_{ij} = P(X = x_i)n_{+j} = P(Y = y_j)n_{i+}$$

N.B.: utilizziamo \hat{n}_{ij} (e non n_{ij}) per indicare che si tratta di una stima della frequenza congiunta, calcolata assumendo l'indipendenza tra le variabili.

Questo ci dice, che date due variabili indipendenti, la frequenza congiunta stimata può essere calcolata a partire dalle frequenze marginali.

4.1.2 Statistica di Pearson

La statistica di Pearson, chiamata χ^2 , misura la discrepanza tra le frequenze congiunte osservate e quelle attese sotto l'ipotesi di indipendenza. Segue la formula:

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Dove k, l sono rispettivamente il numero di categorie delle variabili X e Y .

Questa statistica calcola il quadrato delle differenze (discrepanza) e lo normalizza rispetto alle frequenze attese, in modo da tenere conto delle diverse dimensioni delle categorie. Un valore elevato di χ^2 indica una maggiore discrepanza tra le frequenze osservate e quelle attese, suggerendo una possibile associazione tra le variabili. Al contrario, un valore basso suggerisce che le variabili sono probabilmente indipendenti.

Questa statistica tuttavia, ha dei problemi. Il valore atteso per ogni cella deve essere sufficientemente grande (almeno 5) per garantire l'accuratezza dell'approssimazione della distribuzione χ^2 . Inoltre, il valore di χ^2 dipende dalla dimensione del campione: campioni più grandi tendono a produrre valori più elevati di χ^2 , anche per associazioni deboli.

4.1.3 Statistica di Cramér

Per superare i limiti della statistica di Pearson, si può utilizzare la statistica di Cramér, che normalizza il valore di χ^2 per tener conto della dimensione del campione e del numero di categorie delle variabili. La formula è la seguente:

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, l-1)}}$$

Dove n è la dimensione del campione, e k, l sono rispettivamente il numero di categorie delle variabili X e Y .

Questa statistica ritorna un valore tra 0 e 1, dove 0 indica nessuna associazione tra le variabili e 1 indica un'associazione perfetta. Essendo normalizzata, la statistica di Cramér è meno influenzata dalla dimensione del campione rispetto a χ^2 , rendendola una misura più robusta dell'associazione tra variabili discrete.

4.1.4 Rischio relativo

Il rischio relativo (RR) è una misura utilizzata per quantificare l'associazione tra due variabili discrete, spesso in studi epidemiologici. Esso confronta la probabilità di un evento (ad esempio, una malattia) tra due gruppi distinti (ad esempio, esposti e non esposti a un fattore di rischio). La formula per calcolare il rischio relativo è la seguente:

$$RR = \frac{P(E|A)}{P(E|\neg A)}$$

Dove $P(E|A)$ è la probabilità dell'evento E dato l'esposizione A , e $P(E|\neg A)$ è la probabilità dell'evento E dato la non esposizione $\neg A$.

Per fare un esempio, consideriamo uno studio che esamina l'associazione tra il fumo (esposizione) e lo sviluppo di una malattia polmonare (evento). Supponiamo di avere i seguenti dati:

- Numero di fumatori che sviluppano la malattia: a
- Numero di fumatori che non sviluppano la malattia: b
- Numero di non fumatori che sviluppano la malattia: c
- Numero di non fumatori che non sviluppano la malattia: d

La probabilità di sviluppare la malattia tra i fumatori è:

$$P(E|A) = \frac{a}{a + b}$$

La probabilità di sviluppare la malattia tra i non fumatori è:

$$P(E|\neg A) = \frac{c}{c + d}$$

Pertanto, il rischio relativo è dato da:

$$RR = \frac{P(E|A)}{P(E|\neg A)} = \frac{a/(a + b)}{c/(c + d)}$$

Un valore di RR maggiore di 1 indica che l'esposizione è associata a un aumento del rischio dell'evento, mentre un valore inferiore a 1 indica una riduzione del rischio. Un valore di RR uguale a 1 suggerisce che non vi è alcuna associazione tra l'esposizione e l'evento.

4.1.5 Odds Ratio

L'Odds Ratio (OR) è una misura di associazione utilizzata per quantificare la forza della relazione tra due variabili discrete, spesso in studi caso-controllo. L'OR confronta le probabilità (odds) di

un evento tra due gruppi distinti. La formula per calcolare l’Odds Ratio è la seguente:

$$OR = \frac{P(E|A)/P(\neg E|A)}{P(E|\neg A)/P(\neg E|\neg A)}$$

Per fare lo stesso esempio precedente, consideriamo i seguenti dati:

- Numero di fumatori che sviluppano la malattia: a
- Numero di fumatori che non sviluppano la malattia: b
- Numero di non fumatori che sviluppano la malattia: c
- Numero di non fumatori che non sviluppano la malattia: d

La probabilità di sviluppare la malattia tra i fumatori è:

$$P(E|A) = \frac{a}{a+b}$$

La probabilità di non sviluppare la malattia tra i fumatori è:

$$P(\neg E|A) = \frac{b}{a+b}$$

La probabilità di sviluppare la malattia tra i non fumatori è:

$$P(E|\neg A) = \frac{c}{c+d}$$

Quindi l’Odds Ratio è dato da:

$$OR = \frac{(a/b)}{(c/d)} = \frac{a \cdot d}{b \cdot c}$$

Un valore di OR maggiore di 1 indica che l’esposizione è associata a un aumento delle probabilità dell’evento, mentre un valore inferiore a 1 indica una riduzione delle probabilità. Un valore di OR uguale a 1 suggerisce che non vi è alcuna associazione tra l’esposizione e l’evento.

4.2 Misure di associazioni tra variabili continue

Il problema delle variabili continue, rispetto a quelle discrete, è che non possiamo costruire una tabella di contingenza. Per questo motivo, utilizziamo misure di associazione basate sulla covarianza e sulla correlazione.

4.2.1 Visualizzazione grafica dell’associazione

Esistono modi grafici di visualizzare una possibile associazione tra due variabili continue, come lo **scatter plot** (sezione 17.4) che mostra i punti dati in un piano cartesiano, permettendo di osservare visivamente la relazione tra le due variabili. Si può creare una matrice di scatter plot per visualizzare le relazioni tra più variabili continue contemporaneamente, chiamata **scatter matrix** (sotto-sezione 17.4.1).

Un altro modo è utilizzare gli **hexbin plots** (sezione 17.5), ovvero una forma di istogramma a due dimensioni. In questo tipo di grafico, lo spazio bidimensionale viene suddiviso in esagoni (hexbin) e il colore di ciascun esagono rappresenta la densità dei punti dati in quella regione. Questo

è particolarmente utile per visualizzare grandi quantità di dati, poiché riduce il sovraffollamento e rende più facile identificare le aree con alta concentrazione di punti.

Un ulteriore modo per visualizzare l'associazione tra due variabili continue è attraverso i **grafici di densità e di contorno** (sezione 17.6). Questi grafici mostrano la distribuzione congiunta delle due variabili, evidenziando le aree di maggiore densità dei dati. I grafici di contorno, in particolare, utilizzano linee per collegare punti con la stessa densità, facilitando l'identificazione di pattern e relazioni tra le variabili.

4.2.2 Covarianza

La varianza misura la dispersione di una singola variabile rispetto alla sua media. La covarianza estende questo concetto a due variabili.

La covarianza misura come due variabili X, Y variano insieme rispetto alle loro medie \bar{x}, \bar{y} .

Si definisce come:

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^N (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})$$

Definizione 4.2

Esiste una versione campionaria della covarianza, che utilizza $N - 1$ al denominatore invece di N , per correggere il bias nella stima della covarianza dalla popolazione (molto importante per piccoli campioni):

$$Cov_{campione}(X, Y) = \frac{1}{N-1} \sum_{i=1}^N (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})$$

Notiamo che, in base al valore risultante da $(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})$, possiamo avere tre casi:

- Se il valore è positivo, significa che entrambe le variabili si discostano dalla loro media nello stesso verso (entrambe sopra o entrambe sotto la media), contribuendo positivamente alla covarianza.
- Se il valore è negativo, significa che le variabili si discostano dalla loro media in direzioni opposte (una sopra e l'altra sotto la media), contribuendo negativamente alla covarianza.
- Quanto il valore è più vicino allo zero, tanto meno le due variabili sono associate tra loro.

In un grafico cartesiano, i punti dati possono essere distribuiti in quattro quadranti:

- Primo quadrante (entrambe le variabili sopra la media): contribuisce positivamente alla covarianza.
- Secondo quadrante (una variabile sopra la media e l'altra sotto): contribuisce negativamente alla covarianza.
- Terzo quadrante (entrambe le variabili sotto la media): contribuisce positivamente alla covarianza.
- Quarto quadrante (una variabile sotto la media e l'altra sopra): contribuisce negativamente alla covarianza.

Si noti che, la covarianza di due variabili uguali, è uguale alla varianza:

$$Cov(X, X) = \frac{1}{N} \sum_{i=1}^N (x^{(i)} - \bar{x})^2 = var$$

Il problema principale della covarianza è che il suo valore dipende dalle unità di misura delle variabili. Ad esempio, se una variabile è misurata in metri e l'altra in chilogrammi, la covarianza avrà unità di misura miste (metri·chilogrammi), rendendo difficile l'interpretazione del valore.

4.2.3 Coefficiente di correlazione di Pearson

Per superare il problema delle unità di misura nella covarianza, si utilizza il coefficiente di correlazione di Pearson, che normalizza la covarianza dividendo per il prodotto delle deviazioni standard delle due variabili. La formula è la seguente:

$$p(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{1}{N} \sum_{i=1}^N \frac{(x^{(i)} - \bar{x})}{\sigma_X} \cdot \frac{(y^{(i)} - \bar{y})}{\sigma_Y} = Cov(z(X), z(Y))$$

Dove σ_X e σ_Y sono le deviazioni standard delle variabili X e Y rispettivamente.

Il coefficiente di correlazione di Pearson varia tra -1 e 1, poiché è stato utilizzata la standardizzazione (z-scoring) z per eliminare le unità di misura:

$$z(X) = \frac{X - \bar{X}}{\sigma_X}$$

Un valore di 1 indica una correlazione positiva perfetta, $p(x, x)$, mentre un valore di -1 indica una correlazione negativa perfetta, $p(x, -x)$. Quanto un valore è più vicino a 0, tanto meno le due variabili sono associate tra loro. Si può anche interpretare il segno: un valore positivo indica che le variabili tendono a variare nella stessa direzione, mentre un valore negativo indica che tendono a variare in direzioni opposte.

Coefficiente di correlazione di Pearson per variabili discrete. Anche per le variabili discrete, si può calcolare il coefficiente di correlazione di Pearson, trattando le categorie come valori numerici, ma non ha senso farlo sulle variabili nominali (senza ordine intrinseco). Per le variabili ordinali, invece, si può assegnare un ordine numerico alle categorie e calcolare il coefficiente di correlazione di Pearson sui ranghi delle categorie.

Si noti tuttavia, che se calcolato nel caso specifico di una variabile continua e una dicotomica (0/1), il coefficiente di correlazione di Pearson è equivalente a un valore chiamato **punto-biseriale**.

Alla luce di tutto questo, il principale problema di questo coefficiente di correlazione, è che misura solo relazioni lineari tra le variabili. Se la relazione tra le variabili è non lineare, il coefficiente di correlazione di Pearson potrebbe non catturare adeguatamente l'associazione tra di esse.

4.2.4 Coefficiente di correlazione di Spearman

Il coefficiente di correlazione di Spearman è una misura non parametrica della correlazione tra due variabili. A differenza del coefficiente di Pearson, che si basa sui valori effettivi delle variabili,

il coefficiente di Spearman si basa sui ranghi dei valori. Per rango si intende la posizione di un valore all'interno di un insieme ordinato di valori e l'intuizione alla base di questa misura è che, se due variabili sono correlate in modo monotono (cioè una variabile tende ad aumentare quando l'altra aumenta, indipendentemente dalla forma della relazione), allora i loro ranghi dovrebbero essere correlati. La formula per calcolare il coefficiente di correlazione di Spearman, siano $R(x^{(i)})$, $R(y^{(i)})$ i ranghi associati alle variabili X , Y , è la seguente:

$$R = 1 - \frac{6 \sum_{i=1}^N (R(x^{(i)}) - R(y^{(i)}))^2}{N(N^2 - 1)}$$

Dove N è il numero di coppie di dati.

Questo risultato è normalizzato in $[-1, 1]$ e può essere interpretato in modo simile al coefficiente di Pearson: un valore di 1 indica una correlazione positiva perfetta tra i ranghi, -1 indica una correlazione negativa perfetta, e 0 indica nessuna correlazione tra i ranghi.

4.2.5 Coefficiente di correlazione di Kendall

Il coefficiente di correlazione di Kendall, noto anche come τ di Kendall, è un'altra misura non parametrica della correlazione tra due variabili basata sui ranghi. A differenza del coefficiente di Spearman, che si basa sulle differenze tra i ranghi, il coefficiente di Kendall si basa sul concetto di coppie concordanti e discordanti. Una coppia di osservazioni $(x^{(i)}, y^{(i)})$ e $(x^{(j)}, y^{(j)})$ è considerata concordante se l'ordine dei ranghi è lo stesso per entrambe le variabili:

$$(x^{(i)} - x^{(j)})(y^{(i)} - y^{(j)}) > 0$$

Al contrario, è considerata discordante se l'ordine dei ranghi è opposto:

$$(x^{(i)} - x^{(j)})(y^{(i)} - y^{(j)}) < 0$$

Il coefficiente τ di Kendall è calcolato come:

$$\tau = \frac{(C - D)}{\frac{1}{2}N(N - 1)}$$

Dove C è il numero di coppie concordanti, D è il numero di coppie discordanti, e N è il numero totale di osservazioni.

Questo risultato è normalizzato in $[-1, 1]$ e può essere interpretato in modo simile al coefficiente di Pearson: un valore di 1 indica una correlazione positiva perfetta tra i ranghi, -1 indica una correlazione negativa perfetta, e 0 indica nessuna correlazione tra i ranghi.

Matrice di correlazione. Una matrice di correlazione è una tabella che mostra i coefficienti di correlazione tra più variabili (sezione ??). Ogni cella della matrice rappresenta il coefficiente di correlazione tra due variabili specifiche. La matrice è simmetrica, poiché il coefficiente di correlazione tra la variabile X e la variabile Y è lo stesso del coefficiente tra Y e X . La diagonale principale della matrice contiene sempre il valore 1, poiché ogni variabile è perfettamente correlata con se stessa.

Da questa matrice, può nascere una heatmap (sezione 17.7): una rappresentazione grafica della matrice di correlazione, in cui i valori dei coefficienti di correlazione sono rappresentati da colori. Le heatmap facilitano l'identificazione visiva delle relazioni tra le variabili, evidenziando rapidamente quali coppie di variabili sono fortemente correlate (positive o negative) e quali non lo sono.

4.3 Consigli utili su come scegliere la misura di associazione

Di seguito alcuni suggerimenti pratici, organizzati per tipo di variabili e obiettivo dell'analisi.

Discrete - Discrete: se entrambe le variabili sono categoriali usare tabelle di contingenza, statistica di Pearson (χ^2) per testare indipendenza, e Cramér's V per ottenere una misura normalizzata dell'intensità dell'associazione. Per studi caso-controllo o epidemiologici valutare anche Odds Ratio o Rischio Relativo.

Continuous - Continuous: valutare prima la relazione visiva con uno scatter plot. Usare il coefficiente di Pearson per relazioni lineari (assunzione: approssimativamente normale e assenza di outlier influenti). Se la relazione è monotona ma non lineare usare Spearman o Kendall (ranghi) che sono più robusti agli outlier.

Ordinal - Ordinal: i coefficienti basati sui ranghi (Spearman, Kendall) sono i più appropriati perché rispettano l'ordine senza imporre distanze metriche arbitrarie.

Più variabili / esplorazione: utilizzare matrici di correlazione (per variabili continue), heatmap, scatter-matrix e analisi di regressione (lineare o non lineare) per stimare l'effetto condizionato di una variabile su un'altra mentre si controlla per confondenti.

Capitolo 5

Distribuzione dei dati

Il problema dei dati è che non sempre la singola probabilità di un evento è sufficiente a descrivere il fenomeno che stiamo studiando. Spesso infatti è necessario considerare l'insieme delle possibili occorrenze di un evento.

5.1 Distribuzione di probabilità

La distribuzione di probabilità di una variabile casuale descrive come le probabilità sono distribuite tra i possibili valori che la variabile può assumere.

Definizione 5.1

Nel caso di variabili casuali discrete, la distribuzione di probabilità viene denominata **funzione di massa di probabilità** (pmf, *probability mass function*), mentre nel caso di variabili casuali continue viene chiamata **funzione di densità di probabilità** (pdf, *probability density function*).

Una distribuzione di probabilità caratterizza completamente una variabile casuale, in quanto fornisce tutte le informazioni necessarie per calcolare le probabilità di qualsiasi evento associato a quella variabile. Denotiamo con $P(X)$ la distribuzione di probabilità della variabile casuale X e possiamo scrivere che " **X segue la distribuzione $P(X)$** " come:

$$X \sim P(X)$$

5.2 Distribuzioni discrete

Una distribuzione di probabilità discreta è utilizzata per variabili casuali che possono assumere solo un numero finito o numerabile di valori.

5.2.1 Funzione di massa di probabilità (PMF)

La **funzione di massa di probabilità** (PMF) è la funzione che descrive la probabilità di ciascun valore possibile.

Una PMF sulla variabile casuale X è definita come:

$$P : \Omega \rightarrow [0, 1]$$

La PMF deve soddisfare una proprietà:

$$\sum_{x \in \Omega} P(x) = 1$$

Ipotizzando di simulare il lancio di un dado a sei facce, la PMF associata a questo esperimento è:

$$P(X = x) = \begin{cases} \frac{1}{6} & \text{se } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{altrimenti} \end{cases}$$

Questo significa che ogni faccia del dado ha una probabilità di $\frac{1}{6}$ di essere estratta, e la somma delle probabilità di tutte le facce è uguale a 1, come richiesto dalla proprietà della PMF.

5.2.2 Funzione di distribuzione cumulativa (CDF)

Se consideriamo il caso dove la variabile casuale X può essere ordinata (come nel caso dei numeri interi), possiamo definire la **funzione di distribuzione cumulativa** (CDF, *cumulative distribution function*) come:

$$F(x) = P(X \leq x) = \sum_{y \leq x} P(X = y)$$

La CDF fornisce la probabilità cumulativa che la variabile casuale X assuma un valore minore o uguale a x (per esempio, tutte le persone con altezza minore o uguale a 170 cm).

5.3 Distribuzioni continue

Una distribuzione di probabilità continua è utilizzata per variabili casuali che possono assumere un infinito numero di valori all'interno di un intervallo. Tuttavia, per poter capire il concetto di distribuzione continua, bisogna prima introdurre il concetto di funzione di densità di probabilità (PDF).

5.3.1 Funzione di densità di probabilità (PDF)

La **funzione di densità di probabilità** (PDF) è la funzione che descrive la densità di probabilità in ogni punto dell'intervallo. Una PDF sulla variabile casuale X è definita come:

$$f : \Omega \rightarrow [0, 1]$$

E deve soddisfare la seguente proprietà:

$$\int f(x) dx = 1$$

(*Questa condizione è equivalente alla somma delle probabilità nella distribuzione discreta*).

Si può utilizzare la PDF per calcolare la probabilità che la variabile casuale X assuma un valore all'interno di un intervallo specifico $[a, b]$:

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

La pdf descrive la densità di probabilità in ogni punto, ma non fornisce direttamente la probabilità di un singolo punto, poiché in una distribuzione continua la probabilità di un singolo punto è sempre zero: si parla infatti di densità di probabilità e non di probabilità puntuale. Tanto è più grande il valore della pdf in un punto, tanto più alta è la probabilità che la variabile casuale assuma valori vicini a quel punto.

PDF uniforme. Un esempio di PDF è la distribuzione uniforme continua su un intervallo $[a, b]$, dove la PDF è definita come:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{se } a \leq x \leq b \\ 0 & \text{altrimenti} \end{cases}$$

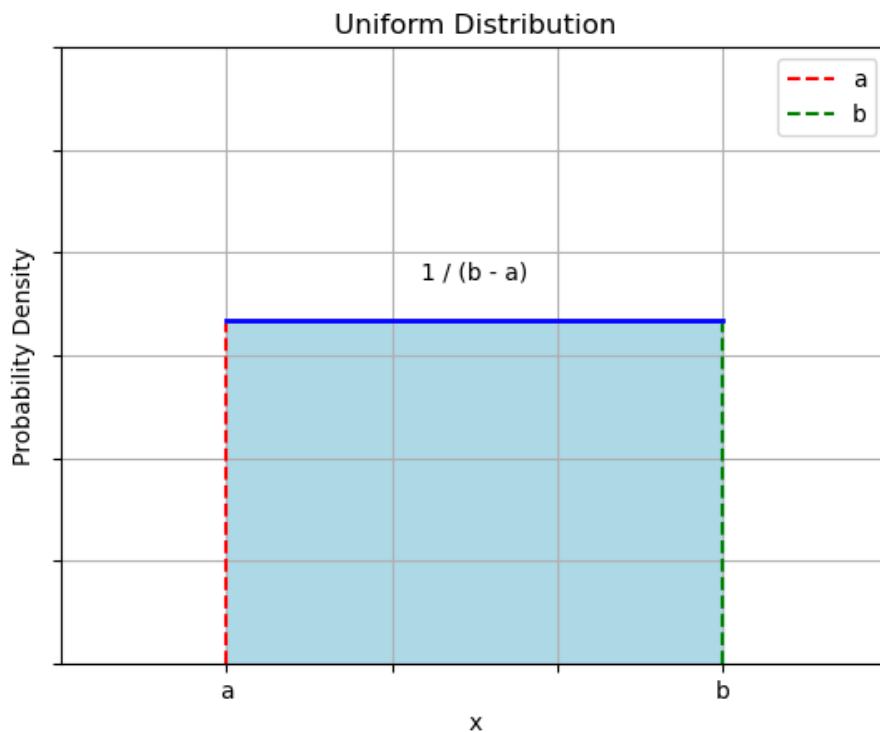


Figura 5.1: Esempio di PDF uniforme continua su un intervallo $[a, b]$. La funzione è costante all'interno dell'intervallo e zero al di fuori.

Approssimare la PDF con istogrammi. In pratica, quando si lavora con dati continui, spesso si cerca di approssimare la PDF utilizzando istogrammi (sezione 17.1).

Per un certo bin b_j sia h_j l'altezza dell'istogramma e sia w_j la larghezza del bin. Allora possiamo dire, per un istogramma normalizzato¹, che l'altezza del bin è:

$$h_j = \frac{c_j}{n}$$

dove c_j è il conteggio degli elementi nel bin b_j e n è il numero totale di elementi.

¹Un istogramma è detto normalizzato quando l'area totale sotto l'istogramma è pari a 1. Ovviamente non soddisfano i requisiti di una PDF, ma possono comunque fornire un'approssimazione utile.

Da questo, possiamo approssimare l'area dell'istogramma nel bin b_j come:

$$A_j = h_j \cdot w_j = \frac{c_j}{n} \cdot w_j$$

Calcolata quest'area per un j -esimo bin, possiamo calcolare l'area dell'istogramma per intero come:

$$\int_{-\infty}^{\infty} H(X) dx = \sum_j A_j = \sum_j \frac{c_j}{n} \cdot w_j = \sum_j w_j \neq 1$$

Per soddisfare la proprietà di una PDF, dobbiamo normalizzare l'istogramma in modo che l'area totale sia pari a 1. Per fare ciò, possiamo dividere l'altezza di ogni bin per l'area totale dell'istogramma:

$$h'_j = \frac{c_j}{n \cdot \sum_j w_j}$$

e da qui riscrivere l'area del bin normalizzato come:

$$\int_{-\infty}^{\infty} H'(X) dx = \sum_j A'_j = \sum_j h'_j \cdot w_j = \sum_j \frac{c_j}{n \cdot \sum_j w_j} \cdot w_j = 1$$

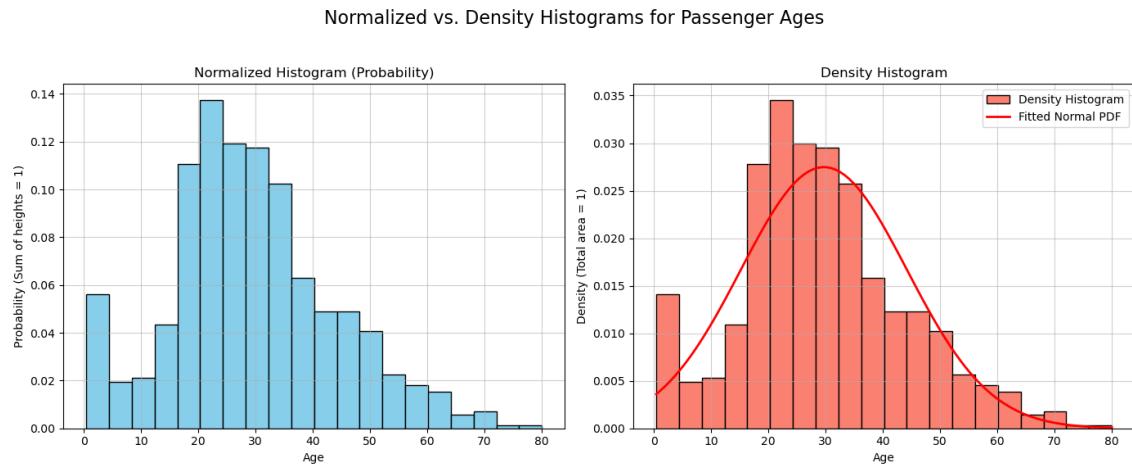


Figura 5.2: Confronto tra due rappresentazioni dell'istogramma delle età dei passeggeri. A sinistra: istogramma normalizzato, in cui l'altezza di ciascun bin è pari alla frequenza relativa (la somma delle altezze dei bin è 1); l'asse y è interpretato come probabilità. A destra: istogramma di densità, in cui l'area totale sotto le barre è 1 e quindi le altezze approssimano la funzione di densità di probabilità continua. In rosso è mostrata la PDF normale stimata sui dati.

5.3.2 Funzione di distribuzione cumulativa (CDF)

Ricordando che la CDF è definita come la probabilità cumulativa che la variabile casuale X assuma un valore minore o uguale a x , possiamo estendere questa definizione anche alle distribuzioni continue:

$$F(x) = P(X \leq x)$$

La CDF fornisce la probabilità cumulativa che la variabile casuale X assuma un valore minore

o uguale a x . In una distribuzione continua, la CDF è ottenuta integrando la PDF:

$$F(x) = \int_{-\infty}^x f(t) dt$$

Dove $f(t)$ è la PDF della variabile casuale X .

Possiamo trarre diverse conclusioni da qui:

- La CDF è una funzione non decrescente, poiché le probabilità cumulative non possono diminuire al crescere di x .
- Conoscendo la CDF, possiamo ottenere la PDF derivando la CDF:

$$f(x) = \frac{d}{dx} F(x)$$

- La CDF di una distribuzione continua è continua, a differenza della CDF di una distribuzione discreta che può presentare salti.

5.4 Distribuzioni di probabilità comuni

Ci sono diverse distribuzioni di probabilità comuni che vengono spesso utilizzate in statistica e machine learning. Molto spesso si utilizzano perché si nota che la distribuzione che stiamo studiando si avvicina a una di queste distribuzioni standard e in quel caso si possono sfruttare le loro proprietà per fare inferenza statistica.

5.4.1 Distribuzione uniforme discreta

La distribuzione uniforme discreta è una distribuzione di probabilità in cui tutti i valori possibili di una variabile casuale discreta hanno la stessa probabilità di verificarsi al variare di un parametro k :

$$P(X = a_i) = \frac{1}{k}$$

Dove $\Omega = \{a_1, \dots, a_k\}$.

5.4.2 Distribuzione di Bernoulli

La distribuzione di Bernoulli è una distribuzione di probabilità discreta che descrive un esperimento con due possibili esiti: successo e fallimento. La distribuzione viene definita come un singolo parametro ϕ , che rappresenta la probabilità di successo. Quindi possiamo formulare la distribuzione di Bernoulli come:

$$P(X = x) = \begin{cases} \phi & \text{se } x = 1 \\ 1 - \phi & \text{se } x = 0 \end{cases}$$

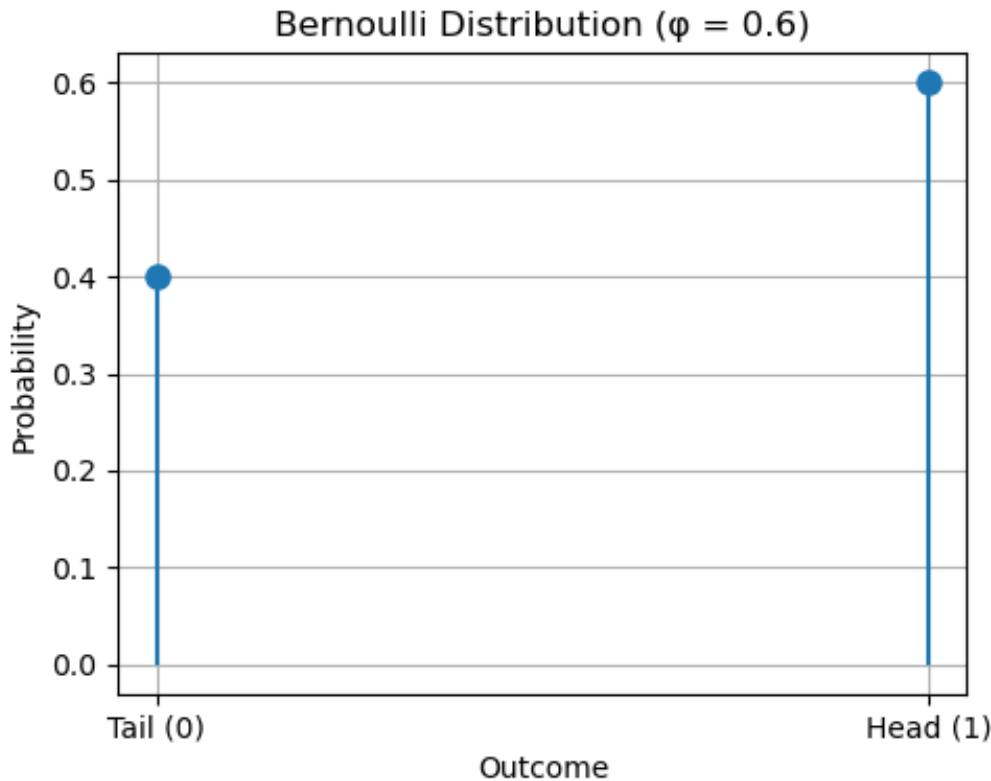


Figura 5.3: Esempio di distribuzione di Bernoulli su una moneta truccata con parametro $\phi = 0.6$. La probabilità di successo (ovvero fare testa) è 0.6, mentre la probabilità di fallimento (fare croce) è 0.4

5.4.3 Distribuzione binomiale

La distribuzione binomiale è una distribuzione di probabilità discreta (PMF) su numeri naturali con parametri n e p . Essa modella la probabilità di ottenere k successi in una sequenza di n esperimenti indipendenti che seguono una distribuzione di Bernoulli con parametro p .

La funzione di massa di probabilità è data da:

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Dove:

- k è il numero di successi
- n è il numero di prove indipendenti
- p è la probabilità di successo in una singola prova

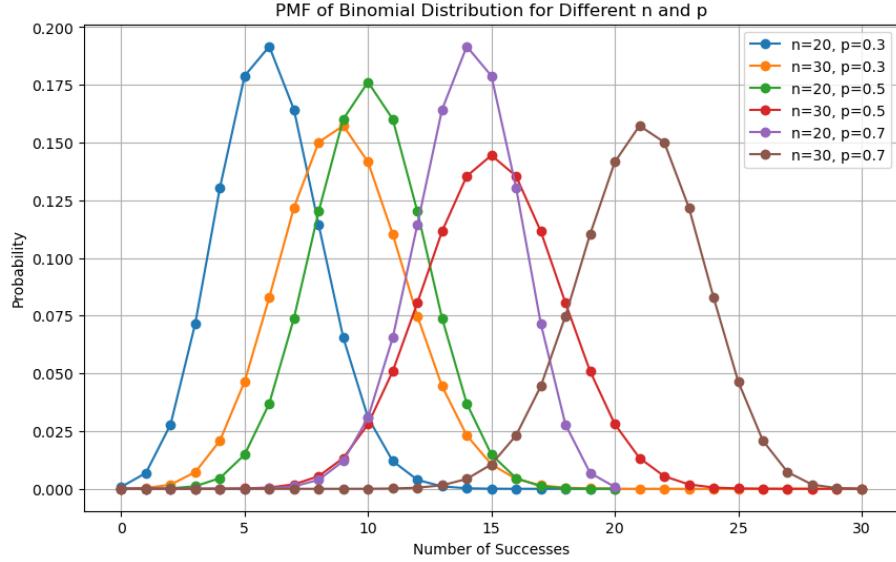


Figura 5.4: Funzioni di massa di probabilità (PMF) della distribuzione binomiale per diversi valori di n (numero di prove) e p (probabilità di successo in ciascuna prova). Ogni curva mostra la probabilità di ottenere k successi su n prove indipendenti: si osserva che il picco della distribuzione si sposta attorno al valore atteso np e che, al crescere di n , la distribuzione tende a diventare più concentrata e più simile a una forma "gaussiana".

Lancio di una moneta. Un esempio comune di distribuzione binomiale è il lancio di una moneta più volte, dove si vuole calcolare la probabilità di ottenere un certo numero di teste (successi) in un numero totale di lanci (prove):

$$P(k) = \binom{n}{k} \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{n-k} = \binom{n}{k} \left(\frac{1}{2}\right)^n$$

Dove n è il numero totale di lanci e k è il numero di teste ottenute. Se prendessimo $n = 3$ numero di lanci, $k = 3$ numero di successi (teste) e $p = 0.5$ probabilità di successo in ogni lancio, avremmo:

$$P(3) = \binom{3}{3} (0.5)^3 (1 - 0.5)^{3-3} = 1 \cdot (0.5)^3 \cdot 1 = 0.125$$

5.4.4 Distribuzione categorica

La distribuzione categorica è una generalizzazione della distribuzione di Bernoulli per variabili casuali che possono assumere k categorie distinte per un certo k finito.

La distribuzione categorica è definita da un vettore di probabilità $\mathbf{p} \in [0, 1]^k$, dove ogni p_i ci dà la probabilità di essere nell' i -esima categoria, e deve soddisfare la condizione:

$$\sum_{i=1}^k p_i = 1$$

La forma analitica è data da:

$$P(x = i) = p_i \quad \text{per } i = 1, 2, \dots, k$$

Esempio: lancio di un dado truccato. Consideriamo un singolo lancio di un dado a sei facce truccato. Gli esiti possibili sono:

$$X \in \{1, 2, 3, 4, 5, 6\}.$$

Se il dado è truccato in modo che le probabilità siano:

- $P(X = 1) = 0.10$
- $P(X = 2) = 0.15$
- $P(X = 3) = 0.20$
- $P(X = 4) = 0.20$
- $P(X = 5) = 0.10$
- $P(X = 6) = 0.25$

Allora l'esito segue una distribuzione categorica con $k = 6$ e probabilità specificate $\{p_1, p_2, \dots, p_6\} = \{0.1, 0.15, 0.2, 0.2, 0.1, 0.25\}$.

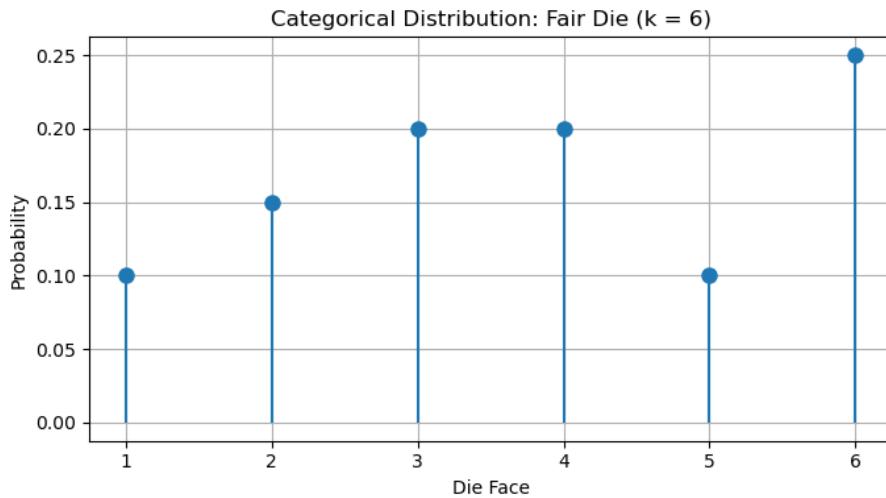


Figura 5.5: Esempio di distribuzione categorica per il lancio di un dado truccato. Le altezze delle barre rappresentano le probabilità associate a ciascuna faccia del dado.

5.4.5 Distribuzione multinomiale

La distribuzione multinomiale è una generalizzazione della distribuzione binomiale per esperimenti con più di due possibili esiti. Viene utilizzata per modellare il numero di occorrenze di ciascuna categoria in una serie di prove indipendenti. In particolare, la distribuzione multinomiale descrive la probabilità di ottenere esattamente (n_1, \dots, n_k) occorrenze per ciascuna delle k categorie in una sequenza di n esperimenti indipendenti che seguono una distribuzione categorica con probabilità (p_1, \dots, p_k) .

Possiamo definire i parametri della distribuzione multinomiale come:

- n : numero totale di prove indipendenti
- k : numero di categorie

- p_i : probabilità di successo per la categoria i (con $i = 1, 2, \dots, k$)²

Da questo, la funzione di massa di probabilità (PMF) della distribuzione multinomiale è data da:

$$P(n_1, n_2, \dots, n_k) = \frac{n!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}$$

Lancio di un dado più volte. Un esempio comune di distribuzione multinomiale è il lancio di un dado più volte, dove si vuole calcolare la probabilità di ottenere un certo numero di occorrenze per ciascuna faccia del dado in un numero totale di lanci. Considerato un dado a sei facce, qual è la probabilità di ottenere:

- 3 volte il numero 1, $n_1 = 3$
- 2 volte il numero 2, $n_2 = 2$
- 4 volte il numero 3, $n_3 = 4$
- 5 volte il numero 4, $n_4 = 5$
- 0 volte il numero 5, $n_5 = 0$
- 1 volta il numero 6, $n_6 = 1$

Sapendo che la probabilità di ottenere ciascuna faccia del dado è $p_i = \frac{1}{6}$ per ogni $i = 1, 2, \dots, 6$. Abbiamo $k = 6$ categorie e $n = 15$ lanci, quindi possiamo calcolare la probabilità come:

$$P(3, 2, 4, 5, 0, 1) = \frac{15!}{3!2!4!5!0!1!} \left(\frac{1}{6}\right)^3 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^4 \left(\frac{1}{6}\right)^5 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^1 = 8.04 \times 10^{-5}$$

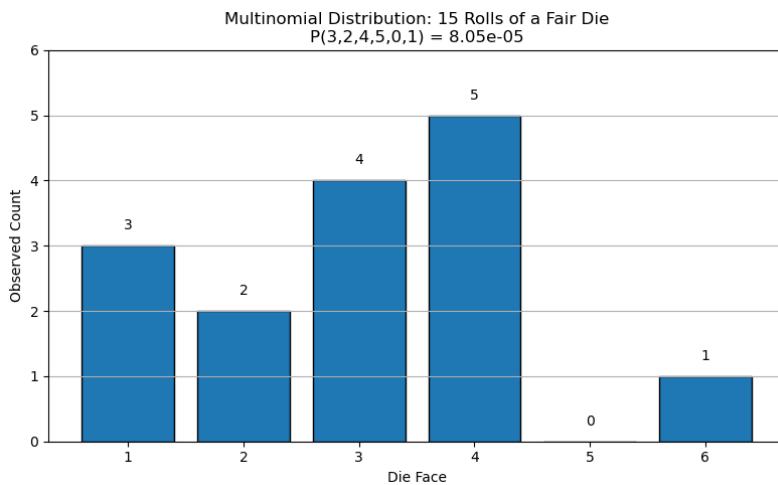


Figura 5.6: Esempio di distribuzione multinomiale per il lancio di un dado a sei facce 15 volte. Le altezze delle barre rappresentano le probabilità associate a ciascuna combinazione di occorrenze delle facce del dado.

²La somma totale di ogni p_i deve fare 1.

5.4.6 Distribuzione Gaussiana (Normale)

La distribuzione Gaussiana, o distribuzione normale, è una delle distribuzioni di probabilità più importanti e ampiamente utilizzate in statistica e machine learning. È molto importante perché, è una funzione di densità molto comune che descrive molti fenomeni naturali e processi casuali. La distribuzione Gaussiana è caratterizzata da due parametri principali:

- La media $\mu \in \mathbb{R}$ che rappresenta il centro della distribuzione.
- La deviazione standard $\sigma > 0$ che misura la dispersione dei dati intorno alla media.

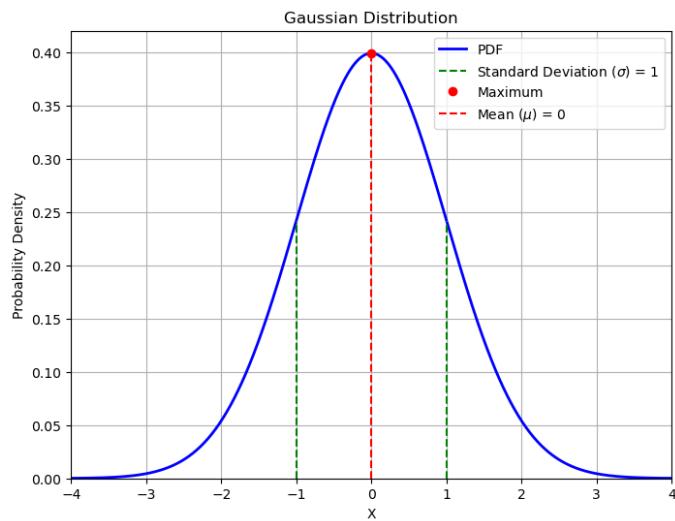


Figura 5.7: Esempio di distribuzione Gaussiana (Normale) con media $\mu = 0$ e deviazione standard $\sigma = 1$. La curva a campana rappresenta la funzione di densità di probabilità (PDF) della distribuzione.

La formula analitica della distribuzione normale è data da:

$$N(x, \mu, \sigma) = \sqrt{\frac{1}{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Questa è la **funzione di densità di probabilità** (PDF) della normale. Ogni pezzo della formula ha un significato preciso:

- Il termine μ è la media. Controlla *dove* è centrata la campana: il punto in cui la curva raggiunge il valore massimo è proprio $x = \mu$. Spostare μ verso destra o verso sinistra sposta tutta la distribuzione senza cambiarne la forma.
- Il termine σ è la deviazione standard. Controlla *quanto è larga o stretta* la campana. Se σ è piccolo, la distribuzione è molto concentrata attorno alla media (picco alto e stretto). Se σ è grande, la distribuzione è più sparsa (curva più bassa e larga). La varianza della distribuzione è σ^2 .
- Il fattore $\frac{1}{\sigma\sqrt{2\pi}}$, è un fattore di normalizzazione. Serve a garantire che l'area totale sotto la

curva valga 1, cioè che la funzione sia davvero una densità di probabilità valida:

$$\int_{-\infty}^{+\infty} N(x, \mu, \sigma) dx = 1.$$

Senza questo termine la forma sarebbe "a campana", ma non rappresenterebbe una distribuzione di probabilità corretta.

- Il termine esponenziale $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ determina la forma a campana. Il numeratore $(x - \mu)^2$ misura quanto x è lontano dalla media: più x è lontano, più questo quadrato cresce, e quindi più l'esponenziale crolla verso 0. Il denominatore $2\sigma^2$ controlla quanto velocemente crolla: con una σ grande, si scende più lentamente (code più larghe); con una σ piccola, si scende molto in fretta.

Possiamo notare una cosa molto interessante della distribuzione normale: calcolando l'intervallo di valori che si trovano entro una certa distanza dalla media, possiamo osservare che:

- Circa il 68% dei valori si trova entro una deviazione standard dalla media $(\mu - \sigma, \mu + \sigma)$.
- Circa il 95% dei valori si trova entro due deviazioni standard dalla media $(\mu - 2\sigma, \mu + 2\sigma)$.
- Circa il 99.7% dei valori si trova entro tre deviazioni standard dalla media $(\mu - 3\sigma, \mu + 3\sigma)$.

Questa cosa è utile per notare che, se una certa distribuzione segue la normale, allora possiamo fare delle stime sulla probabilità che un certo valore cada entro un certo intervallo dalla media. In pratica, la PDF ci dice quanto è "densa" la probabilità attorno a ciascun punto x , mentre l'area sotto la curva tra due punti a e b ci dice la probabilità che la variabile aleatoria cada proprio tra a e b . Questo è esattamente ciò che rende la normale così comoda per stimare intervalli di confidenza, fare assunzioni sui dati e modellare rumore nei modelli di machine learning.

5.4.7 Teorema del limite centrale

Il teorema del limite centrale è uno dei risultati più importanti in statistica.

Il teorema del limite centrale afferma che la distribuzione della somma (o della media) di un gran numero di variabili casuali $\{X_i\}_{i=1}^n$ indipendenti e identicamente distribuite tende a una distribuzione normale per $n \rightarrow \infty$, indipendentemente dalla distribuzione originale delle variabili casuali X_i .

Definizione 5.2

Questo risultato è fondamentale perché giustifica la diffusione della distribuzione normale in molti contesti statistici e applicazioni pratiche. In particolare, il teorema del limite centrale consente di utilizzare la distribuzione normale come approssimazione per la distribuzione di campioni di grandi dimensioni, anche quando la distribuzione originale non è normale.

Esempio. Consideriamo un dado equo a sei facce, in cui ogni faccia da 1 a 6 ha la stessa probabilità $\frac{1}{6}$. Se lanciamo il dado una sola volta, il risultato è chiaramente discreto (1, 2, 3, 4, 5 o 6) e la distribuzione non assomiglia affatto a una Gaussiana.

Adesso però facciamo così: lanciamo il dado n volte, calcoliamo la *media* dei risultati ottenuti, e ripetiamo questo esperimento molte volte. Per $n = 1$ le medie possibili coincidono con i valori

del dado, quindi la distribuzione è piatta e discreta. Per $n = 2, 5, 10$ la distribuzione delle medie campionarie comincia a diventare più "a campana". Per $n = 50$ e soprattutto $n = 5000$, le medie campionarie si concentrano attorno a circa 3.5 (che è il valore atteso di un dado equo) e la loro distribuzione è ormai molto simile a una distribuzione normale.

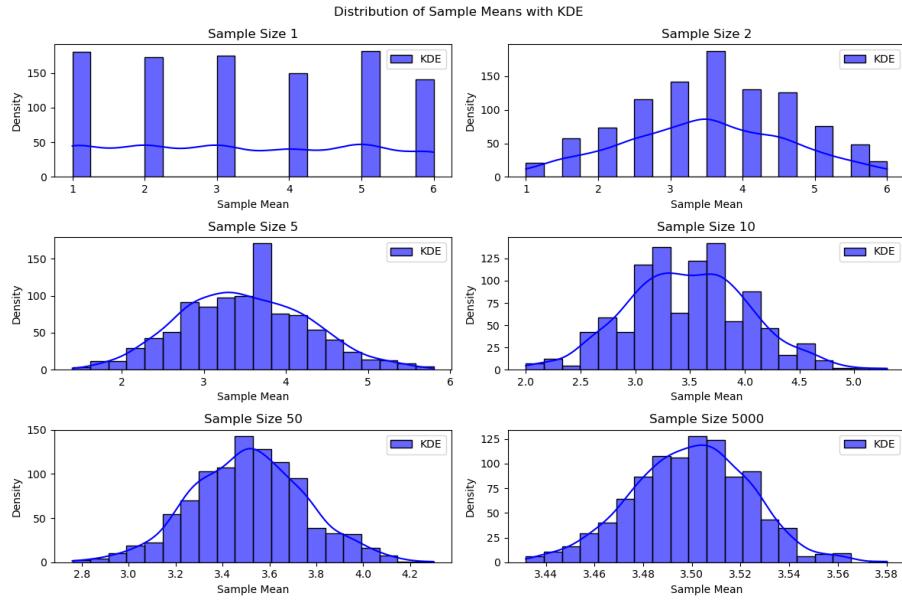


Figura 5.8: Esempio del teorema del limite centrale con il lancio di un dado equo. La distribuzione delle medie campionarie tende a una distribuzione normale.

5.4.8 Distribuzione Gaussiana Multivariata

La distribuzione Gaussiana multivariata è un'estensione della distribuzione normale a più dimensioni. Viene utilizzata per modellare vettori di variabili casuali continue che possono essere correlate tra loro. La distribuzione è caratterizzata da due parametri principali:

- Il vettore di medie $\mu \in \mathbb{R}^d$, che rappresenta il centro della distribuzione.
- La matrice di covarianza $\Sigma \in \mathbb{R}^{d \times d}$, che misura la dispersione e la correlazione tra le variabili.

La formulazione analitica della distribuzione Gaussiana multivariata è data da:

$$N(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{\frac{1}{(2\pi)^d |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

A due dimensioni, ovviamente, il vettore μ rappresenta il centro della distribuzione nel piano XY, mentre la matrice di covarianza Σ determina la forma e l'orientamento delle ellissi di densità di probabilità.

Spesso è difficile visualizzare la gaussiana multivariata in un grafico 3D, ma possiamo rappresentarla tramite curve di livello (linee di uguale densità) che mostrano come la densità di probabilità varia nello spazio bidimensionale.

Come si può notare nella figura 5.9, la distribuzione Gaussiana multivariata con media $(0, 0)$ e correlazione positiva tra le due variabili mostra un picco al centro (la media) e le linee di livello formano ellissi orientate, indicando sia la varianza lungo le direzioni principali sia la dipendenza lineare tra le due componenti.

Sotto la superficie 3D, le curve di livello rappresentano le linee di uguale densità sul piano XY. L'orientamento e la forma di queste ellissi sono determinate dalla matrice di covarianza Σ .

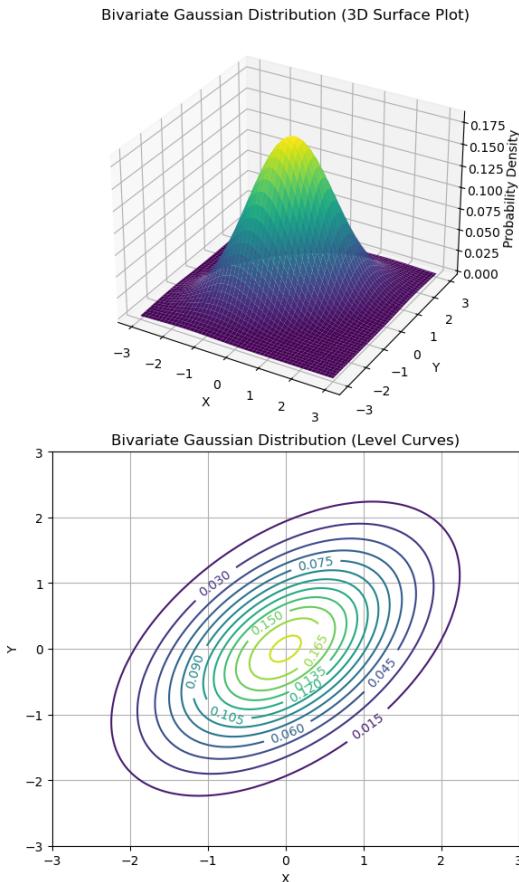


Figura 5.9: Distribuzione Gaussiana bivariata con media $(0, 0)$. La superficie 3D mostra la densità congiunta, mentre le curve di livello (ellissi) mostrano linee di uguale densità sul piano XY . L'orientamento delle ellissi è determinato dalla matrice di covarianza Σ .

Effetti di Σ . In una distribuzione normale a singola dimensione, l'effetto della dispersione è controllato dalla deviazione standard σ . Nella distribuzione Gaussiana multivariata, la matrice di covarianza Σ svolge un ruolo simile, ma in modo più complesso, in quanto determina **forma**, **orientamento e scala** della distribuzione nello spazio multidimensionale:

- Quando Σ è una matrice diagonale³, le variabili sono indipendenti tra loro (in quanto $\sigma_{xy} = \sigma_{yx} = 0$), e la distribuzione appare come un "ellissoide" allineato con gli assi.
 - Se la varianza lungo le assi è diversa (cioè gli elementi diagonali di Σ sono diversi), la distribuzione sarà più "allungata" in alcune direzioni rispetto ad altre.
 - Gli elementi fuori diagonale di Σ rappresentano la covarianza tra le variabili. Se questi valori sono positivi, indica una correlazione positiva (le variabili tendono a crescere insieme); se sono negativi, indica una correlazione negativa (una variabile tende a crescere mentre l'altra diminuisce). Questo si riflette nell'orientamento delle ellissi di densità di probabilità.

³Una matrice diagonale è una matrice quadrata in cui tutti gli elementi al di fuori della diagonale principale sono zero.

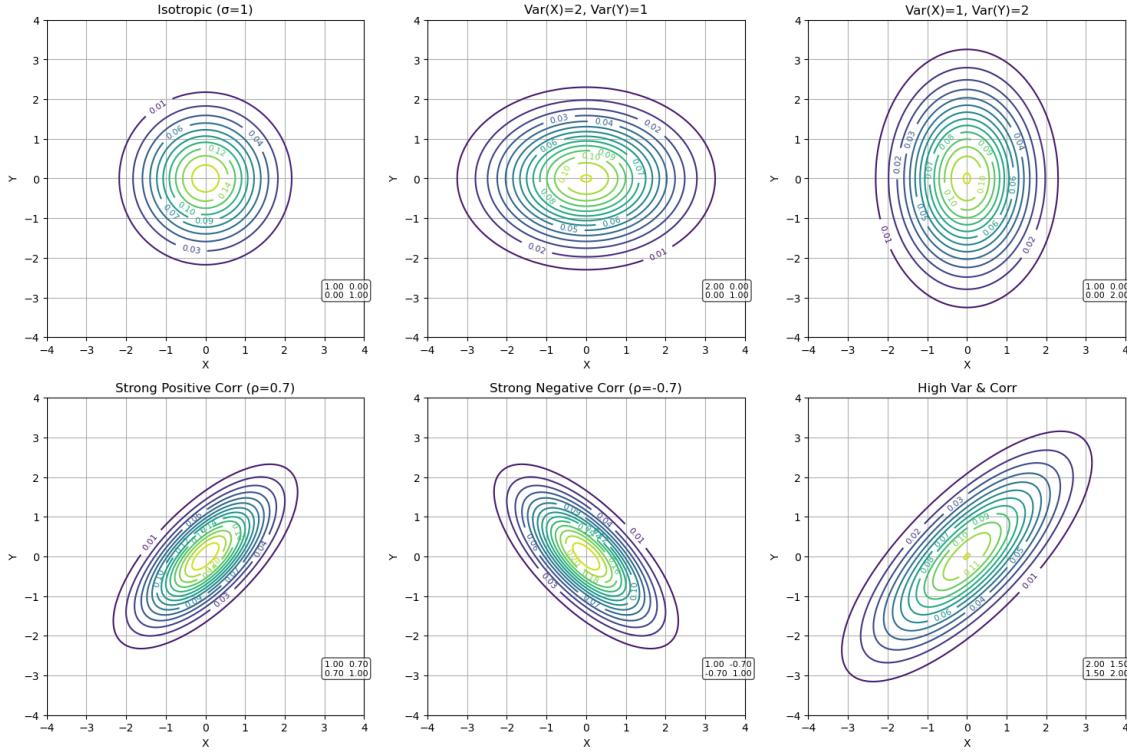


Figura 5.10: Effetto della matrice di covarianza Σ sulla distribuzione Gaussiana bivariata con media $(0, 0)$. Ogni pannello mostra le curve di livello (isodensità), che evidenziano forma, orientamento e scala della distribuzione. Nella riga superiore: caso isotropico (varianze uguali su X e Y), poi varianza di X maggiore di quella di Y , poi varianza di Y maggiore di quella di X ; in questi casi Σ è diagonale, quindi non c'è covarianza e le ellissi sono allineate con gli assi. Nella riga inferiore: presenza di termini fuori diagonale in Σ (covarianza non nulla), che introduce correlazione tra le variabili. Con correlazione positiva le ellissi ruotano lungo la diagonale crescente, con correlazione negativa lungo la diagonale decrescente. Questo mostra che Σ controlla non solo quanto è larga la distribuzione in ciascuna direzione, ma anche come è orientata nello spazio.

Stime dei parametri nella distribuzione Gaussiana multivariata. Per stimare i parametri della distribuzione Gaussiana multivariata, ovvero il vettore di medie μ e la matrice di covarianza Σ possono essere ottenuti grazie alla massima verosimiglianza (MLE, *maximum likelihood estimation*) sui dati osservati.

Nel caso **univariato**:

- La stima della media μ è data dalla media campionaria:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

- La stima della varianza σ^2 è data dalla varianza campionaria:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

Dove x_i sono i dati osservati e n è il numero totale di osservazioni.

Nel caso **multivariato**:

- La stima del vettore di medie μ è data dalla media campionaria vettoriale:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$$

- La stima della matrice di covarianza Σ è data dalla matrice di covarianza in relazione alla variabile casuale X :

$$\hat{\Sigma} = Cov(X) \Rightarrow \hat{\Sigma}_{ij} = Cov(X_i, X_j)$$

5.5 Descrivere una distribuzione di probabilità

Esistono diverse misure per descrivere una distribuzione di probabilità, che si basano su concetti statistici come la media, la varianza e la forma della distribuzione stessa.

5.5.1 Aspettativa (media)

L'aspettativa, o valore atteso, di una variabile casuale X è una misura della tendenza centrale della distribuzione di probabilità. È simile alla media aritmetica, perché quando computiamo una media non facciamo altro che sommare tutti i valori di un insieme e dividere per il numero totale di valori. L'aspettativa fa la stessa cosa, ma tiene conto delle probabilità associate a ciascun valore.

In particolare, possiamo definire per una variabile casuale discreta l'aspettativa come se fosse una media pesata:

$$E[X]_{X \sim P} = \sum_{x \in \Omega} x \cdot P(x)$$

Nel caso di variabili continue, l'aspettativa è definita come:

$$E[X]_{X \sim P} = \int_{x \in \Omega} x \cdot f(x) dx$$

Dove $f(x)$ è la funzione di densità di probabilità (PDF) della variabile casuale X .

5.5.2 Varianza e deviazione standard

La varianza è una misura della dispersione dei valori di una variabile casuale rispetto alla sua media. Indica quanto i valori si discostano in media dall'aspettativa. La varianza di una variabile casuale X è definita come:

$$Var_{X \sim P}[X] = E[(X - E_{X \sim P}[X])^2]$$

Si utilizza spesso, però, la deviazione standard, che è la radice quadrata della varianza in quanto è più interpretabile:

$$\sigma = \sqrt{Var_{X \sim P}[X]}$$

5.5.3 Covarianza

La covarianza è una misura della relazione lineare tra due variabili casuali X e Y . Indica come le due variabili variano insieme. La covarianza è definita come:

$$Cov_{X \sim P_X, Y \sim P_Y}[X, Y] = E[(X - E_{X \sim P_X}[X])(Y - E_{Y \sim P_Y}[Y])]$$

Si possono distinguere i termini:

- $E[X], E[Y]$ sono le aspettative (medie) delle variabili casuali X e Y .
- $(X - E_{X \sim P_X}[X])$ e $(Y - E_{Y \sim P_Y}[Y])$ sono le deviazioni delle variabili casuali X e Y rispetto alle loro aspettative.
- $(X - E[X])(Y - E[Y])$ rappresenta il prodotto delle deviazioni, che indica come le due variabili variano insieme.

Possiamo distinguere tre casi:

- Se la covarianza è positiva, significa che quando una variabile aumenta, l'altra tende ad aumentare anch'essa.
- Se la covarianza è negativa, significa che quando una variabile aumenta, l'altra tende a diminuire.
- Tanto la covarianza è vicina a zero, tanto meno le due variabili sono correlate linearmente.

Se X è una variabile multidimensionale con d componenti, la covarianza può essere rappresentata come una matrice di covarianza $\Sigma \in \mathbb{R}^{d \times d}$, dove ogni elemento Σ_{ij} rappresenta la covarianza tra la i -esima e la j -esima componente di X :

$$\Sigma_{ij} = Cov[X_i, X_j]$$

5.5.4 Entropia

L'entropia è una misura della quantità di incertezza o imprevedibilità associata a una distribuzione di probabilità. In altre parole, l'entropia quantifica quanto "disordinata" o "casuale" è una variabile casuale.

Self-information. Prima di definire l'entropia, è utile introdurre il concetto di **self-information** (o informazione auto-riferita) di un evento x , che misura la quantità di informazione associata al verificarsi di quell'evento. La self-information è definita come:

$$I(x) = -\log P(x)$$

Il logaritmo viene solitamente calcolato in base 2 (bit) o in base e (nat).

Questa definizione ha senso perché eventi con bassa probabilità (cioè eventi rari) forniscono più informazione quando si verificano, mentre eventi con alta probabilità (eventi comuni) forniscono meno informazione. Possiamo anche notare che:

- Il logaritmo rende la self-information additiva per eventi indipendenti. Se due eventi x e y sono indipendenti, allora:

$$I(X = x, Y = y) = -\log[P(X = x) \cdot P(Y = y)] = I(x) + I(y)$$

- Inoltre il logaritmo negativo riflette il fatto che eventi più probabili (con $P(x)$ vicino a 1) hanno self-information più bassa, mentre eventi meno probabili (con $P(x)$ vicino a 0) hanno self-information più alta.

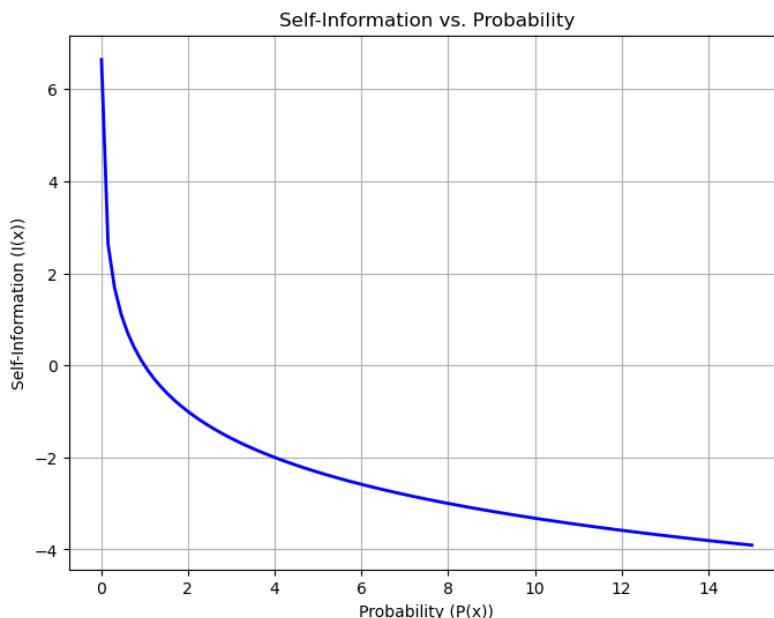


Figura 5.11: Esempio di self-information $I(x) = -\log P(x)$ per diversi valori di probabilità $P(x)$. Si nota che eventi con bassa probabilità (vicino a 0) hanno alta self-information, mentre eventi con alta probabilità (vicino a 1) hanno bassa self-information.

Entropia di una distribuzione. Per una certa variabile casuale X con distribuzione di probabilità $P(X)$, l'entropia $H(X)$ è definita come l'aspettativa della self-information:

- Per una variabile casuale discreta:

$$H(X) = E_{X \sim P}[I(X)] = - \sum_{x \in \Omega} P(x) \log P(x)$$

- Per una variabile casuale continua:

$$H(X) = E_{X \sim P}[I(X)] = - \int_{x \in \Omega} f(x) \log f(x) dx$$

Entropia di una variabile di Bernoulli Un esempio semplice è l'entropia di una variabile casuale di Bernoulli con parametro ϕ (probabilità di successo):

$$H(X) = -[\phi \log \phi + (1 - \phi) \log(1 - \phi)]$$

Si può notare che l'entropia è massima quando $\phi = 0.5$ (massima incertezza) e minima quando $\phi = 0$ o $\phi = 1$ (nessuna incertezza), come si vede nell'immagine 5.12.

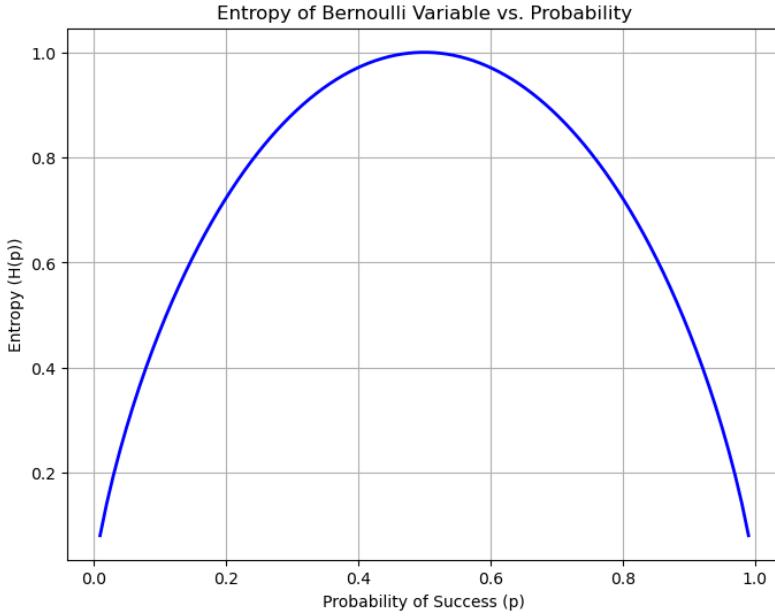


Figura 5.12: Entropia $H(X)$ di una variabile casuale di Bernoulli in funzione del parametro ϕ . L'entropia è massima a $\phi = 0.5$ e minima a $\phi = 0$ o $\phi = 1$.

5.5.5 Standardizzazione

La standardizzazione è una tecnica utilizzata per trasformare una variabile casuale in modo che abbia media zero e deviazione standard uno. Questo processo è utile per confrontare variabili con scale diverse o per preparare i dati per algoritmi di machine learning che sono sensibili alla scala delle caratteristiche. La standardizzazione di una variabile casuale X con media μ e deviazione standard σ è data dalla formula:

$$Z = \frac{X - \mu_X}{\sigma_X} = \frac{X - E[X]}{\sqrt{Var[X]}}$$

Dove Z è la variabile standardizzata. Dopo la standardizzazione, Z avrà una media di 0 e una deviazione standard di 1. Questa tecnica è anche chiamata z-scoring.

Come si evince dalla figura 5.13, la standardizzazione consente di confrontare direttamente la distribuzione della variabile casuale con una distribuzione normale standard, facilitando l'analisi statistica.

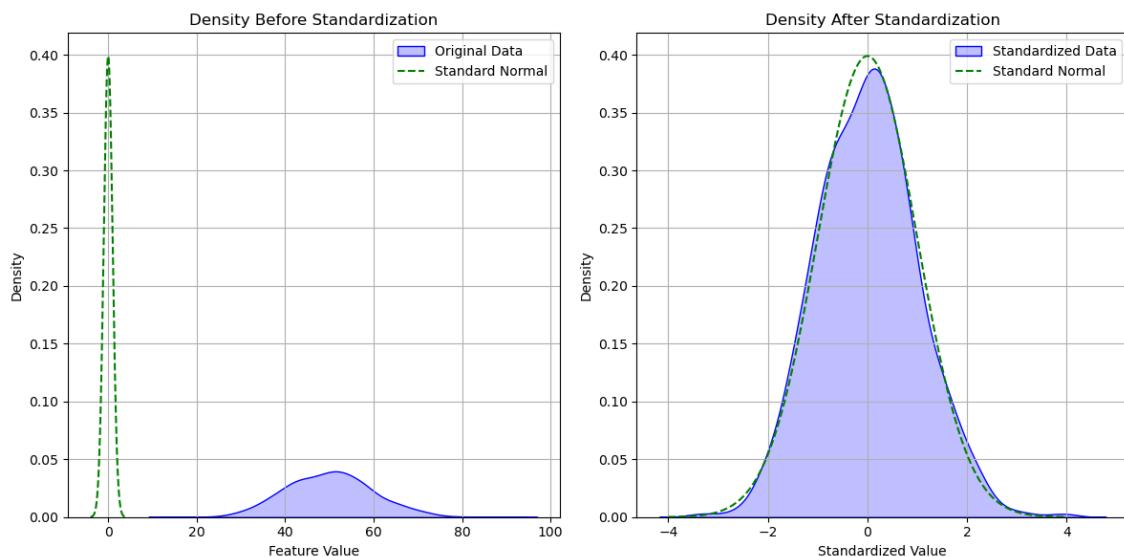


Figura 5.13: Effetto della standardizzazione su una variabile casuale. A sinistra: distribuzione originale della variabile (in blu), confrontata con una normale standard (media 0, deviazione standard 1) mostrata in verde tratteggiato: le scale sono diverse, quindi le curve non sono confrontabili direttamente. A destra: la stessa variabile dopo standardizzazione $Z = \frac{X - \mu_X}{\sigma_X}$; ora i dati trasformati hanno media 0 e deviazione standard 1 e risultano allineati alla distribuzione normale standard.

Capitolo 6

Inferenza Statistica

Molto spesso, in statistica, ci si trova nella situazione di dover prendere decisioni o fare previsioni basate su dati campionari. L'inferenza statistica fornisce gli strumenti necessari per trarre conclusioni riguardo a una popolazione più ampia a partire da un campione limitato di dati.

6.1 Campionamento

Il campionamento è il processo di selezione di un sottoinsieme di individui, oggetti o osservazioni da una popolazione più grande. Quando scarichiamo un dataset, il campionamento è stato già effettuato per noi, mentre se collezionassimo i dati dovremmo campionare dalla popolazione per intero.

6.1.1 Campionamento casuale semplice

Il modo più facile di selezionare un campione dalla popolazione è in **modo casuale**. Questo tipo di campionamento è detto **campionamento casuale semplice** (simple random sampling). Si fanno due assunzioni principali:

- Ogni elemento ha la stessa probabilità di essere selezionato. (selezione equi-probabile).
- La selezione di un elemento non influenza la selezione di un altro elemento (selezione indipendente).

Grazie a questo approccio garantiamo che, per un grande numero di campioni, le proprietà del campione riflettano quelle della popolazione.

Un problema di questo tipo di campionamento è che, in pratica, è difficile da realizzare. Inoltre non è sempre detto che le assunzioni di selezione equi-probabile e indipendente siano soddisfatte: Ipotizziamo di chiedere agli abitanti di una città se sono soddisfatti dei servizi pubblici, ma lo facciamo solo in centro città. In questo caso, la selezione non è equi-probabile, in quanto gli abitanti delle periferie non hanno la stessa probabilità di essere selezionati rispetto a quelli del centro città.

Inoltre, in questo tipo di campionamento è molto importante il numero di campioni che si estraggono: più campioni si estraggono, più le proprietà del campione tenderanno a riflettere quelle della popolazione.

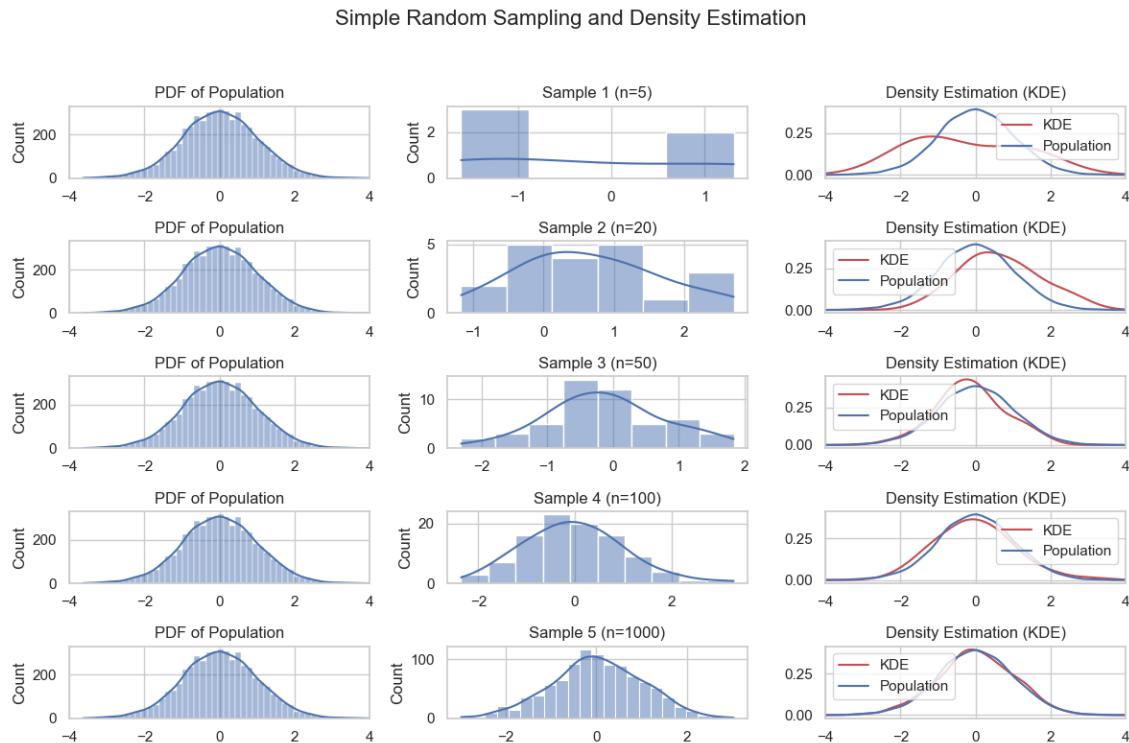


Figura 6.1: Illustrazione del campionamento casuale semplice da una popolazione normale. Ogni riga corrisponde a un diverso numero di osservazioni campionate dalla stessa popolazione ($n = 5, 20, 50, 100, 1000$). Nella prima colonna è mostrata la distribuzione della popolazione (istogramma ad alta risoluzione con la sua densità teorica). Nella seconda colonna è mostrato l'istogramma del singolo campione estratto a quella dimensione n , con una stima di densità sovrapposta. Nella terza colonna è mostrata la stima di densità (KDE, in rosso) del campione confrontata con la densità della popolazione reale (in blu). All'aumentare della dimensione del campione, l'istogramma e la densità stimata del campione diventano via via più simili alla distribuzione originale della popolazione: questo evidenzia che campioni più grandi approssimano meglio la popolazione.

6.1.2 Campionamento stratificato

Uno dei problemi riscontrabili durante il campionamento è che la popolazione è **eterogenea**, ovvero è composta da sottogruppi con caratteristiche diverse. In questi casi, il campionamento casuale semplice potrebbe non essere rappresentativo della popolazione intera, poiché alcuni sottogruppi potrebbero essere sottorappresentati o sovrarappresentati nel campione.

Per risolvere questa problematica si usa il **campionamento stratificato** (stratified sampling). In questo approccio, la popolazione viene suddivisa in sottogruppi omogenei chiamati **strati** (strata) basati su caratteristiche rilevanti (ad esempio età, genere, reddito). Successivamente, si esegue un campionamento casuale semplice all'interno di ciascuno strato.

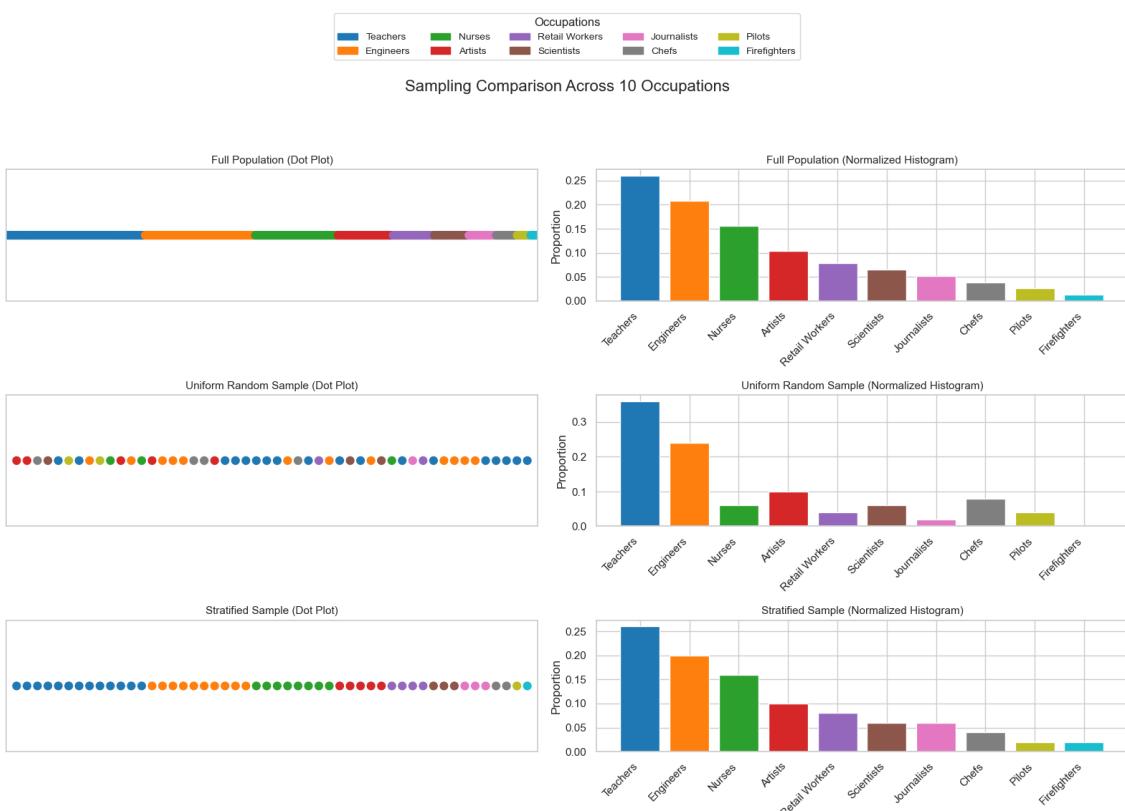


Figura 6.2: Confronto tra diversi metodi di campionamento su una popolazione suddivisa in 10 professioni. Riga superiore: popolazione completa, mostrata sia come dot plot (a sinistra, ogni punto è un individuo colorato per professione) sia come istogramma normalizzato (a destra, proporzioni reali di ciascuna professione nella popolazione). Riga centrale: campione ottenuto tramite campionamento casuale semplice (uniform random sample). Le proporzioni osservate nel campione possono discostarsi da quelle reali della popolazione, specialmente per le categorie meno frequenti. Riga inferiore: campione stratificato (stratified sample), in cui si forza la presenza di ogni categoria in proporzione alla popolazione. In questo caso, l'istogramma delle proporzioni nel campione è molto più fedele alla distribuzione originale.

Come si vede nella figura 6.2, il campionamento stratificato garantisce che ogni strato sia rappresentato nel campione in proporzione alla sua presenza nella popolazione, migliorando così la rappresentatività del campione sulla popolazione complessiva.

6.2 Campionare la distribuzione della media

Uno degli obiettivi principali dell'inferenza statistica è stimare parametri della popolazione, come la media o la varianza, a partire dai dati campionari. Un concetto fondamentale in questo contesto è la **distribuzione campionaria** (sampling distribution) di una statistica, che descrive come quella statistica varia da un campione all'altro. In particolare, la distribuzione campionaria della media campionaria è di grande interesse. La media campionaria è la media calcolata su un campione estratto dalla popolazione. Questo viene fatto perché spesso non si ha accesso all'intera popolazione, ma solo a un campione di essa.

Consideriamo questo esempio: un panificio pacchi di biscotti da 1kg ciascuno. Ogni pacco ha un peso che varia leggermente a causa delle variazioni nel processo di produzione. Supponiamo di prendere un campione randomico di $n = 1000$ pacchi e misurare il loro peso medio:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

Troviamo tuttavia, che il peso medio è uguale a 1000.2g, leggermente superiore al peso nominale di 1000g. Ci rendiamo conto che se ripetessimo l'estrazione di un campione di 1000 pacchi e calcolassimo nuovamente la media, otterremmo un valore leggermente diverso. Questo accade perché ogni campione è diverso e quindi la media campionaria varia da campione a campione.

Questo fenomeno ripetuto è descritto dalla **distribuzione campionaria della media** (sampling distribution of the sample mean). La distribuzione campionaria della media descrive come la media campionaria varia quando si estraggono ripetutamente campioni dalla popolazione. Trattando ogni pacco come una variabile casuale X_i che ha $E[X_i] = \mu$ e $Var(X_i) = \sigma^2$, la media campionaria \bar{X} è anch'essa una variabile casuale con:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Dal limite del teorema centrale (sotto-sezione 5.4.7) si sa che, per campioni sufficientemente grandi, la distribuzione campionaria della media tende a una distribuzione normale con media μ e varianza $\frac{\sigma^2}{n}$. Quindi possiamo scrivere:

$$\begin{aligned}\mathbb{E}[\bar{X}] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \mu \\ \text{Var}[\bar{X}] &= \text{Var}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}[X_i] = \frac{\sigma^2}{n} \\ \text{Std}[\bar{X}] &= \sqrt{\text{Var}[\bar{X}]} = \frac{\sigma}{\sqrt{n}}\end{aligned}$$

Possiamo notare diverse cose:

- La media della distribuzione campionaria della media tende alla media della popolazione μ . Questo significa che la media campionaria è uno stimatore non distorto della media della popolazione.
- La deviazione standard della distribuzione campionaria della media quantifica la precisione

della stima della media campionaria, in quanto un campione più grande (maggior n) riduce la variabilità della media campionaria intorno alla media della popolazione.

Tuttavia persiste un problema: nella maggior parte dei casi, non conosciamo la varianza della popolazione σ^2 .

6.2.1 Errore standard

Per risolvere il problema della varianza sconosciuta, si può stimare la varianza della popolazione utilizzando quella che si chiama "varianza campionaria" (sample variance):

$$s_{n-1} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Da questa misura, possiamo definire l'errore standard (standard error) della media campionaria come:

$$SE_{\bar{X}} = \frac{s_{n-1}}{\sqrt{n}}$$

L'errore standard fornisce una stima della variabilità della media campionaria intorno alla media della popolazione. È molto simile alla deviazione standard della distribuzione campionaria della media, ma utilizza la varianza stimata dal campione invece della varianza reale della popolazione.

Notiamo anche una cosa: ridurre l'errore standard è molto costoso, in quanto per dimezzare l'errore standard bisogna quadruplicare la dimensione del campione n perché l'errore standard decresce con la radice quadrata di n .

Questo può essere descritto bene in una figura del genere (figura 6.3):

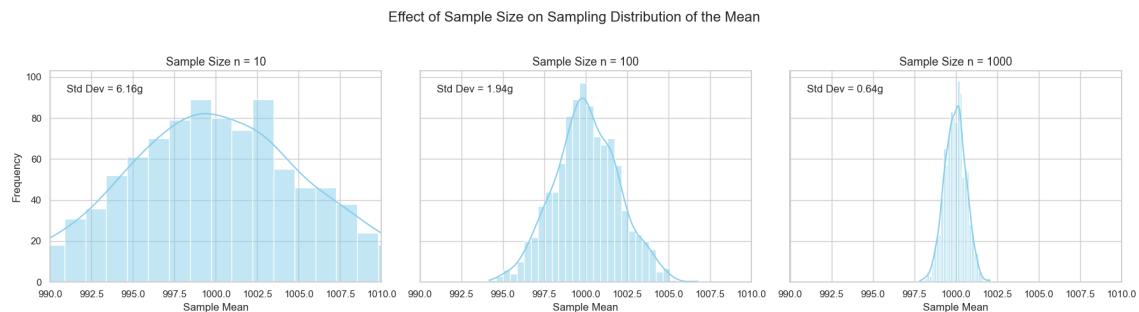


Figura 6.3: Distribuzione delle medie campionarie per diversi valori di n (10, 100, 1000) ottenute simulando il campionamento ripetuto dalla stessa popolazione. Ogni pannello mostra l'istogramma delle medie dei campioni e una stima di densità sovrapposta. All'aumentare della dimensione del campione, la distribuzione delle medie diventa più stretta attorno al valore medio della popolazione e la deviazione standard della media campionaria (errore standard) diminuisce in accordo con $Std[\bar{X}] = \sigma/\sqrt{n}$.

6.2.2 Distribuzione t-Student

Per risolvere il problema della stima su una piccola dimensione del campione, si può utilizzare la distribuzione t di Student (Student's t-distribution). Questa distribuzione è simile alla distribuzione

normale, ma ha code più pesanti, il che significa che c'è una maggiore probabilità di osservare valori estremi. Viene definita come:

$$t_{n-1} = \frac{\bar{X} - \mu}{SE_{\bar{X}}}$$

Dove $n - 1$ sono i gradi di libertà¹ (degrees of freedom). Al crescere di n la distribuzione t di Student si avvicina sempre più alla distribuzione normale (figura 6.4).

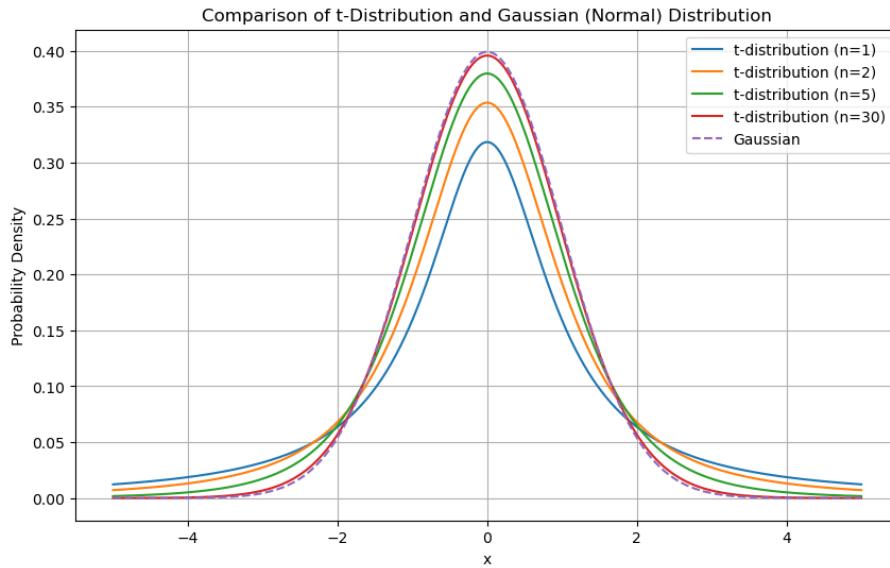


Figura 6.4: Confronto tra la distribuzione t di Student e la distribuzione normale standard. Ogni curva t corrisponde a un diverso numero di gradi di libertà ($n = 1, 2, 5, 30$). Per valori piccoli di n la distribuzione t ha code più pesanti (maggior probabilità di valori estremi) e un picco più basso rispetto alla Gaussiana. All'aumentare di n , la distribuzione t si avvicina alla distribuzione normale standard, fino a diventare praticamente indistinguibile per n grandi.

6.2.3 Intervallo di confidenza

Nel nostro esempio del panificio, se dovessimo riportare il peso medio dei pacchi di biscotti, potremmo voler includere una misura della nostra incertezza riguardo a questa stima. Un modo comune per farlo è attraverso un **intervallo di confidenza** (confidence interval). Un intervallo di confidenza fornisce un range di valori all'interno del quale ci aspettiamo che il vero parametro della popolazione (in questo caso, la media del peso dei pacchi) cada con una certa probabilità (ad esempio, il 95%).

Ricordando che la media campionaria \bar{X} segue una distribuzione normale centrata sulla media della popolazione μ e ricordando che per una distribuzione normale il 68% dei valori è compreso in $\mu \pm \sigma$ possiamo scrivere:

$$P(\mu - \sigma \leq \bar{X} \leq \mu + \sigma) = 0.68$$

¹I gradi di libertà rappresentano il numero di valori indipendenti che possono variare in un'analisi statistica. Nel caso della distribuzione t di Student, i gradi di libertà sono pari a $n - 1$ perché stiamo stimando la media della popolazione a partire da un campione di dimensione n .

E ovviamente segue che:

$$\bar{x} \in [\mu - \sigma, \mu + \sigma] \Leftrightarrow \mu \in [\bar{x} - \sigma, \bar{x} + \sigma]$$

E quindi possiamo stimare un intervallo di confidenza al 68% per la media della popolazione come:

$$P(\bar{x} - \sigma \leq \mu \leq \bar{x} + \sigma) = 0.68$$

Questo risultato può anche essere visto nel grafico della figura 6.5.

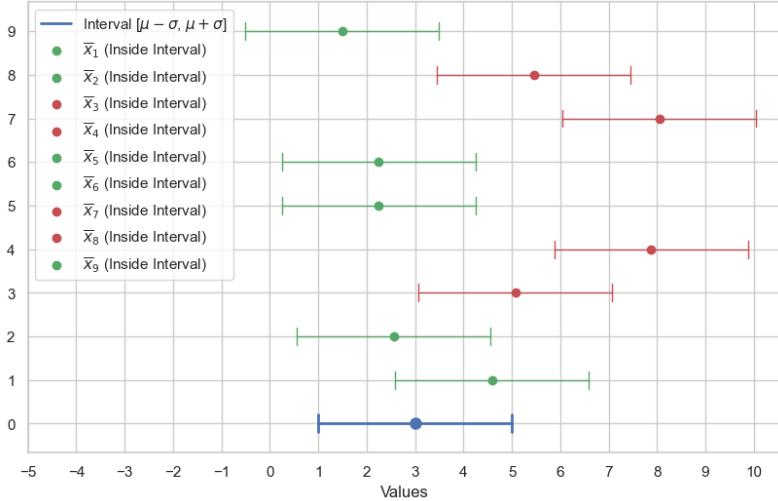


Figura 6.5: Visualizzazione dell'intervallo $[\mu - \sigma, \mu + \sigma]$ (barra orizzontale blu) e delle medie campionarie $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_9$ ottenute da campioni diversi. Per ciascun campione è mostrato il suo intervallo $\bar{X}_i \pm \sigma$ (barra orizzontale). I punti verdi indicano le medie campionarie i cui intervalli contengono la media reale μ , mentre i punti rossi indicano quelle che non la contengono. L'idea è che, ripetendo il campionamento, la maggior parte degli intervalli stimati copre il valore vero del parametro, ma non tutti.

Questo è un risultato molto potente, in quanto ci permette di quantificare l'incertezza associata alla nostra stima della media della popolazione. Tuttavia, è importante notare che l'intervallo di confidenza dipende dalla dimensione del campione e dalla variabilità dei dati: campioni più grandi e dati meno variabili portano a intervalli di confidenza più stretti, indicando una maggiore precisione nella stima della media della popolazione.

Chiamiamo quindi, in questo contesto, l'intervallo $[\bar{x} - \sigma, \bar{x} + \sigma]$ di confidenza al 68%.

Il problema però rimane che molto spesso non abbiamo a disposizione la deviazione standard della popolazione σ , ma solo quella del campione s_{n-1} , da cui ricaviamo lo standard error e lo sostituiamo nell'intervallo di confidenza:

$$[\bar{x} - SE_{\bar{X}}, \bar{x} + SE_{\bar{X}}]$$

Generalizzazione per altri livelli di confidenza. In generale, si può generalizzare questa formulazione per un certa percentuale di confidenza p chiamata **livello di confidenza** (confidence level). Da questo, formuliamo la probabilità come:

$$p = P(\bar{x} - \beta\sigma \leq \mu \leq \bar{x} + \beta\sigma)$$

Che porta a:

$$[\bar{x} - \beta \cdot SE_{\bar{X}}, \bar{x} + \beta \cdot SE_{\bar{X}}]$$

Livello di significatività. Possiamo scrivere una formulazione alternativa basata su un parametro α definito come **livello di significatività** (significance level). In questo contesto α rappresenta la probabilità che l'intervallo di confidenza non contenga il vero parametro della popolazione. Quindi, se vogliamo un intervallo di confidenza del 95%, il livello di significatività sarà $\alpha = 0.05$. Quindi:

- Con probabilità $1 - \alpha$, l'intervallo di confidenza contiene il vero parametro della popolazione, quindi "cattura" μ .
- Con probabilità α , l'intervallo di confidenza non contiene il vero parametro della popolazione, quindi "manca" μ .

Se scegliessimo $\alpha = 0.05$, avremmo un intervallo di confidenza del 95% dato da:

$$[\bar{x} - \beta \cdot SE_{\bar{X}}, \bar{x} + \beta \cdot SE_{\bar{X}}]$$

Dove β è scelto in modo tale che l'area sotto la curva della distribuzione normale tra $-\beta$ e $+\beta$ sia pari a $1 - \alpha$. Per $\alpha = 0.05$, β è approssimativamente uguale a 1.96.

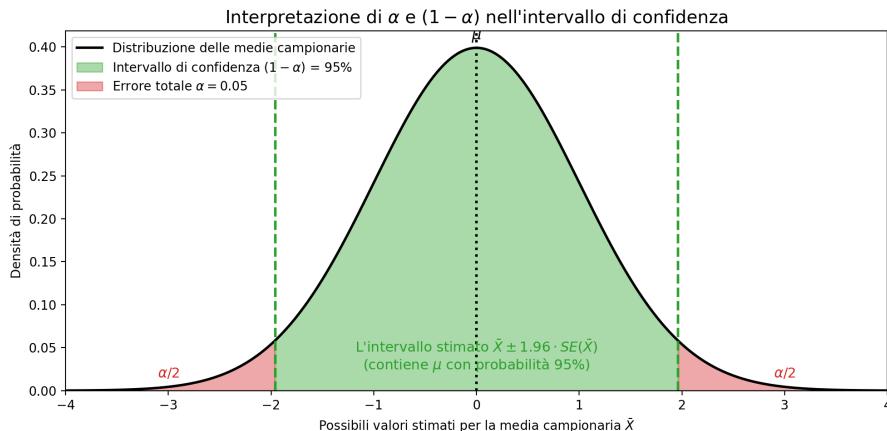


Figura 6.6: Interpretazione dell'intervallo di confidenza al 95%. L'area verde rappresenta l'intervallo di confidenza $(1 - \alpha) = 95\%$, che corrisponde ai valori della media campionaria \bar{X} che, una volta stimati sul campione, "catturano" la vera media μ . Le aree rosse nelle code (ciascuna di area $\alpha/2$) rappresentano i casi in cui l'intervallo stimato non contiene la media reale: la probabilità complessiva di errore è $\alpha = 0.05$. Le linee tratteggiate verdi indicano i limiti $\bar{X} \pm 1.96 \cdot SE(\bar{X})$, mentre la linea tratteggiata nera indica la vera media μ .

6.3 Bootstrapping

Il bootstrapping è una tecnica di inferenza statistica che consente di stimare la distribuzione di una statistica campionaria quando è piccola e non segue una distribuzione normale. Questa tecnica si basa sul concetto di campionamento con reinserimento dal campione originale per creare nuovi campioni chiamati **campioni bootstrap**.

L'idea alla base del bootstrapping è semplice:

1. Si parte con il campione di dimensione n .
2. Si creano B nuovi campioni bootstrap, ciascuno di dimensione n , estraendo casualmente con reinserimento dal campione originale. Il nuovo campione avrà dimensione originale ma alcuni elementi potrebbero essere ripetuti, mentre altri potrebbero non essere selezionati.
3. Si calcola la statistica di interesse (ad esempio, la media, la mediana, la varianza) per ciascun campione bootstrap.
4. Si ripetono gli step 2 e 3 per un numero elevato di volte per ottenere una distribuzione della statistica di interesse.
5. Si ricostruisce la distribuzione della statistica di interesse dai valori calcolati sui campioni bootstrap.

6.4 Stimatori

Nell'inferenza statistica, uno degli obiettivi principali è stimare i parametri della popolazione a partire dai dati campionari. Per fare questo, si introducono gli **stimatori** (estimators), che sono funzioni dei dati campionari utilizzate per stimare i parametri della popolazione. Un esempio è la media, che è uno stimatore della media della popolazione.

6.4.1 Stimatore del bias

Sia X una variabile casuale e siano $x = (x_1, x_2, \dots, x_n)$ un campione di dimensione n estratto da X . Sia $T(X)$ uno stimatore della quantità di popolazione ϕ :

$$T(X) = \frac{1}{n} \sum_{i=1}^n x_i$$

Poiché il campione cambia, anche il valore di $T(X)$ cambia. Quindi $T(X)$ è anch'esso una variabile casuale con una sua distribuzione.

Possiamo quindi definire il **bias** (bias) dello stimatore $T(X)$ come la differenza tra il valore atteso dello stimatore e il vero valore del parametro della popolazione:

$$\text{Bias}(T) = \mathbb{E}[T(X)] - \phi$$

Che è una misura di quanto lo stimatore si discosta in media dal vero parametro della popolazione. Se il bias è zero, lo stimatore è detto **non distorto** (unbiased), altrimenti è **distorto** (biased).

6.4.2 Stimatore della varianza

Varianza distorta. Se volessimo riprendere l'esempio dei biscotti, potremmo usare la varianza campionaria per stimare la varianza della popolazione:

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Il problema di questa stima è che è uno stimatore distorto della varianza della popolazione, in quanto il valore atteso è:

$$\mathbb{E}[s_n^2] = \frac{n-1}{n} \sigma^2$$

che è sempre minore della varianza reale della popolazione σ^2 :

$$\frac{n-1}{n} < 1 \quad \Rightarrow \quad \mathbb{E}[s_n^2] < \sigma^2$$

Quindi lo stimatore s_n^2 è distorto, con un bias negativo e va a sottostimare la varianza della popolazione. Per risolvere si può usare lo stimatore della varianza non distorto:

$$s_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

risolvendo il problema del bias perché ha valore atteso $\mathbb{E}[s_{n-1}^2] = \sigma^2$.

6.4.3 Varianza di uno stimatore

Grazie alla varianza possiamo misurare la precisione di uno stimatore. La **varianza di uno stimatore** è definita come:

$$Var(T(X)) = \mathbb{E}[(T(X) - \mathbb{E}[T(X)])^2]$$

Anche qui una bassa misura di varianza indica che lo stimatore è preciso, mentre una varianza alta indica che lo stimatore è meno preciso.

6.4.4 Bias-Varianza Tradeoff

Nella scelta di uno stimatore, spesso si deve affrontare un compromesso tra bias e varianza, noto come **bias-variance tradeoff**. Uno stimatore con un bias basso potrebbe avere una varianza elevata, mentre uno stimatore con una varianza bassa potrebbe avere un bias elevato. Si possono distinguere quattro casi principali, (illustrati nella figura 6.7):

6.5 Test statistici

I test statistici sono procedure utilizzate per prendere decisioni riguardo a una popolazione basandosi su dati campionari. Questi test permettono di valutare ipotesi specifiche riguardo a parametri della popolazione, come la media o la varianza, e di determinare se le osservazioni campionarie forniscono prove sufficienti per accettare o rifiutare tali ipotesi.

6.5.1 Test di ipotesi

Gli intervalli di confidenza e gli stimatori sono strettamente legati ai **test di ipotesi** (hypothesis testing), che sono procedure statistiche utilizzate per prendere decisioni riguardo a una popolazione basandosi su dati campionari. La differenza sta nel fatto che mentre gli intervalli di confidenza forniscono un range di valori plausibili per un parametro della popolazione, i test di ipotesi valutano la validità di una specifica affermazione riguardo a quel parametro.

- **Basso bias, bassa varianza:** Le stime sono vicine tra loro e vicine al valore vero. Lo stimatore è sia accurato che stabile (caso ideale).
- **Basso bias, alta varianza:** Le stime sono in media corrette (attorno al valore vero), ma molto disperse. Lo stimatore è accurato in media, ma instabile.
- **Alto bias, bassa varianza:** Le stime sono tutte raggruppate, ma lontane dal valore vero. Lo stimatore è sistematicamente sbilanciato, ma coerente.
- **Alto bias, alta varianza:** Le stime sono lontane dal valore vero e molto disperse. È il caso peggiore: impreciso e inaccurato.

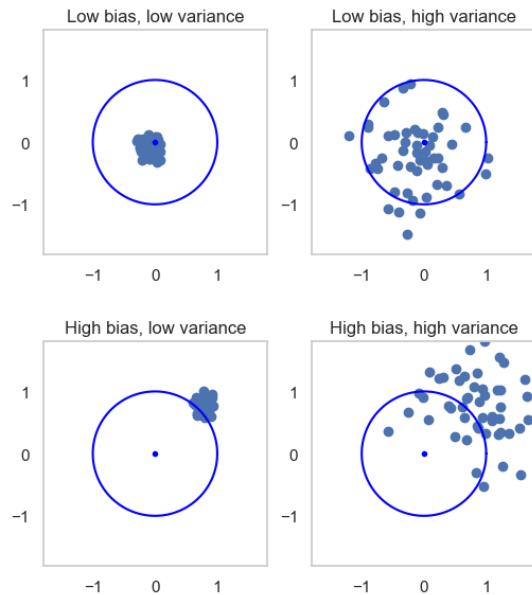


Figura 6.7: Illustrazione concettuale di bias e varianza con l'analogia del bersaglio. Ogni pannello dell'immagine mostra una serie di stime (punti blu) rispetto al valore vero (centro del bersaglio). Il modo in cui i punti si distribuiscono rispetto al centro riflette combinazioni diverse di bias (quanto siamo lontani dal valore vero) e varianza (quanto le stime sono stabili tra loro).

Si definisce:

- **Ipotesi nulla** (null hypothesis, H_0): è l'ipotesi di base che si vuole testare. Spesso rappresenta uno stato di "nessun effetto" o "nessuna differenza".
- **Ipotesi alternativa** (alternative hypothesis, H_a): è l'ipotesi che si vuole sostenere se i dati forniscono prove sufficienti contro l'ipotesi nulla.

L'ipotesi nulla è l'ipotesi che stiamo cercando di smentire² e usiamo l'ipotesi alternativa come supporto per la nostra affermazione.

Prima di procedere con il testo, si sceglie un **livello di significatività** α , che rappresenta la probabilità di rifiutare l'ipotesi nulla H_0 quando essa è vera. Dopo ci chiediamo qual è la probabilità di osservare i dati campionari (o qualcosa di più estremo) sotto l'assunzione che l'ipotesi nulla sia vera. Questa probabilità è chiamata **valore p** (p-value).

Per fare questo, utilizziamo la statistica di test usando la distribuzione t di Student (sezione 6.2.2) perché ci dice come si comporta la media campionaria rispetto alla media della popolazione sotto l'ipotesi nulla.

Dopo aver calcolato il valore p, lo confrontiamo con il livello di significatività α :

- Se il valore p è minore o uguale a α , rifiutiamo l'ipotesi nulla H_0 a favore dell'ipotesi alternativa H_a .
- Se il valore p è maggiore di α , non rifiutiamo l'ipotesi nulla H_0 .

Esempio: una moneta è truccata? Immaginiamo di avere una moneta e di voler capire se è equa (ossia, se la probabilità di ottenere testa è uguale a quella di ottenere croce). Formuliamo le

²Un po' come avviene con le dimostrazioni per assurdo.

ipotesi:

- Ipotesi nulla H_0 : la moneta è equa, quindi $p = 0.5$.
- Ipotesi alternativa H_a : la moneta è truccata, quindi $p \neq 0.5$.

Facciamo l'esperimento: lanciamo la moneta 10 volte e otteniamo 9 teste e 1 croce. Fissiamo un livello di significatività $\alpha = 0.05$, ovvero siamo disposti ad accettare un 5% di rischio di rifiutare l'ipotesi nulla quando essa è vera (ovvero dire "la moneta è truccata" quando in realtà è equa).

Calcoliamo il valore p (che indica la probabilità di ottenere 9 o più teste in 10 lanci se la moneta fosse equa):

$$\text{Valore } p = P(X \geq 9) = P(X = 9) + P(X = 10) = \binom{10}{9} (0.5)^{10} + \binom{10}{10} (0.5)^{10} \approx 1.6\%$$

Con una moneta onesta ($P(\text{Testa}) = P(\text{Croce}) = 0.5$), la probabilità di ottenere 9 o più teste in 10 lanci è:

$$\binom{10}{9} (0.5)^{10} \cdot (0.5)^{10} + \binom{10}{10} (0.5)^{10} = \frac{10}{1024} + \frac{1}{1024} = \frac{11}{1024} \approx 1.07\%$$

Ma siccome stiamo testando $p \neq 0.5$, dobbiamo considerare anche l'altra coda della distribuzione (ovvero ottenere 1 o meno teste in 10 lanci):

$$\binom{10}{0} (0.5)^{10} + \binom{10}{1} (0.5)^{10} = \frac{1}{1024} + \frac{10}{1024} = \frac{11}{1024} \approx 1.07\%$$

Quindi se la moneta fosse onesta, un risultato così estremo capitrebbe solo nel:

$$p - \text{value} \approx 2 \cdot 1.07\% = 2.14\%$$

Adesso confrontiamo il p-value con α :

$$2.14\% < 5\%$$

Poiché il p-value è minore di α , rifiutiamo l'ipotesi nulla H_0 e concludiamo che c'è evidenza sufficiente per suggerire che la moneta è truccata (ipotesi alternativa H_a).

Tipi di errori nei test statistici. Il test di ipotesi è una tipologia di test che può portare a due tipi di errori: rigettare l'ipotesi nulla oppure non rigettarla e questa è una tipologia di classificazione binaria.

Come in tutte le classificazioni binarie, non si ha mai la probabilità del 100% di prendere la decisione giusta. Si possono quindi commettere due tipi di errori:

- **Errore di tipo I** (Type I error): si verifica quando si rifiuta l'ipotesi nulla H_0 quando essa è vera. La probabilità di commettere un errore di tipo I è pari al livello di significatività α .
- **Errore di tipo II** (Type II error): si verifica quando non si rifiuta l'ipotesi nulla H_0 quando essa è falsa. La probabilità di commettere un errore di tipo II è indicata con β .

Da questo, si può costruire una tabella di contingenza che riassume i possibili esiti di una classificazione binaria:

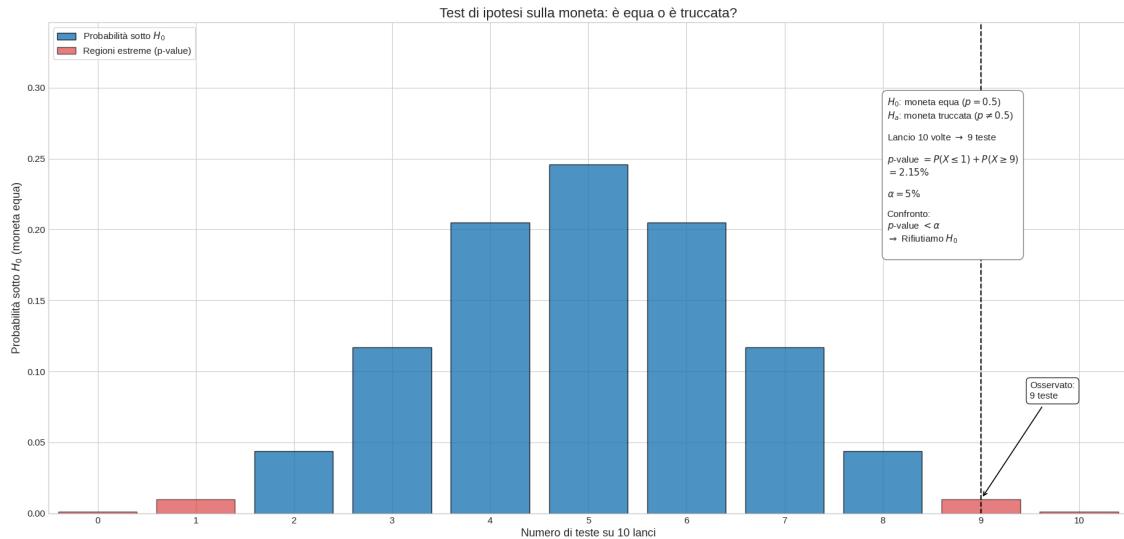


Figura 6.8: Esempio di test di ipotesi su una moneta. Le barre mostrano la probabilità di ottenere k teste su 10 lanci se la moneta fosse equa ($H_0 : p = 0.5$). Le barre rosse evidenziano gli esiti “estremi” ($k \leq 1$ o $k \geq 9$) che contribuiscono al valore p in un test bilaterale. Nell'esperimento osserviamo 9 teste (linea tratteggiata): questo caso cade in zona estrema. Il valore $p \approx 2.15\%$ è minore di $\alpha = 5\%$, quindi rifiutiamo H_0 e concludiamo che ci sono evidenze per dire che la moneta è truccata ($H_a : p \neq 0.5$).

	Ipotesi nulla vera	Ipotesi nulla falsa
Rifiuto	Errore di tipo I (α)	Vero positivo
Non rifiuto	Vero negativo	Errore di tipo II (β)

Tabella 6.1: Tabella di contingenza per i test di ipotesi.

6.5.2 T-test a un campione

Il t-test a un campione (one-sample t-test) è un test statistico utilizzato per determinare se la media di un campione differisce significativamente da un valore specifico della popolazione. Questo test è particolarmente utile quando la varianza della popolazione è sconosciuta e il campione è di dimensioni ridotte.

6.5.3 T-test a due campioni

Il t-test a due campioni (two-sample t-test) è un test statistico utilizzato per confrontare le medie di due gruppi indipendenti e determinare se esistono differenze significative tra di esse. Questo test è utile quando si vuole valutare l'effetto di un trattamento o di una condizione su due gruppi distinti.

6.5.4 Test χ^2 per indipendenza

Il test χ^2 per indipendenza è un test statistico utilizzato per determinare se esiste una relazione significativa tra due variabili categoriali. Questo test confronta le frequenze osservate in un campione con le frequenze attese se le due variabili fossero indipendenti. Utilizza il test di ipotesi, formulando come ipotesi nulla H_0 l'indipendenza tra le due variabili e come ipotesi alternativa H_a la dipendenza tra di esse.

È spesso accompagnato dalla statistica di Cramér, che misura la forza dell'associazione tra le due variabili categoriali.

6.5.5 Test χ^2 di bontà di adattamento

Il test χ^2 di bontà di adattamento è un test statistico utilizzato per determinare se un insieme di dati osservati si discosta significativamente da un modello teorico atteso. Questo test è spesso utilizzato per verificare se una distribuzione di frequenze osservate si adatta a una distribuzione attesa, come la distribuzione uniforme o la distribuzione normale. L'ipotesi nulla H_0 in questo caso afferma che non ci sono differenze significative tra le frequenze osservate e quelle attese, mentre l'ipotesi alternativa H_a suggerisce che ci sono differenze significative.

6.5.6 Test di correlazione di Pearson

Il test di correlazione di Pearson è un test statistico utilizzato per misurare la forza e la direzione della relazione lineare tra due variabili continue. Il coefficiente di correlazione di Pearson, denotato come r , varia tra -1 e 1, dove valori vicini a 1 indicano una forte correlazione positiva, valori vicini a -1 indicano una forte correlazione negativa, e valori vicini a 0 indicano nessuna correlazione lineare (sotto-sezione 4.2.3). L'ipotesi nulla H_0 afferma che non esiste una correlazione significativa tra le due variabili, mentre l'ipotesi alternativa H_a suggerisce che esiste una correlazione significativa.

6.5.7 Test di correlazione di Spearman

Il test di correlazione di Spearman, d'altra parte, è un test non parametrico utilizzato per misurare la forza e la direzione della relazione monotona tra due variabili ordinali o continue. Il coefficiente di correlazione di Spearman, denotato come ρ (rho), varia anch'esso tra -1 e 1, con interpretazioni simili a quelle del coefficiente di Pearson (sotto-sezione 4.2.4). L'ipotesi nulla H_0 afferma che non

esiste una correlazione monotona significativa tra le due variabili, mentre l'ipotesi alternativa H_a suggerisce che esiste una correlazione monotona significativa.

6.6 Valutare quando un campione è distribuito normalmente

Per valutare se un campione di dati segue una distribuzione normale, si possono utilizzare diversi metodi statistici e grafici.

6.6.1 Grafici Q-Q

I grafici Q-Q (quantile-quantile) sono strumenti grafici per confrontare la distribuzione di un campione con una distribuzione teorica (ad esempio una normale). Si mettono sull'asse orizzontale i quantili teorici e sull'asse verticale i quantili osservati nel campione.

- Se i punti stanno vicino a una linea retta, il campione segue bene la distribuzione teorica.
- Se i punti si allontanano dalla linea in modo sistematico, i dati non seguono quella distribuzione.

Un esempio di Q-Q plot rispetto alla normale è mostrato in figura 6.9.

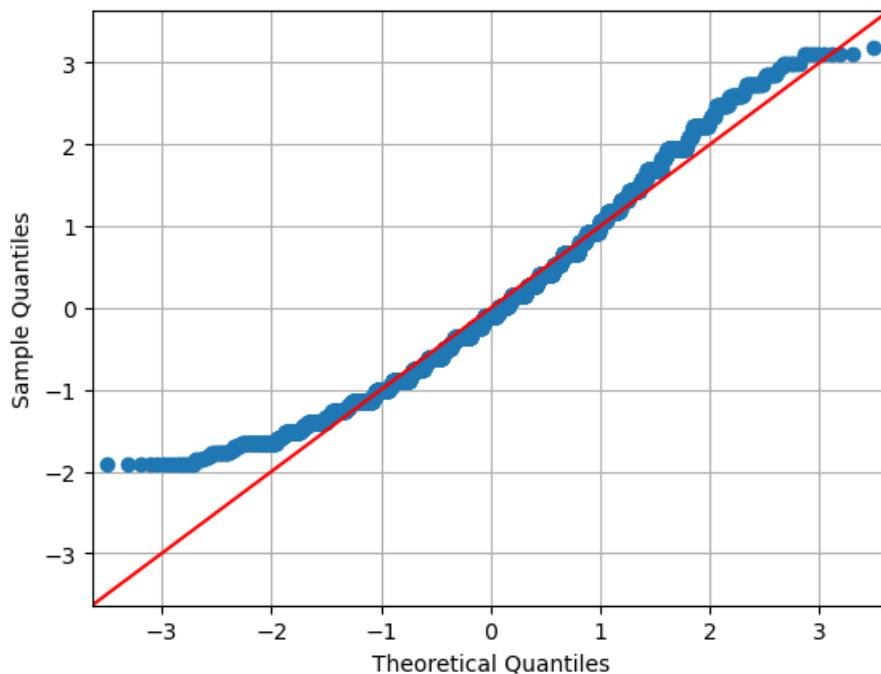


Figura 6.9: Q-Q plot rispetto alla normale: più i punti seguono la linea rossa, più i dati possono essere considerati circa normali.

All'inizio interpretare un Q-Q plot non è sempre immediato. Per questo spesso si confrontano diversi casi tipici, così da riconoscere pattern ricorrenti (figura 6.10).

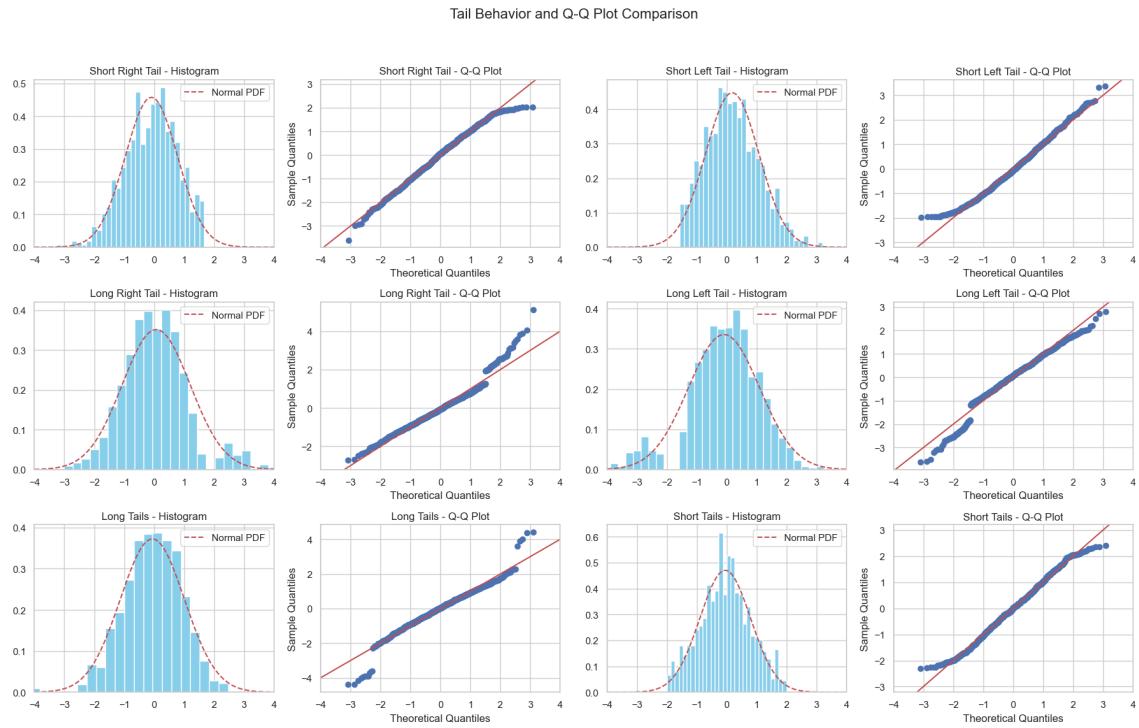


Figura 6.10: Relazione tra forma delle code e Q-Q plot. Code troppo leggere (*short tails*) piegano verso l'interno; code pesanti (*long tails*) si incurvano verso l'esterno; una coda destra lunga fa salire la parte destra del Q-Q plot sopra la linea; una coda sinistra lunga fa scendere la parte sinistra sotto la linea.

6.6.2 Test di normalità di Shapiro-Wilk

Il test di Shapiro-Wilk è un test statistico utilizzato per valutare se un campione di dati segue una distribuzione normale. È usato principalmente quando si ha un campione di dimensioni ridotte (tipicamente $n \leq 2000$). Il test funziona calcolando una statistica W che confronta l'ordine dei dati osservati con l'ordine atteso se i dati fossero normalmente distribuiti. Se il valore di W è significativamente basso, si rifiuta l'ipotesi nulla H_0 che i dati seguono una distribuzione normale.

6.6.3 Test K^2 di D'Agostino

Per campioni grandi ($n \geq 50$) si può utilizzare il test K^2 di D'Agostino, che valuta la normalità basandosi su due misure: la skewness (asimmetria) e la kurtosis (appiattimento). Il test calcola una statistica K^2 combinando queste due misure e confronta il risultato con una distribuzione χ^2 . Se il valore di K^2 è significativamente alto, si rifiuta l'ipotesi nulla H_0 che i dati seguono una distribuzione normale.

Capitolo 7

Analisi predittiva

L'analisi predittiva è una branca dell'analisi dei dati che utilizza tecniche statistiche, di machine learning e di data mining per prevedere eventi futuri basandosi su dati storici.

I principali obiettivi sono:

- Fare inferenza sulle relazioni tra variabili.
- Costruire modelli che possono essere usati per fare previsioni su nuovi dati.

7.1 Modello

Prima di continuare, è necessario dare la definizione di "modello".

Un modello è una rappresentazione semplificata di un sistema complesso, che cattura le caratteristiche essenziali del sistema per permettere l'analisi e la previsione del suo comportamento.

Definizione 7.1

Si può vedere una analogia di un modello come una cartina geografica: non rappresenta ogni dettaglio del territorio, ma fornisce informazioni sufficienti per orientarsi e pianificare un percorso. Ma questo porta a una conclusione, quella cartina è tecnicamente **sbagliata**, poiché non rappresenta a pieno il territorio.

Lo statistico George Box ha detto:

"All models are wrong, but some are useful."

Ovvero, tutti i modelli sono sbagliati in quanto semplificazioni della realtà, ma alcuni possono essere utili per fare previsioni accurate e prendere decisioni informate. Per esempio, utilizzando un modello che predice il BMI (Body Mass Index) di una popolazione possiamo fare delle previsioni solo su una parte della popolazione, ma non su ogni singolo individuo (come si vede in figura 7.1).

7.1.1 Modelli predittivi

Quando facciamo analisi predittiva possiamo identificare:

- Una variabile Y detta **Variabile dipendente** o *variabile di risposta*, che rappresenta l'output che vogliamo prevedere.

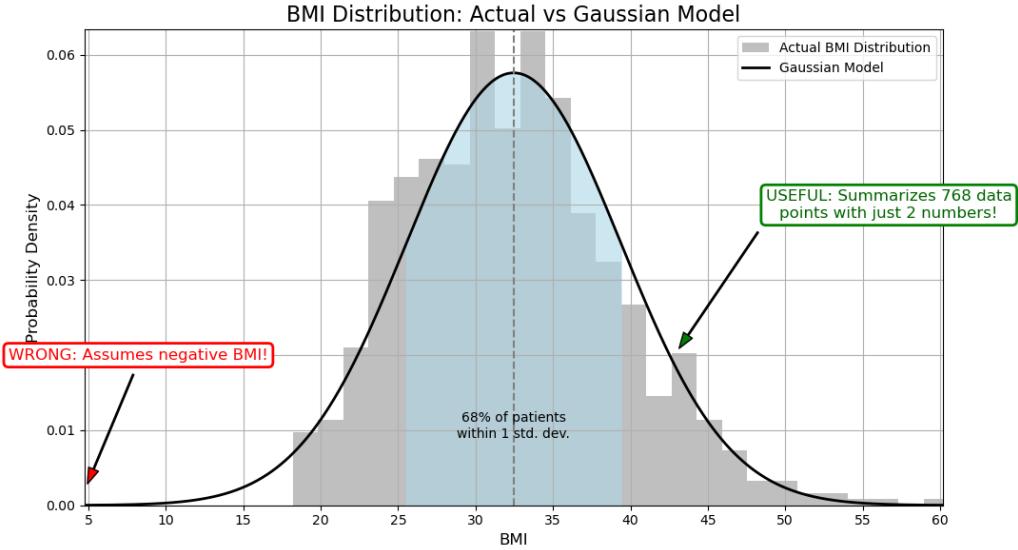


Figura 7.1: Distribuzione del BMI in una popolazione. Istogramma reale composto da $n = 768$ individui (barre grigie) a confronto con un modello gaussiano (linea nera). La Gaussiana riassume con media e deviazione una parte della popolazione in modo corretto, ma assegna probabilità anche a BMI negativi.

- Un vettore di variabili $X = (X_1, X_2, \dots, X_p)$ dette **Variabili indipendenti** o *predittori*, che rappresentano gli input utilizzati per fare la previsione.

Possiamo esprimere il modello predittivo come una funzione matematica, con una forma generale:

$$Y = f(X) + \epsilon$$

dove f è il modello che descrive la relazione tra le variabili indipendenti e la variabile dipendente, ed ϵ è un termine di errore che rappresenta la variabilità non spiegata dal modello.

7.2 Predizione vs Spiegazione

Un aspetto importante dell'analisi predittiva è la distinzione tra **previsione** e **spiegazione**.

7.2.1 Predizione

La previsione si concentra sulla capacità del modello di fare previsioni accurate su nuovi dati, indipendentemente dal fatto che il modello sia interpretabile o meno.

Definizione 7.2

Quindi risponde a una semplice domanda: "Cosa avverrà?". Nel caso particolare della predizione, l'accuratezza del modello è la metrica più importante.

Per fare un esempio, si pensi a un modello per prevedere il prezzo delle case basandosi su caratteristiche come la posizione, la dimensione e il numero di stanze. Anche se la rete neurale può essere complessa e difficile da interpretare, se riesce a fare previsioni accurate sui prezzi delle case, allora è considerata un buon modello predittivo.

7.2.2 Spiegazione

La spiegazione si concentra sulla comprensione delle relazioni tra le variabili e sull'interpretabilità del modello.

Definizione 7.3

In questo caso il modello deve essere al 100% interpretabile, rispondendo alla domanda: "Perché avverrà?". Qui l'accuratezza del modello è meno importante rispetto alla capacità di spiegare i fenomeni osservati.

Per fare un esempio, si consideri un modello per analizzare come una malattia e da quali fattori essa dipende (età, stile di vita, genetica, ecc.). In questo caso, un modello semplice è preferibile, anche se meno accurato, perché permette ai medici di comprendere i fattori di rischio e di prendere decisioni informate sui trattamenti.

7.2.3 Compromesso tra Predizione e Spiegazione

Spesso esiste un compromesso tra predizione e spiegazione, in quanto i due obiettivi non sono spesso mutuamente esclusivi. Un modello potente infatti, se il problema lo permette, deve sia fare inferenza che predizione.

Si pensi a un dataset di rischio del diabete di tipo 1: un modello potrebbe essere utilizzato per prevedere la probabilità che un individuo sviluppi la malattia (predizione), ma potrebbe anche essere utile per identificare i fattori di rischio associati alla malattia (spiegazione).

7.3 Statistica vs Machine Learning

Un altro aspetto importante dell'analisi predittiva è la distinzione tra **statistica** e **machine learning**.

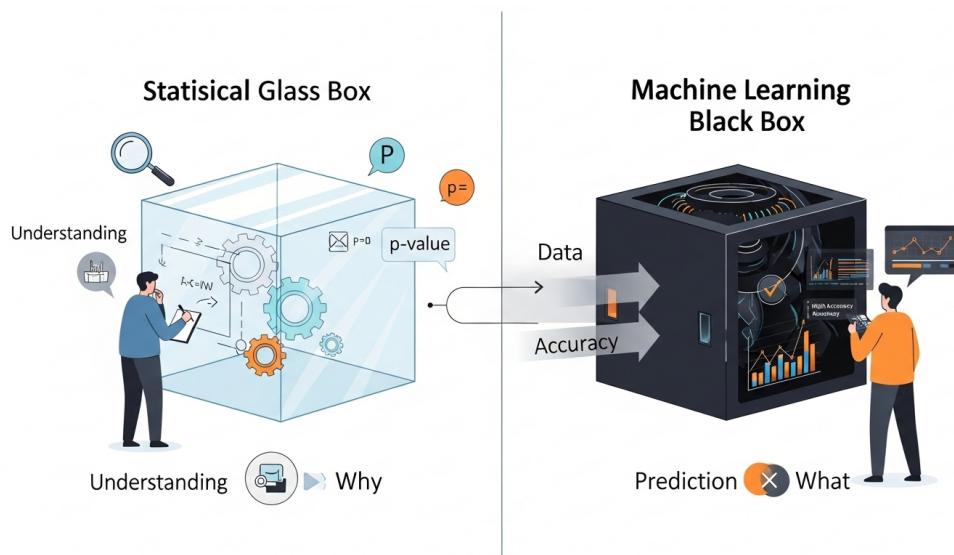


Figura 7.2: Confronto tra *glass box* statistica e *black box* di ML: la statistica privilegia interpretabilità e spiegazione del *perché*, mentre il machine learning privilegia accuratezza predittiva sul *che cosa* a partire dai dati, spesso con modelli opachi.

7.3.1 Approccio statistico

L'approccio statistico è concentrato sul capire il modello e sull'inferenza. Si comporta come in figura 7.2, come una "glass box", dove il funzionamento interno del modello è trasparente e interpretabile. Le sue metodologie includono:

- Affidarsi alle assunzioni fatte sui dati (distribuzioni, linearità, indipendenza, ecc.).
- Utilizzare molto il testo di ipotesi, con p-values e intervalli di confidenza.
- Prediligere modelli semplici e interpretabili.

7.3.2 Approccio di Machine Learning

L'approccio di Machine Learning ha un focus primario sulla predizione accurata. Si comporta come in figura 7.2, come una "black box", dove il funzionamento interno del modello può essere complesso e difficile da interpretare. Le sue metodologie includono:

- Le performance del modello si vedono sui dati, senza fare molte assunzioni a priori.
- Si affida allo split dei dati in training, validation e test set per simulare la generalizzazione.
- Non vengono misurate le performance con p-values, ma con metriche di accuratezza predittiva sul test-set.
- L'intepretabilità è una cosa in più, non un requisito.

7.3.3 Trade-Off di Complessità-Interpretabilità

Quando si lavora su modelli predittivi, statistici o di machine learning, spesso si deve affrontare un trade-off tra complessità e interpetabilità del modello:

Modelli semplici : un modello semplice (come la regressione lineare) è facile da interpretare e spiegare, ma potrebbe non catturare tutte le complessità dei dati, portando a una minore accuratezza predittiva. Come già detto prima, è esattamente la prerogativa della statistica.

Modelli complessi : un modello complesso (come le reti neurali profonde) può catturare meglio le complessità dei dati e fornire previsioni più accurate, ma spesso è difficile da interpretare e spiegare. Questo è il punto di forza del machine learning.

7.4 Tipologie di problema

Una volta stabilito se vogliamo capire i dati o fare predizione, dobbiamo identificare il tipo di problema che stiamo affrontando. I principali tipi di problemi nel machine learning sono:

- **Supervised learning**: il modello viene addestrato su un dataset etichettato, dove ogni esempio di input ha una corrispondente etichetta di output. L'obiettivo è imparare una funzione che mappa gli input agli output corretti.
- **Unsupervised learning**: il modello viene addestrato su un dataset non etichettato, dove non ci sono etichette di output. L'obiettivo è trovare strutture nascoste nei dati, come cluster o pattern.

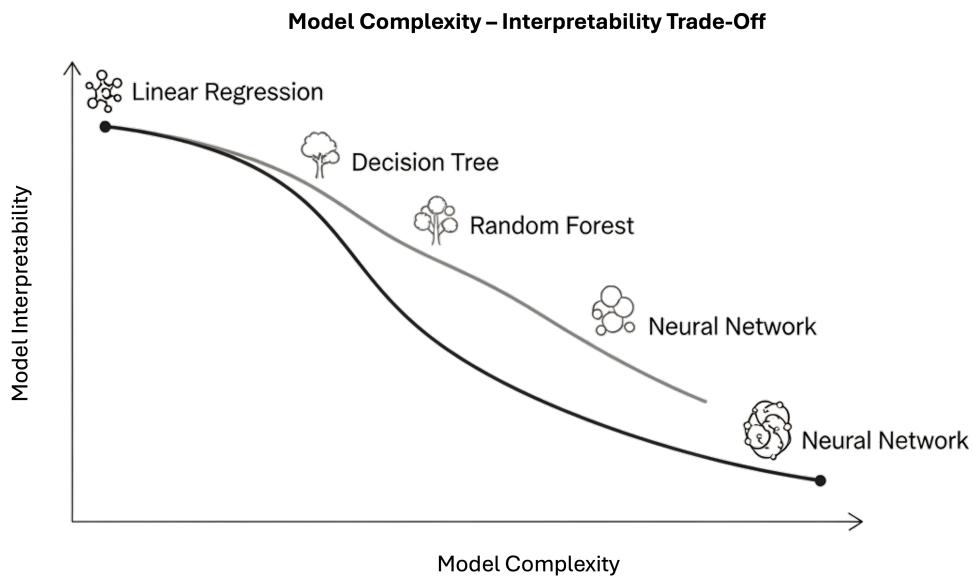


Figura 7.3: Accuratezza vs complessità del modello: dalla Regressione Lineare a Decision Tree e Random Forest fino alle Reti Neurali, l'errore tende a diminuire man mano che cresce la complessità (ma aumenta il costo/opaquezza del modello).

7.4.1 Regressione

La regressione è un tipo di problema di supervised learning in cui l'obiettivo è prevedere un valore y continuo. Dobbiamo trovare quindi il miglior "fit" di una funzione all'interno dei nostri dati, per descrivere a pieno la relazione tra le variabili indipendenti e la variabile dipendente.

Per fare un esempio legato alla figura 7.4, si consideri un dataset che contiene informazioni sulle case, come la dimensione in metri quadri, il numero di stanze e il prezzo di vendita. L'obiettivo è prevedere il prezzo di una casa basandosi sulle sue caratteristiche. Un modello di regressione potrebbe essere utilizzato per trovare la relazione tra la dimensione della casa e il suo prezzo, permettendo di fare previsioni sui prezzi delle case in base alle loro dimensioni.

7.4.2 Classificazione

La classificazione è un altro tipo di problema di supervised learning in cui l'obiettivo è prevedere una categoria o classe discreta y (etichetta). In questo caso, durante il training, il modello impara a mappare gli input alle classi corrette.

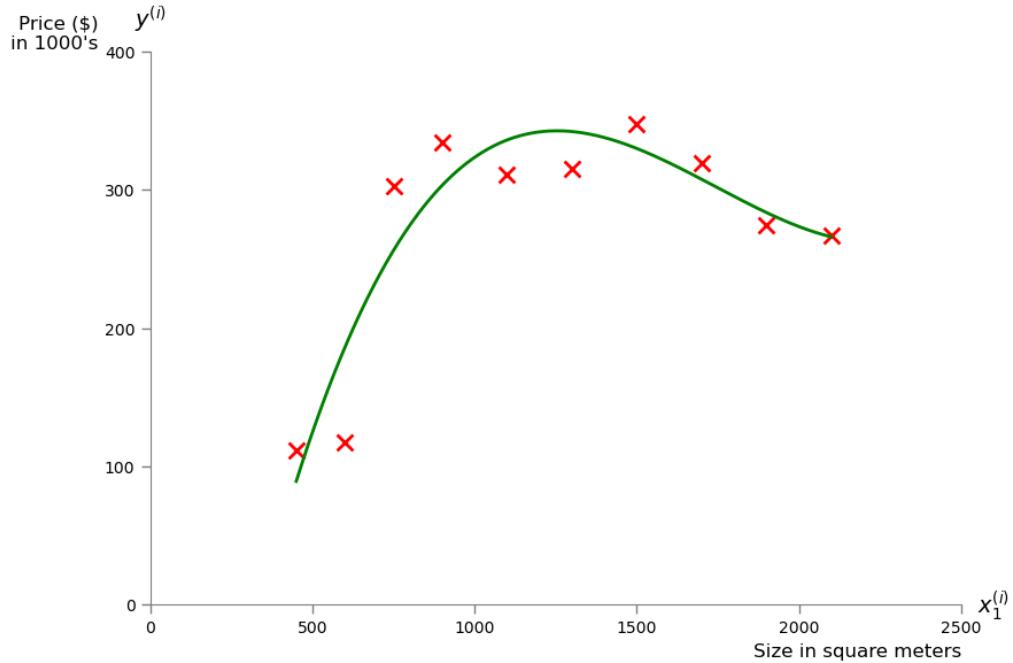


Figura 7.4: Esempio di problema di regressione: prevedere il prezzo di una casa in base ai metri quadri.

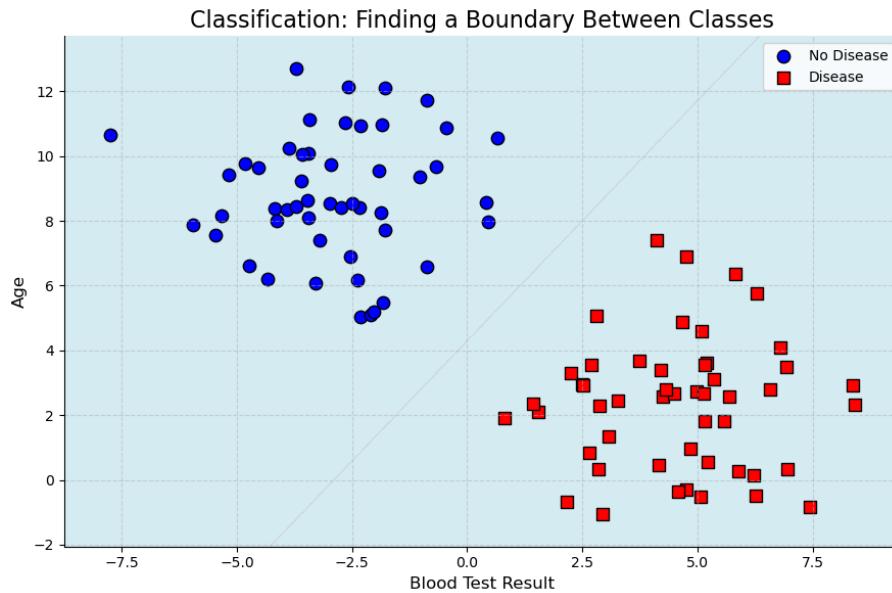


Figura 7.5: Esempio di problema di classificazione: capire se una malattia è presente oppure no in base all'età e a un'analisi del sangue.

Continuando l'esempio in figura 7.5, si consideri un dataset che contiene informazioni sui pazienti, come l'età, il sesso e i risultati di un'analisi del sangue. L'obiettivo è prevedere se un paziente ha una certa malattia (ad esempio, diabete) basandosi sulle sue caratteristiche. Nel caso particolare dell'immagine, notiamo che può essere separata da una retta per risolvere il problema.

7.4.3 Clustering

Il clustering è un problema di unsupervised learning in cui l'obiettivo è raggruppare i dati in cluster basandosi sulla somiglianza tra gli esempi. In questo caso, il modello cerca di identificare strutture nascoste nei dati senza l'uso di etichette di output.



Figura 7.6: Esempio di problema di clustering: raggruppare i clienti in segmenti basandosi sui loro comportamenti di acquisto.

Il modello, unsupervised, viene addestrato ad identificare gruppi con caratteristiche simili. Nella figura 7.6 si considera un esempio dove si vogliono raggruppare quei gruppi di persone in base all'età e a una feature chiamata "Spending Score", che indica quanto una persona spende in un negozio.

7.5 Modelli parametrici vs Modelli non parametrici

I modelli possono differirsi anche in base a come rappresentano la funzione $f(X)$ che mappa gli input agli output. Possiamo distinguere tra modelli parametrici, che assumono una forma funzionale specifica, e modelli non parametrici, che non fanno assunzioni rigide sulla forma della funzione.

7.5.1 Modelli parametrici

Nello specifico, un modello parametrico è caratterizzato da un numero fisso di parametri che definiscono la funzione $f(X)$. Questi modelli sono spesso più semplici e veloci da addestrare, ma possono essere limitati nella loro capacità di catturare la complessità dei dati. Qui il compito principale è quello di modificare i valori dei parametri per adattare il modello ai dati di training.

Un esempio comune di modello parametrico è la regressione lineare, dove la funzione $f(X)$ è una combinazione lineare delle variabili indipendenti:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

dove $\beta_0, \beta_1, \dots, \beta_p$ sono i parametri del modello.

Ha come pro il fatto che è semplice, veloce e richiede meno dati. Tuttavia questi modelli sono "biased": le assunzioni iniziali possono portare a errori sistematici se i dati non seguono quelle assunzioni, inoltre sono più propensi all'*underfitting*¹.

7.5.2 Modelli non parametrici

Un modello non parametrico, invece, non assume una forma funzionale specifica per la funzione $f(X)$. Questi modelli sono più flessibili e possono adattarsi meglio alla complessità dei dati, ma possono essere più lenti da addestrare e richiedere più dati. Qui il compito principale è quello di memorizzare i dati di training e utilizzare questi dati per fare predizioni sui nuovi input.

Un esempio comune di modello non parametrico è il k-Nearest Neighbors (k-NN), dove la predizione per un nuovo input viene fatta basandosi sui k esempi più vicini nel dataset di training.

Ha come pro il fatto che è flessibile e può catturare relazioni complesse nei dati. Tuttavia questi modelli sono "low-bias": non fanno assunzioni rigide sui dati, ma possono essere più propensi all'*overfitting*².

7.6 Learning

Alla base dei problemi di Machine Learning c'è il concetto di **learning**, ovvero l'apprendimento di una funzione $f(X)$ dai dati.

7.6.1 Definizione formale

Siano \mathcal{X} e \mathcal{Y} gli spazi, rispettivamente, di input e di output al modello. Possono essere visti come spazi di variabili aleatorie, in particolare una variabile X e una Y con osservazioni $x \in \mathcal{X}, y \in \mathcal{Y}$. L'obiettivo è trovare una funzione che risponde alla nostra ipotesi $h \in \mathcal{H}$, dove \mathcal{H} è lo spazio delle ipotesi (funzioni candidate), ovvero una funzione:

$$h : \mathcal{X} \rightarrow \mathcal{Y}$$

che approssima la relazione tra X e Y . Possiamo scrivere una funzione che lo fa non in modo preciso, ma approssimando a una predizione $\hat{y} \approx y \in \mathcal{Y}$:

$$\hat{y} = h(x)$$

Definizione 7.4

¹L'*underfitting* è un fenomeno causato da un modello troppo "semplice", che non riesce a catturare la complessità dei dati.

²L'*overfitting* è un fenomeno causato da un modello troppo "complesso", che si adatta troppo bene ai dati di training, ma non generalizza bene su nuovi dati.

Esempio 1: non parametrico. Per fare un esempio, immaginiamo $\mathcal{X} = \mathbb{R}^m$ e $\mathcal{Y} = \{0, 1\}$, ovvero uno spazio di input a m variabili che rappresentano lo spazio dei risultati di un'analisi del sangue e uno spazio di output binario che rappresenta la presenza o meno di una malattia. L'obiettivo è trovare una funzione h che mappa i risultati dell'analisi del sangue alla presenza o meno della malattia.

Esempio 2: parametrico. In un altro esempio, immaginiamo di dover classificare le email in spam e non: in questo caso possiamo definire \mathcal{X} come lo spazio delle email rappresentate da vettori di caratteristiche (come la frequenza di certe parole chiave) e $\mathcal{Y} = \{0, 1\}$ come lo spazio delle etichette (spam o non spam). L'obiettivo è trovare una funzione h che mappa le caratteristiche delle email alle etichette corrette:

$$h(x) = \begin{cases} 1 & f(x) > \theta \\ 0 & f(x) \leq \theta \end{cases}$$

Dove θ è esattamente il nostro parametro, in questo caso una *soglia*.

7.6.2 Il processo di Learning

Per trovare la funzione h che meglio approssima la relazione tra X e Y , dobbiamo avere un modo per valutare quanto bene una funzione h si adatta ai dati. Questo viene fatto utilizzando una funzione di perdita (loss function) $L(y, \hat{y})$, che misura l'errore tra il valore reale y e la predizione \hat{y} fatta dal modello.

In realtà vorremmo calcolare l'errore atteso (o rischio) del modello su tutta la distribuzione dei dati, ma non avendo accesso a questa distribuzione, possiamo solo stimare l'errore empirico $R(h)$ definita sulla nostra ipotesi h^* ³ come **l'errore atteso** sotto la distribuzione di dati $P(X, Y)$:

$$R(h) = \mathbb{E}_{(X,Y) \sim P}[L(Y, h(X))]$$

In questo caso, si definisce **obiettivo del learning statistico** come la risoluzione del problema di ottimizzazione seguente:

$$h^* = \arg \min_{h \in \mathcal{H}} R(h)$$

7.6.3 ERM: Empirical Risk Minimization

Esiste un problema: non possiamo calcolare la funzione di rischio $R(h)$ perché non conosciamo la vera distribuzione $P(X, Y)$. Tuttavia, possiamo stimare il rischio empirico $R_{emp}(h)$ utilizzando un dataset di training⁴ di N esempi che sono un campione rappresentativo della popolazione. Si definisce TR training set, come l'insieme delle coppie:

$$\text{TR} = \{(x_i, y_i)\}_{i=1}^N$$

³Indichiamo la nostra ipotesi con un asterisco h^* perché rappresenta la migliore funzione di approssimazione possibile dati i dati a disposizione.

⁴Il dataset di training è un insieme di dati utilizzati per addestrare il modello.

Grazie a questo e alla **legge dei grandi numeri**⁵, possiamo stimare il rischio empirico come:

$$R_{\text{emp}}(h) = \frac{1}{N} \sum_{i=1}^N L(y_i, h(x_i))$$

Quindi, possiamo risolvere il problema di ottimizzazione empirica:

$$h_{\text{emp}} = \arg \min_{h \in \mathcal{H}} R_{\text{emp}}(h)$$

Questo approccio è noto come **Empirical Risk Minimization** (ERM), ovvero la minimizzazione del rischio empirico.

7.7 Capacità del modello

Un aspetto cruciale nell'analisi predittiva è la capacità del modello, ovvero la sua capacità di adattarsi ai dati di training e di generalizzare a nuovi dati. Possiamo definire formalmente la capacità come una relazione allo spazio e alla "ricchezza" (intesa come complessità) dello spazio delle ipotesi \mathcal{H} dal quale il modello può scegliere la funzione h . Esistono modelli:

A bassa capacità : modelli con uno spazio delle ipotesi limitato, che possono adattarsi solo a funzioni semplici. Questi modelli sono meno propensi all'overfitting, ma possono soffrire di underfitting. Un esempio è una classificazione lineare con una retta: sicuramente andrà bene per dati linearmente separabili, ma nei dati più complessi non riuscirà a catturare le relazioni tra le variabili.

Ad alta capacità : modelli con uno spazio delle ipotesi ampio, che possono adattarsi a funzioni complesse. Questi modelli sono più propensi all'overfitting, ma possono catturare meglio le relazioni nei dati. Un esempio è una classificazione non lineare con un polinomio di grado 10: sicuramente riuscirà a catturare le relazioni nei dati complessi, ma rischia di adattarsi troppo ai dati di training e di non generalizzare bene su nuovi dati.

7.7.1 Misurare la capacità del modello

Per misurare la capacità de modello si deve misurare quanto bene il modello generalizza sui nuovi dati. Questa metrica è solitamente una funzione che valuta l'accuratezza del modello su un dataset di test separato dal training set. Un modo comune, nel caso della regressione, è utilizzare l'errore quadratrico medio (Mean Squared Error, MSE):

$$R_{\text{emp}}(h) = \frac{1}{M} \sum_{j=1}^M (y_j - h(x_j))^2$$

Generalmente però, questa metrica, non basta. Perché si potrebbe pensare di utilizzare come mappatura di una regressione il valore del training set:

$$\hat{h}(x) = y \quad (x, y) \in \text{TR}$$

⁵La legge dei grandi numeri afferma che, al crescere del numero di osservazioni, la media campionaria converge alla media della popolazione.

ottenendo un errore di 0 sul training set, ma un errore altissimo sul test set. Per questo motivo si usano tecniche di validazione incrociata (cross-validation) per stimare la capacità del modello in modo più robusto.

Questo porta alla definizione di due concetti già visti in breve, **underfitting** e **overfitting**:

Underfitting Si verifica quando un modello è troppo semplice per catturare le relazioni nei dati, portando a prestazioni scadenti sia sul training set che sul test set. Per esempio, dei dati che seguono una distribuzione quadratica vengono approssimati con una retta (7.7, sinistra).

Overfitting Si verifica quando un modello è troppo complesso e si adatta troppo ai dati di training, catturando il rumore invece della vera relazione tra le variabili. Questo porta a buone prestazioni sul training set ma scarse sul test set. Per esempio, dei dati che seguono una distribuzione quadratica vengono approssimati con un polinomio di grado 5 (7.7, destra).

7.7.2 Bias e Varianza

Per valutare le prestazioni di un modello possiamo usare i valori del bias e della varianza per fare delle stime su come il modello si comporta sui dati.

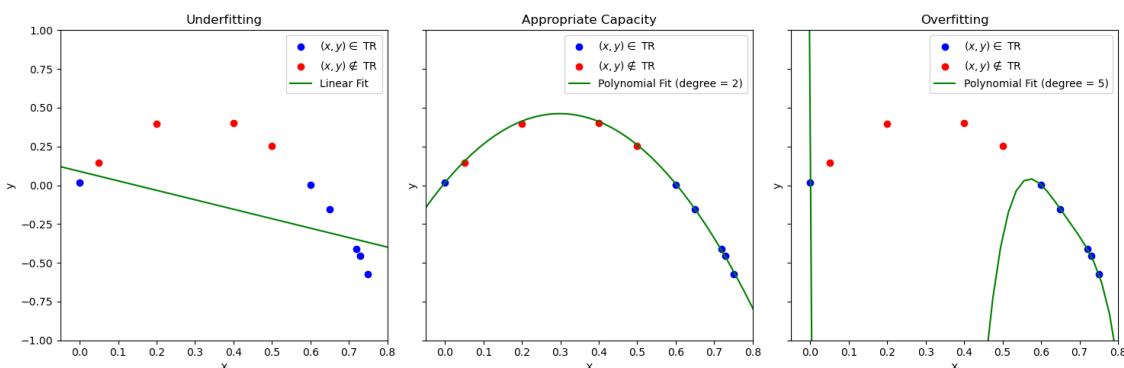


Figura 7.7: Underfitting—capacità adeguata—overfitting in regressione polinomiale: punti blu = dati di training (TR), punti rossi = fuori da TR; la linea verde mostra il fit del modello: lineare (sinistra), polinomio di grado 2 (centro) e polinomio di grado 5 (destra).

Bias. Il bias è l'**errore sistematico** che il modello commette sui dati. Un modello con alto bias tende a sottostimare la complessità del problema, portando a errori elevati sia sul training set che sul test set. Questo fenomeno è noto come **underfitting**. Nell'immagine 7.7, il grafico a sinistra mostra un esempio di underfitting, dove il modello lineare non riesce a catturare la relazione tra le variabili e commette un errore sistematico sia sul training set (punti blu) che sul test set (punti rossi).

Varianza. La varianza rappresenta, invece, la sensibilità del modello alle variazioni nei dati di training. Un modello con alta varianza tende a sovrardattarsi ai dati di training, catturando il rumore invece della vera relazione tra le variabili. Questo porta a errori bassi sul training set ma elevati sul test set, fenomeno noto come **overfitting**. Nell'immagine 7.7, il grafico a destra

mostra un esempio di overfitting, dove il modello polinomiale è troppo complesso e si adatta troppo strettamente ai dati di training, risultando in prestazioni scadenti sui dati di test⁶.

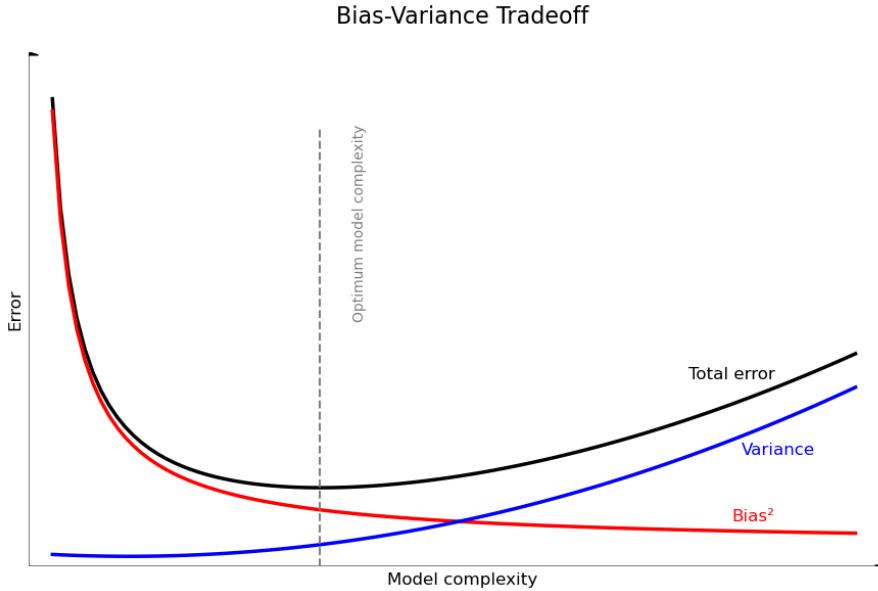


Figura 7.8: Trade-off tra bias e varianza in funzione della complessità del modello.

7.7.3 Parametri vs Iperparametri

Per controllare la capacità del modello, possiamo agire su due tipi di parametri:

Parametri I parametri sono i valori che il modello impara durante il processo di training. Questi parametri definiscono la funzione h che mappa gli input agli output. Per esempio, in una regressione lineare, i parametri sono i coefficienti della retta.

Iperparametri Gli iperparametri sono i valori che vengono impostati prima del processo di training e controllano il comportamento del modello. Questi iperparametri influenzano la capacità del modello e il modo in cui viene addestrato. Sono coefficienti che hanno fattori esterni, come il tasso di apprendimento, la profondità di un albero decisionale o il grado del polinomio nella funzione di regressione.

7.8 Selezione del modello

Il problema a questo punto diventa: come facciamo a far sì che l'algoritmo di learning (ERM) trovi i migliori parametri per il modello fornito un set di iperparametri.

7.8.1 Approccio 1: selezione statistica

Quando il nostro obiettivo principale è l'*inferenza* (il modello glass-box), selezionamo un modello in base a quanto spiega bene i dati in favore della semplicità. Questo approccio utilizza l'intero

⁶Si noti che basterebbe rimuovere un punto del training set per far cambiare completamente il modello, in quanto altamente sensibile ai dati di input.

dataset **in una volta**, perché non cerchiamo di "predire" il futuro, ma di trovare la migliore spiegazione per i dati che osserviamo.

Si usano misure statistiche che bilanciano bontà di fit e complessità del modello, come:

- **p-values**: Facciamo un test sul valore di ogni variabile e potremmo rimuovere quelle con un p-value alto (ovvero quelle che creano rumore).
- **R²**: Questa misura indica la proporzione di varianza nella variabile dipendente che è spiegata dalle variabili indipendenti nel modello. Un valore di R^2 più alto indica un modello che spiega meglio i dati.

7.8.2 Approccio 2: selezione predittiva

Quando il nostro obiettivo principale è la *predizione* (il modello black-box), selezioniamo un modello in base a quanto bene predice nuovi dati, bilanciando accuratezza e complessità. Questo approccio utilizza tecniche di validazione incrociata per stimare la capacità del modello di generalizzare a nuovi dati.

Esiste infatti una "golden rule" (regeola d'oro) nei modelli predittivi:

"The performance of a model on the data it was trained on is *irrelevant*. The only measure that matters is its performance on new, unseen data."

Ovvero, è inutile valutare un modello in base a quanto bene si adatta ai dati di training: l'unica metrica che conta è quanto bene si comporta su nuovi dati mai visti prima.

Per valutare le prestazioni dobbiamo ricordare che il rischio non è calcolabile, quindi usiamo il rischio empirico sul test set. Generalmente si utilizzano *misure di performance* oppure *misura di errori* (loss function):

- **Loss Function**: la loss function è molto utile nei casi di training, in quanto misura l'errore tra la predizione del modello e il valore reale. Per esempio, nella regressione si può usare l'errore quadratico medio (MSE), mentre nella classificazione si può usare la log-loss.
- **Metriche di performance**: le metriche di performance sono utilizzate per valutare le prestazioni del modello su un dataset di test.

Una misura di performance p misura l'insieme delle verità Y e delle predizioni corrispondenti \hat{Y} per un certo dataset:

$$p : \mathcal{Y}^N \times \mathcal{Y}^N \rightarrow \mathbb{R}$$

La differenza con il rischio empirico è che questa misura valuta direttamente le prestazioni del modello, senza passare per una funzione di perdita.

7.8.3 Validazione Holdout

La validazione holdout (o single split) è una tecnica semplice per stimare la capacità di generalizzazione di un modello. Consiste nel dividere il dataset in due parti: un training set e un test set. Il modello viene addestrato sul training set e poi valutato sul test set.

Questo paradigma è utile per modelli semplici e dataset grandi, in quanto un dataset più piccolo restituirebbe con qualità inferiore le stime delle prestazioni del modello.

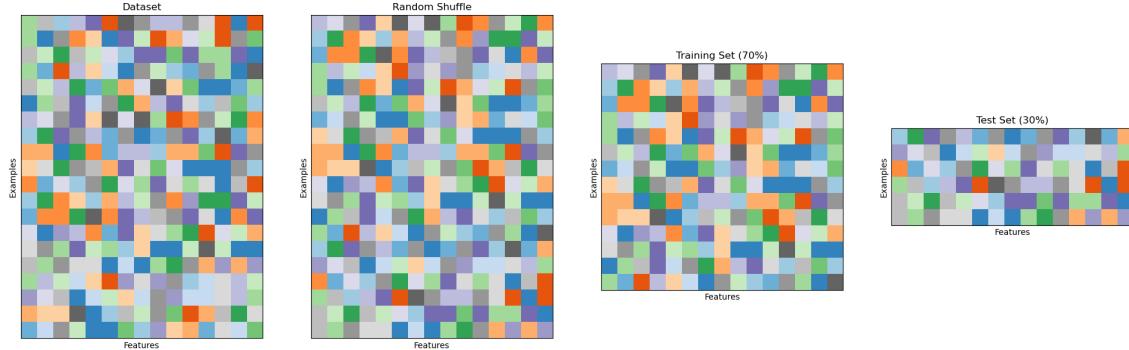


Figura 7.9: Validazione Holdout: il dataset viene prima mescolato, poi diviso in training set (70%) e test set (30%). Il modello viene addestrato sul training set e valutato sul test set per stimare la capacità di generalizzazione.

7.8.4 K-Fold Cross-Validation

La K-Fold Cross-Validation risolve il problema di dataset piccoli: divide il dataset in diversi sottogruppi (folds) e utilizza ogni sottogruppo sia come test che come training set in diverse iterazioni. In particolare, il dataset viene diviso in K folds di dimensioni approssimativamente uguali. In ogni iterazione, uno dei folds viene utilizzato come test set, mentre gli altri $K - 1$ folds vengono utilizzati come training set. Questo processo viene ripetuto K volte, in modo che ogni sottogruppo venga utilizzato come test set una volta.

Alla fine del processo, le prestazioni del modello vengono mediate su tutte le iterazioni per ottenere una stima più robusta della capacità di generalizzazione del modello. Questo metodo è particolarmente utile quando si lavora con dataset di dimensioni limitate, in quanto consente di utilizzare tutti i dati disponibili sia per l'addestramento che per la valutazione del modello.

Rimangono due problemi:

- È una tecnica computazionalmente costosa, in quanto se il training di un modello richiede 1 giorno, questo paradigma ne fa impiegare 4.
- Gli iperparametri non vengono ottimizzati durante il processo di training.

7.8.5 Leave-One-Out Cross-Validation (LOOCV)

La Leave-One-Out Cross Validation (LOOCV) è una variante estrema della K-Fold Cross-Validation, in cui il numero di folds K è uguale al numero di esempi nel dataset. In altre parole, in ogni iterazione, un singolo esempio viene utilizzato come test set, mentre tutti gli altri esempi vengono utilizzati come training set. Questo processo viene ripetuto per ogni esempio nel dataset.

Questo approccio è molto utile quando si lavora con dataset molto piccoli, in quanto consente di utilizzare quasi tutti i dati disponibili per l'addestramento del modello in ogni iterazione. Tuttavia, la LOOCV soffre degli stessi problemi della K-Fold Cross Validation.

7.8.6 Ottimizzazione degli iperparametri

Il problema degli iperparametri, con questi paradigmi, comunque persiste.

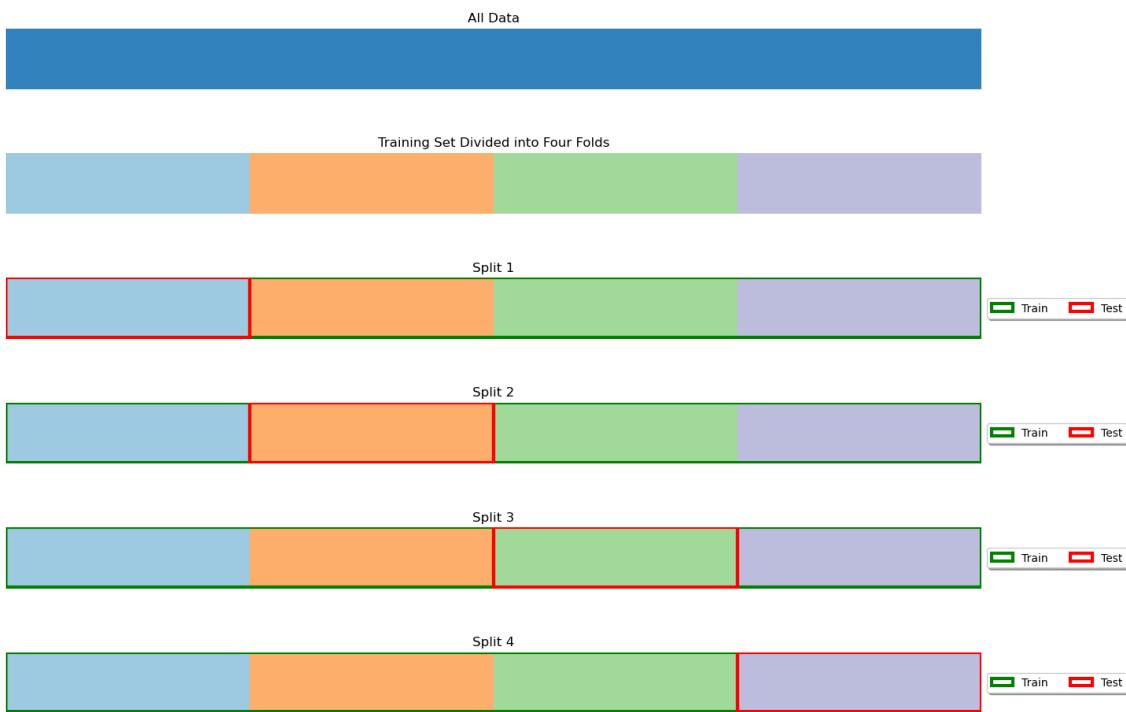


Figura 7.10: K-Fold Cross-Validation: il dataset viene diviso in K folds. In ogni iterazione, un fold viene utilizzato come test set e gli altri $K - 1$ folds come training set. Questo processo viene ripetuto K volte, e le prestazioni del modello vengono mediate su tutte le iterazioni per ottenere una stima più robusta della capacità di generalizzazione.

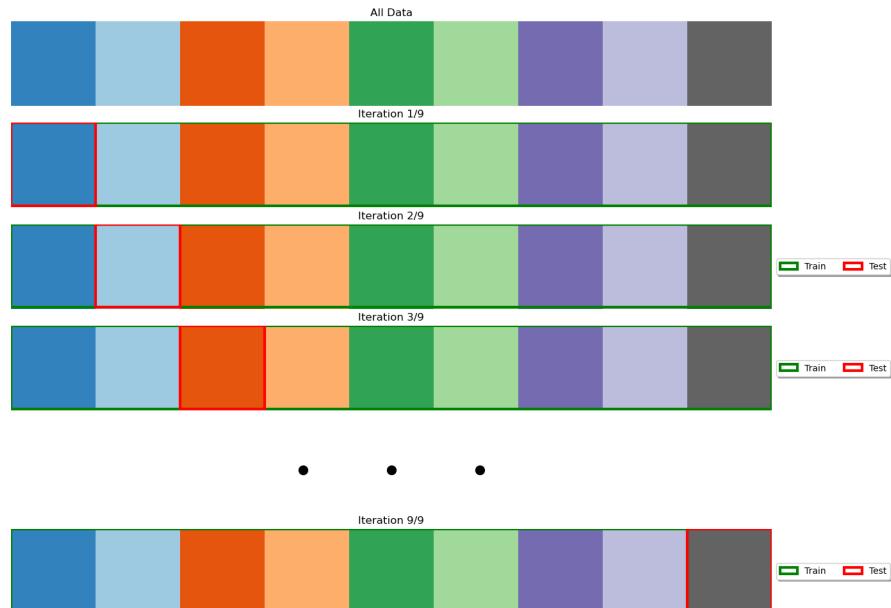


Figura 7.11: Leave-One-Out Cross-Validation (LOOCV): in ogni iterazione, un singolo esempio viene utilizzato come test set, mentre tutti gli altri esempi vengono utilizzati come training set. Questo processo viene ripetuto per ogni esempio nel dataset.

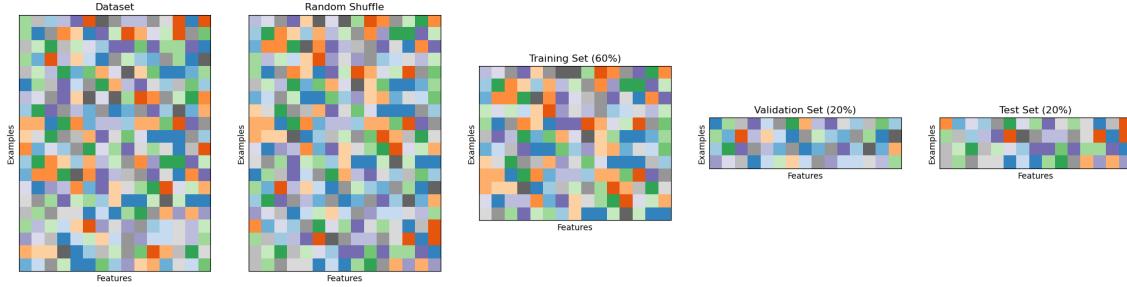


Figura 7.12: Utilizzo di un validation set per l’ottimizzazione degli iperparametri: il dataset viene diviso in training set (60%), validation set (20%) e test set (20%). Il modello viene addestrato sul training set, gli iperparametri vengono ottimizzati sul validation set e infine le prestazioni del modello vengono valutate sul test set.

Grid Search. Alcuni algoritmi suggeriscono di utilizzare una **grid search**: un tipo di ricerca esaustiva sugli iperparametri, in cui si definisce una griglia di valori possibili per ogni iperparametro e si valuta il modello per ogni combinazione di valori nella griglia. Questo approccio risolve indubbiamente l’ottimizzazione degli iperparametri, ma può essere computazionalmente costoso, specialmente con un gran numero di iperparametri e valori da esplorare.

Validation set. Un altro approccio è quello di utilizzare un **validation set**: un sottoinsieme del dataset separato dal training set e dal test set, utilizzato per ottimizzare gli iperparametri. Per esempio, in una Validazione Holdout, si potrebbe dividere il dataset in tre parti: training set, validation set e test set. Il modello viene addestrato sul training set, gli iperparametri vengono ottimizzati sul validation set e infine le prestazioni del modello vengono valutate sul test set.

Ci sono tuttavia, combinazioni migliori. Si potrebbe effettuare, per esempio, una K-Fold Cross Validation dove si divide inizialmente il dataset in training/validation set e test set: il training/validation set dopo è sottoposto a una variante di K-Fold Cross Validation per ottimizzare gli iperparametri, mentre il test set viene usato solo alla fine per valutare le prestazioni finali del modello.

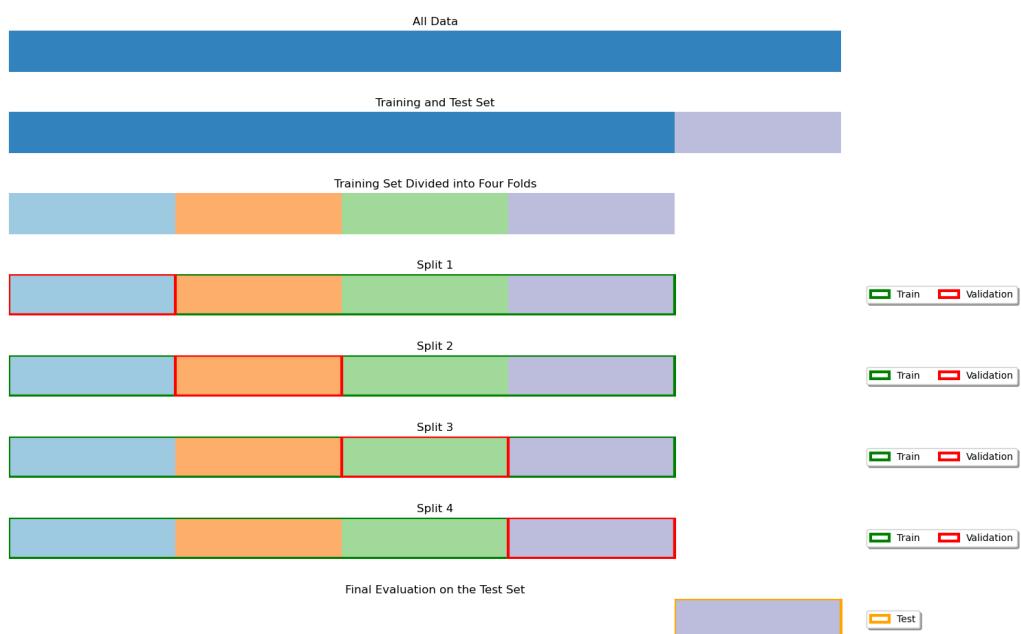


Figura 7.13: K-Fold Cross Validation con ottimizzazione degli iperparametri: il dataset viene diviso in training/validation set e test set. Il training/validation set viene sottoposto a K-Fold Cross Validation per ottimizzare gli iperparametri, mentre il test set viene utilizzato solo alla fine per valutare le prestazioni finali del modello.

Capitolo 8

Regressione lineare

La regressione lineare è una tecnica utilizzata per modellare la relazione tra una variabile dipendente e una o più variabili indipendenti. In particolare si rileva utile in tutti quei casi dove si vuole studiare l'effetto di una o più variabili esplicative su una variabile di interesse, oppure si vogliono fare previsioni sui valori futuri della variabile di interesse basandosi sui valori delle variabili esplicative.

8.1 Formalizzazione della regressione

La regressione lineare mira a studiare l'associazione tra una variabile dipendente Y e una o più variabili indipendenti X_1, X_2, \dots, X_p definendo un modello matematico f tale che:

$$Y = f(X) + \epsilon$$

*Dove ϵ rappresenta l'errore o il **rumore** nel modello. L'obiettivo della regressione è stimare la funzione f in modo da minimizzare l'errore tra i valori osservati di Y e i valori predetti dal modello.*

Definizione 8.1

Si parla di *rumore* con il valore di ϵ per indicare tutte quelle variabili che influenzano Y ma che non sono state incluse nel modello. Anche perché basti pensare al fatto che f non è un modello deterministico ma probabilistico, quindi non è possibile prevedere con certezza il valore di Y dato X .

8.2 Regressione lineare semplice

Si parla di **regressione lineare semplice** quando si ha una sola variabile indipendente X . In questo caso, il modello di regressione lineare può essere espresso come:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Esempio. Ipotizziamo di voler prendere in considerazione un dataset che contiene informazioni sul numero di ore di studio e i voti ottenuti dagli studenti in un esame. Vogliamo capire se esiste

una relazione tra il numero di ore di studio (variabile indipendente X) e il voto ottenuto (variabile dipendente Y), con un modello circa così:

$$\text{voto} = \beta_0 + \beta_1 \cdot \text{ore_studio} + \epsilon$$

8.2.1 Analogia geometrica con una retta

In realtà la forma del regressore lineare è, in questo caso, quella di una retta:

$$y = mx + q \longrightarrow Y = \beta_1 X + \beta_0$$

Dove m rappresenta la pendenza della retta (coefficiente angolare) e q rappresenta l'intercetta con l'asse delle ordinate (coefficiente lineare).

Si può facilmente notare, ricordando un po' di algebra (figura 8.1), che:

- Al variare del parametro β_1 si forma un fascio di rette proprio¹ passanti per il punto $(0, \beta_0)$, ovvero l'intercetta con l'asse delle ordinate (a sinistra in figura 8.1).
- Al variare del parametro β_0 si forma un fascio di rette improprio² con pendenza β_1 (a destra in figura 8.1).

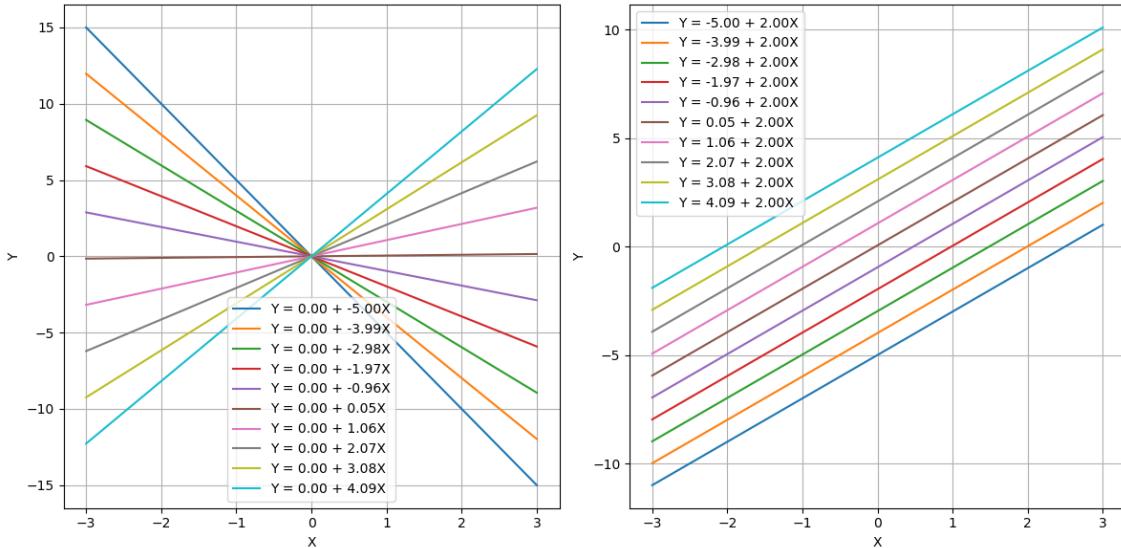


Figura 8.1: Esempio di rette di regressione al variare del coefficiente: a sinistra varia β_1 mantenendo fisso β_0 , a destra varia β_0 mantenendo fisso β_1 .

8.2.2 OLS: Ordinary Least Squares

Per stimare i coefficienti β_0 e β_1 del modello "ottimale" bisogna definire cos'è un modello *ottimale*. Si può dire ottimale un modello se predice bene la variabile Y dalla variabile X . Per misurare questo, partiamo con il definire il nostro insieme di osservazioni:

$$\{(x_i, y_i)\}_{i=1}^n$$

¹Un fascio di rette proprio è un insieme di rette che passano tutte per uno stesso punto.

²Un fascio di rette improprio è un insieme di rette parallele tra loro.

Dove n è il numero di osservazioni, x_i è il valore della variabile indipendente per l'osservazione i e y_i è il valore della variabile dipendente per l'osservazione i .

Sia:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

il valore predetto dal modello per l'osservazione i , dove $\hat{\beta}_0$ e $\hat{\beta}_1$ sono le stime dei coefficienti. Per ogni punto dati (x_i, y_i) , definiamo la deviazione della predizione dal valore corretto y_i come:

$$e_i = y_i - \hat{y}_i$$

Questa quantità è chiamata **residuo** o **errore di predizione**. Ovviamente questo valore può essere positivo o negativo, a seconda che la predizione sia inferiore o superiore al valore reale. Da questo valore definiamo la **RSS: Residual Sum of Squares** (Somma dei quadrati dei residui) come:

$$RSS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Con questo valore possiamo avere una misura di un modello ottimale: un modello è tanto più ottimale quanto più piccolo è il valore di RSS. Quindi l'obiettivo diventa minimizzare il valore di RSS trovando i valori ottimali di $\hat{\beta}_0$ e $\hat{\beta}_1$:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

che viene chiamata **loss (o cost) function**, ovvero la funzione che misura l'errore del modello ed è quella da minimizzare.

Per minimizzare questa funzione si può usare il calcolo differenziale ricordando che per trovare i punti di minimo di una funzione basta trovare i punti in cui la derivata prima è nulla³. Calcoliamo le derivate parziali⁴ della funzione di loss rispetto a β_0 e β_1 e le poniamo uguali a zero:

$$\frac{\partial RSS}{\partial \beta_0} = 0, \quad \frac{\partial RSS}{\partial \beta_1} = 0$$

Risolvendo il sistema di equazioni si ottengono le seguenti formule per i coefficienti stimati:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Dove \bar{x} e \bar{y} sono le medie dei valori di X e Y rispettivamente:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

³Quando la derivata prima è nulla, la crescenza della curva cambia e quindi si ha un punto di massimo o minimo locale.

⁴I parametri sono due, non si può derivare rispetto a entrambi contemporaneamente, quindi si calcolano le derivate parziali.

Varianza dalla popolazione ideale. Considerando una popolazione ideale dove:

$$Y = 2x + 1$$

ci si aspetta che i campioni estratti da questa popolazione abbiano coefficienti $\hat{\beta}_0 \approx 1$ e $\hat{\beta}_1 \approx 2$, ma a causa del rumore e della variabilità dei dati, i valori stimati possono differire leggermente da questi valori ideali.

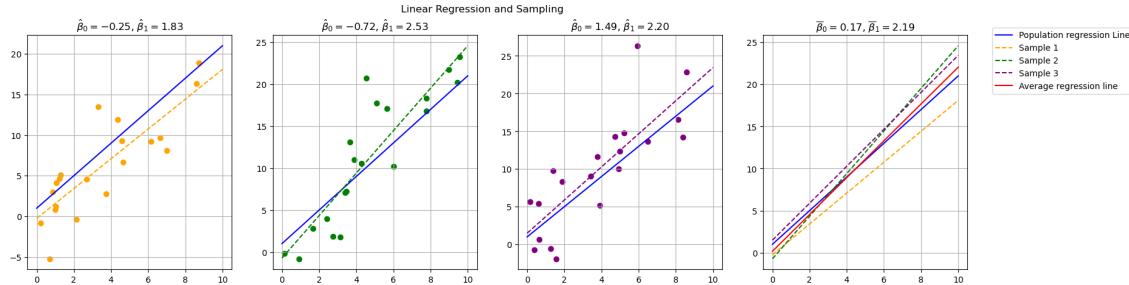


Figura 8.2: Esempio di regressione lineare e variabilità campionaria. Ogni pannello mostra un campione diverso estratto dalla stessa popolazione con la rispettiva retta di regressione stimata ($\hat{\beta}_0, \hat{\beta}_1$). La linea blu rappresenta la retta di regressione della popolazione, mentre l'ultimo grafico illustra la media delle rette di regressione campionarie (in rosso), evidenziando come, in media, essa coincida con la vera retta di regressione della popolazione.

Dalla figura 8.2 si può notare come, nonostante la variabilità dei campioni, la media delle rette di regressione campionarie (in rosso nell'ultimo grafico) tende a coincidere con la vera retta di regressione della popolazione (in blu), indicando che il metodo OLS potrebbe avere una varianza bassa, ma rimane comunque uno stimatore⁵ *unbiased*.

8.2.3 Intervalli di confidenza per i coefficienti

Poiché i coefficienti stimati $\hat{\beta}_0$ e $\hat{\beta}_1$ sono calcolati su un campione di dati, possono essere visti come *stimatori* dei veri coefficienti della popolazione. Da qui possiamo calcolare gli **intervalli di confidenza** e gli **standard errors** (SE) per questi coefficienti, che ci danno un'idea della precisione delle nostre stime.

Esempio. Prendiamo per esempio un dataset così formato:

⁵Ricordiamo che uno stimatore è qualcosa che ci fornisce una stima dei parametri della popolazione basata sui dati campionari.

	displacement	cylinders	horsepower	weight	acceleration	model_year	origin	mpg
0	307.0	8	130.0	3504	12.0	70	1	18.0
1	350.0	8	165.0	3693	11.5	70	1	15.0
2	318.0	8	150.0	3436	11.0	70	1	18.0
3	304.0	8	150.0	3433	12.0	70	1	16.0
4	302.0	8	140.0	3449	10.5	70	1	17.0
:	:	:	:	:	:	:	:	:
393	140.0	4	86.0	2790	15.6	82	1	27.0
394	97.0	4	52.0	2130	24.6	82	2	44.0
395	135.0	4	84.0	2295	11.6	82	1	32.0
396	120.0	4	79.0	2625	18.6	82	1	28.0
397	119.0	4	82.0	2720	19.4	82	1	31.0

298 rows × 8 columns

Tabella 8.1: Esempio di dataset con caratteristiche delle automobili

Da qui creiamo il modello:

$$\text{mpg} \approx \beta_0 + \text{horsepower} \cdot \beta_1 + \epsilon$$

E otteniamo i seguenti risultati:

COEFFICIENT	SE	INTERVALLI DI CONFIDENZA
$\hat{\beta}_0 = 39.94$	0.717	[38.53, 41.35]
$\hat{\beta}_1 = -0.1578$	0.006	[-0.17, -0.15]

Tabella 8.2: Stima dello standard error e degli intervalli di confidenza per i coefficienti della regressione

Da gli intervalli di confidenza nella tabella 8.2 si può notare come il coefficiente $\hat{\beta}_1$ abbia un intervallo di confidenza che non include lo zero, suggerendo che esiste una relazione significativa tra la variabile indipendente (horsepower) e la variabile dipendente (mpg). Inoltre, lo standard error relativamente basso indica che la stima del coefficiente è precisa.

Un tipo di plot a cui si può fare riferimento a uno scatter plot con la retta di regressione e gli intervalli di confidenza intorno alla retta stessa, come mostrato in figura 8.3 (più approfondito nella sotto-sezione 17.4.2).

8.2.4 Test statistici per la significatività dei coefficienti

Nell'esempio di prima abbiamo considerato i coefficienti come stimatori della popolazione. Quindi possiamo eseguire dei test statistici per verificare se questi coefficienti sono significativamente diversi da zero. Questo perché per $\beta_1 = 0$ non esiste correlazione tra le variabili X e Y e il regressore diventa:

$$Y = \beta_0 + \epsilon$$

che non dipende da X . Eseguiamo dunque un test di ipotesi, definendo l'*ipotesi nulla*:

$$H_0 : \text{Non c'è associazione tra } X \text{ e } Y \Leftrightarrow \beta_1 = 0$$

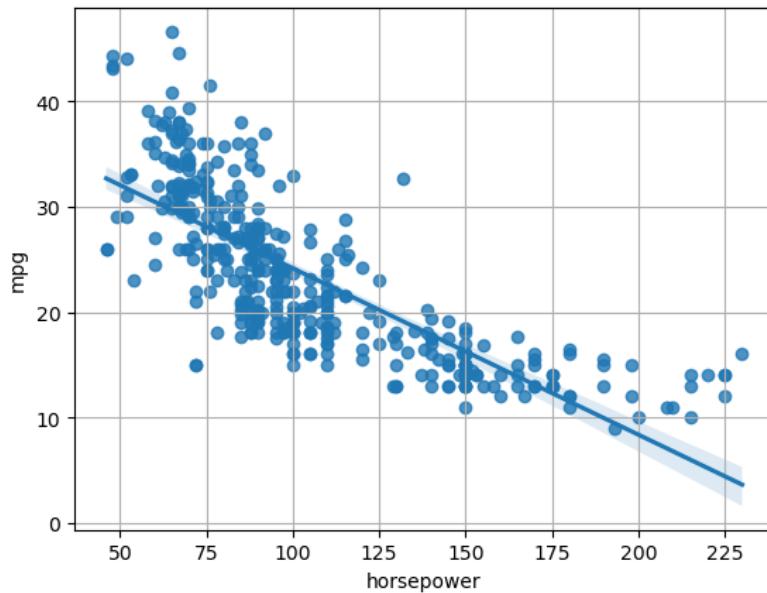


Figura 8.3: Scatter plot della relazione tra *horsepower* e *mpg* con retta di regressione stimata e relativo intervallo di confidenza.

e l'*ipotesi alternativa*:

$$H_a : \text{C'è associazione tra } X \text{ e } Y \Leftrightarrow \beta_1 \neq 0$$

Per testare queste ipotesi, si può utilizzare il test *t*-Student, calcolando il valore *t* come:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Dove $SE(\hat{\beta}_1)$ è lo standard error del coefficiente stimato. Confrontando il valore *t* calcolato con la distribuzione *t*-Student con $n - 2$ gradi di libertà (dove n è il numero di osservazioni), si può determinare il valore *p* associato. Se il *p*-value è inferiore a una soglia di significatività predefinita (ad esempio, 0.05), si rifiuta l'ipotesi nulla, suggerendo che esiste una relazione significativa tra X e Y .

Si può eseguire lo stesso test anche per l'intercetta β_0 , anche se il modello non cambierà molto perché non è un coefficiente di una variabile indipendente. Si possono comunque aggiornare i risultati:

COEFFICIENT	SE	<i>t</i>	$P > t $	INTERVALLI DI CONFIDENZA
$\hat{\beta}_0 = 39.94$	0.717	55.66	0.000	[38.53, 41.35]
$\hat{\beta}_1 = -0.1578$	0.006	-26.30	0.000	[-0.17, -0.15]

Tabella 8.3: Stima dello standard error e degli intervalli di confidenza per i coefficienti della regressione

8.3 Valutazione del modello di regressione

Dopo aver stimato i coefficienti del modello di regressione, è importante valutare la bontà del modello. I test statistici ci dicono se una relazione statistica significativa esiste, ma non ci dicono quanto bene il modello si **adatta** ai dati o quanto sono **accurate** le predizioni del modello, per le quali si usano metriche diverse:

Metriche per la bontà del modello. Queste metriche misurano quanto bene il modello si adatta ai dati osservati. Un esempio comune è il coefficiente di determinazione R^2 , che indica la proporzione della varianza nella variabile dipendente che è spiegata dalle variabili indipendenti nel modello. Un valore di R^2 vicino a 1 indica un buon adattamento del modello ai dati.

Metriche per l'accuratezza delle predizioni. Queste metriche valutano quanto accurate sono le predizioni del modello sui dati nuovi o non visti. Esempi comuni includono l'Errore Quadratico Medio (MSE) e l'Errore Assoluto Medio (MAE). Queste metriche misurano la differenza tra i valori predetti dal modello e i valori effettivi osservati. Un valore più basso di MSE o MAE indica una maggiore accuratezza delle predizioni del modello.

Residui e RSS. Tutte le metriche di regressione sono costruite sui **residui** (sotto-sezione 8.2.2), che rappresentano la differenza tra i valori osservati e i valori predetti dal modello. La somma dei quadrati dei residui (RSS) è una misura comune della bontà del modello, che quantifica la quantità totale di errore nelle predizioni del modello. Un valore più basso di RSS indica un miglior adattamento del modello ai dati.

8.3.1 Metriche per la bontà del modello

Generalmente il nostro obiettivo in queste metriche è capire⁶ quanto il modello spiega i dati osservati, ovvero quanto della variabilità totale della variabile dipendente Y è spiegata dalle variabili indipendenti X .

RSE: Residual Standard Error. Il Residual Standard Error (RSE) è una misura della dispersione dei residui intorno alla retta di regressione stimata. Si calcola come la radice quadrata della somma dei quadrati dei residui divisa per i gradi di libertà:

$$RSE = \sqrt{\frac{RSS}{n - p - 1}} = \sqrt{\frac{1}{n - p - 1} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

dove n è il numero di osservazioni e p è il numero di variabili indipendenti nel modello, ovvero i "gradi di libertà" del modello, nel costro caso di regressione semplice $p = 1 \Rightarrow$ il denominatore diventa $n - 2$.

Poiché l'RSE è una misura della dispersione dei residui, un valore più basso di RSE indica che i residui sono più vicini alla retta di regressione stimata, suggerendo un miglior adattamento del modello ai dati. Inoltre, l'RSE è espresso nelle stesse unità della variabile dipendente Y , rendendolo interpretabile nel contesto del problema di regressione. Ha tuttavia un difetto: non è

⁶Nel caso della statistica, capire spesso significa *inferire*: ovvero stimare parametri e testare ipotesi

normalizzato, quindi non permette di confrontare modelli con variabili dipendenti diverse e per capire se è buono bisogna confrontarlo con la scala dei valori di Y .

Statistica di R^2 . La statistica di R^2 (coefficiente di determinazione) è una misura assoluta della bontà del modello di regressione. Essa quantifica la proporzione della varianza nella variabile dipendente Y che è spiegata dalle variabili indipendenti X nel modello. Per calcolare, dobbiamo prima calcolare la somma totale dei quadrati (TSS), che rappresenta la variabilità totale nella variabile dipendente:

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

Dove \bar{y} è la media dei valori osservati di Y . A questo punto, possiamo calcolare R^2 come:

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{TSS - RSS}{TSS}$$

Il valore di R^2 varia tra 0 e 1, dove un valore di 0 indica che il modello non spiega alcuna variabilità nella variabile dipendente, mentre un valore di 1 indica che il modello spiega tutta la variabilità. In generale, un valore più alto di R^2 indica un miglior adattamento del modello ai dati. Tuttavia, è importante notare che un alto valore di R^2 non implica necessariamente che il modello sia appropriato o che le variabili indipendenti siano causalmente correlate alla variabile dipendente, perché questa misura spiega la correlazione ma non la causalità.

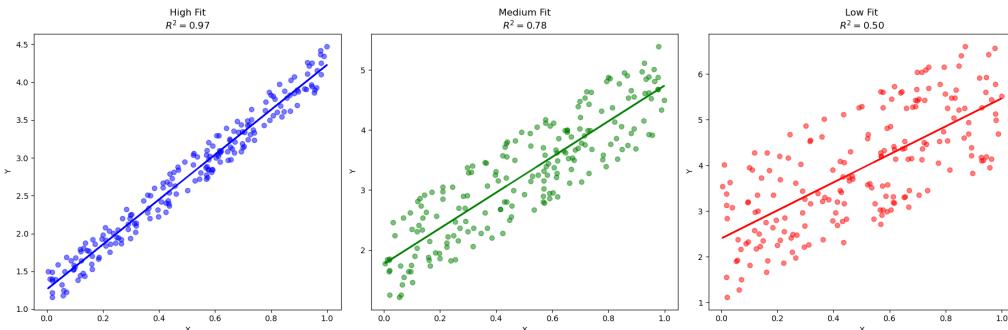


Figura 8.4: Confronto tra tre modelli di regressione lineare con livelli diversi di varianza del rumore: *High Fit* con $R^2 = 0.97$, *Medium Fit* con $R^2 = 0.78$ e *Low Fit* con $R^2 = 0.50$. All'aumentare della variabilità dei dati attorno alla retta di regressione, la bontà dell'adattamento diminuisce.

8.3.2 Grafici di diagnostica

Le metriche sono unicamente numeriche e quantificano l'errore, senza però spiegare *perché* l'errore esiste. Anche qui, si utilizzano i residui per fare dei grafici di diagnostica che aiutano a capire se il modello è adatto o meno ai dati.

Grafico dei residui vs valori predetti. Un grafico comune è il grafico dei residui contro i valori predetti. In questo grafico, i residui e_i sono tracciati sull'asse delle ordinate contro i valori predetti \hat{y}_i sull'asse delle ascisse. Questo grafico aiuta a identificare eventuali pattern nei residui che potrebbero indicare problemi con il modello di regressione. Esistono tre casi possibili:

- I residui sono distribuiti casualmente intorno allo zero, senza pattern evidente. Questo indica che il modello di regressione è appropriato per i dati.
- Si forma una "U" o una \cap nei residui, suggerendo che il modello di regressione lineare non cattura una relazione non lineare tra X e Y . In questo caso, potrebbe essere necessario considerare un modello di regressione non lineare.
- La dispersione dei residui aumenta o diminuisce con i valori predetti, indicando eteroschedasticità⁷. Questo suggerisce che la variabilità dei residui non è costante e potrebbe richiedere una trasformazione delle variabili o l'uso di un modello di regressione più robusto.

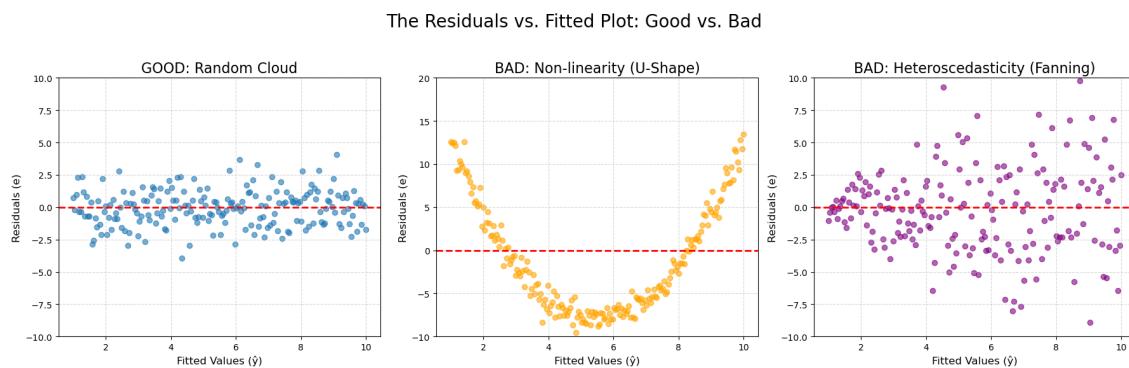


Figura 8.5: Confronto tra diversi pattern nei grafici Residuals vs. Fitted. A sinistra un caso “buono”, in cui i residui formano una nube casuale attorno allo zero, indicando che il modello lineare è appropriato. Al centro un caso “cattivo”, dove i residui mostrano una chiara struttura a U, segnalando non-linearità. A destra un altro caso “cattivo”, in cui la dispersione dei residui aumenta con i valori stimati, evidenziando eteroschedasticità (effetto “fanning”).

Grafici Q-Q (Quantile-Quantile). Un altro grafico utile è il grafico Q-Q (Quantile-Quantile), che confronta la distribuzione dei residui con una distribuzione normale teorica. In questo grafico, i quantili dei residui sono tracciati contro i quantili di una distribuzione normale. Se i residui seguono una distribuzione normale, i punti nel grafico Q-Q dovrebbero allinearsi lungo una linea retta diagonale. Eventuali deviazioni significative da questa linea indicano che i residui non seguono una distribuzione normale, suggerendo che il modello di regressione potrebbe non essere appropriato per i dati.

8.4 Regressione lineare multivariata

La regressione lineare multivariata estende il concetto di regressione lineare semplice includendo più variabili indipendenti. In questo caso, il modello di regressione può essere espresso come:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

Dove X_1, X_2, \dots, X_p sono le variabili indipendenti e p è il numero di variabili indipendenti nel modello. I coefficienti $\beta_1, \beta_2, \dots, \beta_p$ rappresentano l'effetto di ciascuna variabile indipendente

⁷L'eteroschedasticità è una forma di non costanza della varianza degli errori.

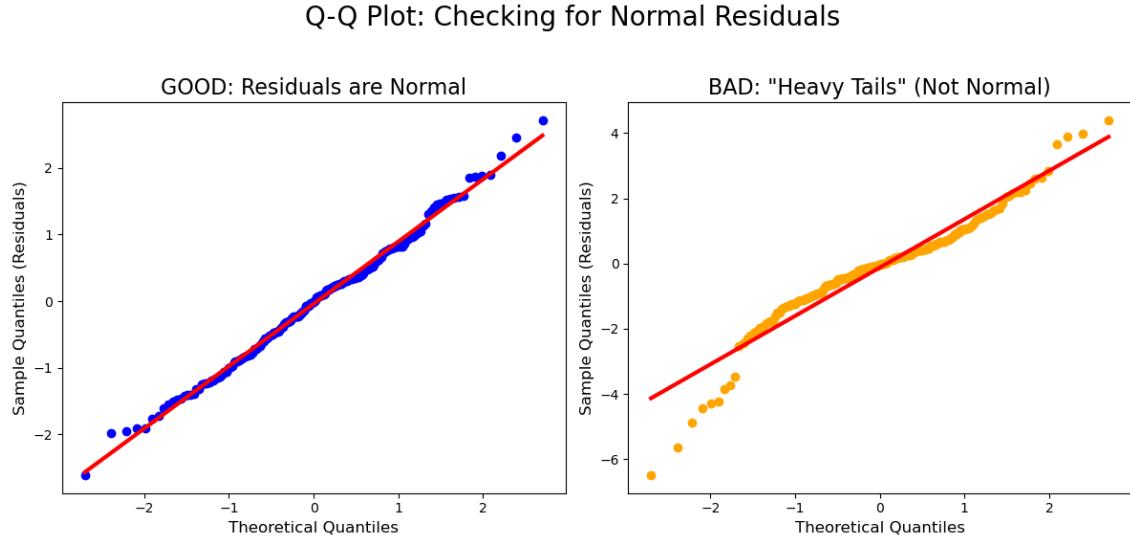


Figura 8.6: Confronto tra due Q–Q plot dei residui. A sinistra un caso “buono”, in cui i punti seguono approssimativamente la linea rossa, indicando che i residui possono essere considerati normalmente distribuiti. A destra un caso “cattivo”, dove le code pesanti (heavy tails) producono deviazioni marcate dalla linea teorica, mostrando che i residui non seguono una distribuzione normale.

sulla variabile dipendente Y , tenendo conto dell’effetto delle altre variabili indipendenti nel modello.

Questa generalizzazione viene utilizzata quando si vuole studiare l’effetto di più variabili esplicative su una variabile di interesse, permettendo di controllare per l’influenza di altre variabili e di ottenere stime più accurate degli effetti delle singole variabili indipendenti.

Nell’esempio della tabella 8.1, si potrebbe costruire un modello:

$$mpg \approx \beta_0 + \beta_1 \cdot \text{horsepower} + \beta_2 \cdot \text{weight} + \beta_3 \cdot \text{model_year} + \epsilon$$

e ottenere un valore di $R^2 = 0.808$ rispetto a $R^2 = 0.682$ del modello semplice con solo *horsepower* come variabile indipendente, suggerendo che l’inclusione di più variabili indipendenti migliora la capacità del modello di spiegare la variabilità nella variabile dipendente *mpg*.

8.4.1 Interpretazione geometrica

In un contesto multivariato, la regressione lineare può essere interpretata geometricamente come la ricerca di un iperpiano che meglio si adatta ai dati in uno spazio a più dimensioni. Ogni variabile indipendente X_j rappresenta una dimensione nello spazio, e la variabile dipendente Y rappresenta l’altezza dell’iperpiano in quella posizione.

8.4.2 Interpretazione statistica

Generalmente interpretare statisticamente un modello di regressione lineare multivariata è simile all’interpretazione della regressione lineare semplice. Dato un modello:

$$Y = \beta_0 + \beta \cdot x$$

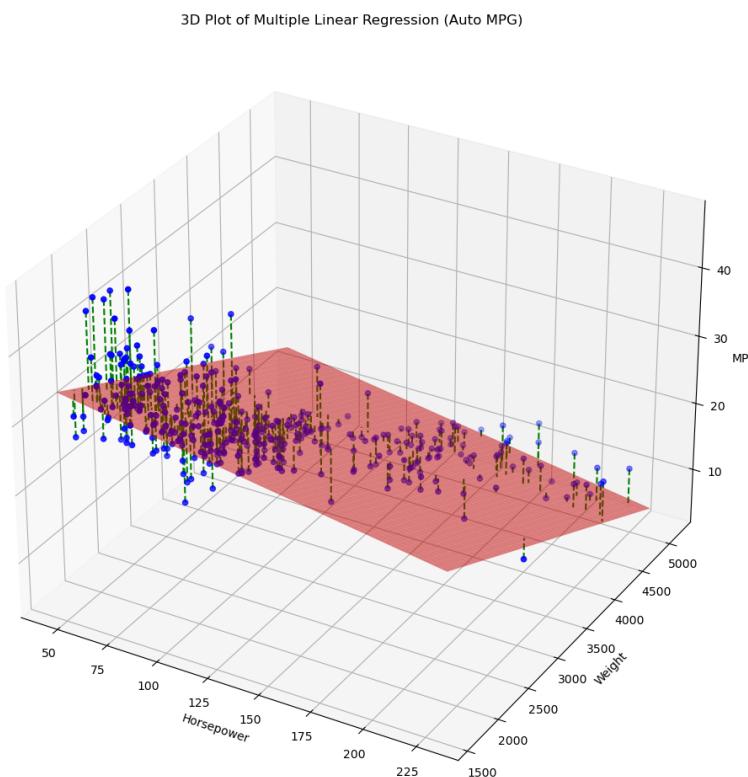


Figura 8.7: Rappresentazione tridimensionale della regressione lineare multipla sul dataset Auto MPG. I punti indicano i valori osservati di *mpg* in funzione di *horsepower* e *weight*, mentre il piano rosso rappresenta il piano di regressione stimato. Le linee verticali mostrano la distanza tra i valori osservati e quelli predetti dal modello, evidenziando l'errore di regressione.

dove x è un vettore di variabili indipendenti e β è un vettore di coefficienti associati a ciascuna variabile indipendente. Si può interpretare:

- Il valore di β_0 come l'intercetta del modello, ovvero il valore atteso di Y quando tutte le variabili indipendenti sono uguali a zero.
- Il valore di un certo β_i , con $i \in [1, p]$, come il cambiamento atteso in Y associato a un aumento unitario della variabile indipendente X_i , mantenendo costanti tutte le altre variabili indipendenti nel modello. Questo permette di isolare l'effetto di ciascuna variabile indipendente sulla variabile dipendente, tenendo conto dell'influenza delle altre variabili nel modello.

Come variano i coefficienti al variare delle variabili indipendenti. Da questa interpretazione statistica, possiamo dire che: aggiungere nuove variabili al regressore cambia il modo in cui la varianza di Y viene “spiegata” dai coefficienti. Nel modello semplice

$$mpg = \beta_0 + \beta_1 \cdot \text{horsepower}$$

otteniamo ad esempio $\hat{\beta}_0 \approx 39.94$ e $\hat{\beta}_1 \approx -0.16$. Quando introduciamo anche il *weight*,

$$mpg = \beta_0 + \beta_1 \cdot \text{horsepower} + \beta_2 \cdot \text{weight},$$

le stime diventano circa $\hat{\beta}_0 \approx 45.64$ e $\hat{\beta}_1 \approx -0.05$. Questo non significa che uno dei due modelli sia “sbagliato”, ma che nel primo caso l'effetto di *horsepower* ingloba anche parte della variabilità dovuta al *weight*⁸ (e ad altre variabili non osservate), mentre nel secondo caso ogni coefficiente descrive l'effetto della propria variabile a parità delle altre.

8.4.3 Stima dei coefficienti di regressione

Dato il modello generale di regressione lineare multivariata:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

Possiamo definire la **funzione di costo** come la somma dei quadrati dei residui (RSS):

$$RSS(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}))^2$$

dove n è il numero di osservazioni, y_i è il valore osservato della variabile dipendente per l'osservazione i , e x_{ij} è il valore della variabile indipendente X_j per l'osservazione i .

I valori $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ che minimizzano la funzione di costo RSS possono essere trovati

⁸Ecco perché il test R^2 non spiega completamente la bontà del modello.

risolvendo il seguente sistema di equazioni:

$$\begin{cases} y^{(1)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(1)} + \hat{\beta}_2 x_2^{(1)} + \dots + \hat{\beta}_p x_p^{(1)} \\ y^{(2)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(2)} + \hat{\beta}_2 x_2^{(2)} + \dots + \hat{\beta}_p x_p^{(2)} \\ \vdots \\ y^{(n)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(n)} + \hat{\beta}_2 x_2^{(n)} + \dots + \hat{\beta}_p x_p^{(n)} \end{cases}$$

trasformato in forma matriciale come:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Dove:

- \mathbf{Y} è il vettore colonna dei valori osservati della variabile dipendente.
- \mathbf{X} è la matrice delle variabili indipendenti, con una colonna di 1s per l'intercetta, chiamata anche **design matrix**.
- $\boldsymbol{\beta}$ è il vettore colonna dei coefficienti di regressione da stimare.
- $\boldsymbol{\epsilon}$ è il vettore colonna degli errori.

Dalla notazione sopra, possiamo derivare la soluzione per i coefficienti stimati ricordando che la somma dei quadrati dei residui può essere espressa come:

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^n (e^{(i)})^2 = \mathbf{e}^T \mathbf{e}$$

Minimizzando la funzione di costo, utilizzando sempre il metodo dei *minimi quadrati ordinari* (OLS), otteniamo la seguente formula per i coefficienti stimati:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

8.4.4 F-Test

Per valutare la significatività complessiva di un modello di regressione lineare multivariata, si può utilizzare il test F. Questo test confronta un modello completo (con tutte le variabili indipendenti) con un modello ridotto (senza alcune o tutte le variabili indipendenti) per determinare se l'inclusione delle variabili indipendenti migliora significativamente l'adattamento del modello ai dati. Nella realtà, non è altro che definire un test d'ipotesi:

$$H_0 : \text{Tutte le variabili indipendenti non hanno effetto su } Y \Leftrightarrow \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_a : \text{Almeno una variabile indipendente ha un effetto su } Y \Leftrightarrow \exists i : \beta_i \neq 0$$

8.4.5 Eliminazione backward delle variabili

L'eliminazione backward è una tecnica di selezione delle variabili utilizzata nella regressione lineare multivariata per identificare un sottoinsieme ottimale di variabili indipendenti da includere nel modello. Questo metodo inizia con un modello completo che include tutte le variabili

indipendenti disponibili e procede rimuovendo iterativamente le variabili meno significative fino a quando non si raggiunge un criterio di arresto predefinito. In particolare, si elimina quella variabile il cui valore p associato al test t è il più alto (quindi la meno significativa) e si ripete il processo finché tutte le variabili rimanenti sono statisticamente significative secondo una soglia di significatività scelta (ad esempio, 0.05).

8.4.6 Colinearità e instabilità di OLS

La colinearità si verifica quando due o più variabili indipendenti in un modello di regressione lineare sono altamente correlate tra loro. Questo può portare a problemi nell'interpretazione dei coefficienti di regressione e a instabilità nelle stime dei coefficienti stessi. In presenza di colinearità, le stime dei coefficienti possono diventare molto sensibili a piccole variazioni nei dati, rendendo difficile determinare l'effetto individuale di ciascuna variabile indipendente sulla variabile dipendente. Inoltre, la colinearità può aumentare lo standard error delle stime dei coefficienti, riducendo la precisione delle stime e rendendo più difficile rilevare relazioni significative tra le variabili indipendenti e la variabile dipendente.

Definizione 8.2

Prendiamo per esempio la soluzione di OLS per i coefficienti stimati:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Se due o più variabili indipendenti sono altamente correlate, la matrice $X^T X$ può diventare prossima alla **singolarità**, ovvero una matrice non invertibile poiché le righe sono linearmente *dipendenti* e non è applicabile il teorema della matrice inversa⁹, rendendo il calcolo dell'inversa numericamente instabile. Questo può portare a stime dei coefficienti che variano notevolmente con piccole modifiche nei dati, rendendo difficile interpretare i risultati del modello di regressione.

Da qui, nasce il problema su OLS: in presenza di colinearità, le stime dei coefficienti possono essere instabili e poco affidabili. Per affrontare questo problema si può fare la computazione della matrice **pseudo-inversa** di Moore-Penrose $(X^T X)^+$, che permette di calcolare una soluzione approssimata anche quando la matrice non è invertibile. Tuttavia, questa soluzione può ancora essere sensibile alla colinearità e potrebbe non fornire stime affidabili dei coefficienti.

8.4.7 Adjusted R^2

Nel caso della regressione lineare semplice $R^2 = \pi(x, y)^2$ (con π coefficiente di correlazione di Pearson). Nel caso della multivariata, si può scrivere come:

$$R^2 = \pi(Y, \hat{Y})^2$$

dove \hat{Y} sono i valori predetti dal modello di regressione. Tuttavia, un problema con R^2 nella regressione multivariata è che tende ad aumentare con l'aggiunta di più variabili indipendenti, anche se queste variabili non migliorano realmente il modello. Per affrontare questo problema, si utilizza una versione modificata di R^2 chiamata **Adjusted R^2** (o R^2 aggiustato), che tiene conto

⁹Il teorema della matrice inversa afferma che una matrice quadrata è invertibile se e solo se ha rango massimo, ovvero il rango è uguale alla dimensione della matrice.

del numero di variabili indipendenti nel modello. Da notare che anche questo è un problema di **bias-varianza trade-off**, in quanto aggiungere più variabili riduce il bias ma aumenta la varianza.

Si può esprimere come:

$$\bar{R}^2 = 1 - (1 - R^2) \frac{m - 1}{m - p - 1}$$

dove m è il numero di osservazioni e p è il numero di variabili indipendenti nel modello. L'Adjusted R^2 penalizza l'aggiunta di variabili indipendenti che non migliorano significativamente il modello, fornendo una misura più accurata della bontà del modello in presenza di più variabili indipendenti.

8.5 Predittori qualitativi

In molti casi, le variabili indipendenti in un modello di regressione possono essere di natura qualitativa (categorica) piuttosto che quantitativa (numerica). Ad esempio, una variabile qualitativa potrebbe rappresentare il genere (maschio/femmina), il colore (rosso/blu/verde) o la categoria di un prodotto (A/B/C). Per includere queste variabili qualitative in un modello di regressione lineare, è necessario convertirle in una forma numerica.

8.5.1 Variabili dummy

Un metodo comune per includere variabili qualitative in un modello di regressione è l'uso di **variabili dummy**. Le variabili dummy sono variabili binarie (0 o 1) che indicano la presenza o l'assenza di una particolare categoria di una variabile qualitativa. Per una variabile qualitativa con k categorie, si creano $k - 1$ variabili dummy, dove ciascuna variabile dummy rappresenta una categoria specifica, e la categoria rimanente viene utilizzata come categoria di riferimento (baseline). Continuiamo l'esempio del dataset 8.1 e supponiamo di voler introdurre la variabile qualitativa:

$$\text{fuel_type[T.gas]} = \begin{cases} 1 & \text{se il tipo di carburante è gas} \\ 0 & \text{altrimenti} \end{cases}$$

da questo poi fittiamo il modello:

$$mpg = \beta_0 + \beta_1 \cdot \text{horsepower} + \beta_2 \cdot \text{fuel_type[T.gas]} + \epsilon$$

ottenendo i seguenti risultati: Da qui si può notare come il coefficiente associato alla variabile

	coef	std err	t	P> t	[0.025	0.975]
Intercept	41.2379	1.039	39.705	0.000	39.190	43.286
fueltype[T.gas]	-2.7658	0.918	-3.013	0.003	-4.576	-0.956
horsepower	-0.1295	0.007	-18.758	0.000	-0.143	-0.116

Tabella 8.4: Risultati della regressione

dummy fuel_type[T.gas] sia negativo, suggerendo che, a parità di horsepower , le automobili che utilizzano gas come carburante tendono ad avere un valore di mpg inferiore rispetto a quelle che utilizzano altri tipi di carburante (categoria di riferimento). Inoltre, il valore p associato a questo coefficiente è inferiore a 0.05, indicando che l'effetto del tipo di carburante sulla variabile dipendente è statisticamente significativo.

Predittori con più di due categorie Considerato un predittore a n categorie, si possono creare $n - 1$ variabili dummy per rappresentare le categorie. Per esempio, si potrebbe avere una variabile qualitativa *color* con tre categorie: rosso, blu e verde. Si possono creare due variabili dummy:

$$\text{color_red} = \begin{cases} 1 & \text{se il colore è rosso} \\ 0 & \text{altrimenti} \end{cases}$$

$$\text{color_blue} = \begin{cases} 1 & \text{se il colore è blu} \\ 0 & \text{altrimenti} \end{cases}$$

La categoria verde viene utilizzata come categoria di riferimento. Includendo queste variabili dummy in un modello di regressione, si può valutare l'effetto di ciascun colore sulla variabile dipendente rispetto alla categoria di riferimento (verde).

Capitolo 9

Oltre la regressione lineare

Un problema evidente nell'esempio del dataset 8.1 è che la relazione tra le variabili non è lineare. In questi casi, l'uso di una regressione lineare semplice non è adatto per modellare i dati.

9.1 Interazione tra variabili

Un modo per affrontare la non linearità è introdurre termini di interazione tra le variabili. Un modello del dataset così formato:

$$\text{mpg} = \beta_0 + \beta_1 \cdot \text{horsepower} + \beta_2 \cdot \text{weight}$$

può essere esteso includendo un termine di interazione:

$$\text{mpg} = \beta_0 + \beta_1 \cdot \text{horsepower} + \beta_2 \cdot \text{weight} + \beta_3 \cdot (\text{horsepower} \times \text{weight})$$

Questo modello consente di catturare l'effetto combinato di *horsepower* e *weight* sulla variabile dipendente *mpg*. I risultati di questo modello sono:

Additive Model (R-squared) Adjusted $R^2 = 0.7049$

Interaction Model (R-squared) Adjusted $R^2 = 0.7465$

	coef	std err	t	P> t	[0.025, 0.975]
Intercept	63.5579	2.343	27.127	0.000	58.951 – 68.164
horsepower	-0.2508	0.027	-9.195	0.000	-0.304 – -0.197
weight	-0.0108	0.001	-13.921	0.000	-0.012 – -0.009
horsepower:weight	5.355e-05	6.65e-06	8.054	0.000	4.05e-05 – 6.66e-05

Tabella 9.1: Coefficients for the Interaction Regression Model

9.1.1 Interpretazione dei coefficienti

Dai risultati, si evince che il termine di interazione tra *horsepower* e *weight* è statisticamente significativo (*p-value* < 0.05). Questo indica che l'effetto di *horsepower* sul *mpg* dipende dal valore di *weight*, e viceversa.

Possiamo riscrivere l'equazione del modello come:

$$\text{mpg} = \beta_0 + (\beta_1 + \beta_3 \cdot \text{weight}) \cdot \text{horsepower} + \beta_2 \cdot \text{weight}$$

Ciò significa che l'effetto di horsepower non è più il valore fisso di β_1 , ma varia in funzione del peso dell'auto. Ad esempio, per un'auto più pesante, l'effetto negativo di un aumento di horsepower sul mpg sarà maggiore rispetto a un'auto più leggera. L'effetto può essere visto nella figura 9.1, da cui si nota come il modello con interazione migliori la distribuzione dei residui e del Q-Q plot rispetto al modello additivo.

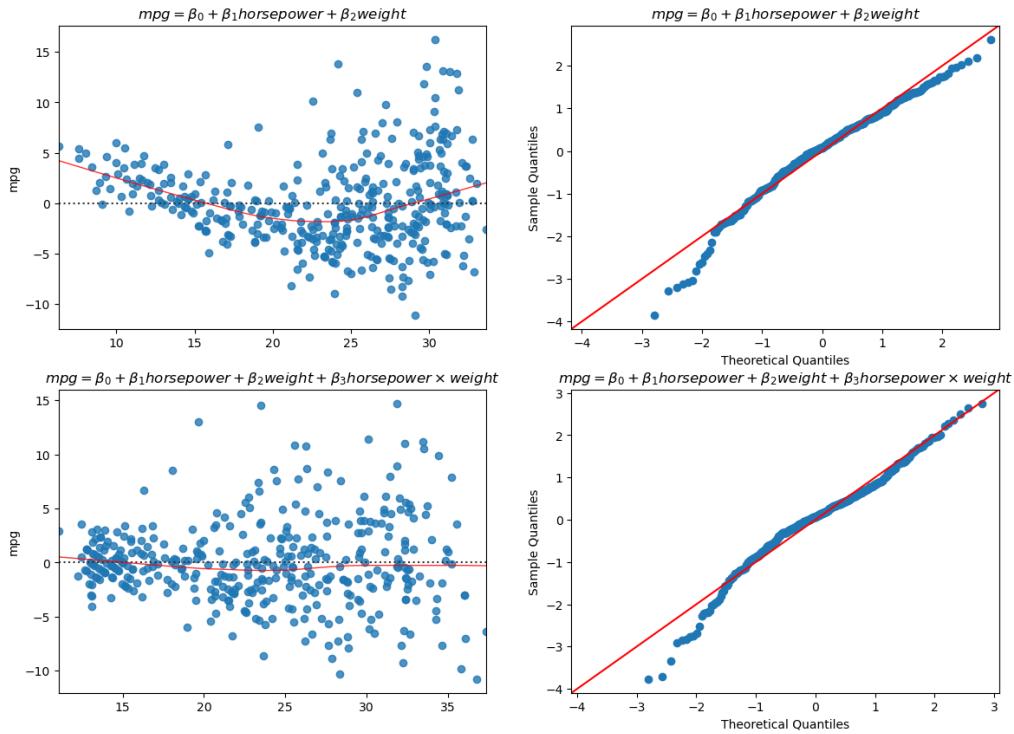


Figura 9.1: Confronto tra il modello additivo $\text{mpg} = \beta_0 + \beta_1 \text{horsepower} + \beta_2 \text{weight}$ (prima riga) e il modello con interazione $\text{mpg} = \beta_0 + \beta_1 \text{horsepower} + \beta_2 \text{weight} + \beta_3 \text{horsepower} \times \text{weight}$ (seconda riga). A sinistra sono riportati i grafici dei residui, a destra i Q-Q plot dei residui rispetto alla distribuzione normale teorica.

9.2 Regressione polinomiale

Un altro approccio per modellare relazioni non lineari è l'uso di regressioni polinomiali, ovvero modelli che includono termini di potenza delle variabili indipendenti.

9.2.1 Regressione quadratica

Ad esempio, possiamo estendere il modello originale includendo un termine quadratico per *horsepower*:

$$\text{mpg} = \beta_0 + \beta_1 \cdot \text{horsepower} + \beta_2 \cdot \text{horsepower}^2$$

e notando che l'aggiunta del termine quadratico migliora l'adattamento del modello ai dati. I risultati di questo modello sono:

	coef	std err	t	P> t	[0.025, 0.975]
Intercept	56.9001	1.800	31.604	0.000	53.360 – 60.440
horsepower	-0.4662	0.031	-14.978	0.000	-0.527 – -0.405
I(horsepower ²)	0.0012	0.000	10.080	0.000	0.001 – 0.001

Tabella 9.2: Risultati della regressione con termine quadratico

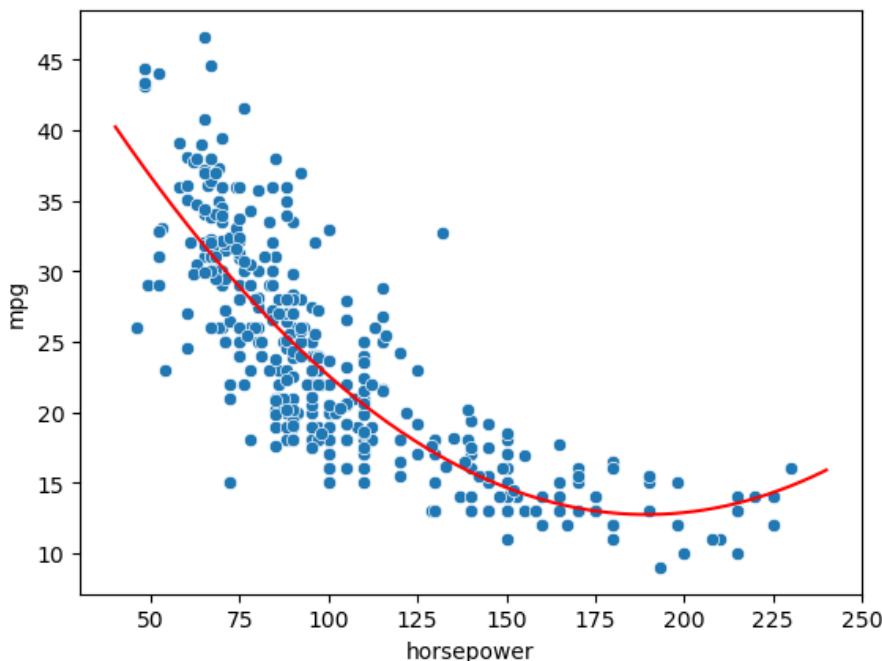


Figura 9.2: Curva del modello di regressione polinomiale di secondo grado per il dataset delle auto. Si nota come la curva si adatti meglio ai dati rispetto a una retta.

9.2.2 Polinomi di grado superiore

Possiamo anche considerare polinomi di grado superiore, come un modello di grado 4:

$$\text{mpg} = \beta_0 + \beta_1 \cdot \text{horsepower} + \beta_2 \cdot \text{horsepower}^2 + \beta_3 \cdot \text{horsepower}^3 + \beta_4 \cdot \text{horsepower}^4$$

Questo modello fitta i dati in modo ancora più accurato, come mostrato nella figura 9.3.

9.2.3 Problema dell'interpretabilità

Un aspetto importante da considerare quando si utilizzano modelli polinomiali è l'interpretabilità dei coefficienti. A differenza della regressione lineare semplice, dove ogni coefficiente rappresenta l'effetto marginale di una variabile indipendente sulla variabile dipendente, nei modelli polinomiali i coefficienti delle potenze superiori non hanno un'interpretazione diretta.

Proviamo a interpretare il modello con grado 4. In questo caso i coefficienti indicano l'effetto complessivo delle variazioni di *horsepower* sulla variabile *mpg*. Provando a isolarli, si può descri-

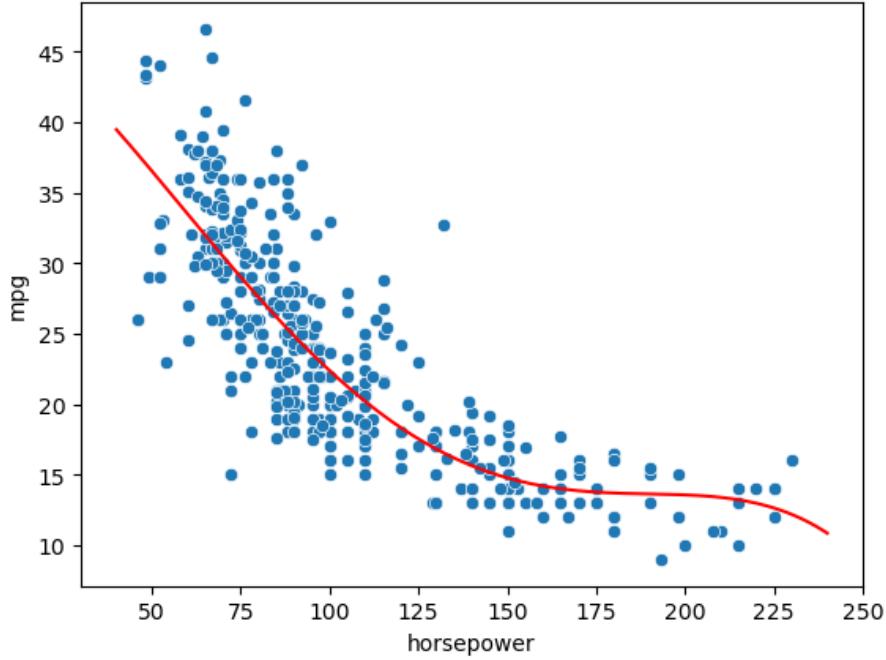


Figura 9.3: Curva del modello di regressione polinomiale di quarto grado per il dataset delle auto. Si nota come la curva si adatti molto bene ai dati.

vere β_1 come l'effetto marginale iniziale di un aumento di *horsepower*, ma gli effetti di β_2 , β_3 e β_4 dipendono dai valori specifici di *horsepower*. Ad esempio, l'effetto di un aumento di 1 unità di *horsepower* quando *horsepower* è basso sarà diverso rispetto a quando *horsepower* è alto, a causa dei termini di potenza superiore.

Notiamo subito una cosa: l'interpretazione dei coefficienti diventa complessa e meno intuitiva man mano che complichiamo il modello. Da qui nasce il paradigma *machine learning*, che si concentra più sulla capacità predittiva del modello piuttosto che sull'interpretabilità dei coefficienti.

9.2.4 Metriche di predizione

Poiché l'obiettivo principale dei modelli più complessi è la predizione accurata, è importante utilizzare metriche appropriate per valutare le prestazioni del modello sul test set (ovvero valutare l'errore).

9.2.5 MSE: Mean Squared Error

Una metrica comune per valutare le prestazioni di un modello di regressione è il Mean Squared Error (MSE), che misura la media dei quadrati degli errori tra i valori previsti e i valori effettivi:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

dove y_i sono i valori effettivi, \hat{y}_i sono i valori previsti dal modello, e n è il numero di osservazioni.

L'MSE ha un problema: le unità sono quadratiche rispetto alla variabile di interesse. Per esempio, se la variabile di interesse è espressa in miglia per gallone (mpg), l'MSE sarà espresso in $(\text{mpg})^2$, rendendo difficile l'interpretazione diretta dell'errore, inoltre l'MSE è sensibile ai valori anomali (outliers) a causa della natura quadratica della formula.

9.2.6 RMSE: Root Mean Squared Error

Per risolvere il problema delle unità, si può utilizzare la radice quadrata dell'MSE, nota come Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Il RMSE riporta l'errore nella stessa unità della variabile di interesse, rendendo più facile l'interpretazione. Tuttavia, rimane sensibile ai valori anomali.

9.2.7 MAE: Mean Absolute Error

La metrica che risolve entrambi i problemi precedenti è il Mean Absolute Error (MAE), che misura la media degli errori assoluti tra i valori previsti e i valori effettivi:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Il MAE è espresso nella stessa unità della variabile di interesse e non è influenzato in modo significativo dai valori anomali, poiché utilizza il valore assoluto invece del quadrato degli errori.

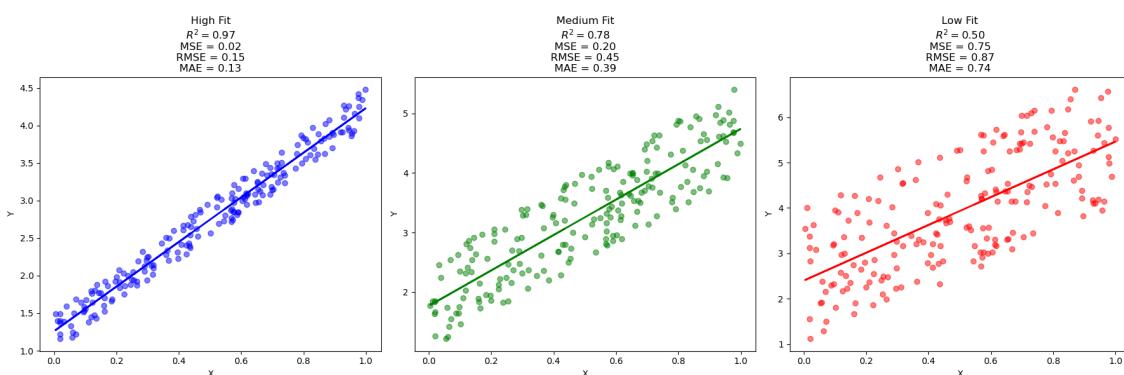


Figura 9.4: Confronto visivo tra tre modelli di regressione con diverso livello di adattamento ai dati. Da sinistra a destra: *High Fit*, *Medium Fit* e *Low Fit*. Per ciascun modello sono riportati i punti osservati, la retta di regressione stimata e le principali metriche di performance (R^2 , MSE, RMSE e MAE). Si osserva come la dispersione crescente dei dati attorno alla retta comporti un peggioramento sistematico di tutte le metriche di errore.

9.3 Overfitting

Il problema dell'*underfitting* è stato risolto con l'introduzione di modello più complessi, quadratico nel caso del dataset di auto. Ma sussiste ancora un altro problema: **l'overfitting**. L'overfitting si verifica quando un modello si adatta troppo strettamente ai dati di addestramento, catturando il rumore invece del segnale sottostante. Questo porta a una scarsa capacità di generalizzazione su nuovi dati. Per esempio, un modello di regressione polinomiale di grado molto alto potrebbe adattarsi perfettamente ai dati di addestramento, ma fallire nel prevedere correttamente i valori su un set di test.

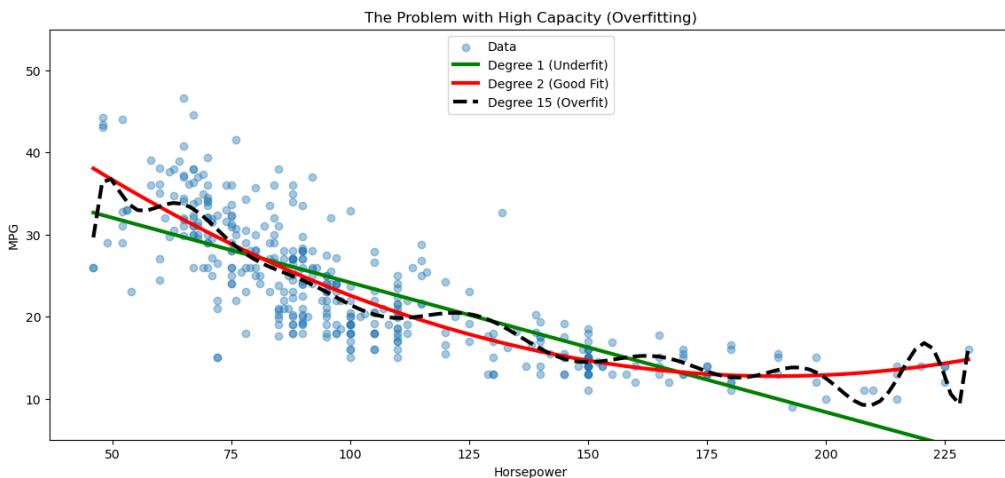


Figura 9.5: Esempio di underfitting, good fit e overfitting nel rapporto tra *horsepower* e *MPG*. La curva di grado 1 (verde) sottostima la complessità della relazione (*underfit*), quella di grado 2 (rossa) fornisce un buon compromesso tra bias e varianza, mentre il modello di grado 15 (linea tratteggiata nera) si adatta eccessivamente al rumore dei dati (*overfit*), mostrando forti oscillazioni prive di significato inferenziale.

Generalmente questo è un problema di *alta varianza*, ovvero il modello è troppo sensibile alle variazioni nei dati di addestramento (come si vede in figura 9.5)

9.3.1 Regolarizzazione

Per mitigare l'overfitting, si possono utilizzare tecniche di regolarizzazione che penalizzano la complessità del modello. La regolarizzazione aggiunge un termine di **penalità** alla funzione di costo del modello, scoraggiando coefficienti troppo grandi. Questo termine è progettato per forzare il vettore dei coefficienti β ad essere più piccolo, riducendo così la complessità del modello e migliorando la sua capacità di generalizzazione.

Ridge Regression (L2 Regularization) La Ridge Regression aggiunge una penalità basata sulla somma dei quadrati dei coefficienti:

$$\text{RSS}_{L2} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

dove λ è un iperparametro che controlla la forza della penalità. Un valore più alto di λ porta a coefficienti più piccoli.

Da qui possiamo scrivere OLS con regolarizzazione L2 come:

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

dove I è la matrice identità, quindi λI aggiunge un termine positivo alla diagonale di $X^T X$, migliorando la stabilità numerica dell'inversione della matrice.

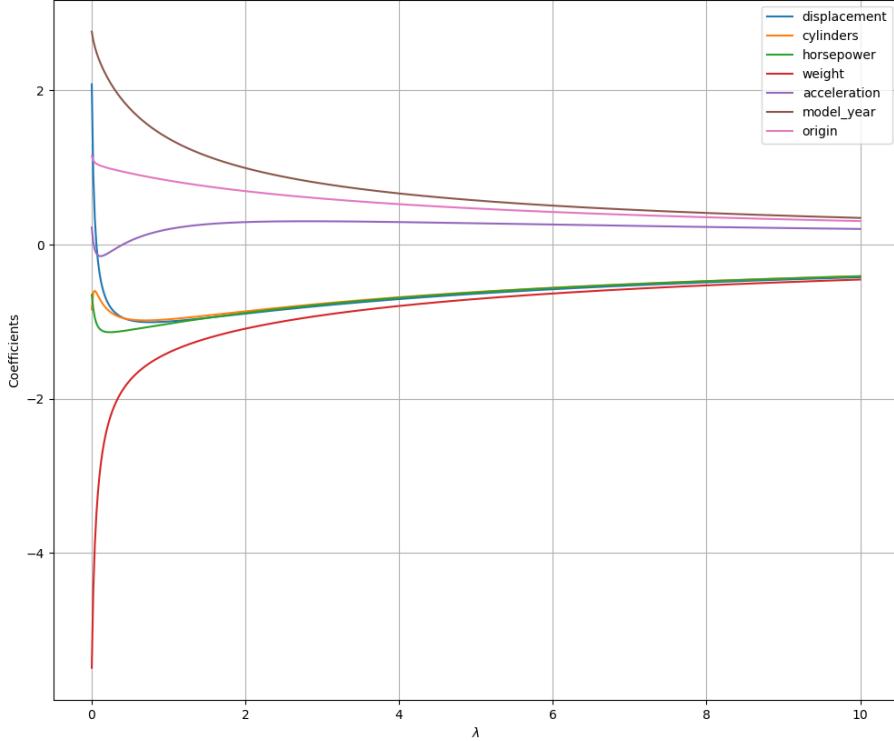


Figura 9.6: Percorsi dei coefficienti del modello Ridge al variare del parametro di penalizzazione λ . All'aumentare di λ , i coefficienti vengono spinti verso valori più piccoli in modulo, riducendo la varianza del modello. Si osserva come le variabili con minore rilevanza predittiva vengano contratte più rapidamente, mentre i predittori più informativi rimangano relativamente stabili per valori più elevati di λ .

Come si vede in figura 9.6, all'aumentare di λ , i coefficienti del modello vengono spinti verso zero, riducendo la varianza del modello e migliorando la sua capacità di generalizzazione.

Lasso Regression (L1 Regularization) La Lasso Regression utilizza una penalità basata sulla somma dei valori assoluti dei coefficienti:

$$RSS_{L1} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

Anche in questo caso, λ controlla la forza della penalità. Un aspetto interessante della Lasso è che può portare a coefficienti esattamente pari a zero, effettuando una **selezione automatica** delle

variabili. Questo rende la Lasso particolarmente utile quando si sospetta che solo un sottoinsieme delle variabili predittive sia rilevante per il modello.

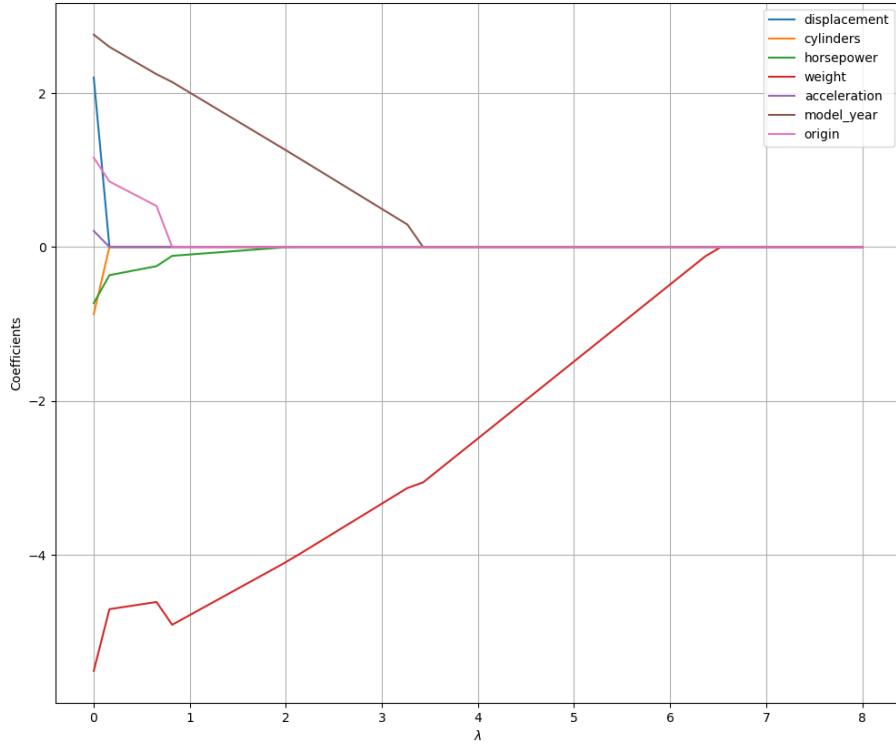


Figura 9.7: Percorsi dei coefficienti del modello Lasso al variare del parametro di penalizzazione λ . A differenza della regressione Ridge, la penalizzazione L1 forza progressivamente alcuni coefficienti esattamente a zero, effettuando una vera e propria selezione delle variabili. Si osserva che, per valori crescenti di λ , solo pochi predittori mantengono un coefficiente diverso da zero, mentre gli altri vengono eliminati dal modello.

9.3.2 Bias-Varianza trade-off con la regolarizzazione

La regolarizzazione introduce un compromesso tra bias e varianza nel modello. Aumentando la penalità (cioè aumentando λ), si riduce la varianza del modello, ma si introduce anche un certo bias, poiché il modello potrebbe non adattarsi perfettamente ai dati di addestramento. Tuttavia, questo compromesso spesso porta a una migliore performance complessiva su nuovi dati, riducendo l'overfitting. In particolare, l'iperparametro λ può essere visto come una "manopola" che regola il bilanciamento tra bias e varianza:

- Quando $\lambda = 0$, non c'è regolarizzazione e il modello può avere alta varianza (overfitting).
- Man mano che λ aumenta, la varianza diminuisce ma il bias aumenta.
- Esiste un valore ottimale di λ che bilancia bias e varianza, minimizzando l'errore di generalizzazione.
- Per $\lambda \rightarrow \infty$, tutti i coefficienti tendono a zero, portando a un modello con alto bias e bassa varianza (underfitting).

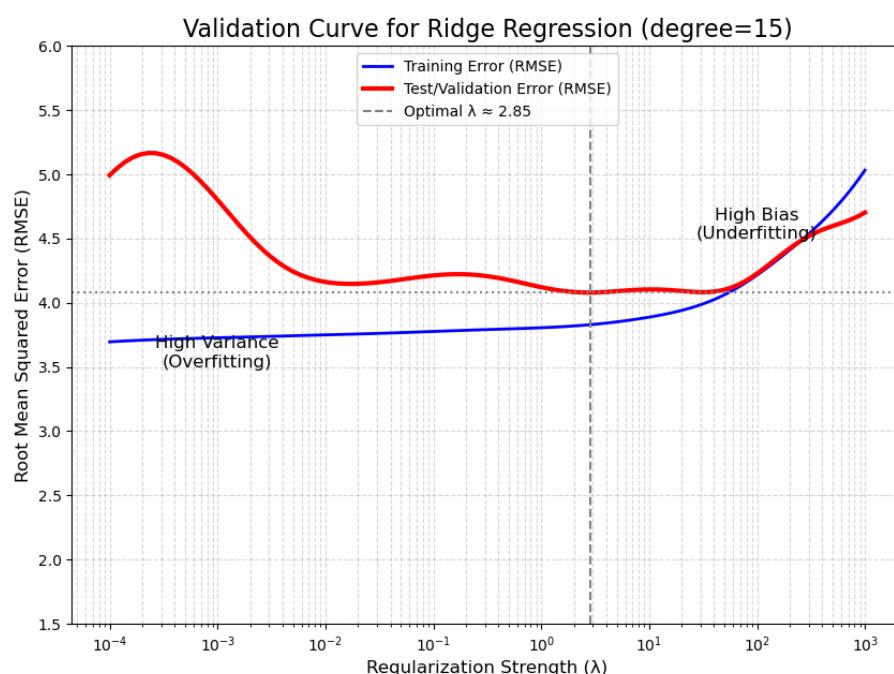


Figura 9.8: Curva di validazione per la regressione Ridge con modello polinomiale di grado 15. La penalizzazione λ controlla il compromesso tra bias e varianza: per valori molto piccoli di λ il modello soffre di alta varianza (overfitting), mentre per valori molto grandi mostra alto bias (underfitting). L'errore di validazione (curva rossa) raggiunge il minimo attorno a $\lambda \approx 2.85$, indicato dalla linea tratteggiata, che rappresenta il valore ottimale di regolarizzazione.

Capitolo 10

Classificazione

La classificazione è un altro tipo di modello *predittivo*. In particolare, la classificazione viene utilizzata quando l'obiettivo è prevedere una categoria o una classe a cui appartiene un'osservazione basandosi su un insieme di caratteristiche o attributi. A differenza della regressione, che prevede valori continui, la classificazione si occupa di variabili categoriche.

10.1 Definizione formale

Sia dato un insieme di dati D , dove ogni osservazione x_i è rappresentata da un vettore di caratteristiche $(x_{i1}, x_{i2}, \dots, x_{in})$ e associata a una classe y_i appartenente a un insieme finito di classi $C = \{c_1, c_2, \dots, c_k\}$. Il modello di classificazione può essere definito come:

$$h : \mathbb{R}^n \rightarrow C = \{c_1, c_2, \dots, c_k\}$$

dove h è la funzione di classificazione che mappa le caratteristiche dell'osservazione alla sua classe corrispondente.

Definizione 10.1

Anche in questo caso, l'insieme dei dati D viene suddiviso in training set e test set per addestrare e valutare il modello di classificazione:

$$D = \{(x_i, y_i)\}_{i=1}^N = D_{train} \cup D_{test}$$

Da questo capiamo che $x_i \in \mathbb{R}^n$ rappresenta il vettore delle caratteristiche dell'osservazione i -esima, mentre $y_i \in C$ rappresenta la classe associata a quell'osservazione, ovvero l'etichetta che vogliamo prevedere. Possiamo utilizzare il **rischio empirico** per valutare le prestazioni del modello, utilizzando come *loss function*:

$$L(y, \hat{y}) = \begin{cases} 1 & \text{se } \hat{y} \neq y \\ 0 & \text{se } \hat{y} = y \end{cases}$$

e da qui scriviamo la formula del rischio empirico:

$$R_{\text{emp}}(h) = \frac{1}{N} \sum_{i=1}^N L(\hat{y}_i, y_i) = \frac{\text{numero di classificazioni errate}}{N}$$

Il rischio empirico calcolato in questo modo si chiama **tasso di errore** (error rate) e rappresenta la proporzione di osservazioni che sono state classificate in modo errato dal modello. Il classificatore ideale è quello che minimizza il rischio empirico, quindi:

$$\hat{h} = \arg \min_{h \in \mathcal{H}} R_{\text{emp}}(h)$$

dove \mathcal{H} è l'insieme di tutti i possibili classificatori.

10.2 Misure di valutazione

Come nel caso della regressione, è necessario valutare il modello di classificazione durante il training, per fare *fine tuning*¹ e per valutare le sue prestazioni sul test set. Come nel caso della regressione consideriamo come input di valutazione i risultati corretti e le etichette predette dal modello.

10.2.1 Accuratezza

L'**accuratezza** (accuracy) è una delle misure più comuni per valutare le prestazioni di un modello di classificazione. Essa rappresenta la proporzione di osservazioni correttamente classificate rispetto al totale delle osservazioni:

$$\text{Accuratezza}(Y_{\text{TE}}, \hat{Y}_{\text{TE}}) = \frac{\text{Numero di classificazioni corrette}}{\text{Totale delle osservazioni}} = \frac{|y^{(i)} : y^{(i)} = \hat{y}^{(i)}|}{|Y_{\text{TE}}|}$$

Dove Y_{TE} è l'insieme delle etichette vere del test set e \hat{Y}_{TE} è l'insieme delle etichette predette dal modello.

Calcolata in questo modo l'accuratezza è complementare al tasso di errore, quindi se avessimo il 30% di accuracy avremo il 70% di tasso di errore.

10.2.2 Tipi di errori

In un problema di classificazione distinguiamo due tipi di errori principali:

- **Falsi positivi (FP):** Si verificano quando il modello predice una classe positiva (ad esempio, la presenza di una malattia), ma l'osservazione appartiene effettivamente alla classe negativa (assenza della malattia).
- **Falsi negativi (FN):** Si verificano quando il modello predice una classe negativa, ma l'osservazione appartiene effettivamente alla classe positiva.

Dalla lista dei tipi di errori si può estrapolare la lista delle predizioni corrette:

¹Il *fine tuning* è il processo di ottimizzazione dei parametri del modello per migliorare le sue prestazioni.

- **Vero positivo (TP):** Si verificano quando il modello predice correttamente una classe positiva.
- **Vero negativo (TN):** Si verificano quando il modello predice correttamente una classe negativa.

10.2.3 Matrice di confusione

La **matrice di confusione** (confusion matrix) è uno strumento utile per visualizzare le prestazioni di un modello di classificazione. Essa mostra il numero di predizioni corrette e errate suddivise per ciascuna classe. La matrice di confusione per un problema di classificazione binaria è rappresentata come segue:

Matrice di confusione	Predetto Positivo	Predetto Negativo
Reale Positivo	TP	FN
Reale Negativo	FP	TN

Tabella 10.1: Matrice di confusione per un problema di classificazione binaria

In questa matrice abbiamo che le righe rappresentano le classi reali, mentre le colonne rappresentano le classi predette dal modello. Questo strumento ci permette di calcolare diverse metriche di valutazione utili per analizzare le prestazioni del modello.

Dalla matrice, per esempio, possiamo calcolare l'accuratezza come:

$$\text{Accuratezza} = \frac{TP + TN}{TP + TN + FP + FN}$$

Si può notare che l'accuratezza è la somma delle predizioni positive (la diagonale principale della matrice) divisa per il totale delle osservazioni. Questa misura non è sempre utile, specialmente in presenza di classi sbilanciate, dove una classe è molto più rappresentata dell'altra.

Esempio. Supponiamo di avere un dataset con 1000 osservazioni, di cui 900 appartengono alla classe negativa e 100 alla classe positiva. Un modello *naive* $f(x) = 1$ che predice sempre la classe negativa, avrà una matrice di confusione del tipo:

Matrice di confusione	Predetto Positivo	Predetto Negativo
Reale Positivo	0	100
Reale Negativo	0	900

Tabella 10.2: Matrice di confusione per il modello naive

In questo caso, l'accuratezza del modello sarà:

$$\text{Accuratezza} = \frac{0 + 900}{0 + 900 + 0 + 100} = \frac{900}{1000} = 0.9$$

Quindi, nonostante il modello abbia un'accuratezza del 90%, non è in grado di identificare correttamente nessuna delle osservazioni della classe positiva, il che lo rende inefficace per questo tipo di problema e dimostra che l'accuratezza da sola non è sufficiente per valutare le prestazioni di un modello di classificazione, specialmente in presenza di classi sbilanciate.

10.2.4 Precision e recall

Per risolvere i problemi legati all'accuratezza in presenza di classi sbilanciate, possiamo utilizzare altre metriche come la **precision** (precisione) e il **recall** (richiamo). In particolare, si definisce la precisione come:

$$\text{Precisione} = \frac{TP}{TP + FP}$$

ovvero la proporzione di predizioni positive corrette rispetto al totale delle predizioni positive effettuate dal modello.

Il recall, invece, si definisce come:

$$\text{Recall} = \frac{TP}{TP + FN}$$

ovvero la proporzione di osservazioni positive correttamente identificate dal modello rispetto al totale delle osservazioni positive reali.

High precision vs high recall. In alcuni casi, potrebbe essere più importante avere un'alta precisione, mentre in altri casi potrebbe essere più importante avere un alto recall. Ad esempio, in un sistema di rilevamento delle frodi, potrebbe essere più importante avere un alto recall per identificare il maggior numero possibile di transazioni fraudolente, anche a costo di avere qualche falso positivo in più. Al contrario, in un sistema di raccomandazione di prodotti, potrebbe essere più importante avere un'alta precisione per evitare di raccomandare prodotti non rilevanti agli utenti. Si può anche pensare a un test medico per una malattia grave: in questo caso, un alto recall è cruciale per assicurarsi che il maggior numero possibile di casi veri venga identificato, anche se ciò comporta alcuni falsi positivi. D'altra parte, in uno screening di massa per una malattia rara, un'alta precisione potrebbe essere preferibile per evitare di causare ansia inutile ai pazienti con falsi positivi.

10.2.5 F_1 -score

Per bilanciare l'importanza di precisione e recall, possiamo utilizzare l' **F_1 -score**, che è la media armonica² tra precisione e recall. L' F_1 -score viene calcolato come:

$$F_1 = 2 \cdot \frac{\text{Precisione} \cdot \text{Recall}}{\text{Precisione} + \text{Recall}}$$

L' F_1 -score fornisce una singola misura che bilancia sia la precisione che il recall, ed è particolarmente utile quando si desidera un compromesso tra i due.

Nella figura 10.1 sono rappresentate le curve di livello dell' F_1 -score (a sinistra) e della media aritmetica tra precisione e recall, $(\text{precision} + \text{recall})/2$ (a destra), al variare di precisione e recall. Si può notare come l' F_1 -score penalizzi maggiormente gli squilibri tra precisione e recall rispetto alla media aritmetica, evidenziando l'importanza di mantenere un equilibrio tra le due metriche per ottenere buone prestazioni complessive del modello di classificazione.

²La media armonica di due numeri a e b è definita come $\frac{2ab}{a+b}$ ed indica un tipo di media che tende a penalizzare valori molto bassi rispetto alla media aritmetica.

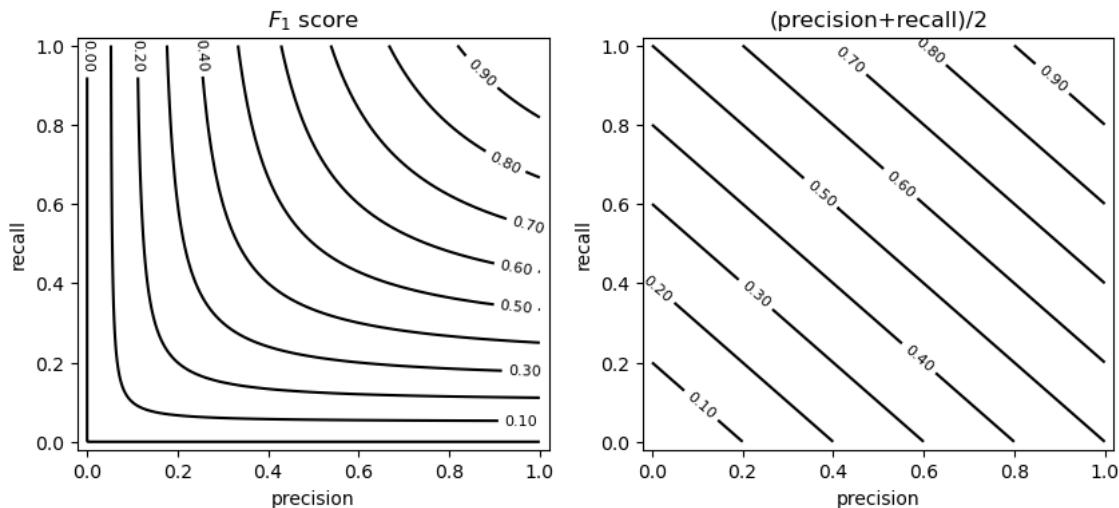


Figura 10.1: Curve di livello dell’F₁-score (a sinistra) e della media aritmetica tra precision e recall, (precision+recall)/2 (a destra), al variare di precision e recall.

10.2.6 Matrice di confusione multiclasse

La matrice di confusione per un problema di classificazione multiclasse è una generalizzazione della matrice di confusione binaria. In questo caso, la matrice avrà dimensioni $k \times k$, dove k è il numero di classi. Ogni cella (i, j) della matrice rappresenta il numero di osservazioni della classe i che sono state classificate come classe j e la diagonale principale rappresenta le predizioni corrette per ciascuna classe.

10.2.7 ROC e AUC

La curva ROC (Receiver Operating Characteristic) è uno strumento grafico molto utile per analizzare le prestazioni di un classificatore binario in modo indipendente dalla soglia di decisione utilizzata. L’idea è la seguente: invece di fissare una soglia specifica per stabilire se un’osservazione appartiene alla classe positiva, valutiamo come cambiano il tasso di veri positivi (True Positive Rate, TPR) e il tasso di falsi positivi (False Positive Rate, FPR) al variare della soglia stessa.

Sia quindi $c(\mathbf{x})$ la funzione che assegna a ciascuna osservazione \mathbf{x} un valore di confidenza, tipicamente interpretato come la probabilità che \mathbf{x} appartenga alla classe positiva. A partire da questo valore definiamo un classificatore binario dipendente dalla soglia θ :

$$h_\theta(\mathbf{x}) = [c(\mathbf{x}) \geq \theta]$$

dove le parentesi di Iverson (le parentesi quadre) indicano che l’espressione vale 1 se la condizione è vera e 0 altrimenti. Variando θ da 0 a 1 otteniamo una famiglia di classificatori, ognuno dei quali produce una diversa matrice di confusione. In particolare i valori di θ producono:

$$TP_\theta, FP_\theta, TN_\theta, FN_\theta$$

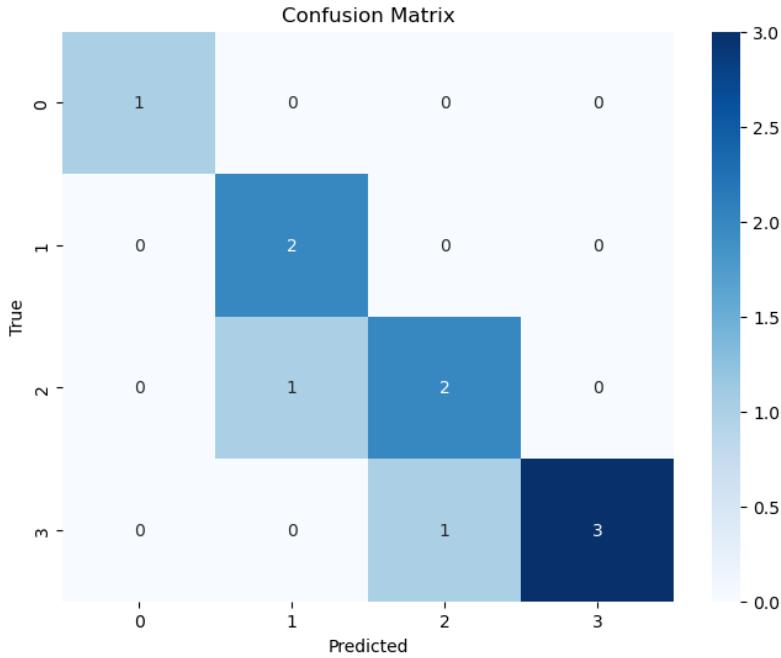


Figura 10.2: Matrice di confusione per un problema di classificazione a quattro classi: sulla diagonale sono visibili le predizioni corrette, mentre i valori fuori diagonale rappresentano gli errori di classificazione tra le diverse classi.

da cui possiamo calcolare il TPR e l'FPR, ovvero il tasso di veri positivi e il tasso di falsi positivi:

$$TPR(\theta) = \frac{TP_\theta}{TP_\theta + FN_\theta} \quad , \quad FPR(\theta) = \frac{FP_\theta}{FP_\theta + TN_\theta}$$

Notiamo che il TPR è equivalente al recall, mentre il FPR rappresenta la proporzione di osservazioni negative che sono state erroneamente classificate come positive. Da questo possiamo dire che:

- Per una soglia θ bassa, il classificatore tenderà a predire più osservazioni come positive, aumentando sia il TPR che l'FPR.
- Per una soglia θ alta, il classificatore tenderà a predire meno osservazioni come positive, riducendo sia il TPR che l'FPR.

Si può mostrare graficamente la curva ROC tracciando il TPR in funzione dell'FPR, come mostrato in figura 10.3.

Un classificatore ideale raggiungerebbe il punto più in alto a sinistra della curva ROC, dove il TPR è 1 (tutte le osservazioni positive sono correttamente identificate) e l'FPR è 0 (nessuna osservazione negativa è erroneamente classificata come positiva). Al contrario, un classificatore casuale produrrebbe una curva ROC che si avvicina alla diagonale del grafico, rappresentando una performance equivalente al caso casuale.

Come misura, possiamo utilizzare l'**AUC** (Area Under the Curve), che rappresenta l'area sotto la curva ROC. L'AUC varia tra 0 e 1, dove un valore di 0.5 indica una performance equivalente al caso casuale, mentre un valore di 1 indica un classificatore perfetto. In generale, un AUC più alto indica una migliore capacità del modello di distinguere tra le classi positive e negative.

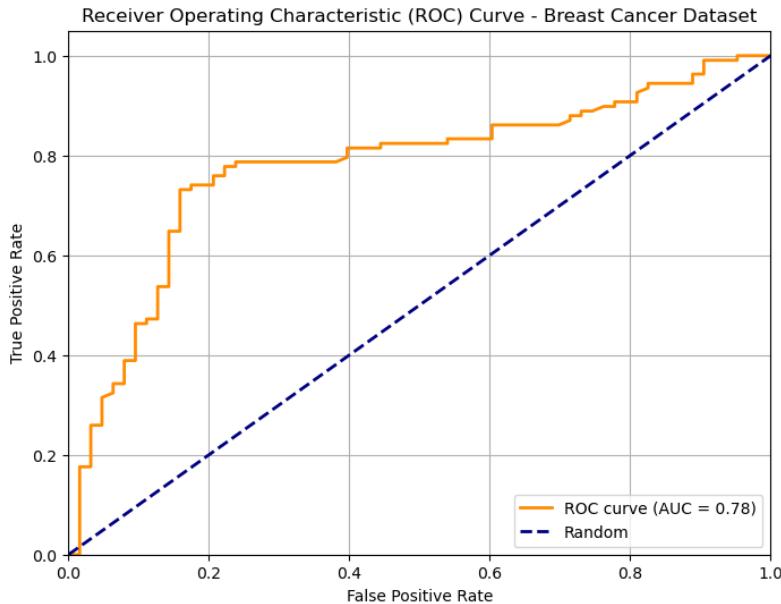


Figura 10.3: Curva ROC ottenuta sul dataset Breast Cancer: la curva mostra il compromesso tra TPR e FPR al variare della soglia di decisione, mentre l'AUC pari a 0.78 indica una buona capacità discriminante rispetto alla classificazione casuale (linea tratteggiata).

Statistica J di Youden. La statistica J di youden è una misura utilizzata per massimizzare la somma della sensibilità (TPR) e della specificità ($1 - FPR$) di un classificatore. La statistica J è definita come:

$$J = TPR + (1 - FPR)$$

L'obiettivo è trovare la soglia di decisione che massimizza il valore di J, ovvero:

$$\theta^* = \arg \max_{\theta} J(\theta)$$

Questa soglia ottimale θ^* rappresenta il punto in cui il classificatore bilancia al meglio la capacità di identificare correttamente le osservazioni positive (sensibilità) e di evitare falsi positivi (specificità).

Esempio: classificatore a soglia. Consideriamo un classificatore che assegna una classe positiva se la misurazione stimata $c(x)$ supera una soglia θ . Consideriamo quindi un classificatore che assegna la classe *uomo* se l'altezza stimata supera la soglia θ , e la classe *donna* altrimenti. Ipotizziamo che la distribuzione nel nostro insieme dei dati sia quella in figura 10.4.

Da qui capiamo che, generalmente, gli uomini, sono più alti delle donne. Supponiamo di scegliere una soglia $\theta = 170$ cm per classificare le persone in uomini e donne. In questo caso, tutte le persone con altezza superiore a 170 cm saranno classificate come uomini, mentre tutte le persone con altezza inferiore o uguale a 170 cm saranno classificate come donne producendo i seguenti risultati:

I risultati sono abbastanza buoni, con un'accuratezza complessiva dell'83%. Tuttavia, possiamo notare che il recall per la classe *True* (donne) è più alto rispetto alla precisione, indicando che il modello è più efficace nell'identificare le donne rispetto a quanti uomini classifica correttamente. Al contrario, per la classe *False* (uomini), la precisione è più alta del recall, suggerendo che

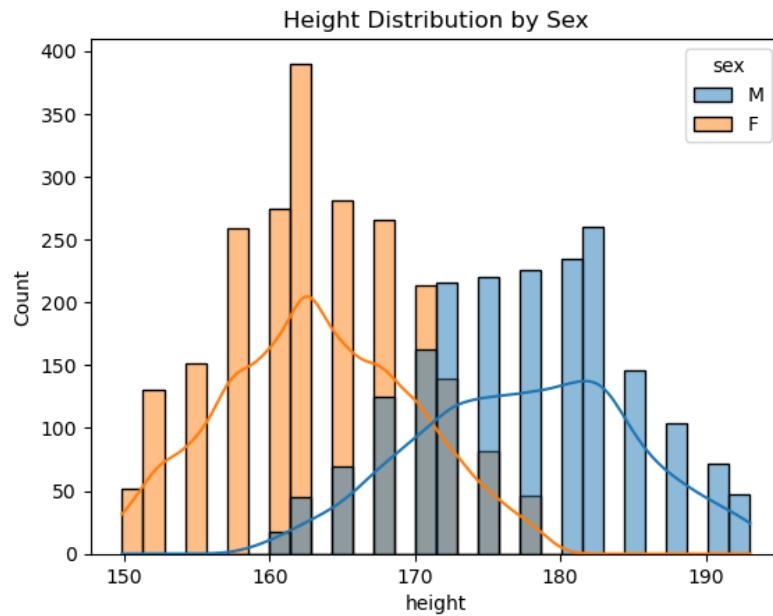


Figura 10.4: Distribuzione dell'altezza per uomini e donne.

Matrice di confusione	Predetto Uomo	Predetto Donna
Reale Uomo	1804	481
Reale Donna	257	1689

Tabella 10.3: Matrice di confusione per il classificatore a soglia $\theta = 170$ cm

Class	Precision	Recall	F1-score	Support
False	0.88	0.79	0.83	2285
True	0.78	0.87	0.82	1946
Accuracy		0.83		4231
Macro Avg	0.83	0.83	0.83	4231
Weighted Avg	0.83	0.83	0.83	4231

Tabella 10.4: Metriche di classificazione per le classi False e True.

il modello tende a classificare correttamente gli uomini, ma perde alcune donne che vengono erroneamente classificate come uomini.

Si potrebbe utilizzare la statistica J di Youden per trovare una soglia ottimale che bilanci meglio precisione e recall per entrambe le classi, migliorando così le prestazioni complessive del classificatore.

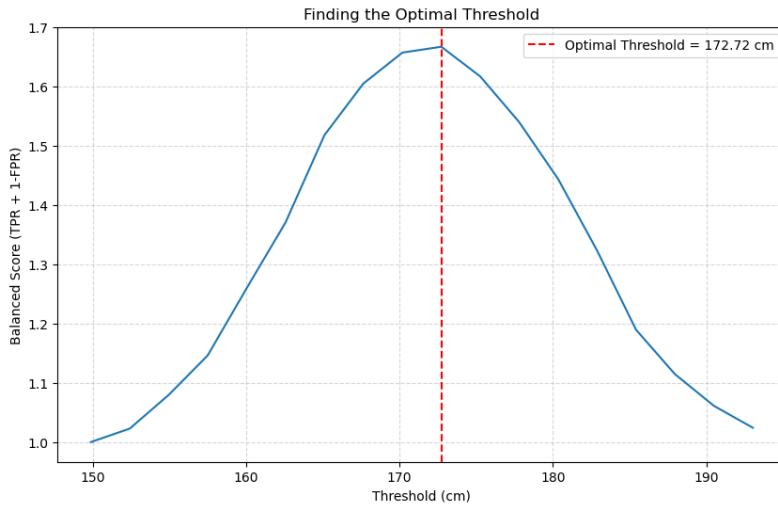


Figura 10.5: Statistica J di Youden: la curva mostra il compromesso tra sensibilità e specificità al variare della soglia di decisione, mentre il punto ottimale (massimo della curva) indica la soglia che bilancia meglio i due aspetti.

Dalla figura 10.5 possiamo notare come la migliore soglia di decisione sia 172.72 cm, che maximizza la statistica J di Youden, bilanciando al meglio la sensibilità e la specificità del classificatore basato sull'altezza.

10.3 K-Nearest Neighbors (KNN)

Il problema dei classificatori a soglia è che spesso non è possibile trovare una soglia che separi perfettamente le classi, specialmente in presenza di dati rumorosi o sovrapposti. Inoltre funzionano solamente con una feature alla volta. L'algoritmo KNN, invece, va a trasformare le caratteristiche di un'osservazione in uno spazio multidimensionale e utilizza la distanza tra le osservazioni per effettuare la classificazione.

10.3.1 1-NN

L'algoritmo più semplice della famiglia KNN è il **1-NN** (1-Nearest Neighbor). In questo caso, per classificare una nuova osservazione, l'algoritmo cerca l'osservazione più vicina nel training set e assegna la stessa classe dell'osservazione trovata. La distanza tra le osservazioni può essere calcolata utilizzando diverse metriche, come la distanza euclidea, la distanza di Manhattan o altre metriche appropriate per il tipo di dati.

In particolare, possiamo misurare la distanza tra due osservazioni x e x' con la distanza

euclidea:

$$d(x, x') = \|x - x'\|_2 = \sqrt{\sum_{i=1}^n (x_i - x'_i)^2}$$

Dalla distanza euclidea poi, si cerca di minimizzare la distanza tra l'osservazione da classificare e le osservazioni del training set:

$$h(x') = \arg \min_y \{d(x, x') : (x, y) \in \text{TR}\}$$

dove TR è il training set. L'algoritmo 1-NN è molto semplice per il resto e funziona in due fasi principali:

- Trova l'elemento \bar{x} nel training set che minimizza la distanza $d(x, x')$ con l'osservazione x' da classificare.
- Assegna a x' la stessa etichetta di \bar{x} , ovvero $h(x') = y$ dove $(\bar{x}, y) \in \text{TR}$.

10.3.2 K-NN

L'algoritmo **K-NN** (K-Nearest Neighbors) è una generalizzazione del 1-NN che considera i K vicini più prossimi invece di uno solo. In questo caso, per classificare una nuova osservazione, l'algoritmo trova i K osservazioni più vicine nel training set e assegna la classe più comune tra queste osservazioni. Il valore di K è un **iperparametro** che deve essere scelto in base al problema specifico e può influenzare significativamente le prestazioni del modello.

In modo analogo a quanto visto nel caso della stima di densità, nel K-NN definiamo, per un punto di test x' , un intorno di ampiezza K centrato proprio in x' . Indichiamo tale insieme con:

$$N_K(x') = N(x', R_K(x'))$$

dove $N(x, r)$ rappresenta l'intorno centrato in x con raggio r . Il raggio $R_K(x')$ è definito come il più grande valore tale per cui l'intorno contiene *al massimo* K punti del training set (escludendo eventualmente x' stesso). Formalmente:

$$R_K(x') = \sup \{ r : |N(x', r) \setminus \{x'\}| \leq K \}$$

In questo modo, l'intorno $N_K(x')$ contiene esattamente i K punti del training set più vicini a x' , sui quali verrà poi applicata la regola di maggioranza per determinare la classe predetta.

10.4 Curse of dimensionality

Uno dei principali problemi che si incontrano con gli algoritmi basati sulla distanza, come K-NN, è la **maledizione della dimensionalità** (curse of dimensionality). Questo fenomeno si verifica quando il numero di caratteristiche (dimensioni) aumenta, rendendo difficile per l'algoritmo distinguere tra le osservazioni. In spazi ad alta dimensionalità, tutte le osservazioni tendono a essere equidistanti tra loro, rendendo la nozione di vicinanza meno significativa.

10.4.1 Spazio vuoto

Il concetto di spazio vuoto si riferisce a quella parte di uno spazio multidimensionale che non contiene alcuna osservazione del dataset. In spazi con bassa dimensione, questo problema è meno evidente, perché le osservazioni tendono a essere più vicine tra loro. Tuttavia, man mano che la dimensionalità aumenta, lo spazio vuoto diventa predominante, rendendo difficile per gli algoritmi basati sulla distanza trovare vicini significativi.

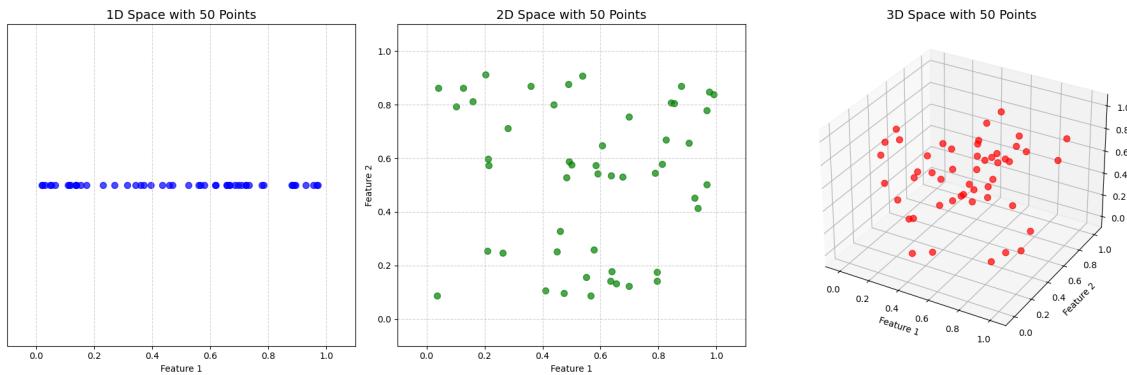


Figura 10.6: Distribuzione di 50 punti in spazi di diversa dimensionalità: a sinistra uno spazio 1D, al centro uno spazio 2D e a destra uno spazio 3D. Al crescere della dimensionalità, i punti si disperdonno in regioni sempre più ampie dello spazio delle caratteristiche.

10.4.2 Impatto su K-NN

L'impatto della maledizione della dimensionalità sugli algoritmi K-NN è significativo. In spazi ad alta dimensionalità, la distanza tra le osservazioni tende a diventare simile, rendendo difficile per l'algoritmo identificare i vicini più prossimi in modo efficace. Questo può portare a una riduzione delle prestazioni del modello, poiché le predizioni basate sui vicini più prossimi diventano meno affidabili.

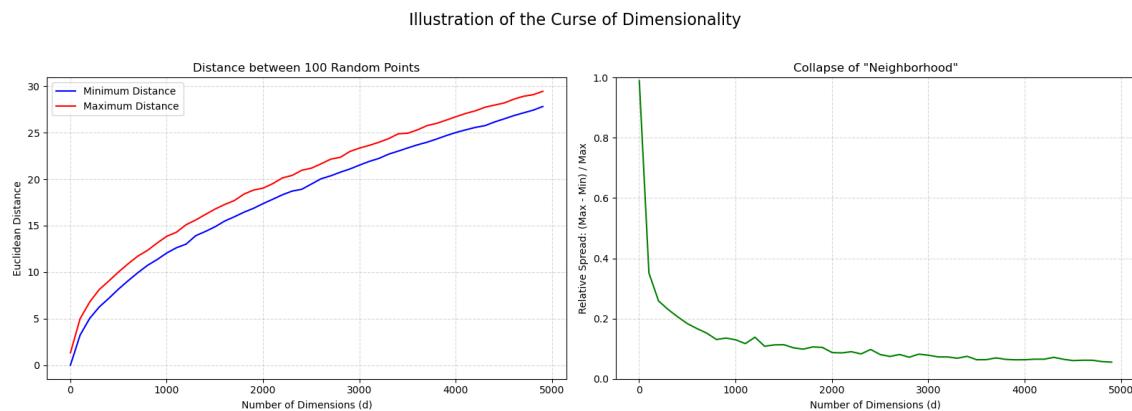


Figura 10.7: Illustrazione del *curse of dimensionality*: a sinistra sono riportate la distanza minima e massima tra 100 punti generati casualmente al crescere della dimensionalità dello spazio; a destra è mostrata la riduzione relativa della variabilità delle distanze, evidenziando come, in spazi ad alta dimensionalità, le distanze tra punti tendano a diventare sempre più simili, causando il collasso del concetto di “vicinato”.

Dalla figura 10.7 possiamo notare come, al crescere della dimensionalità dello spazio, la distanza minima e massima tra i punti generati casualmente si avvicinano sempre di più. Questo fenomeno porta a una riduzione relativa della variabilità delle distanze, evidenziando come, in spazi ad alta dimensionalità, le distanze tra punti tendano a diventare sempre più simili, causando il collasso del concetto di “vicinato”. Di conseguenza, gli algoritmi K-NN possono perdere efficacia, poiché la distinzione tra vicini prossimi e lontani diventa meno significativa.

Si può inoltre calcolare lo **spread relativo**, ovvero la differenza tra la distanza massima e minima normalizzata rispetto alla distanza minima:

$$\text{Spread relativo} = \frac{d_{\max} - d_{\min}}{d_{\min}}$$

10.5 Classificatori discriminativi e generativi

Nella classificazione, i modelli possono essere suddivisi in due categorie principali: **classificatori discriminativi** e **classificatori generativi**. La differenza è nel modo in cui i modelli apprendono a classificare le osservazioni.

10.5.1 Classificatori discriminativi

I classificatori discriminativi si concentrano direttamente sulla modellazione della frontiera decisionale tra le classi. In altre parole, questi modelli cercano di imparare la funzione che mappa le caratteristiche delle osservazioni alle loro classi senza fare assunzioni esplicite sulla distribuzione dei dati. Rispondono alla domanda "Qual è la differenza tra le classi dato un insieme di caratteristiche?".

Esempio. Un esempio di classificatore discriminativo è la **regressione logistica**, che modella la probabilità condizionata della classe data le caratteristiche delle osservazioni. La regressione logistica utilizza una funzione sigmoide per mappare le caratteristiche a una probabilità compresa tra 0 e 1, permettendo di classificare le osservazioni in base a una soglia predefinita.

10.5.2 Classificatori generativi

I classificatori generativi, d'altra parte, cercano di modellare la distribuzione congiunta delle caratteristiche e delle classi. Questi modelli fanno assunzioni sulla distribuzione dei dati e cercano di imparare come le osservazioni vengono generate per ciascuna classe. Rispondono alla domanda "Come vengono generate le osservazioni per ciascuna classe?".

Esempio. Un esempio di classificatore generativo è il **Naive Bayes**, che assume che le caratteristiche siano condizionalmente indipendenti dato la classe. Il modello stima la probabilità congiunta delle caratteristiche e delle classi e utilizza il teorema di Bayes per calcolare la probabilità condizionata della classe data le caratteristiche, permettendo di classificare le osservazioni in base alla massima probabilità a posteriori.

10.5.3 Macro, Micro e Weighted Averaging

Quando valutiamo un modello di classificazione multi-classe, le metriche come *precision*, *recall* e *F1-score* vengono inizialmente calcolate per ciascuna classe in modo indipendente. Tuttavia, spesso abbiamo bisogno di un singolo valore che riassume le prestazioni globali del modello. Per ottenere questa misura complessiva, è necessario definire una strategia di aggregazione delle metriche tra le diverse classi. Le tre più utilizzate sono: *macro averaging*, *weighted averaging* e *micro averaging*.

Macro averaging. In questo caso calcoliamo la metrifica (ad esempio la precision) separatamente per ogni classe e, alla fine, ne facciamo la media semplice:

$$\text{Macro Precision} = \frac{1}{K} \sum_{i=1}^K \text{Precision}_i.$$

Ogni classe contribuisce allo stesso modo, indipendentemente dal numero di esempi che la compongono. È quindi una misura particolarmente utile quando ci interessa valutare le prestazioni su tutte le classi, incluse quelle rare.

Weighted averaging. Anche qui la metrifica viene calcolata per ciascuna classe, ma la media finale assegna a ogni classe un peso proporzionale alla sua numerosità (il cosiddetto *support*):

$$\text{Weighted Precision} = \frac{\sum_{i=1}^K \text{Precision}_i \cdot \text{Support}_i}{\sum_{i=1}^K \text{Support}_i}.$$

In questo modo, le classi più frequenti hanno un'influenza maggiore sul risultato finale. Si tratta di una media “equilibrata”, particolarmente appropriata in presenza di dataset sbilanciati.

Micro averaging. Diversamente dai metodi precedenti, qui non si calcola la metrifica per classe: si sommano invece tutte le quantità globali (TP, FP, FN) su tutte le classi e si applica la formula della metrifica una sola volta:

$$\text{Micro Precision} = \frac{\sum_i TP_i}{\sum_i (TP_i + FP_i)}.$$

Ogni singolo *campione* ha lo stesso peso, indipendentemente dalla classe a cui appartiene. Da notare che, nelle classificazioni multi-classe, la micro-precisione (così come micro-recall e micro-F1) coincide esattamente con l'*accuracy*.

10.5.4 Decision Boundary

Un classificatore f assegna una classe a ciascun punto dello spazio delle caratteristiche. La **frontiera decisionale** (decision boundary) è l'insieme dei punti in cui il classificatore cambia la sua predizione da una classe all'altra. In altre parole, è la superficie che separa le regioni dello spazio delle caratteristiche in cui il classificatore assegna classi diverse.

Come si può notare dalla figura 10.8, la forma della frontiera decisionale dipende fortemente dal valore di K scelto nell'algoritmo K-NN. Con un valore di K basso, la frontiera tende a essere

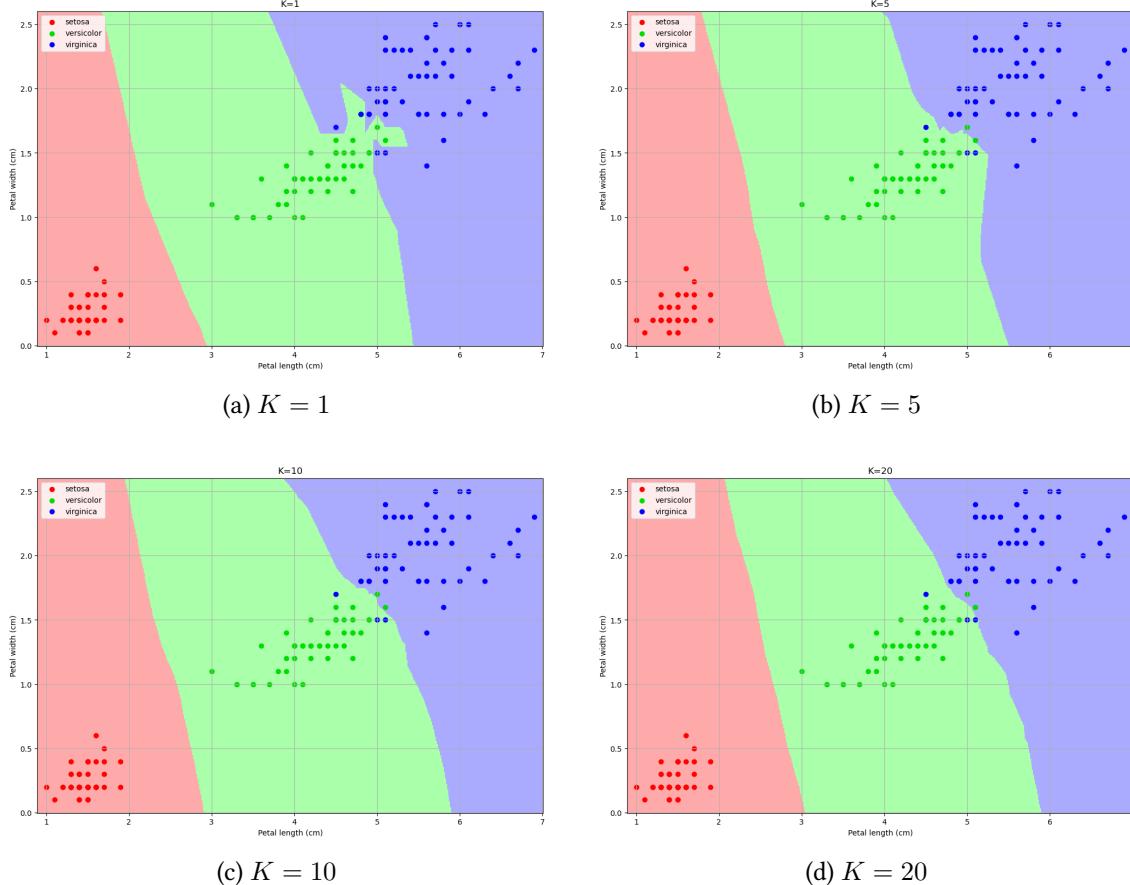


Figura 10.8: Confronto dei confini decisionali del classificatore K-NN sul dataset Iris al variare del numero di vicini K . Per $K = 1$ (in alto a sinistra) la frontiera segue in modo molto dettagliato i singoli punti di training, risultando irregolare e potenzialmente soggetto a overfitting. Aumentando K a 5 e 10 (in alto a destra e in basso a sinistra) le regioni decisionali diventano via via più lisce, riducendo la sensibilità al rumore ma smussando alcune strutture locali, in particolare nella zona di separazione tra *versicolor* e *virginica*. Per $K = 20$ (in basso a destra) la frontiera è molto regolare: il modello è più stabile, ma rischia di sottostimare le differenze tra le due classi non linearmente separabili, evidenziando il classico compromesso tra complessità del modello e capacità di generalizzazione.

molto irregolare e sensibile al rumore nei dati, mentre con un valore di K più alto, la frontiera diventa più liscia e meno influenzata da singoli punti anomali. Tuttavia, un valore di K troppo alto può portare a una perdita di dettagli importanti nella struttura dei dati, risultando in una sottostima delle differenze tra le classi.

Questo è un trade-off che rispecchia il bilanciamento tra **bias** e **varianza** nel modello, dove un valore di K basso tende a ridurre il bias ma aumentare la varianza, mentre un valore di K alto tende a ridurre la varianza ma aumentare il bias.

10.6 K-NN per regressione

Sebbene l'algoritmo K-NN sia principalmente utilizzato per la classificazione, può essere adattato anche per problemi di regressione. In questo caso, invece di assegnare una classe basata sui vicini più prossimi, l'algoritmo calcola la media (o la mediana) dei valori target dei K vicini più prossimi per fare una predizione continua.

10.6.1 Bias-Varianza trade-off

Come nel caso della classificazione, la scelta di k controlla il compromesso tra bias e varianza nel modello di regressione K-NN. Un valore di k basso tende a ridurre il bias, permettendo al modello di adattarsi più strettamente ai dati di training, ma aumenta la varianza, rendendo il modello più sensibile al rumore nei dati. Al contrario, un valore di k alto tende a ridurre la varianza, rendendo il modello più stabile, ma aumenta il bias, poiché il modello può perdere dettagli importanti nella struttura dei dati.

Trovare il miglior valore di k . Anche in questo caso si può utilizzare la cross-validation per trovare il valore ottimale di k che minimizza l'errore di predizione su dati non visti.

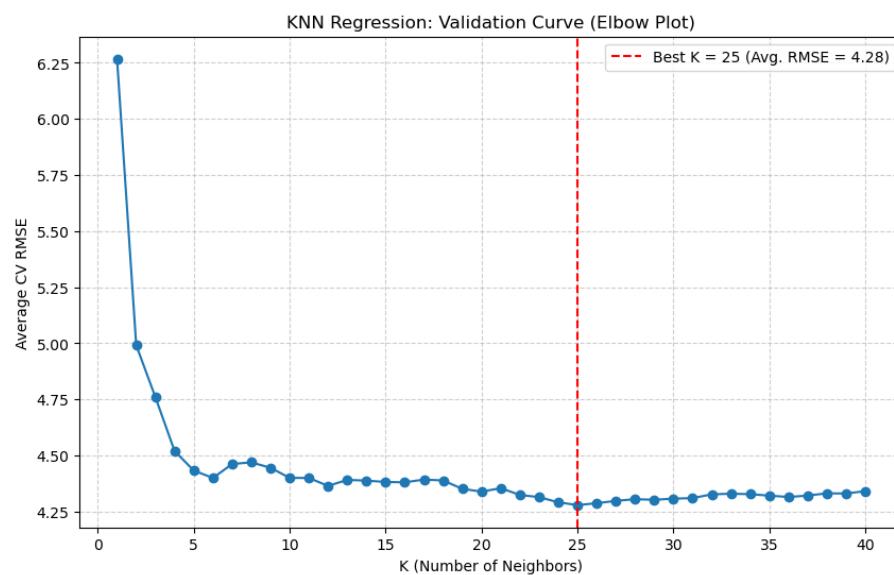


Figura 10.9: Curva di validazione per il modello K-NN in regressione, ottenuta tramite cross-validation. L'asse delle ascisse rappresenta il numero di vicini K , mentre l'asse delle ordinate mostra l'errore medio (RMSE) sui diversi fold. Si osserva che l'errore diminuisce rapidamente per valori piccoli di K , per poi stabilizzarsi. Il valore ottimale individuato è $K = 25$, che corrisponde al minimo RMSE medio, rappresentato dalla linea tratteggiata rossa.

Capitolo 11

Regressione logistica

La regressione logistica è un modello statistico utilizzato per prevedere la probabilità di un evento binario (ad esempio, successo/fallimento, sì/no) in base a una o più variabili indipendenti. A differenza di K-NN, che è *non parametrizzato*, la regressione logistica è un modello *parametrizzato*, il che significa che assume una forma specifica per la relazione tra le variabili indipendenti e la probabilità dell'evento.

La differenza principale tra la regressione logistica e la regressione lineare risiede nella natura della variabile dipendente. Mentre la regressione lineare è utilizzata per prevedere valori continui, la regressione logistica è progettata per gestire variabili dipendenti categoriali, in particolare binarie. Si può evincere bene questo limite quando andiamo a utilizzare un modello di regressione logistica su una variabile binaria come uomo o donna, come mostrato in figura 11.1.

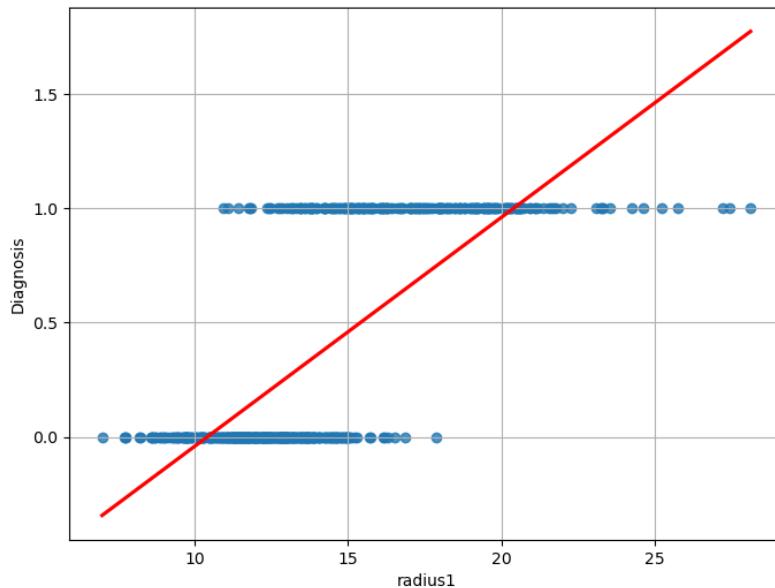


Figura 11.1: Esempio di regressione lineare su variabile binaria.

11.1 Modello di regressione logistica

Il problema principale è: come trovare quella funzione che permette di fare regressione su una variabile binaria? Si potrebbe utilizzare un classificatore a soglia, ma questo avrebbe dei problemi quando ci sono delle zone di **overlap** tra le classi.

11.1.1 Funzione logistica

La *funzione logistica* (o funzione sigmoide) è una soluzione al problema, in quanto crea una curva a forma di S tra due asintoti orizzontali (0 e 1), che rappresentano le probabilità delle due classi. La funzione logistica è definita come:

$$P(Y = 1|X) = \frac{1}{1 + e^{-x}}$$

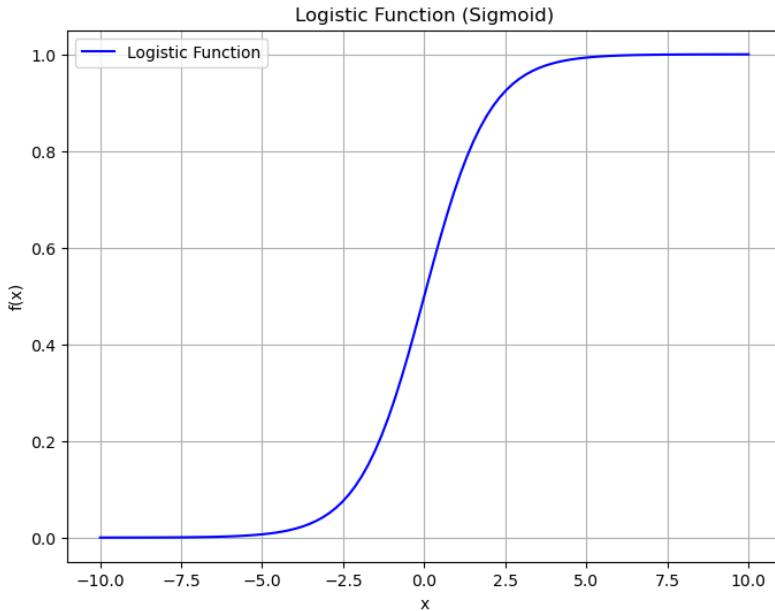


Figura 11.2: Esempio di funzione logistica. Come si evince dalla figura, la funzione logistica è limitata tra 0 e 1, rendendola adatta per modellare probabilità.

11.1.2 Modello di regressione logistica

Si può definire il modello f di regressione logistica come:

$$P(Y = 1|X) = f(\beta_0 + \beta_1 X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

o nel caso più generale della regressione logistica multipla:

$$P(Y = 1|X) = f(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}}$$

dove β_0 è l'intercetta, il vettore $\beta = (\beta_1, \beta_2, \dots, \beta_n)$ rappresenta i coefficienti delle variabili indipendenti $X = (X_1, X_2, \dots, X_n)$.

11.1.3 Odds

Un concetto importante nella regressione logistica è quello degli *odds*, ovvero il rapporto tra la probabilità di successo e la probabilità di insuccesso. Gli odds sono definiti come:

$$\text{Odds} = \frac{P(Y = 1|X)}{P(Y = 0|X)} = \frac{P(Y = 1|X)}{1 - P(Y = 1|X)}$$

L'odd si può interpretare come la probabilità che un evento si verifichi rispetto alla probabilità che non si verifichi. Ad esempio, se la probabilità di successo è 0.8, allora la probabilità di insuccesso è 0.2, e gli odds sono:

$$\text{Odds} = \frac{0.8}{0.2} = 4$$

Ciò significa che l'evento di successo è quattro volte più probabile dell'evento di insuccesso.

La funzione logistica può essere riscritta in termini di odds:

$$p = \frac{1}{1 + e^{-x}} \Rightarrow p + pe^{-x} = 1 \Rightarrow pe^{-x} = 1 - p \Rightarrow \frac{p}{1 - p} = e^x$$

e da qui:

$$e^{\beta_0 + \beta \cdot \mathbf{x}} = \frac{P(Y = 1|X)}{1 - P(Y = 1|X)}$$

Il termine a destra rappresenta gli odds, mentre il termine a sinistra è l'esponenziale della combinazione lineare delle variabili indipendenti. Questo mostra come la regressione logistica modella gli odds in funzione delle variabili indipendenti.

11.1.4 Log-odds

Prendendo il logaritmo naturale degli odds, otteniamo i *log-odds* (o logit):

$$\log \left(\frac{P(Y = 1|X)}{1 - P(Y = 1|X)} \right) = \beta_0 + \beta \cdot \mathbf{x}$$

I log-odds rappresentano una trasformazione lineare della probabilità, rendendo più semplice l'interpretazione dei coefficienti β . Ogni incremento unitario in una variabile indipendente X_i comporta un cambiamento di β_i nei log-odds.

Quindi, la regressione logistica può essere vista come un modello di regressione lineare applicato ai log-odds della probabilità dell'evento di interesse. Questo consente di stimare l'effetto delle variabili indipendenti sulla probabilità dell'evento, facilitando l'interpretazione e l'analisi dei risultati.

11.2 Stima dei parametri

Per stimare i parametri del modello di regressione logistica, dobbiamo definire una loss function adeguata, analoga a quella utilizzata per la regressione lineare. Tuttavia, poiché la regressione logistica ha una natura probabilistica e l'output desiderato è una variabile binaria, l'errore quadratico non è una scelta appropriata. Il modello produce infatti una probabilità tramite la funzione sigmoide, e ciò richiede una funzione di costo coerente con tale interpretazione.

11.2.1 Interpretazione probabilistica del modello

Ricordiamo che il modello di regressione logistica associa a ciascuna osservazione \mathbf{x} la probabilità stimata che appartenga alla classe positiva:

$$f(\mathbf{x}) = \sigma(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n),$$

dove σ è la funzione sigmoide. Possiamo quindi interpretare $f(\mathbf{x}_i)$ come la probabilità che $y_i = 1$. Di conseguenza, la probabilità di osservare un'etichetta $y_i \in \{0, 1\}$ è:

$$P(y_i | \mathbf{x}_i, \beta) = f(\mathbf{x}_i)^{y_i} (1 - f(\mathbf{x}_i))^{1-y_i}.$$

11.2.2 Cross-entropy loss

La funzione di costo più comune per la regressione logistica è la *cross-entropy loss*, che misura la differenza tra le probabilità previste dal modello e le etichette osservate.

Possiamo stimare i parametri utilizzando la log-likelihood, ovvero scegliamo i parametri che massimizzano la probabilità di osservare i dati secondo il modello definito dai parametri stessi:

$$L(\beta) = P(Y | X; \beta)$$

dove Y è il vettore delle etichette osservate e X è la matrice delle caratteristiche. Questo significa che scegliamo gli elementi $\beta_i \in \beta$ tali che il modello assegna probabilità alte ai valori realmente osservati.

Si può trasformare la log-likelihood in una somma logaritmica per semplificare i calcoli:

$$L(\beta) = \prod_{i=1}^N P(y^{(i)} | \mathbf{x}^{(i)}; \beta) = \prod_{i=1}^N f_\beta(\mathbf{x}^{(i)})^{y^{(i)}} (1 - f_\beta(\mathbf{x}^{(i)}))^{1-y^{(i)}}$$

Massimizzare questa espressione è equivalente a minimizzare il logaritmo negativo di $L(\beta)$ (nll), che porta:

$$\begin{aligned} nll(\beta) &= -\log L(\beta) \\ &= -\sum_{i=1}^N \log \left[f_\beta(\mathbf{x}^{(i)})^{y^{(i)}} (1 - f_\beta(\mathbf{x}^{(i)}))^{1-y^{(i)}} \right] \\ &= -\sum_{i=1}^N \left[y^{(i)} \log f_\beta(\mathbf{x}^{(i)}) + (1 - y^{(i)}) \log (1 - f_\beta(\mathbf{x}^{(i)})) \right] \end{aligned}$$

Da qui, possiamo definire la nostra funzione di costo come:

$$J(\beta) = -\sum_{i=1}^N \left[y^{(i)} \log f_\beta(\mathbf{x}^{(i)}) + (1 - y^{(i)}) \log (1 - f_\beta(\mathbf{x}^{(i)})) \right]$$

Ora abbiamo una funzione di costo che possiamo minimizzare utilizzando metodi di ottimizzazione numerica, come la discesa del gradiente. La minimizzazione di questa funzione di costo ci permetterà di trovare i valori ottimali dei parametri β che meglio si adattano ai dati osservati.

11.2.3 Visualizzazione della cross-entropy loss

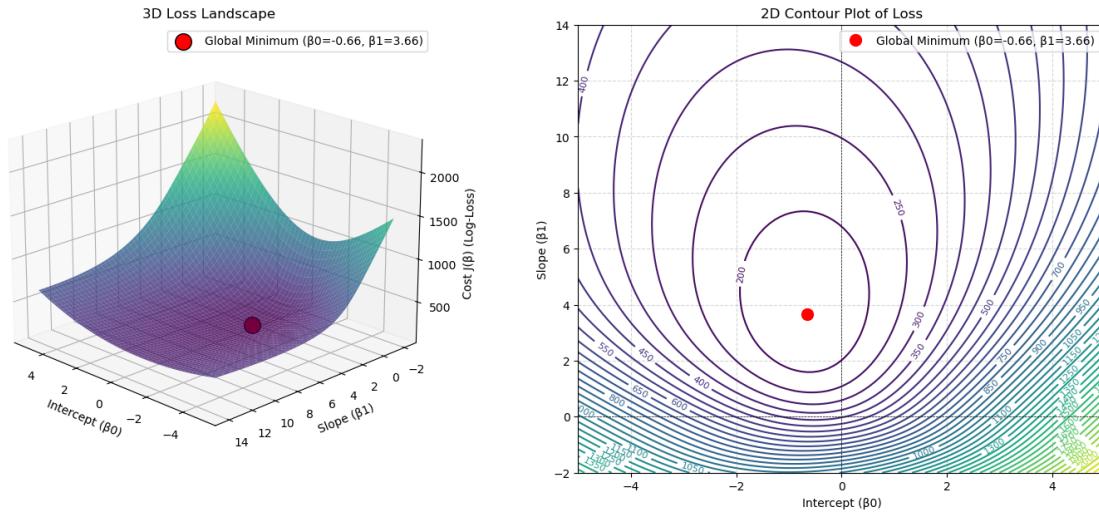


Figura 11.3: Visualizzazione della superficie della cross-entropy loss per un modello di regressione logistica con due parametri (β_0 e β_1). A sinistra è rappresentata la superficie tridimensionale della loss, che evidenzia la presenza di un unico minimo globale. A destra sono riportate le curve di livello della stessa funzione, che mostrano come la loss aumenti progressivamente man mano che ci si allontana dal minimo. Il punto rosso indica la combinazione di parametri che minimizza la loss, corrispondente alla stima $\hat{\beta}$ ottenuta tramite massima verosimiglianza.

Per comprendere meglio il comportamento della funzione di costo della regressione logistica, è utile analizzare come la *cross-entropy loss* vari al variare dei parametri del modello. Nel caso di un modello con una sola variabile esplicativa, la funzione di costo dipende da due soli parametri, β_0 (intercetta) e β_1 (coefficiente), e può quindi essere rappresentata graficamente nello spazio bidimensionale dei parametri. La superficie tridimensionale mostra come la loss cresca rapidamente quando i parametri si allontanano dalla combinazione ottimale, mentre il grafico delle curve di livello (contour plot) permette di individuare più chiaramente la regione in cui la funzione assume i valori minimi.

Questa rappresentazione aiuta a visualizzare un concetto fondamentale: la funzione di costo della regressione logistica è *convessa* rispetto ai parametri (in particolare alla massima verosimiglianza), il che garantisce l'esistenza di un unico minimo globale. Di conseguenza, metodi iterativi come la *gradient descent* o il *Newton-Raphson* convergeranno sempre alla stessa soluzione ottimale, rendendo l'ottimizzazione stabile ed efficace. La forma "a scodella" della superficie conferma che piccoli cambiamenti nei parametri vicino al minimo producono variazioni limitate della loss, mentre modifiche più grandi portano rapidamente ad un aumento significativo dell'errore.

11.3 Interpretazione statistica dei coefficienti

Come detto in precedenza, i coefficienti della regressione logistica possono essere interpretati in termini di log-odds. In particolare, ogni coefficiente β_i rappresenta il cambiamento nei log-odds associato a un incremento unitario nella variabile indipendente X_i , mantenendo costanti tutte le altre variabili. Questo perché la regressione logistica modella i log-odds come una combinazione lineare delle variabili indipendenti.

11.3.1 Interpretazione dell'intercetta

L'intercetta β_0 può essere interpretata come i log-odds quando tutte le variabili indipendenti sono uguali a zero. Infatti, ponendo tutte le variabili indipendenti a zero, possiamo scrivere:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta \cdot \mathbf{x}$$

da qui:

$$x = 0 \Rightarrow \log\left(\frac{p}{1-p}\right) = \beta_0 \Rightarrow \frac{p}{1-p} = e^{\beta_0}$$

Ricordando che $\frac{p}{1-p}$ sono gli odds, possiamo affermare che e^{β_0} rappresenta gli odds quando tutte le variabili indipendenti sono uguali a zero. Quindi, per $x = 0$ possiamo dire che sono e^{β_0} volte più probabili gli eventi di successo rispetto agli eventi di insuccesso.

11.3.2 Interpretazione dei coefficienti delle variabili indipendenti

Ogni coefficiente β_i rappresenta il cambiamento nei log-odds associato a un incremento unitario nella variabile indipendente X_i , mantenendo costanti tutte le altre variabili. Possiamo scrivere che:

$$\text{odds}(p | x) = \frac{P(Y = 1 | X = x)}{1 - P(Y = 1 | X = x)} = e^{\beta_0 + \beta \mathbf{x}}$$

quindi

$$\log(\text{odds}(p | x + 1)) - \log(\text{odds}(p | x)) = \beta_i$$

e da qui:

$$\begin{aligned} e^{\log \text{odds}(p|x+\mathbf{e}_i) - \log \text{odds}(p|x)} &= e^{\beta_i} \Rightarrow \frac{e^{\log \text{odds}(p|x+\mathbf{e}_i)}}{e^{\log \text{odds}(p|x)}} = e^{\beta_i} \\ &\Rightarrow \frac{\text{odds}(p | x + \mathbf{e}_i)}{\text{odds}(p | x)} = e^{\beta_i} \Rightarrow \text{odds}(p | x + \mathbf{e}_i) = e^{\beta_i} \text{odds}(p | x) \end{aligned}$$

Questo significa che per ogni incremento unitario nella variabile indipendente X_i , gli odds di successo vengono moltiplicati per un fattore di e^{β_i} . Se β_i è positivo, allora un aumento di X_i aumenta gli odds di successo, mentre se β_i è negativo, un aumento di X_i diminuisce gli odds di successo.

Nel caso della regressione logistica semplice (con una sola variabile indipendente), l'aumento di un'unità in X comporta un cambiamento di e^{β_1} negli odds di successo. Ad esempio, se $\beta_1 = 0.5$, allora un incremento unitario in X moltiplica gli odds di successo per $e^{0.5} \approx 1.65$, indicando che gli odds aumentano del 65%.

11.4 Valutazione della regressione logistica

Per valutare un modello di regressione lineare si utilizzava R^2 , ma questo non è adatto per la regressione logistica, in quanto la variabile dipendente è binaria. Oltretutto noi non andiamo a minimizzare l'errore quadratico medio, ma la cross-entropy loss.

11.4.1 Pseudo R^2

Per valutare la bontà di adattamento di un modello di regressione logistica, si possono utilizzare diverse misure di pseudo R^2 . Una delle più comuni è il *McFadden's R²*, definito come:

$$R_{\text{McFadden}}^2 = 1 - \frac{\log L_{\text{model}}}{\log L_{\text{null}}}$$

dove $\log L_{\text{model}}$ è il logaritmo della verosimiglianza del modello stimato e $\log L_{\text{null}}$ è il logaritmo della verosimiglianza del modello nullo (un modello senza variabili indipendenti).

11.5 Regressione logistica multiclass

Esistono dei casi dove la variabile dipendente non è binaria, ma può assumere più di due valori (ad esempio, classificazione di immagini in più categorie). In questi casi, si può utilizzare la *regressione logistica multiclass* (o multinomiale), che estende il concetto di regressione logistica a più classi.

11.5.1 Modello di regressione logistica multiclass

Nel caso della regressione logistica multiclass per K classi, il modello stima la probabilità che un'osservazione appartenga a ciascuna delle K classi. La probabilità che un'osservazione x appartenga alla classe k è data da:

$$P(Y = k | X = x) = \frac{e^{\beta_k^T x}}{\sum_{j=1}^K e^{\beta_j^T x}}$$

per $k = 1, 2, \dots, K$, dove β_k sono i coefficienti associati alla classe k .

11.5.2 Interpretazione geometrica dei coefficienti

In un modello di regressione logistica multiclass, ogni classe k è associata a un vettore di coefficienti β_k . Questi vettori definiscono delle iperpiani nello spazio delle caratteristiche che separano le diverse classi. La decisione su quale classe assegnare a una nuova osservazione dipende dalla distanza dell'osservazione da questi iperpiani. In particolare, per una nuova osservazione x , si calcolano i prodotti scalari $\beta_k^T x$ per ogni classe k . L'osservazione viene assegnata alla classe con il valore massimo di questo prodotto scalare, che corrisponde alla massima probabilità stimata.

Decision boundary. La *decision boundary* tra due classi k e j è definita dall'insieme dei punti x per i quali le probabilità stimate delle due classi sono uguali, ovvero quando è indeciso (quindi quando la probabilità è 0.5):

$$P(Y = k | X = x) = P(Y = j | X = x)$$

Ricordando che la regressione logistica definisce iperpiani nello spazio delle caratteristiche, la decision boundary tra due classi sarà un iperpiano che separa le regioni dello spazio in cui il modello predice ciascuna delle due classi:

$$P(y = 1 | x) = 0.5 \Leftrightarrow e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)} = 1 \Leftrightarrow \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n = 0$$

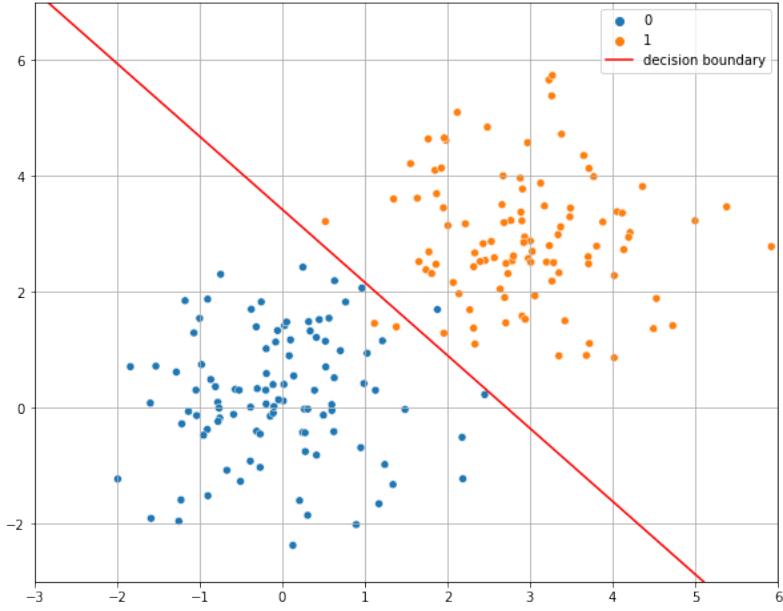


Figura 11.4: Esempio di classificazione binaria ottenuto tramite regressione logistica lineare. I punti rappresentano le osservazioni appartenenti alle due classi, mentre la linea rossa indica la *decision boundary*, ovvero l’insieme dei punti per i quali il modello assegna una probabilità pari a 0.5. Tale frontiera separa le regioni dello spazio delle caratteristiche in cui il modello predice rispettivamente la classe 0 o la classe 1.

e questa è l’equazione di un iperpiano nello spazio delle caratteristiche. Si può rendere in forma esplicita come:

$$x_n = -\frac{\beta_0}{\beta_n} - \frac{\beta_1}{\beta_n}x_1 - \dots - \frac{\beta_{n-1}}{\beta_n}x_{n-1}$$

dove il coefficiente angolare di ciascuna variabile x_i è dato da $-\frac{\beta_i}{\beta_n}$ e l’intercetta è $-\frac{\beta_0}{\beta_n}$.

11.6 Regressione softmax

La regressione softmax è un’estensione della regressione logistica multiclassa che viene utilizzata quando la variabile dipendente può assumere più di due classi. Mentre la regressione logistica binaria utilizza la funzione sigmoide per modellare la probabilità di appartenenza a una classe, la regressione softmax utilizza la funzione softmax per modellare le probabilità di appartenenza a ciascuna delle K classi.

11.6.1 Funzione softmax

A differenza della regressione logistica multiclassa, che può essere vista come una serie di modelli binari indipendenti, la regressione softmax considera tutte le classi contemporaneamente, garantendo che le probabilità stimate per tutte le classi sommino a 1. La funzione softmax è definita come:

$$P(Y = k \mid X = x) = \frac{e^{\beta_k^T x}}{\sum_{j=1}^K e^{\beta_j^T x}}$$

per $k = 1, 2, \dots, K$, dove β_k sono i coefficienti associati alla classe k .

11.6.2 Interpretazione geometrica dei coefficienti

La regressione Softmax estende la regressione logistica al caso multiclasse, e ciò comporta una serie di implicazioni geometriche particolarmente interessanti. In questo modello, a ciascuna classe k è associato un vettore di coefficienti β_k , che definisce una direzione nello spazio delle caratteristiche. L'osservazione viene assegnata alla classe per la quale il prodotto scalare $\beta_k^\top \mathbf{x}$ risulta maggiore, poiché questo valore determina la probabilità stimata per la classe stessa.

Un concetto centrale è quello della *decision boundary*. La frontiera di decisione tra due classi i e j è formata dall'insieme dei punti \mathbf{x} per cui il modello risulta perfettamente indeciso, cioè quando la probabilità di appartenere alla classe i è esattamente uguale a quella della classe j . In questo caso, l'equazione che determina la decision boundary assume una forma molto semplice:

$$\beta_i^\top \mathbf{x} = \beta_j^\top \mathbf{x}$$

che può essere riscritta come:

$$(\beta_i - \beta_j)^\top \mathbf{x} = 0.$$

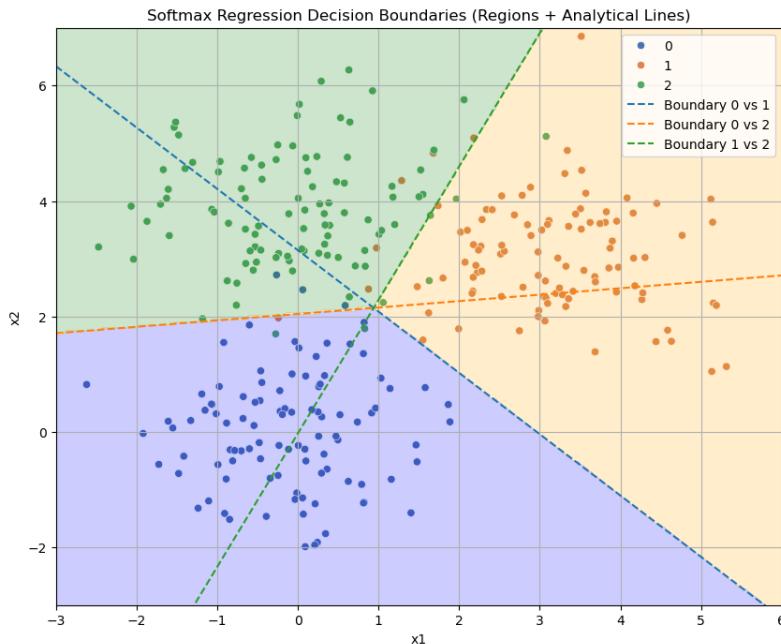


Figura 11.5: Esempio di regressione logistica multiclasse con tre classi, ottenuta tramite softmax regression. Le regioni colorate rappresentano le aree dello spazio delle caratteristiche in cui il modello assegna la probabilità più alta rispettivamente alla classe 0 (blu), alla classe 1 (arancione) e alla classe 2 (verde). Le linee tratteggiate rappresentano le *decision boundary* analitiche tra ciascuna coppia di classi (0 vs 1, 0 vs 2, 1 vs 2), ottenute imponendo l'uguaglianza delle probabilità predette $P(Y = i | x) = P(Y = j | x)$. Le tre frontiere, combinate, suddividono lo spazio in tre regioni lineari, ciascuna associata alla classe con il valore massimo di $\beta_k^\top x$.

Questa relazione mostra chiaramente che la frontiera tra due classi è un *iperpiano* nello spazio delle caratteristiche. Dal punto di vista geometrico, la regressione Softmax suddivide lo spazio in regioni lineari, ognuna delle quali è associata a una delle classi. Ogni iperpiano separa due regioni adiacenti, definendo quindi la struttura complessiva del processo di classificazione.

In questo contesto, i vettori dei coefficienti β_k non vanno interpretati solo come parametri numerici, ma come vere e proprie “direzioni preferenziali”: un valore elevato di β_k in una certa direzione indica che, muovendosi lungo quella direzione, aumenta la probabilità di assegnare l’osservazione alla classe k . Allo stesso modo, la posizione relativa degli iperpiani determinati dai vari vettori β_k stabilisce come lo spazio venga partizionato. Pertanto, variazioni nei coefficienti influiscono direttamente sulla forma e sull’inclinazione delle regioni decisionali.

11.7 Altri modi di regressione multiclass

Esistono altri modi per estendere la regressione logistica al caso multiclass, come il *one-vs-rest* (OvR) e il *one-vs-one* (OvO). Questi approcci suddividono il problema multiclass in una serie di problemi binari, che possono essere risolti utilizzando la regressione logistica standard.

11.7.1 One-vs-Rest (OvR)

Nell’approccio One-vs-Rest, per ogni classe k , si addestra un modello di regressione logistica binaria che distingue la classe k da tutte le altre classi. In questo modo, si ottengono K modelli binari, uno per ciascuna classe. Durante la fase di predizione, si calcolano le probabilità per ciascun modello e si assegna l’osservazione alla classe con la probabilità più alta.

Per esempio, se abbiamo tre classi (A, B, C), si addestrano tre modelli:

- Modello 1: Classe A vs Non-A (B e C)
- Modello 2: Classe B vs Non-B (A e C)
- Modello 3: Classe C vs Non-C (A e B)

11.7.2 One-vs-One (OvO)

Nell’approccio One-vs-One, si addestra un modello di regressione logistica binaria per ogni coppia di classi. Questo significa che per K classi, si addestrano $\frac{K(K-1)}{2}$ modelli binari. Durante la fase di predizione, si utilizza un sistema di voto: ogni modello vota per una delle due classi che distingue, e l’osservazione viene assegnata alla classe che riceve il maggior numero di voti.

Per esempio, se abbiamo tre classi (A, B, C), si addestrano tre modelli:

- Modello 1: Classe A vs Classe B
- Modello 2: Classe A vs Classe C
- Modello 3: Classe B vs Classe C

Capitolo 12

Classificatori generativi

I classificatori generativi sono modelli probabilistici che cercano di modellare la distribuzione congiunta delle caratteristiche e delle etichette di classe, ovvero $P(X, Y)$. Questi modelli assumono che i dati siano generati da un processo probabilistico e cercano di stimare questa distribuzione per effettuare previsioni sulle etichette di classe.

A differenza dei modelli discriminativi, che si concentrano direttamente sulla modellazione della probabilità condizionata $P(Y|X)$, i classificatori generativi cercano di comprendere come i dati vengono generati. Una volta che la distribuzione congiunta è stata stimata, è possibile utilizzare il teorema di Bayes per calcolare la probabilità condizionata delle etichette di classe dato le caratteristiche.

12.1 MAP: Maximum A Posteriori

Il criterio principale utilizzato nei classificatori generativi è il criterio MAP (Maximum A Posteriori). Questo criterio mira a massimizzare la probabilità a posteriori delle etichette di classe dato le caratteristiche osservate. Dobbiamo trovare la classe k che è più probabile dopo aver osservato i dati \mathbf{x} :

$$h(\mathbf{x}) = \arg \max_k P(Y = k | X = \mathbf{x})$$

Utilizzando il teorema di Bayes, possiamo riscrivere questa espressione come:

$$h(\mathbf{x}) = \arg \max_k \frac{P(X = \mathbf{x} | Y = k) P(Y = k)}{P(X = \mathbf{x})}$$

Visto che siamo interessati a massimizzare rispetto a k , possiamo ignorare il denominatore $P(X = \mathbf{x})$ che è costante per tutte le classi. Quindi, il criterio MAP diventa:

$$h(\mathbf{x}) \propto \arg \max_k P(X = \mathbf{x} | Y = k) P(Y = k)$$

da cui abbiamo:

$$h(\mathbf{x}) = \arg \max_k \underbrace{P(X | Y = k)}_{\text{Likelihood}} \underbrace{P(Y = k)}_{\text{Prior}}$$

ovvero scegliamo la classe che massimizza il prodotto tra la probabilità di osservare i dati dato la classe (likelihood) e la probabilità a priori della classe (prior).

12.1.1 Probabilità a priori

Ricordando, dal *teorema di Bayes*, che la probabilità a priori $P(Y = k)$ rappresenta la nostra conoscenza iniziale sulla distribuzione delle classi prima di osservare i dati. Questa probabilità può essere stimata dalla frequenza relativa delle classi nel set di addestramento:

$$P(Y = k) = \frac{\text{Numero di istanze della classe } k}{\text{Numero totale di istanze}}$$

12.1.2 Likelihood

Sempre dal *teorema di Bayes*, la likelihood $P(X = \mathbf{x}|Y = k)$ rappresenta la probabilità di osservare i dati \mathbf{x} dato che sappiamo che appartengono alla classe k . Se ripetiamo questo processo per tutte le classi k , possiamo confrontare queste probabilità per determinare quale classe è più probabile dato i dati osservati abbiamo stimato la probabilità $P(X | Y = k)$ come:

$$P(X = \mathbf{x} | Y = k) = P(X_k)$$

Per stimare questa probabilità $P(X_k)$ si assumono diverse assunzioni sulla distribuzione dei dati all'interno di ciascuna classe. In base a queste assunzioni, creiamo dei modelli differenti per rappresentare la distribuzione dei dati.

12.2 Il problema della likelihood

Un problema comune nei classificatori generativi è la stima accurata della likelihood $P(X = \mathbf{x}|Y = k)$, specialmente quando le caratteristiche sono numerose o quando i dati sono scarsi. Infatti anche questa stima soffre della *curse of dimensionality* (maledizione della dimensionalità), che rende difficile ottenere stime affidabili delle distribuzioni di probabilità in spazi ad alta dimensione.

Pensiamo all'esempio dello *spam*: ipotizziamo di avere un dizionario composto da 10.000 parole e di voler stimare la probabilità che una email sia spam o meno. Per fare questo, dovremmo stimare la probabilità di tutte le possibili combinazioni di parole, ovvero 2^{10000} combinazioni, che diventa computazionalmente costoso. Inoltre, con un numero limitato di esempi di addestramento, molte di queste combinazioni potrebbero non essere mai state osservate, portando a stime di probabilità nulle o inaffidabili.

12.2.1 Il modello ideale nei dati discreti

Se le feature da analizzare sono discrete e assumono un numero limitato di valori, possiamo provare a costruire una tabella di contingenza per ogni classe k . Questa tabella conterrebbe la frequenza di ogni combinazione possibile di valori delle caratteristiche all'interno della classe k . La likelihood $P(X = \mathbf{x}|Y = k)$ può essere stimata come la frequenza relativa di ciascuna combinazione di caratteristiche nella tabella di contingenza.

Il problema principale di questo approccio è che il numero di combinazioni possibili cresce esponenzialmente con il numero di caratteristiche. Ad esempio, se abbiamo d caratteristiche binarie, ci sono 2^d combinazioni possibili. Questo rende impraticabile la costruzione di tabelle di

contingenza per dataset con molte caratteristiche, poiché richiederebbe una quantità enorme di dati per ottenere stime affidabili delle probabilità.

12.2.2 Il modello ideale nei dati continui

Nel caso di feature continue, non si può costruire una tabella di contingenza. L'approccio "ideale" sarebbe stimare la probabilità $P(X = \mathbf{x}|Y = k)$ utilizzando una **distribuzione gaussiana multivariata N** .

QDA: Quadratic Discriminant Analysis. La distribuzione gaussiana multivariata più flessibile è quella che permette a ogni classe di avere la propria matrice di covarianza. Questo modello è noto come *Quadratic Discriminant Analysis* (QDA). Come assunzioni ogni classe k ha il suo vettore di medie μ_k e la sua matrice di covarianza Σ_k . Il problema però, rimane la stima della matrice di covarianza, che richiede un numero elevato di dati per essere stimata accuratamente, specialmente in spazi ad alta dimensione. In particolare, per d caratteristiche, la matrice di covarianza ha $\frac{d(d+1)}{2}$ parametri da stimare per ogni classe e ne dobbiamo stimare una per ogni classe k . Il risultato, tuttavia, è un classificatore potente che può catturare relazioni complesse tra le caratteristiche.

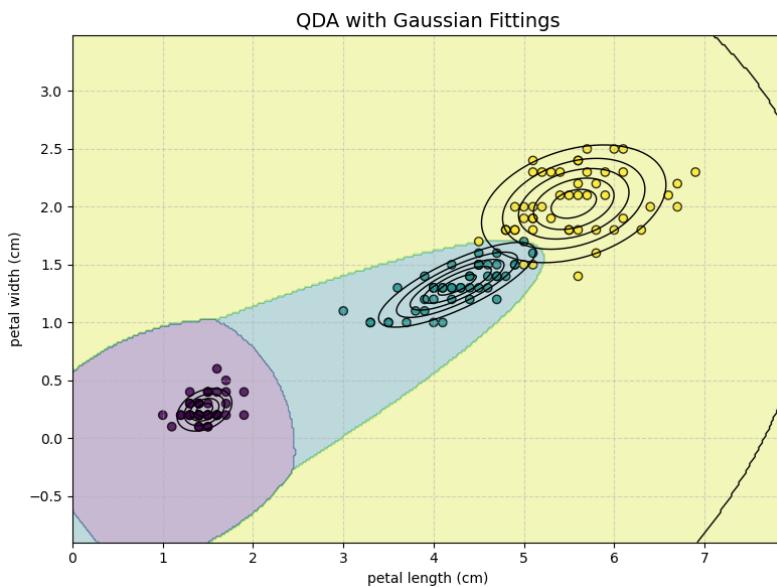


Figura 12.1: Esempio di classificazione tramite *Quadratic Discriminant Analysis* (QDA) applicato alle caratteristiche dei petali del dataset Iris. Le diverse regioni colorate rappresentano le aree dello spazio in cui il classificatore assegna la classe più probabile sulla base delle distribuzioni gaussiane stimate per ciascuna classe. Le ellissi mostrano le isodensity delle gaussiane apprese, che evidenziano le differenti varianze e correlazioni tra le feature: a differenza della LDA, la QDA permette covariance matrices differenti per ciascuna classe, producendo confini decisionali curvi e non lineari.

LDA: Linear Discriminant Analysis. Per semplificare il modello dobbiamo cercare di ridurre il numero di parametri da stimare. Una delle assunzioni più comuni è quella di considerare che tutte le classi condividono la stessa matrice di covarianza, ovvero $\Sigma_k = \Sigma$ per ogni classe k . Questo modello è noto come *Linear Discriminant Analysis* (LDA).

Questo modello è un modello molto più semplice e stabile rispetto al QDA, poiché riduce il numero di parametri da stimare. Comunque computazionalmente rimane costoso, in quanto richiede di stimare una matrice di covarianza di dimensione $d \times d$, che può essere problematico in spazi ad alta dimensione (es. immagini).

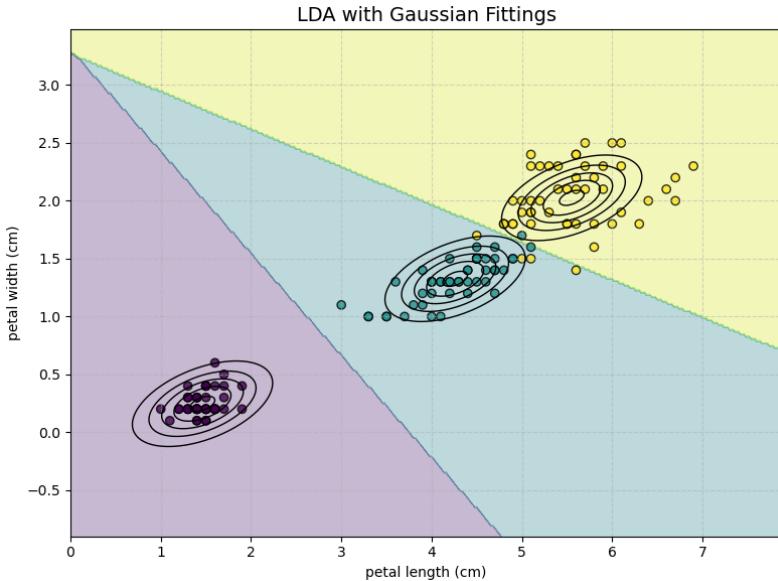


Figura 12.2: Esempio di classificazione ottenuta tramite LDA (Linear Discriminant Analysis) sul dataset Iris utilizzando due caratteristiche dei petali. Le regioni colorate rappresentano le aree dello spazio delle caratteristiche assegnate a ciascuna classe dal modello lineare. Le curve ellittiche mostrano le distribuzioni gaussiane con covarianza condivisa stimate per ciascuna classe, mentre le linee di separazione evidenziano le *decision boundary* lineari tipiche dell'LDA.

12.3 Naive Bayes

Il modello di *Naive Bayes* rappresenta un’ulteriore semplificazione rispetto all’LDA. LDA funziona bene semplificando il modello, ma comunque richiedeva la stima di una matrice di covarianza $d \times d$. Naive Bayes fa un’assunzione: le caratteristiche sono condizionatamente indipendenti dato la classe. Questa assunzione è spesso irrealistica, ma semplifica enormemente il modello e riduce il numero di parametri da stimare.

12.3.1 Assunzione di indipendenza condizionata

Ricordando che il problema è stimare la probabilità $P(X = \mathbf{x} | Y = k)$. Con l’assunzione di indipendenza condizionata, possiamo riscrivere questa probabilità come il prodotto delle probabilità marginali delle singole caratteristiche:

$$\begin{aligned} & P(X = \mathbf{x} | Y = k) \\ &= P(X_1, X_2, \dots, X_n | Y = k) \\ &= P(X_1, \dots, X_n, Y = k) \cdot P(X_2 | X_1, Y = k) \dots P(X_n | X_1, X_2, \dots, X_{n-1}, Y = k) \end{aligned}$$

Dove X_j rappresenta la j -esima caratteristica del vettore \mathbf{x} . Tuttavia, questa fattorizzazione non è sempre utile perché la probabilità $P(X_i \mid X_1, \dots, X_{i-1}, C)$ è condizionata da tutte le altre feature.

Assumiamo quindi **l'indipendenza condizionata** delle features data la classe $Y = k$, ovvero:

$$X_i \perp X_j \mid Y = k \quad \forall i \neq j$$

Sapendo questo, possiamo scrivere:

$$X \perp Y \mid Z \Leftrightarrow P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

quindi segue che, sotto l'assunzione di indipendenza condizionata:

$$P(X_1, \dots, X_n \mid Y = k) = P(X_1 \mid Y) \cdot P(X_2 \mid Y) \dots P(X_n \mid Y) = \prod_{j=1}^n P(X_j \mid Y = k)$$

Quindi, possiamo riscrivere la classificazione MAP come:

$$f(\mathbf{x}) = \arg \max_k P(\mathbf{x}_1 \mid Y = k) \cdot P(\mathbf{x}_2 \mid Y = k) \dots P(\mathbf{x}_n \mid Y = k) \cdot P(Y = k)$$

Varianti del modello Naive Bayes. Il classificatore Naïve Bayes si basa sulla modellazione delle probabilità condizionate dei singoli attributi $P(X_i \mid Y = k)$, che risultano semplici da trattare grazie all'assunzione di indipendenza tra le variabili. A seconda della natura dei dati, queste distribuzioni possono essere modellate in modi differenti. Se si assume che ogni X_i seguia una distribuzione Gaussiana all'interno di ciascuna classe, il modello prende il nome di *Gaussian Naïve Bayes*. Al contrario, quando le variabili rappresentano conteggi o categorie discreti, è più appropriato adottare una distribuzione multinomiale, ottenendo il modello *Multinomial Naïve Bayes*. La scelta della distribuzione dipende quindi dal tipo di dati e dal contesto applicativo.

12.4 Naive Bayes Gaussiano

Nel caso di feature continue, una scelta comune è quella di assumere che ogni caratteristica X_j seguia una distribuzione gaussiana all'interno di ciascuna classe k . In questo caso, la probabilità condizionata $P(X_j \mid Y = k)$ può essere modellata come:

$$P(X_j = x_j \mid Y = k) = \frac{1}{\sqrt{2\pi\sigma_{jk}^2}} \exp\left(-\frac{(x_j - \mu_{jk})^2}{2\sigma_{jk}^2}\right)$$

dove μ_{jk} e σ_{jk}^2 sono la media e la varianza della caratteristica X_j nella classe k .

Ipotizziamo l'esempio di un dover classificare in base a $X = [H, W]$ (altezza e peso) se una persona è maschio o femmina. Se assumiamo che i dati seguono una distribuzione gaussiana, le probabilità $P(H \mid Y = k)$ e $P(W \mid Y = k)$ (ovvero le likelihoods) possono essere modellate con gaussiane separate per ogni classe. Da qui otteniamo:

- H_1 : le altezze di un soggetto quando $k = 1$ (maschio).
- H_0 : le altezze di un soggetto quando $k = 0$ (femmina).

- W_1 : i pesi di un soggetto quando $k = 1$ (maschio).
- W_0 : i pesi di un soggetto quando $k = 0$ (femmina).

Ulteriormente, possiamo definire le gaussiane come:

- $P(H = h | Y = 1) = N(h; \mu_{H1}, \sigma_{H1}^2)$
- $P(H = h | Y = 0) = N(h; \mu_{H0}, \sigma_{H0}^2)$
- $P(W = w | Y = 1) = N(w; \mu_{W1}, \sigma_{W1}^2)$
- $P(W = w | Y = 0) = N(w; \mu_{W0}, \sigma_{W0}^2)$

Ovvero le varie distribuzioni gaussiane per ogni caratteristica e classe. Applicando dopo la regola di classificazione, possiamo dire che un nuovo soggetto con altezza h e peso w viene classificato come maschio se:

$$P(H = h | Y = 1)P(W = w | Y = 1)P(Y = 1) > P(H = h | Y = 0)P(W = w | Y = 0)P(Y = 0)$$

altrimenti viene classificato come femmina.

Questa rappresentazione può essere visualizzata nella figura 12.3.

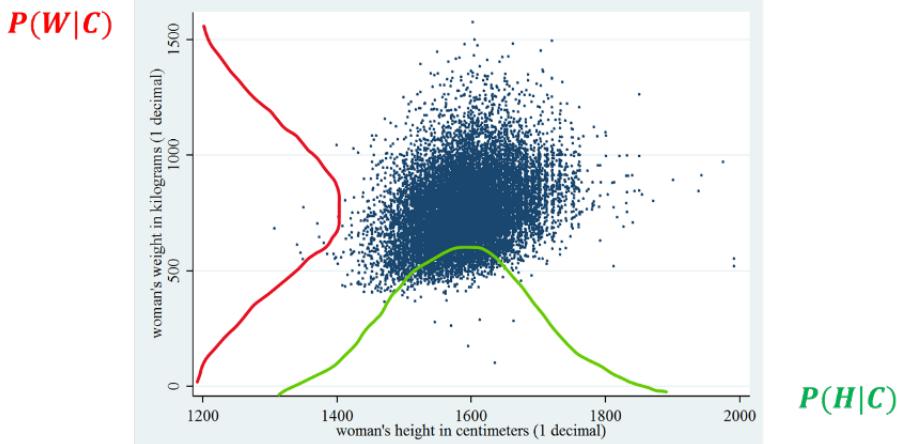


Figura 12.3: Distribuzioni marginali delle caratteristiche utilizzate in un classificatore Naïve Bayes. La curva rossa rappresenta la distribuzione della variabile *peso* condizionata alla classe C ($P(W | C)$), mentre la curva verde rappresenta la distribuzione della variabile *altezza* condizionata alla stessa classe ($P(H | C)$). Il grafico mostra come le due variabili vengano trattate come indipendenti nel modello, permettendo di modellare separatamente le loro densità anche in presenza di una forte correlazione apparente nei dati osservati.

12.4.1 Implicazioni delle assunzioni di Naive Bayes gaussiano

Il modello di Naive Bayes gaussiano, pur essendo semplice ed efficiente, si basa su alcune assunzioni che possono non riflettere accuratamente la realtà dei dati. In particolare, l'assunzione di indipendenza condizionata tra le caratteristiche dato la classe può essere irrealistica in molti contesti, specialmente quando le caratteristiche sono fortemente correlate. Ad esempio, nel caso di dati biometrici come altezza e peso, queste due variabili tendono a essere correlate, e l'assunzione di indipendenza potrebbe portare a stime di probabilità imprecise.

Generalmente questo porta ad avere una matrice di covarianza *diagonale*, dove le covarianze sono 0 e le varianze sono sugli elementi diagonali.

$$\Sigma_k = \begin{bmatrix} \sigma_{1k}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{2k}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{dk}^2 \end{bmatrix}$$

Questo implica che il modello non cattura le relazioni lineari tra le caratteristiche, limitando la sua capacità di modellare dati complessi. Come risultato, abbiamo che le *decision boundary* risultano essere **lineari**.

Questo significa che, pur mantenendo l'assunzione di indipendenza (e dunque una matrice di covarianza priva di termini fuori diagonale), il modello consente comunque a ciascuna classe di avere varianze diverse per ogni caratteristica. In altre parole, nonostante la struttura diagonale sia la stessa per tutte le classi, i valori sulle diagonali possono cambiare da classe a classe. Questa flessibilità fa sì che le distribuzioni gaussiane associate alle diverse classi possano avere “larghezze” differenti lungo ciascun asse, producendo così confini decisionali che non sono più puramente lineari. Le decision boundary risultanti possono quindi curvarsi in base alle differenze di varianza tra le classi, introducendo una forma di non-linearietà, pur rimanendo vincolate dall'impossibilità di modellare covarianze tra le variabili.

12.4.2 Confronto dei decision boundaries

Nella figura 12.4 sono mostrati i confini decisionali (decision boundaries) di tre diversi classificatori generativi: QDA, LDA e Naive Bayes gaussiano.

Si può osservare come le diverse assunzioni sui modelli di covarianza influenzino la forma dei confini decisionali. QDA, con la sua matrice di covarianza completa e differente per ogni classe, produce confini decisionali non lineari e flessibili. LDA, con la sua matrice di covarianza condivisa, genera confini lineari più semplici. Infine, il Naive Bayes gaussiano, con la sua matrice di covarianza diagonale, mantiene confini lineari ma con una struttura che riflette le varianze specifiche di ogni classe.

Perché il decision boundary di Naive Bayes è lineare? Nonostante nell'immagine si veda una certa curvatura nei confini decisionali del Naive Bayes gaussiano, è importante chiarire che, in realtà, i confini decisionali di questo modello sono **linearmente separabili** nello spazio delle caratteristiche originali. Questo significa che, sebbene possano apparire curvi a causa della rappresentazione grafica o della scala degli assi, matematicamente possono essere descritti come iperpiani lineari. Questo avviene perché il modello assume che le caratteristiche siano indipendenti dato la classe, il che porta a una forma di decision boundary che può essere espressa come una combinazione lineare delle caratteristiche.

Inoltre, l'unico modo di rappresentarli linearmente sarebbe quello di imporre che tutte le varianze siano uguali:

$$\sigma_{jk}^2 = \sigma_j \quad \forall k$$

e questo porterebbe a confini decisionali effettivamente lineari e paralleli tra loro, come in LDA (infatti in questo caso la matrice di covarianza sarebbe la stessa per tutte le classi e diagonale).

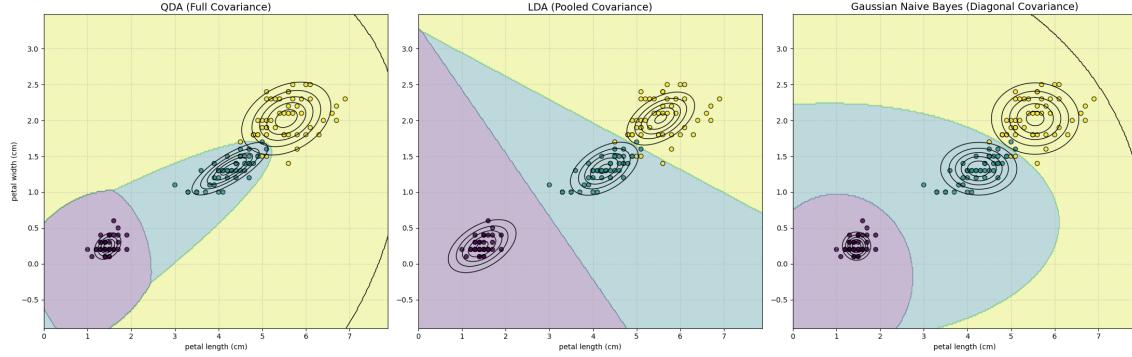


Figura 12.4: Confronto tra tre modelli generativi per la classificazione multiclasse sul dataset *Iris*. A sinistra è mostrato il modello QDA, che utilizza matrici di covarianza *complete* e differenti per ciascuna classe: ciò permette di modellare forme ellittiche con orientamenti e ampiezze diverse, producendo confini decisionali nettamente **non lineari**. Al centro è riportato il modello LDA, che assume una matrice di covarianza *condivisa* tra tutte le classi: le distribuzioni risultano quindi ellissi con la stessa forma e orientamento, e i confini decisionali diventano necessariamente **lineari**. A destra è illustrato il Gaussian Naive Bayes, che assume una matrice di covarianza *diagonale* e diversa per ciascuna classe: ogni classe presenta ellissi allineate con gli assi, e i confini decisionali rimangono **lineari**, ma organizzati secondo regioni che riflettono le varianze specifiche di ogni classe. Questo confronto evidenzia come le diverse assunzioni sulla struttura della covarianza influenzino profondamente la flessibilità del modello e la forma dei confini decisionali.

12.5 Naive Bayes Multinomiale

Nel caso di feature discrete, si utilizza (comunemente) il modello di *Naive Bayes Multinomiale*. Ricordando che la distribuzione multinomiale è una generalizzazione della distribuzione binomiale per più di due categorie, e che modella la probabilità di ottenere esattamente (x_1, \dots, x_d) successi per ogni dei d possibili risultati in una sequenza di n prove indipendenti, dove ogni prova ha una delle d categorie ognuna delle quali con probabilità $(p_{k1}, p_{k2}, \dots, p_{kd})$. Da qui:

- $n = \sum_{i=1}^k x_i$: Il numero totale di prove (o conteggi).
- d : il numero di categorie possibili.
- p_{ki} la probabilità di successo per la categoria i nella classe k .

Usando la forma analitica della distribuzione multinomiale, si può scrivere:

$$P(\mathbf{x} \mid Y = k) = \frac{n!}{x_1! x_2! \dots x_d!} p_{k1}^{x_1} p_{k2}^{x_2} \dots p_{kd}^{x_d}$$

e quindi la regola di classificazione MAP diventa:

$$h(\mathbf{x}) = \arg \max_k P(Y = k) \cdot \frac{(\sum_{i=1}^d x_i)!}{x_1! x_2! \dots x_d!} \prod_{j=1}^d p_{kj}^{x_j}$$

Ma poiché il fattore $\frac{(\sum_{i=1}^d x_i)!}{x_1! x_2! \dots x_d!}$ è costante rispetto a k , possiamo ignorarlo nella massimizzazione,

ottenendo:

$$h(\mathbf{x}) = \arg \max_k P(Y = k) \cdot \prod_{j=1}^d p_{kj}^{x_j}$$

12.5.1 Stima dei parametri

Dobbiamo stimare, per utilizzare il modello, la probabilità a priori $P(Y = k)$ e le probabilità condizionate p_{kj} .

Stimare la probabilità a priori. Stimiamo la probabilità a priori $P(Y = k)$ come la frequenza relativa della classe k nel set di training:

$$P(Y = k) = \frac{\text{Numero di istanze della classe } k}{\text{Numero totale di istanze}} = \frac{\sum_j [y_j = k]}{N}$$

dove N è il numero totale di istanze nel set di training e $[y_j = k]$ sono le parentesi di Iverson, funzione indicatrice che vale 1 se l'istanza j appartiene alla classe k , altrimenti vale 0.

Stimare le probabilità condizionate. Per stimare le probabilità condizionate p_{kj} , possiamo utilizzare la frequenza relativa delle caratteristiche nella classe k :

$$\begin{aligned} p_{kj} &= P(X_j = 1 \mid Y = k) \\ &= \frac{\text{Numero di volte che la caratteristica } j \text{ appare nella classe } k}{\text{Numero totale di caratteristiche nella classe } k} \\ &= \frac{\sum_i x_{ij} [y_i = k]}{\sum_i \sum_j x_{ij} [y_i = k]} \end{aligned}$$

Dove x_{ij} rappresenta il conteggio della caratteristica j nell'istanza i .

12.5.2 Problemi di stima e smoothing

Queste due stime, però, possono causare dei problemi in certe situazioni.

Problema delle probabilità nulle. Un problema comune è quello delle probabilità nulle. Se una caratteristica j non appare mai nella classe k nel set di training, allora la stima di p_{kj} sarà zero. Questo porta a un problema quando si calcola la probabilità condizionata $P(X = \mathbf{x} \mid Y = k)$, poiché il prodotto delle probabilità condizionate includerà un termine zero, rendendo l'intera probabilità nulla, indipendentemente dagli altri termini. Per risolvere questo problema, si può utilizzare una tecnica chiamata *smoothing*, come il *Laplace smoothing*. Questa tecnica consiste nell'aggiungere un piccolo valore costante (solitamente 1) a ciascun conteggio delle caratteristiche, in modo da evitare probabilità nulle. La stima di p_{kj} con Laplace smoothing diventa:

$$p_{kj} = \frac{\sum_i x_{ij} [y_i = k] + 1}{\sum_i \sum_j x_{ij} [y_i = k] + d}$$

dove d è il numero totale di caratteristiche.

Problema dell'underflow numerico. Un altro problema che può sorgere durante il calcolo della probabilità condizionata $P(X = \mathbf{x}|Y = k)$ è l'underflow numerico. Poiché questa probabilità è calcolata come il prodotto di molte probabilità condizionate, il risultato può diventare estremamente piccolo, portando a problemi di precisione numerica. Per evitare questo problema, è comune lavorare con i logaritmi delle probabilità invece delle probabilità stesse. Poiché il logaritmo di un prodotto è la somma dei logaritmi, possiamo riscrivere la regola di classificazione MAP come:

$$h(\mathbf{x}) = \arg \max_k \log P(Y = k) + \sum_{j=1}^d x_j \log p_{kj}$$

Capitolo 13

Rappresentazione dei Dati

La *rappresentazione dei dati* rappresenta il modo in cui le osservazioni vengono codificate per poter essere analizzate, confrontate e trasformate. Un insieme di variabili descrittive può essere interpretato non solo come una collezione di valori eterogenei, ma come una struttura matematica coerente: ogni osservazione diventa un punto in uno **spazio delle features**, le cui dimensioni corrispondono alle variabili considerate. Questa prospettiva consente di tradurre concetti qualitativi come somiglianza, variazione o struttura in relazioni geometriche, aprendo la strada all'uso sistematico di tecniche matematiche e computazionali per l'analisi dei dati.

13.1 Spazio di feature

Da qui, si può estrapolare una definizione più formale di **rappresentazione dei dati**:

La rappresentazione dei dati in uno spazio può essere formalizzata come una funzione di rappresentazione

$$f : \mathcal{X} \rightarrow \mathbb{R}^d,$$

che associa a ogni osservazione $x \in \mathcal{X}$ un vettore $f(x)$ appartenente a uno spazio vettoriale detto spazio delle feature o spazio di rappresentazione.

Definizione 13.1

13.1.1 Feature extraction

Ad esempio, un'immagine digitale può essere inizialmente descritta come una griglia bidimensionale di pixel. Attraverso una rappresentazione vettoriale, tale immagine può essere interpretata come un punto in uno spazio di dimensione pari al numero di pixel, in cui ogni coordinata corrisponde all'intensità di un singolo pixel. In questo caso, la feature extraction coincide con l'operazione di *flattening* dell'immagine in un vettore. In contesti più generali, le feature possono invece catturare proprietà più astratte dei dati, come statistiche locali, pattern strutturali o relazioni semantiche, riducendo la dimensionalità e migliorando la separabilità delle osservazioni nello spazio delle feature.

Un aspetto centrale nella rappresentazione dei dati è la scelta delle *feature*. Spesso, infatti, i dati grezzi non sono immediatamente adatti a essere interpretati come vettori in uno spazio utile per l'analisi: è necessario trasformarli in una forma più strutturata e informativa. Questo processo

prende il nome di **feature extraction** e consiste nel definire una mappatura che, a partire dai dati originali, produca una rappresentazione vettoriale più compatta o più significativa per il compito considerato.

La feature extraction può essere formalizzata come una funzione di rappresentazione

$$f : \mathcal{X} \rightarrow \mathbb{R}^d,$$

che associa a ogni osservazione $x \in \mathcal{X}$ un vettore di feature $f(x)$ appartenente allo spazio delle feature. La scelta della funzione f induce una specifica geometria dello spazio di rappresentazione e determina, di conseguenza, le relazioni di similarità tra le osservazioni.

Definizione 13.2

Gli spazi \mathbb{R}^d (e, più in generale, \mathbb{R}^m) vengono indicati come *feature spaces* o *representation spaces*, in quanto spazi vettoriali a cui appartengono le rappresentazioni dei dati, dette anche *feature vectors*. Il processo di mappatura dei dati mediante una funzione di rappresentazione prende il nome di *data representation*.

13.1.2 Proprietà dello spazio di feature

Poiché le osservazioni x sono vettori che vivono in un certo spazio vettoriale, **tutte le proprietà** di tali vettori e di tali spazi dell'algebra lineare **si applicano anche alle osservazioni**.

Norme

Un esempio di proprietà utili è la possibilità di definire delle *norme* sugli spazi vettoriali, che permettono di misurare la *lunghezza* o la *magnitudine* di un vettore.

Una norma è una funzione che associa a ogni vettore v un numero reale non negativo, denotato come $\|v\|$, che rappresenta la lunghezza del vettore. Le norme più comuni sono:

Norma L_p . - Dato uno spazio vettoriale S , una norma L_p è definita come

$$\|x\|_p = \left(\sum_{i=1}^d |x_i|^p \right)^{1/p},$$

dove x è un vettore a d dimensioni, x_i è la i -esima componente del vettore e $p \geq 1$ è un parametro che determina il tipo di norma.

Norma Euclidea (L_2). - Utilizzando $p = 2$ come parametro della norma L_p , si ottiene la norma Euclidea, definita come

$$\|x\|_2 = \left(\sum_{i=1}^d x_i^2 \right)^{1/2}.$$

Questa norma misura la distanza "diretta" tra l'origine e il punto rappresentato dal vettore x nello spazio Euclideo.

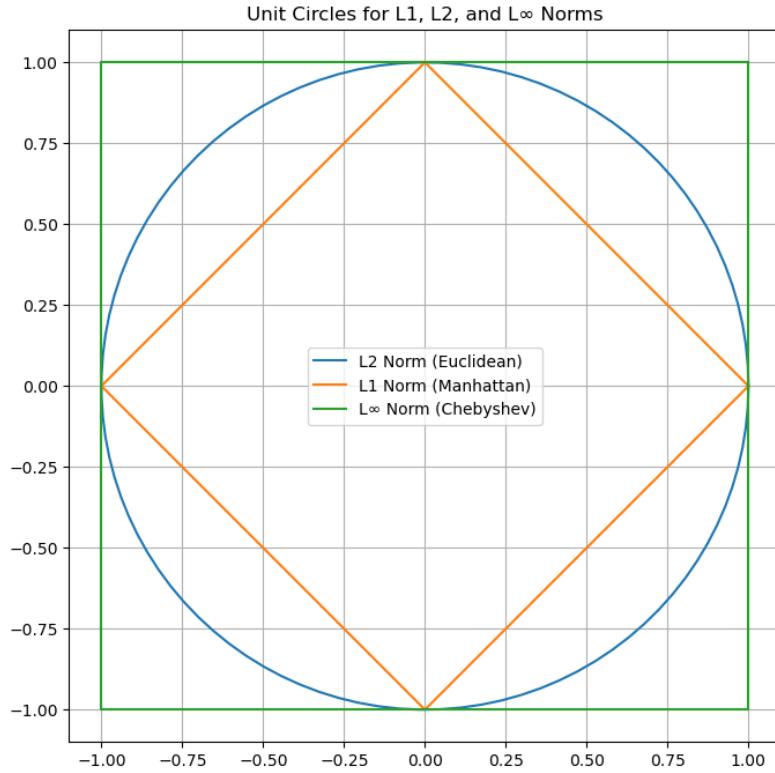


Figura 13.1: Rappresentazione grafica delle norme L_1 , L_2 e L_∞ in una circonferenza unitaria su uno spazio bidimensionale.

Norma Manhattan (L_1). - Utilizzando $p = 1$ come parametro della norma L_p , si ottiene la norma Manhattan, definita come

$$\|x\|_1 = \sum_{i=1}^d |x_i|.$$

Questa norma misura la distanza totale percorsa lungo gli assi cartesiani per raggiungere il punto rappresentato dal vettore x .

Norma di Massimo (L_∞). - Utilizzando $p \rightarrow \infty$ come parametro della norma L_p , si ottiene la norma di Massimo, definita come

$$\|x\|_\infty = \max_{i=1,\dots,d} |x_i|.$$

Questa norma misura la massima distanza lungo una singola dimensione del vettore x .

Si possono visualizzare le differenze tra queste norme utilizzando una **circonferenza unitaria** in uno spazio bidimensionale, come mostrato nella Figura 13.1.

13.2 Metriche

Le norme servono a definire dei modi di misurare la lunghezza dei vettori, chiamate *distanze*. Un problema delle distanze è che funziona solo su spazi euclidei, mentre in molti casi i dati non

vivono in spazi euclidei. Per questo motivo, si utilizzano delle *metriche*, che sono delle funzioni che misurano la distanza tra due punti in uno spazio, indipendentemente dal fatto che lo spazio sia euclideo o meno.

Dato uno spazio S , una funzione

$$m : S \times S \rightarrow \mathbb{R}$$

è una metrica se soddisfa le seguenti proprietà $\forall x, y, z \in S$:

- **Non-negatività:** $m(x, y) \geq 0$ e $m(x, y) = 0$ se e solo se $x = y$.
- **Simmetria:** $m(x, y) = m(y, x)$.
- **Disuguaglianza triangolare:** $m(x, z) \leq m(x, y) + m(y, z)$.

Definizione 13.3

13.2.1 Metriche euclidee

Le metriche più comuni sono quelle basate sulle norme, che misurano la distanza tra due punti in uno spazio euclideo. Dalla definizione di norma L_p , si può derivare la metrica L_p come:

$$L_p(x, y) = \|x - y\|_p = \left(\sum_{i=1}^d |x_i - y_i|^p \right)^{1/p}$$

Da questo si derivano le metriche:

Metrica Euclidea (L_2). - Utilizzando $p = 2$ come parametro della metrica L_p , si ottiene la metrica Euclidea, definita come

$$L_2(x, y) = \left(\sum_{i=1}^d (x_i - y_i)^2 \right)^{1/2}.$$

Questa metrica misura la distanza "diretta" tra i punti rappresentati dai vettori x e y nello spazio Euclideo.

Metrica Manhattan (L_1). - Utilizzando $p = 1$ come parametro della metrica L_p , si ottiene la metrica Manhattan, definita come

$$L_1(x, y) = \sum_{i=1}^d |x_i - y_i|.$$

Questa metrica misura la distanza totale percorsa lungo gli assi cartesiani per raggiungere il punto rappresentato dal vettore y partendo dal punto rappresentato dal vettore x .

Metrica di Massimo (L_∞). - Utilizzando $p \rightarrow \infty$ come parametro della metrica L_p , si ottiene la metrica di Massimo, definita come

$$L_\infty(x, y) = \max_{i=1, \dots, d} |x_i - y_i|.$$

Questa metrica misura la massima distanza lungo una singola dimensione tra i vettori x e y .

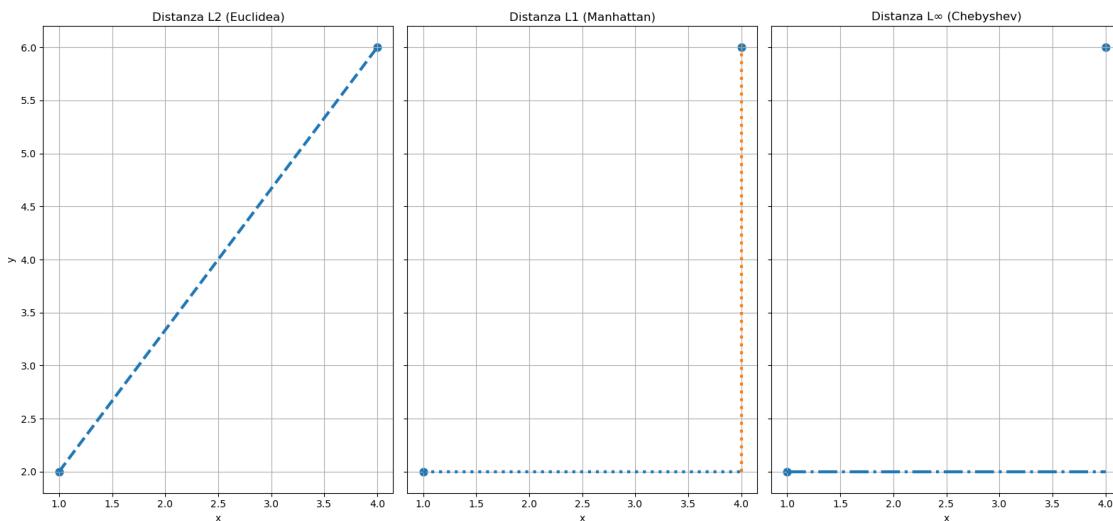


Figura 13.2: Confronto geometrico tra le metriche L_2 , L_1 e L_∞ tra due punti nel piano. La distanza L_2 (Euclidea) corrisponde al segmento rettilineo che li congiunge; la distanza L_1 (Manhattan) è ottenuta come somma di spostamenti lungo gli assi coordinati; la distanza L_∞ (Chebyshev) dipende esclusivamente dalla massima differenza lungo una delle coordinate.

13.2.2 Distanza del coseno

Un'altra metrica molto utilizzata, soprattutto in ambito di elaborazione del linguaggio naturale e recupero delle informazioni, è la **distanza del coseno**. Questa metrica misura la dissimilarità tra due vettori in termini dell'angolo tra di essi, piuttosto che della loro distanza euclidea. Questo viene fatto perché nella distanza euclidea si utilizzano i valori, nella distanza del coseno invece si considerano le loro coordinate

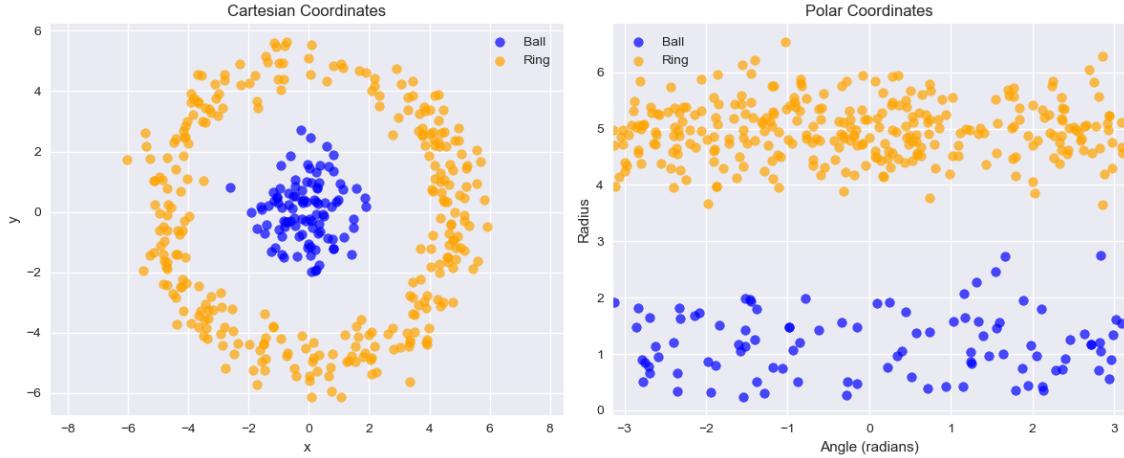
La similarità del coseno tra due vettori x e y è definita come:

$$S_c(x, y) = \frac{x \cdot y}{\|x\| \|y\|},$$

dove $x \cdot y$ è il prodotto scalare tra i vettori x e y , e $\|x\|$ e $\|y\|$ sono le loro norme Euclidee. La distanza del coseno varia tra 0 (quando i vettori sono identici) e 2 (quando i vettori sono opposti).

La distanza del coseno misura l'angolo θ tra i due vettori e restituisce un valore compreso tra -1 e 1 dove:

- 1 indica che i vettori sono identici (angolo di 0 gradi),
- 0 indica che i vettori sono ortogonali (angolo di 90 gradi),
- -1 indica che i vettori sono opposti (angolo di 180 gradi).



Esempio di come una diversa rappresentazione dei dati possa rendere più semplice un problema di separazione. A sinistra, i dati sono rappresentati in coordinate cartesiane (x, y): le due classi (*ball* e *ring*) non risultano linearmente separabili. A destra, la stessa informazione è espressa in coordinate polari (θ, r): utilizzando il raggio come feature principale, le due classi diventano facilmente separabili. Questo esempio mostra come una funzione di rappresentazione appropriata possa semplificare significativamente la struttura geometrica dei dati.

Dalla similarità del coseno, si può derivare la distanza del coseno come:

$$D_c(x, y) = 1 - S_c(x, y) = 1 - \frac{x \cdot y}{\|x\| \|y\|}.$$

Questa misura è utile quando la scala non è importante, ma lo è la direzione. Ad esempio, in applicazioni di elaborazione del linguaggio naturale, due documenti possono essere considerati simili se trattano argomenti simili, indipendentemente dalla loro lunghezza o dal numero di parole utilizzate.

13.3 Feature e funzioni di rappresentazione

Grazie alla rappresentazione dei dati in spazi vettoriali, è possibile applicare una vasta gamma di tecniche matematiche e computazionali per l'analisi dei dati per estrarre informazioni utili, identificare pattern nascosti, applicare modelli di Machine Learning a spazi non lineari e molto altro ancora.

13.3.1 Feature extraction "Black Box"

In molti casi, la funzione di rappresentazione f non è esplicitamente definita, come nell'esempio in figura ??, ma viene appresa automaticamente dai dati stessi attraverso tecniche di *feature learning* o *representation learning*. Questo è il moderno *machine learning*, chiamato **Deep Learning**, che utilizza reti neurali profonde per apprendere rappresentazioni complesse e gerarchiche dei dati, spesso superando le prestazioni delle tecniche di feature extraction manuali.

Alcuni dei (giganteschi) modelli di Deep Learning possono fare feature extraction su:

Testo: I modelli di linguaggio, o LLM (Large Language Models), come GPT-4, BERT e altri, sono in grado di apprendere rappresentazioni semantiche profonde del testo, catturando il

significato, il contesto e le relazioni tra le parole.

Immagini: Le reti neurali convoluzionali (CNN) sono ampiamente utilizzate per l'analisi delle immagini, apprendendo feature gerarchiche che vanno dai bordi e texture di basso livello fino a oggetti e scene complesse. Alcuni esempi sono le ResNet o le Vision Transformer (ViT).

Audio: Le reti neurali ricorrenti (RNN) e le architetture basate su Transformer sono utilizzate per l'elaborazione del segnale audio, catturando pattern temporali e caratteristiche spettrali utili per il riconoscimento vocale, la classificazione musicale e altre applicazioni.

In tutti i casi i principi sono gli stessi elencati sopra: i dati vengono mappati in uno spazio di feature tramite una funzione di rappresentazione, e le proprietà di questo spazio (come le metriche) vengono utilizzate per analizzare e interpretare i dati.

13.3.2 DPI: Data Processing Inequality

La **Data Processing Inequality** (DPI) è un principio fondamentale nella teoria dell'informazione che afferma che l'elaborazione dei dati non può aumentare la quantità di informazione che essi contengono riguardo a una variabile di interesse. In altre parole, qualsiasi trasformazione o manipolazione dei dati non può migliorare la loro capacità di fornire informazioni utili su una variabile target.

Matematicamente, se abbiamo una catena di Markov $X \rightarrow Y \rightarrow Z$, dove Z è condizionalmente indipendente da X dato Y , allora l'informazione mutua¹ soddisfa la seguente diseguaglianza:

$$I(X; Y) \geq I(X; Z)$$

dove $I(X; Y)$ è l'informazione mutua tra le variabili X e Y . Questa diseguaglianza implica che l'informazione che Y fornisce su X è sempre maggiore o uguale a quella che Z fornisce su X .

Questo principio ha importanti implicazioni per la rappresentazione dei dati e la feature engineering:

- **Perdita di informazione:** Ogni trasformazione dei dati, come la riduzione della dimensionalità o la selezione delle feature, può comportare una perdita di informazione. È essenziale valutare attentamente le trasformazioni per garantire che non compromettano la capacità dei dati di rappresentare la variabile di interesse.
- **Bottleneck informativo:** La DPI suggerisce che esiste un limite superiore alla quantità di informazione che può essere estratta dai dati attraverso qualsiasi processo di elaborazione. Questo concetto è cruciale nella progettazione di modelli di machine learning, poiché indica che non è possibile ottenere prestazioni migliori semplicemente aumentando la complessità del modello senza considerare la qualità e la quantità delle informazioni nei dati.
- **I vantaggi delle feature apprese:** Le tecniche di apprendimento delle rappresentazioni, come il deep learning, mirano a trovare trasformazioni dei dati che preservano quanta più informazione possibile riguardo alla variabile target, sfruttando la DPI per guidare la progettazione delle architetture dei modelli.

¹Per informazione mutua si intende una misura della quantità di informazione condivisa tra due variabili casuali.

Capitolo 14

Clustering

Durante l'osservazione di dataset complessi potrebbe essere utile capire le strutture **intrinseche** presenti nei dati. Un approccio per determinare se i dati possono essere suddivisi in gruppi distinti è il *clustering*, una tecnica di unsupervised learning che mira a raggruppare i dati in cluster basati su somiglianze o distanze tra i punti dati.

14.1 Definizione del problema

Il clustering cerca di suddividere un insieme di dati in sottogruppi (cluster) in modo tale che i punti all'interno dello stesso cluster siano più simili tra loro rispetto a quelli appartenenti a cluster diversi. La somiglianza può essere misurata utilizzando varie metriche, come la distanza euclidea, la distanza di Manhattan o altre misure specifiche del dominio.

Sia un set di osservazioni:

$$\mathbf{X} = \{x^{(i)}\}_{i=1}^N, \quad \mathbf{x}^{(i)} \in \mathbb{R}^n$$

Il compito del clustering è quello di dividere \mathbf{X} in K clusters (gruppi):

$$S = \{S_1, S_2, \dots, S_K\}$$

in modo tale che:

$$\forall \mathbf{x}^{(i)} \in \mathbf{X}, \exists! S_j \in S : x^{(i)} \in S_j$$

ovvero ogni punto dati appartiene esattamente a un cluster. Il numero K di cluster è spesso un **iperparametro** che deve essere scelto prima dell'analisi.

14.2 K-means Clustering

Un algoritmo molto popolare di clustering è il *K-means*. L'algoritmo K-means funziona iterativamente per assegnare ogni punto dati al cluster più vicino e aggiornare i centroidi dei cluster fino a quando le assegnazioni non cambiano più.

Per prima cosa si definiscono K differenti vettori $\mu_k \in \mathbb{R}^n$ che rappresentano i centroidi dei cluster. Dopo si definiscono $N \times K$ variabili binarie r_{ij} che ci permettono di mappare ogni punto

dati al cluster più vicino:

$$r_{ij} = \begin{cases} 1 & \text{se } x^{(i)} \text{ è assegnato al cluster } j \\ 0 & \text{altrimenti} \end{cases}$$

Per rendere i cluster il più compatti possibile, l'algoritmo K-means definisce come **funzione di costo**, chiamata anche **distortion function**, la seguente espressione:

$$J = \sum_{j=1}^K \sum_{i=1}^N r_{ij} \|x^{(i)} - \mu_j\|^2$$

La funzione J è anche definita **WCSS**: (*Within-Cluster Sum of Squares*). Per un singolo cluster S_j , il termine interno della somma, $\sum_{i=1}^N r_{ij} \|x^{(i)} - \mu_j\|^2$, rappresenta la somma delle distanze quadratiche tra ogni punto dati assegnato al cluster j e il centroide del cluster μ_j . Questo termine misura quanto i punti all'interno del cluster sono vicini al loro centroide, una misura dell'*inerzia*¹ del cluster, contribuendo così alla compattezza del cluster.

L'obiettivo dell'algoritmo K-means è quello di trovare un set di centroidi $\{\mu_j\}$ e di assegnamenti $\{r_{ij}\}$ che minimizzano l'inerzia J :

$$\hat{S} = \{\hat{r}_{ij}\}_{ij}, \{\hat{\mu}_j\}_j = \arg \min J$$

14.2.1 Ottimizzazione

L'algoritmo K-means utilizza un approccio di ottimizzazione iterativa basato su due passaggi principali: l'assegnazione dei cluster e l'aggiornamento dei centroidi.

Assegnazione. - Per ogni punto dati $x^{(i)}$, si assegna il punto al cluster il cui centroide è più vicino, calcolando la distanza tra il punto e ciascun centroide e scegliendo quello con la distanza minima:

$$r_{ij} = \begin{cases} 1 & \text{se } j = \arg \min_k \|x^{(i)} - \mu_k\|^2 \\ 0 & \text{altrimenti} \end{cases}$$

Questo genererà l'insieme di cluster S_i :

$$S_j = \{\mathbf{x} \in \mathbf{X} : \|\mathbf{x} - \mu_j\|^2 \leq \|\mathbf{x} - \mu_k\|^2, \forall k \neq j\}$$

Aggiornamento. - Una volta che tutti i punti dati sono stati assegnati ai cluster, si aggiornano i centroidi calcolando la media di tutti i punti assegnati a ciascun cluster:

$$\mu_j = \frac{1}{|S_j|} \sum_{x \in S_j} \mathbf{x} = \frac{\sum_{i=1}^N r_{ij} x^{(i)}}{\sum_{i=1}^N r_{ij}}$$

Questo processo di assegnazione e aggiornamento viene ripetuto fino a quando le assegnazioni dei cluster non cambiano più o fino a quando la riduzione della funzione di costo J diventa trascurabile.

¹Per inerzia, in questo contesto, si intende la funzione di costo.

14.2.2 Scegliere il giusto K

Nell'algoritmo K-means, il valore di K (il numero di cluster) è un iperparametro che deve essere scelto prima dell'esecuzione dell'algoritmo. Esistono due tecniche utilizzate comunemente per determinare il valore ottimale di K :

Metodo del gomito (Elbow Method). - Questo metodo prevede di eseguire l'algoritmo K-means per una gamma di valori di K e calcolare la somma delle distanze quadratiche all'interno dei cluster (WCSS) per ciascun valore di K . Si traccia quindi un grafico di WCSS in funzione di K . Il punto in cui la riduzione di WCSS inizia a diminuire significativamente (formando un "gomito" nel grafico) è considerato il valore ottimale di K .

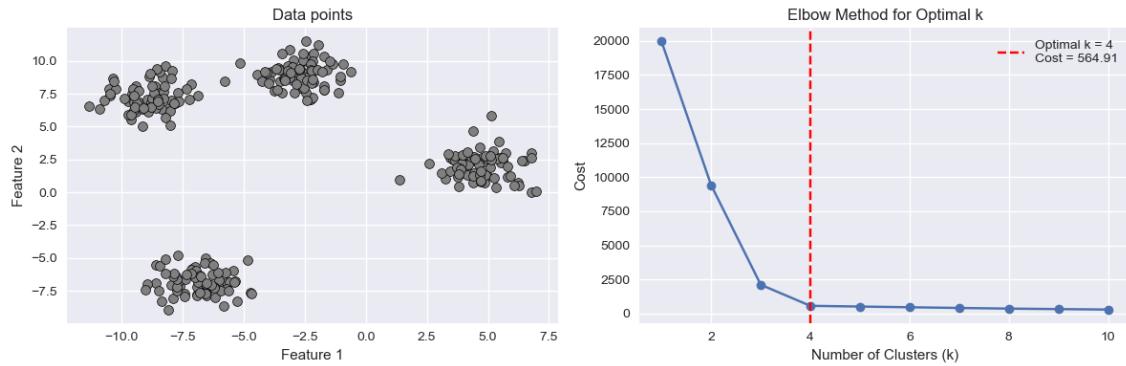


Figura 14.1: Esempio del metodo del gomito per scegliere il numero ottimale di cluster K .

Metodo Silhouette. - Questo metodo valuta la qualità del clustering calcolando il coefficiente di silhouette per ogni punto dati. Il coefficiente di silhouette misura quanto un punto è ben assegnato al suo cluster rispetto ai cluster vicini. Si calcola la silhouette media per diversi valori di K e si sceglie il valore di K che massimizza questa media.

Per fare questo, per un certo punto dati i assegnato al cluster C_I , viene calcolato lo score:

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, j \neq i} d(i, j)$$

dove $d(i, j)$ è la distanza tra i punti i e j . Questo rappresenta la distanza media tra il punto i e tutti gli altri punti nel suo cluster, che deve rimanere piccolo se il cluster ha una poca varianza (cluster compatto).

Successivamente, comparando il risultato $a(i)$ con un cluster diverso C_J , si calcola:

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j)$$

che rappresenta la distanza media tra il punto i e tutti i punti nel cluster più vicino. Infine, il coefficiente di silhouette per il punto i è calcolato come:

$$s(i) = \begin{cases} \frac{b(i) - a(i)}{\max(a(i), b(i))} & \text{se } |C_I| > 1 \\ 0 & \text{altrimenti} \end{cases}$$

Il risultato ottenuto, $-1 \leq s(i) \leq 1$, indica quanto bene il punto è stato assegnato al suo cluster: valori vicini a 1 indicano una buona assegnazione, valori vicini a 0 indicano che il punto è sul confine tra due cluster, e valori negativi indicano che il punto potrebbe essere stato assegnato al cluster sbagliato.

Capitolo 15

Stima della densità

L'obiettivo degli algoritmi di clustering è quello di raggruppare le strutture interne dei dati in insiemi omogenei. K-means, per esempio, assume che i cluster abbiano una forma sferica e che siano separati da confini lineari.

Un altro approccio, più flessibile, è quello basato sulla stima della densità di probabilità che ha generato i dati: i cluster vengono identificati come regioni ad alta densità separate da regioni a bassa densità. Questo approccio è particolarmente utile quando i cluster hanno forme complesse o non lineari.

15.1 Densità di probabilità

La densità di probabilità (PDF, Probability Density Function) è una **funzione continua** che descrive la probabilità relativa di una variabile casuale continua di assumere un certo valore. Per stimarla esistono due metodi principali: Metodi non parametrici e Metodi parametrici.

15.2 Metodi non parametrici

Questi metodi cercano di stimarla direttamente **dai dati**, senza forti assunzioni o la distribuzione di partenza. Il più grande vantaggio di questi metodi è che hanno pochi iperparametri da regolare, uno svantaggio è che, anche se possiamo computare numericamente $f(x)$, non abbiamo una forma chiusa della funzione.

15.2.1 Iistogrammi

Un istogramma è una rappresentazione grafica della distribuzione di un insieme di dati (vedere sezione 17.1). I dati non vengono semplicemente contati, in quanto potrebbe dire che un *bin* 4 volte più largo di un altro contiene 4 volte più dati, ma viene calcolata la PDF $f(x)$ soddisfando una regola: il volume totale sotto la curva deve essere uguale a 1:

$$\int f(x)dx = 1$$

Per fare questo, dobbiamo definire la densità $f(R_i)$, per ogni *bin* R_i , come la probabilità di

quel *bin* diviso il suo *volume* V_i :

$$f(x \in R_i) = \frac{P(R_i)}{V(R_i)} = \frac{|R_i|/N}{V(R_i)}$$

dove $|R_i|$ è il numero di punti nel *bin* R_i e N è il numero totale di punti. Dividendo per il volume ci assicuriamo che un *bin* più largo non abbia una densità più alta solo perché è più largo e ci assicuriamo che l'integrale della densità sia uguale a 1:

$$P(x \in \bigcup_i R_i) = \sum_i P(R_i) = \sum_i f(x \in R_i)V(R_i) = \sum_i \frac{|R_i|/N}{V(R_i)} = \frac{1}{N} \sum_i |R_i| = 1$$

Questo approccio ha 2 *iperparametri* principali:

- Numero di *bin*: Un numero troppo basso di *bin* porta a una stima troppo grossolana della densità, mentre un numero troppo alto porta a una stima troppo rumorosa.
- Posizione della griglia dei *bin*: La posizione della griglia può influenzare significativamente la stima della densità, specialmente se i dati hanno strutture a scale diverse.

Un'alternativa ai bin quadrati sono gli *hexbin* (vedere sezione 17.5), che usano esagoni invece di quadrati per i bin. Gli esagoni hanno il vantaggio di avere una distanza più uniforme tra il centro e i lati rispetto ai quadrati, riducendo così la varianza nella stima della densità.

15.2.2 Kernel Density Estimation (KDE)

Gli istogrammi hanno due problemi:

- I confini rigidi dei *bin* possono introdurre discontinuità nella stima della densità.
- La stima della densità dipende fortemente dalla posizione della griglia dei *bin* e, per il motivo sopra, "a quadretti".

Un'approccio più sofisticato è quello della **stima della densità con kernel** (KDE, Kernel Density Estimation). Invece di contare i punti in ogni *bin*, KDE posiziona una funzione **kernel** centrata su ogni punto dati e somma i contributi di tutti i kernel per ottenere la stima della densità.

Kernel circolare

Un modo "naive" di implementare KDE è quello di centrare una finestra circolare di raggio h , la *bandwidth*, su un certo punto x e calcolare la densità basata su quanti punti $N(x, h)$:

$$N(x, h) = \{x' \in X \text{ s.t. } \|x' - x\|_2 \leq h\}$$

cadono all'interno di questa finestra (la logica è la stessa dell'istogramma):

$$f(x) = \frac{|N(x, h)|/N}{V(h)}$$

dove $V(h)$ è il volume della finestra circolare (in 2D, l'area del cerchio di raggio h : $V(h) = \pi h^2$)

e N è il numero totale di punti. Dalla definizione dell'area del cerchio, possiamo vedere che:

$$\int f(x)dx = \int \frac{|N(x, h)|/N}{V(h)} dx = \frac{1}{NV(h)} \int |N(x, h)|dx = \frac{1}{NV(h)} \cdot N \cdot V(h) = 1$$

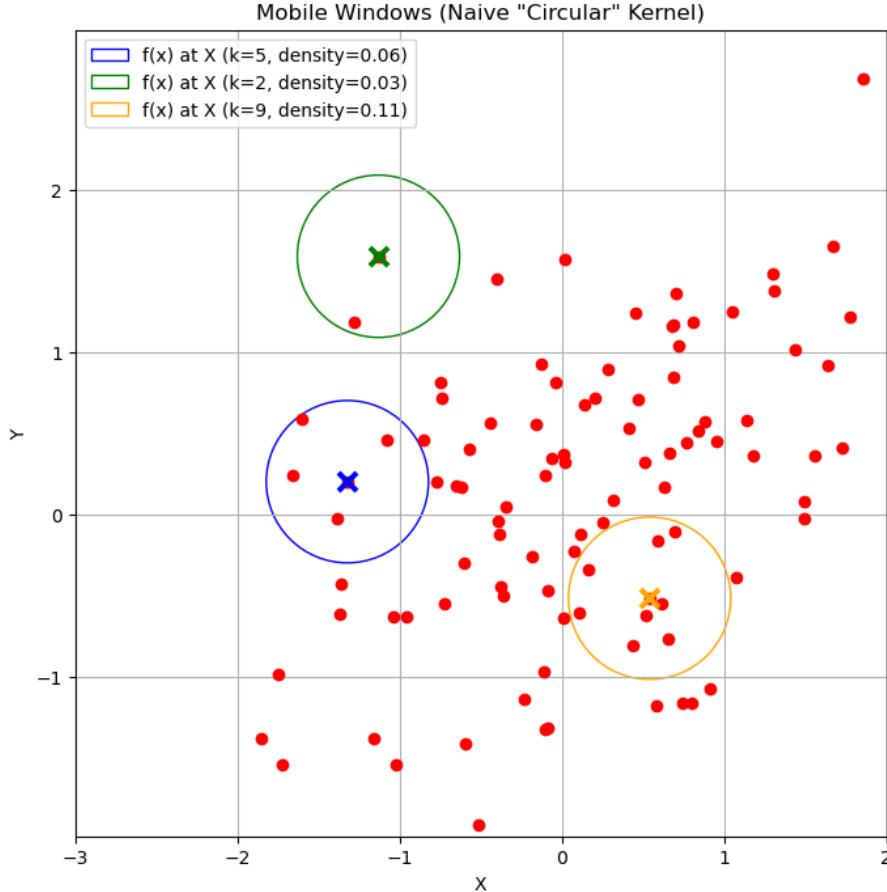


Figura 15.1: Esempio di KDE con kernel circolare. Ogni punto dati contribuisce alla stima della densità all'interno della sua finestra circolare.

Kernel view. Un modo più generale di vedere KDE è quello di scrivere l'espressione $f(x)$ come

$$f(x) = \frac{1}{|X|} \sum_{i=1}^{|X|} K_h(x_i - x)$$

dove x_i sono i punti dati, $|X|$ è il numero totale di punti e K_h è definita come

$$K_h(\mathbf{x}_i - \mathbf{x}) = \begin{cases} \frac{1}{V(h)} & \text{se } \|\mathbf{x}_i - \mathbf{x}\|_2 \leq h \\ 0 & \text{altrimenti} \end{cases}$$

ovvero la funzione kernel che assegna un peso uniforme $\frac{1}{V(h)}$ ai punti all'interno della finestra circolare di raggio h e 0 altrimenti. Questa dipende dal parametro h , che controlla la larghezza della finestra (la *bandwidth*).

Uno dei problemi con questo tipo di kernel è che è molto **sensibile** alla posizione dei punti dati: piccoli cambiamenti nella posizione dei punti possono portare a grandi cambiamenti nella stima della densità. Possiamo notare che, poiché il kernel fa un salto netto da $\frac{1}{V(h)}$ a 0 al bordo della finestra, la stima della densità sarà discontinua in quei punti.

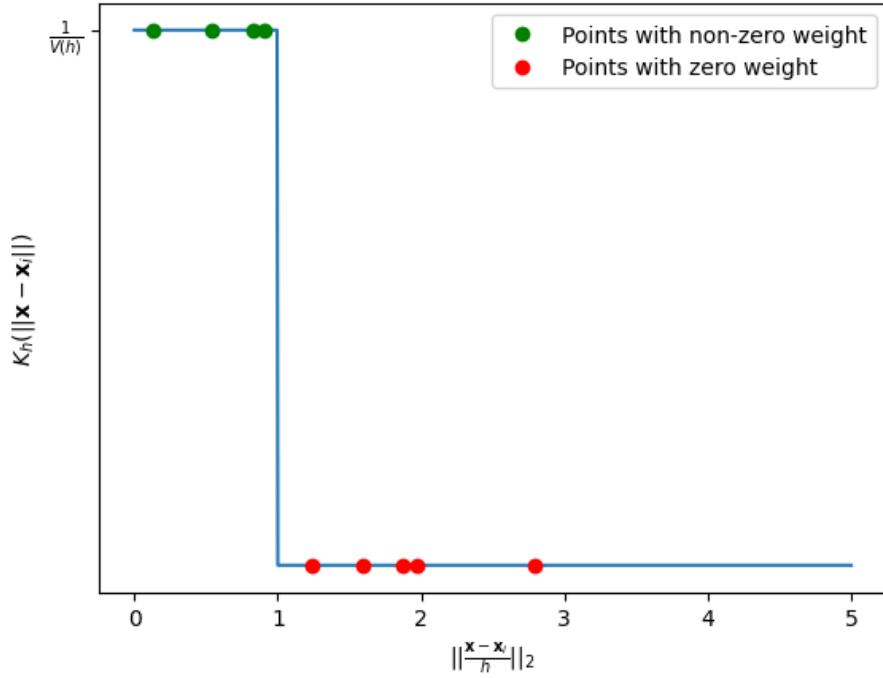


Figura 15.2: Kernel View nella stima di densità tramite Kernel Density Estimation (KDE). Il grafico mostra il valore del kernel $K_h(\|x - x_i\|)$ in funzione della distanza normalizzata $\|(x - x_i)/h\|_2$. I punti verdi contribuiscono alla stima con peso non nullo, mentre i punti rossi, posti oltre il supporto compatto del kernel, hanno peso nullo.

Problemi del kernel circolare. Il kernel circolare ha un problema principale: fa decisioni **nette**, il punto si trova o non si trova all'interno della finestra. Questo porta a stime della densità che sono **discontinue** e **sensibili** alla posizione dei punti dati.

Un modo di risolvere questo è utilizzare dei kernel più *smooth* (morbidi), che assegnano pesi decrescenti ai punti man mano che si allontanano dal centro del kernel, invece di un peso uniforme all'interno di una finestra rigida. Un esempio è il **Kernel Gaussiano**:

$$K_h(x_i - x) \propto \exp\left(-\frac{\|x_i - x\|^2}{2h^2}\right)$$

Questo kernel assegna pesi più alti ai punti vicini al centro e pesi più bassi ai punti lontani, producendo una stima della densità più liscia e meno sensibile alla posizione dei punti dati.

15.2.3 Epanechnikov Kernel

Il kernel guassiano risolve il problema della discontinuità, ma ha un supporto infinito, il che significa che ogni punto dati contribuisce alla stima della densità in ogni punto dello spazio, anche se con un peso molto piccolo. Questo può essere computazionalmente inefficiente.

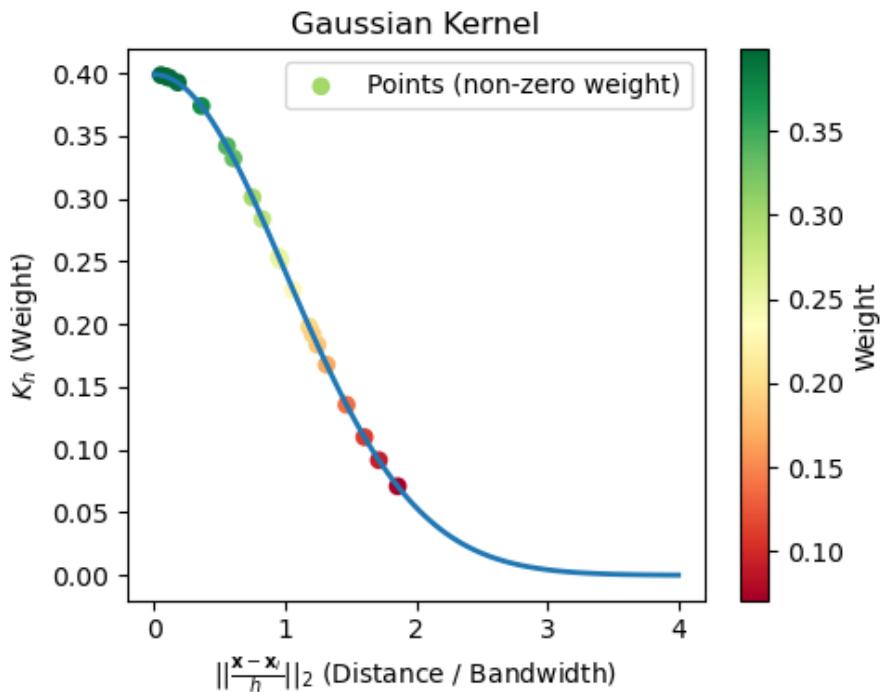


Figura 15.3: Kernel View per la Kernel Density Estimation (KDE) con kernel gaussiano. Il valore del kernel $K_h(\|x - x_i\|)$ decresce in modo continuo all'aumentare della distanza normalizzata $\|(x - x_i)/h\|_2$. A differenza dei kernel a supporto compatto, tutti i punti contribuiscono alla stima con peso non nullo, evidenziato dalla scala cromatica.

Un'alternativa è l'**Epanechnikov Kernel**, che pone a 0 il peso dei punti oltre una certa distanza, mantenendo però una transizione più morbida rispetto al kernel circolare:

$$K_h(x_i - x) = \frac{3}{4h^2} \left(1 - \frac{\|x_i - x\|^2}{h^2} \right) \mathbb{I}(\|x_i - x\|_2 \leq h)$$

dove \mathbb{I} è la funzione indicatrice che vale 1 se la condizione è vera e 0 altrimenti. Questo kernel assegna pesi decrescenti ai punti man mano che si allontanano dal centro, ma pone a 0 il peso dei punti oltre la distanza h .

15.2.4 Tradeoff bias-varianza: scelta della bandwith

La scelta del kernel non è l'unico aspetto importante nella KDE, l'unico iperparametro della KDE è la **bandwith** h , che controlla la larghezza del kernel. La scelta di h ha un impatto significativo sulla stima della densità:

- Un valore di h troppo piccolo porta a una stima della densità rumorosa, con molte fluttuazioni (alta varianza).
- Un valore di h troppo grande porta a una stima della densità troppo liscia, che può perdere dettagli importanti (alto bias).

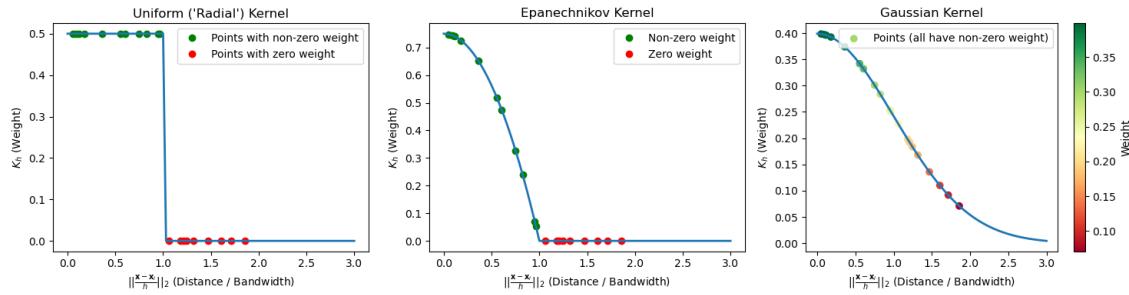


Figura 15.4: Confronto in Kernel View tra diversi kernel utilizzati nella Kernel Density Estimation (KDE). Da sinistra a destra: kernel uniforme (radiale), kernel di Epanechnikov e kernel gaussiano. Il grafico mostra il peso $K_h(\|x - x_i\|)$ in funzione della distanza normalizzata $\|(x - x_i)/h\|_2$. I kernel uniforme ed Epanechnikov presentano supporto compatto, assegnando peso nullo ai punti oltre il raggio unitario, mentre il kernel gaussiano assegna peso non nullo a tutti i punti con contributo decrescente all'aumentare della distanza.

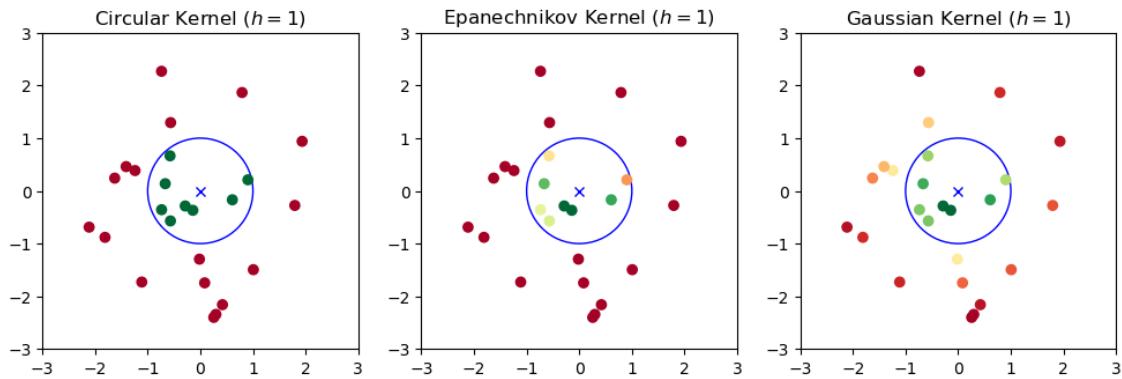


Figura 15.5: Visualizzazione geometrica della Kernel Density Estimation (KDE) in \mathbb{R}^2 per diversi kernel con banda $h = 1$. Da sinistra a destra: kernel circolare (uniforme), kernel di Epanechnikov e kernel gaussiano. Il punto di valutazione x è indicato con una croce blu, mentre il cerchio rappresenta la regione $\|x - x_i\|_2 \leq h$. I punti all'interno del supporto compatto contribuiscono alla stima con peso non nullo per i kernel circolare ed Epanechnikov, mentre il kernel gaussiano assegna un peso decrescente a tutti i punti, inclusi quelli esterni al raggio h .

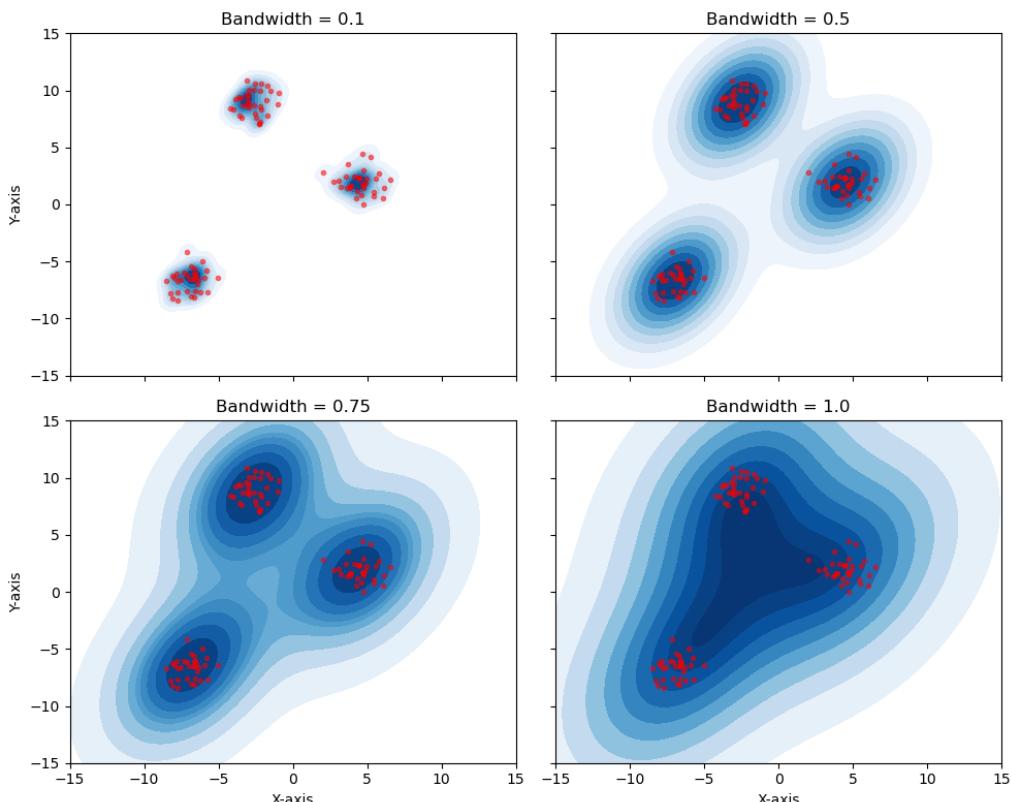


Figura 15.6: Effetto del parametro di banda h nella Kernel Density Estimation (KDE) e relativo trade-off bias-varianza. Per valori piccoli di h (in alto a sinistra) la stima presenta bassa distorsione ma alta varianza, con strutture molto frammentate e sensibili al rumore. All'aumentare di h la varianza diminuisce e la stima diventa più liscia, ma un eccessivo smoothing (in basso a destra) introduce bias, fondendo strutture distinte e perdendo dettagli locali. La scelta del bandwidth governa quindi l'equilibrio tra fedeltà ai dati e regolarità della stima.

15.3 Metodi parametrici

Questi metodi provano a stimare i parametri della **distribuzione di partenza** che si assume abbia generato i dati. Un vantaggio di questi metodi è che, una volta stimati i parametri, abbiamo una forma chiusa della funzione di densità. Uno svantaggio è che spesso richiedono molti iperparametri da regolare e che le assunzioni fatte sulla distribuzione di partenza potrebbero non essere corrette. Ci sono diversi vantaggi per fare questo tipo di stima rispetto a quella non parametrica:

- **Forma analitica:** Una volta stimati i parametri, abbiamo una forma chiusa della funzione di densità, che può essere utile per ulteriori analisi.
- **Modelli compatti:** I modelli parametrici spesso richiedono meno memoria per memorizzare i parametri rispetto a memorizzare tutti i dati.
- **Modelli interpretabili:** I parametri stimati possono avere un significato interpretabile, che può essere utile per comprendere la struttura dei dati.
- **Limiti di dati:** I modelli parametrici possono essere più robusti quando si dispone di pochi dati, poiché fanno assunzioni sulla distribuzione di partenza.

Esistono, tuttavia, anche degli svantaggi:

- **Processo di ottimizzazione:** La stima dei parametri spesso richiede l'ottimizzazione di una funzione obiettivo, che può essere computazionalmente costosa e sensibile ai valori iniziali.
- **Assunzioni sulla distribuzione:** I modelli parametrici fanno assunzioni sulla distribuzione di partenza, che potrebbero non essere corrette per i dati reali.
- **Iperparametri:** I modelli parametrici spesso richiedono la scelta di diversi iperparametri, che possono influenzare significativamente la stima della densità.

15.3.1 Distribuzione Gaussiana Multivariata

Un metodo parametrico molto comune è assumere che i dati seguano una **distribuzione gaussiana multivariata**. Ricordando che la PDF per una gaussiana multivariata a d dimensioni è:

$$N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

quindi questa distribuzione è definita unicamente da due valori: la media $\boldsymbol{\mu}$ e la matrice di covarianza $\boldsymbol{\Sigma}$, i valori da trovare.

MLE: Maximum Likelihood Estimation. Per stimare i parametri della distribuzione, possiamo usare il metodo della **massima verosimiglianza** (MLE, Maximum Likelihood Estimation). L'idea è di trovare i parametri che massimizzano la probabilità di osservare i dati dati i parametri stessi. Sia $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ il nostro insieme di dati, la funzione di verosimiglianza è definita come:

$$P(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^N N(\mathbf{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Per semplificare i calcoli, spesso lavoriamo con il logaritmo della funzione di verosimiglianza, chiamato **log-likelihood** (log-verosimiglianza):

$$\log P(\mathbf{X}|\mu, \Sigma) = \sum_{i=1}^N \log N(\mathbf{x}_i; \mu, \Sigma)$$

Per quanto questa soluzione sembri computazionalmente costosa, per una distribuzione gaussiana esistono delle forme chiuse. Si prende la derivata della log-verosimiglianza rispetto a μ e Σ , la si pone uguale a 0 e si risolve per i parametri.

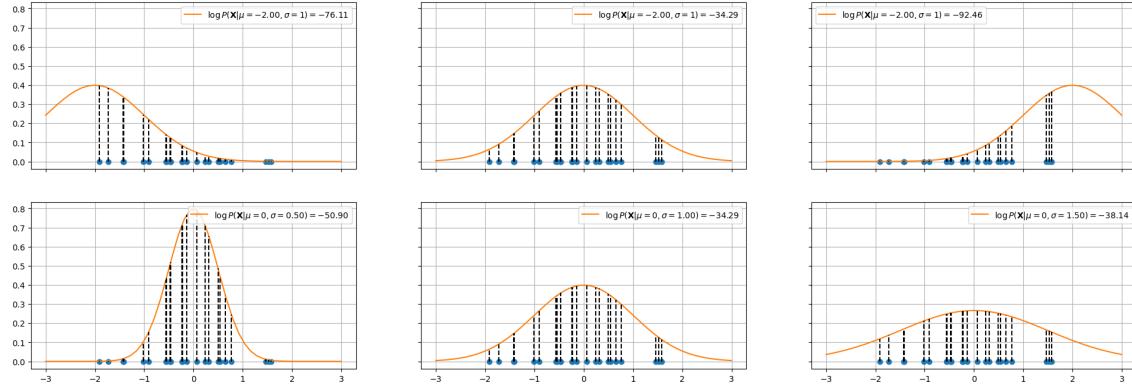
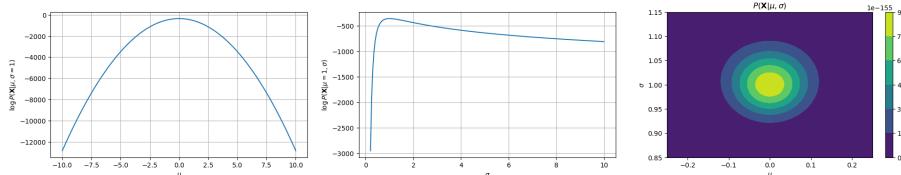


Figura 15.7: Esempio di stima parametrica tramite **Maximum Likelihood Estimation** per una distribuzione gaussiana unidimensionale. In ciascun pannello è mostrata una distribuzione normale $N(\mu, \sigma^2)$ con diversi valori di media μ e deviazione standard σ . I punti blu sull'asse orizzontale rappresentano i dati osservati, mentre le linee verticali tratteggiate indicano il contributo di ciascun campione alla log-verosimiglianza. La curva arancione è la densità di probabilità associata ai parametri considerati e il valore della log-verosimiglianza totale $\log P(\mathbf{X} | \mu, \sigma)$ è riportato in legenda. Il confronto tra i diversi pannelli evidenzia come la log-verosimiglianza vari al cambiare dei parametri e come essa venga massimizzata quando la distribuzione stimata è coerente con i dati osservati.



Confronto della funzione di verosimiglianza per una distribuzione gaussiana unidimensionale in funzione dei parametri. A sinistra è mostrata la log-verosimiglianza $\log P(\mathbf{X} | \mu, \sigma = 1)$ al variare della media μ , che evidenzia un massimo in corrispondenza della media campionaria. Al centro è riportata la log-verosimiglianza $\log P(\mathbf{X} | \mu = 0, \sigma)$ in funzione della deviazione standard σ , che mostra un massimo per un valore ottimale di dispersione dei dati. A destra è visualizzata la superficie di verosimiglianza $P(\mathbf{X} | \mu, \sigma)$ nel piano (μ, σ) , che presenta un unico massimo globale corrispondente alla stima di massima verosimiglianza dei parametri.

Media e covarianza campionaria. Un modo per risolvere il problema di stima dei parametri è massimizzare direttamente la log-likelihood rispetto a μ e Σ . In generale, questo comporta la risoluzione di un problema di ottimizzazione potenzialmente complesso; tuttavia, nel caso

della distribuzione gaussiana, la particolare forma analitica della log-likelihood rende il problema trattabile in forma chiusa.

Sviluppando la log-likelihood della gaussiana multivariata e derivandola rispetto ai parametri, si ottiene che il massimo è raggiunto quando la media della distribuzione coincide con la media empirica dei dati e la covarianza coincide con la covarianza campionaria. In particolare:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i$$

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

Questi stimatori sono noti come **stimatori di massima verosimiglianza** per la media e la covarianza di una distribuzione gaussiana.

Dal punto di vista intuitivo, la MLE sceglie i parametri che rendono i dati osservati il più probabili possibile sotto il modello assunto. Nel caso gaussiano, massimizzare la verosimiglianza equivale a:

- centrare la distribuzione nel punto che minimizza la distanza quadratica media dai dati (la media campionaria);
- scegliere una dispersione che rifletta esattamente la variabilità osservata nei dati (la covarianza campionaria).

Questo risultato evidenzia un aspetto chiave dei metodi parametrici: una volta fissata la forma della distribuzione, la stima dei parametri diventa un problema ben definito e, per alcuni modelli come la gaussiana, ammette soluzioni analitiche semplici e interpretabili.

Limiti dell'approccio MLE. È importante sottolineare che questi risultati valgono *solo* sotto l'assunzione che i dati siano effettivamente generati da una distribuzione gaussiana. Se tale ipotesi è violata, la MLE restituisce comunque una media e una covarianza, ma la densità stimata può rappresentare male la struttura reale dei dati, specialmente in presenza di multimodalità, outlier o distribuzioni fortemente asimmetriche.

15.3.2 Gaussian Mixture Model

La distribuzione gaussiana ha un'assunzione forte: tutti i dati derivano da **un'unico** cluster (è unimodale). Questa assunzione fallisce per molti dataset reali. Una soluzione potrebbe essere utilizzare una **GMM: Gaussian Mixture Model**: invece di utilizzare una gaussiana, modelliamo i dati come una combinazione di K gaussiane, ognuna con la propria media e covarianza. La PDF di un GMM è data da:

$$P(x) = \sum_{k=1}^K \pi_k N(x; \mu_k, \Sigma_k)$$

dove π_k sono i pesi delle singole gaussiane (che sommano a 1), e $N(x; \mu_k, \Sigma_k)$ è la PDF della gaussiana con media μ_k e covarianza Σ_k .

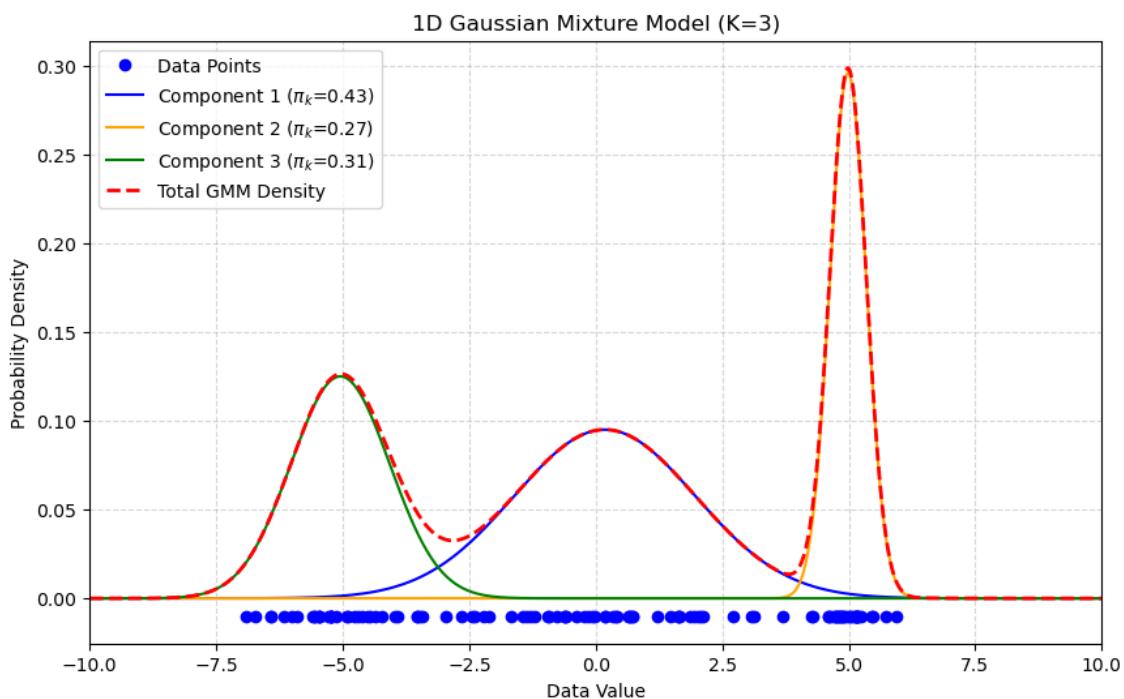


Figura 15.8: Esempio di Gaussian Mixture Model (GMM) unidimensionale con 3 componenti gaussiane. I punti blu sull'asse orizzontale rappresentano i dati osservati. Le curve colorate (blu, arancione e verde) indicano le tre distribuzioni gaussiane componenti con i rispettivi pesi π_k . La curva tratteggiata rossa mostra la densità complessiva risultante dalla somma ponderata delle singole gaussiane. Questo esempio illustra come un GMM possa modellare efficacemente dati multimodali, catturando strutture complesse che una singola gaussiana non potrebbe rappresentare.

Soft clustering con GMM. Un vantaggio importante dei GMM è che permettono un **soft clustering**: ogni punto dati ha una probabilità di appartenere a ciascuna delle K componenti gaussiane, invece di essere assegnato rigidamente a un singolo cluster. Per fare questo, introduce una **variabile latente** Z : una variabile latente è una variabile che non è osservata direttamente nei dati, ma che influenza il processo di generazione dei dati. In questo caso, Z indica quale componente gaussiana ha generato ciascun punto dati.

In GMM, per ogni variabile x_i esiste una variabile latente z_i che indica quale delle K componenti gaussiane ha generato quel punto. Si può utilizzare la probabilità a posteriori, con il teorema di Bayes, per calcolare la probabilità che un punto x_i appartenga alla componente k :

$$P(Z = k|x) = \frac{P(x|Z = k)P(Z = k)}{P(x)}$$

dove $P(x|Z = k)$ è la PDF della componente gaussiana k , $P(Z = k) = \pi_k$ è il peso della componente, e $P(x)$ è la PDF complessiva del GMM. Da questo ricaviamo la **responsabilità** γ_k che un punto x_i appartiene alla componente k :

$$\gamma_k = P(Z = k|\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)}$$

Stima dei parametri con MLE. Per ottimizzare i parametri del GMM (π_k, μ_k, Σ_k), possiamo usare ancora una volta la MLE. Tuttavia, a causa della presenza delle variabili latenti, la log-likelihood diventa più complessa:

$$\log P(\mathbf{X}|\{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K) = \sum_{i=1}^N \log \left(\sum_{k=1}^K \pi_k N(x_i; \mu_k, \Sigma_k) \right)$$

Per massimizzare questa funzione, si usa spesso l'algoritmo di **Expectation-Maximization** (EM).

GMM vs K-means. Un confronto interessante è tra GMM e K-means. Entrambi gli algoritmi cercano di raggruppare i dati in K cluster, ma lo fanno in modi diversi:

- K-means assegna ogni punto dati al cluster più vicino, basandosi sulla distanza euclidea, producendo un **hard clustering**.
- GMM assegna a ogni punto dati una probabilità di appartenenza a ciascun cluster, basandosi sulla stima della densità, producendo un **soft clustering**.

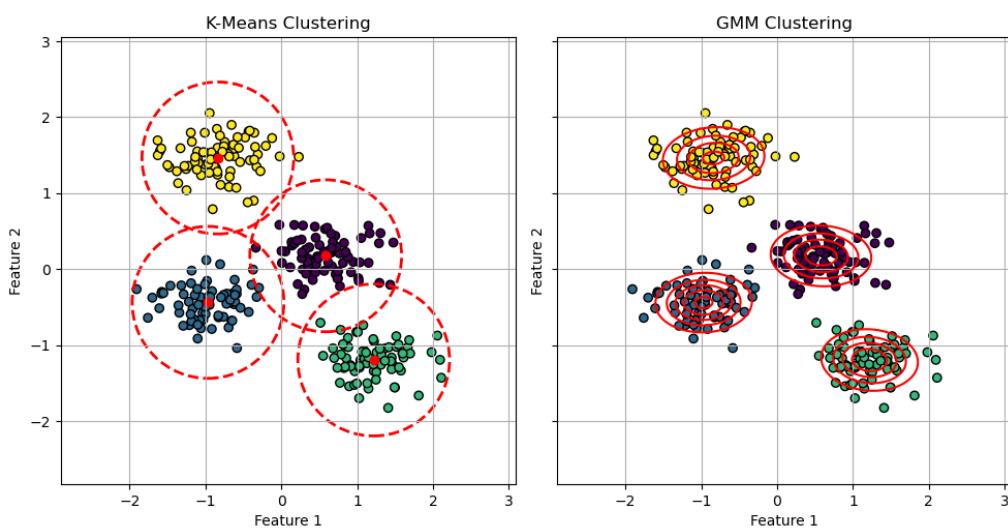


Figura 15.9: Confronto tra clustering con **K-means** e **Modello a miscela di gaussiane (GMM)** su dati bidimensionali. Confronto tra clustering con **K-means** e **Modello a miscela di gaussiane (GMM)** su dati bidimensionali. A sinistra, K-means assegna i punti ai cluster in base alla distanza euclidea dai centroidi, producendo regioni di decisione sferiche e di uguale dimensione (cerchi tratteggiati). A destra, il GMM modella ciascun cluster come una distribuzione gaussiana con media e covarianza proprie, permettendo regioni ellittiche orientate (contorni rossi) e una rappresentazione più flessibile della struttura dei dati.

Capitolo 16

Riduzione della dimensionalità

Finché i dataset con piccole dimensioni erano la norma, l'analisi dei dati era relativamente semplice. Un problema che sussiste tuttavia, è che ad oggi esistono dati rappresentati da dataset con moltissime dimensioni, basti pensare ad un'immagine a colori di risoluzione $800 \times 600 = 1.440.000$ dimensioni. Situazioni come queste possono creare due problemi principali:

- **Maledizione della dimensionalità**: con l'aumento del numero di dimensioni, lo spazio diventa sempre più vuoto e i dati diventano sempre più sparsi. Questo rende difficile trovare pattern significativi nei dati, poiché la distanza tra i punti diventa meno significativa.
- **Multicollinearità**: in dataset con molte dimensioni, è comune che alcune caratteristiche siano altamente correlate tra loro. Questo può portare a problemi di ridondanza e può complicare l'analisi dei dati, rendendo difficile identificare le caratteristiche più importanti.

16.1 Feature Selection vs Feature Extraction

Per affrontare i problemi legati all'alta dimensionalità, esistono due approcci principali:

- **Feature selection**
- **Feature extraction**

Sebbene entrambi mirino a ridurre la dimensionalità dei dati, essi si basano su principi concettualmente diversi.

16.1.1 Feature Selection

La **feature selection** consiste nel selezionare un sottoinsieme delle feature originali, eliminando quelle ritenute ridondanti o poco informative. In questo caso, lo spazio delle feature non viene trasformato: le variabili selezionate sono un sottoinsieme di quelle originali.

Questo approccio presenta il vantaggio di mantenere l'interpretabilità delle feature, ma può risultare limitato quando l'informazione rilevante è distribuita su molte variabili fortemente correlate.

16.1.2 Feature Extraction

La **feature extraction** (o riduzione della dimensionalità) consiste nel costruire nuove feature come combinazioni delle feature originali, ottenendo una rappresentazione dei dati in uno spazio

di dimensione inferiore. A differenza della feature selection, la feature extraction modifica lo spazio delle feature, proiettando i dati in un nuovo spazio che cerca di preservare le proprietà più rilevanti del dataset.

Tra le tecniche di feature extraction, una delle più utilizzate è la **Principal Component Analysis (PCA)**, che consente di ottenere una rappresentazione a dimensionalità ridotta preservando la massima varianza dei dati.

16.2 PCA: Principal Component Analysis

La **Principal Component Analysis (PCA)** è una tecnica di *feature extraction* non supervisionata che consente di ridurre la dimensionalità di un dataset proiettando i dati in uno spazio a dimensione inferiore, preservando quanta più informazione possibile.

In particolare, la PCA cerca una nuova base ortogonale dello spazio delle feature tale che le nuove variabili, dette *componenti principali*, siano ordinate in base alla varianza dei dati lungo tali direzioni.

16.2.1 Interpretazione geometrica

Un dataset può essere visto come un insieme di punti in uno spazio D -dimensionale. In presenza di correlazioni tra le feature, i dati tendono a distribuirsi lungo direzioni privilegiate, occupando solo una porzione dello spazio delle feature.

La Principal Component Analysis individua un nuovo sistema di riferimento in cui:

- gli assi sono ortogonali tra loro;
- le coordinate risultano non correlate;
- il primo asse corrisponde alla direzione lungo cui la varianza dei dati è massima;
- gli assi successivi descrivono direzioni di varianza decrescente.

Nel nuovo sistema di riferimento, la maggior parte della variabilità dei dati è concentrata nelle prime coordinate, rendendo possibile una rappresentazione a dimensionalità ridotta.

16.2.2 Massimizzazione della varianza

Sia $\mathbf{x} \in \mathbb{R}^D$ una variabile aleatoria che rappresenta un punto del dataset. La PCA cerca una direzione unitaria $\mathbf{u} \in \mathbb{R}^D$ tale che la varianza dei dati proiettati lungo \mathbf{u} sia massima:

$$\max_{\mathbf{u}} \text{Var}(\mathbf{u}^T \mathbf{x}) \quad \text{soggetto a} \quad \|\mathbf{u}\| = 1.$$

Il vincolo di norma evita soluzioni banali in cui la varianza cresce indefinitamente.

16.2.3 Matrice di covarianza

La varianza dei dati proiettati lungo una direzione può essere espressa in termini della matrice di covarianza del dataset. Sia $\mathbf{x} \in \mathbb{R}^D$ una variabile aleatoria con media μ . La matrice di covarianza è definita come:

$$\mathbf{S} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$$

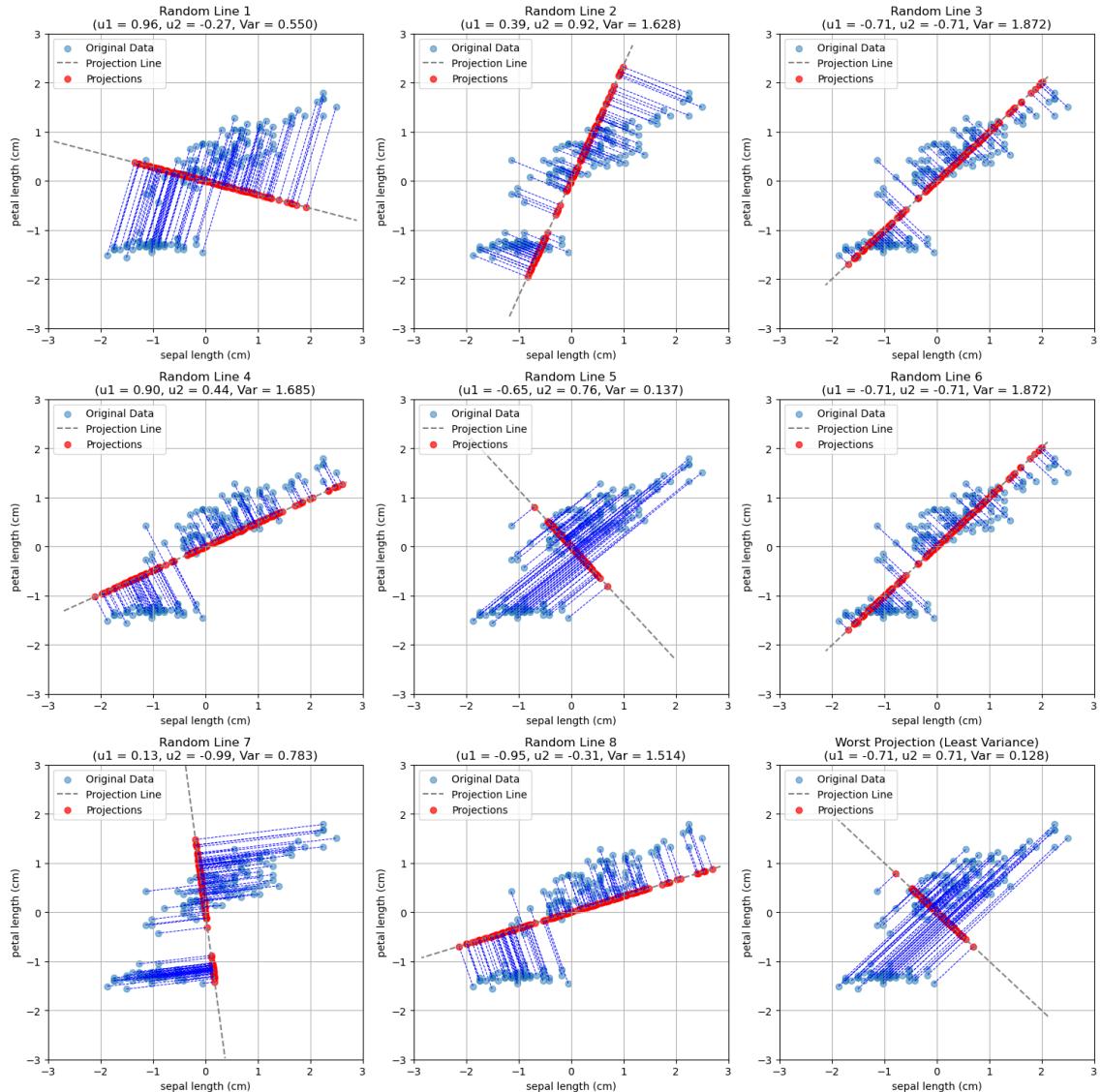


Figura 16.1: Proiezione dei dati lungo diverse direzioni nello spazio delle feature. La varianza delle proiezioni dipende dalla direzione scelta: la PCA individua la direzione che massimizza la varianza dei dati proiettati.

La matrice \mathbf{S} è simmetrica e semidefinita positiva e descrive le relazioni di correlazione tra le feature del dataset.

16.2.4 Varianza della proiezione

La varianza dei dati proiettati lungo una direzione unitaria \mathbf{u} può essere scritta come:

$$\begin{aligned} & \frac{1}{N} \sum_{n=1}^N (\mathbf{u}_1^T \mathbf{x}_n - \mathbf{u}_1^T \bar{\mathbf{x}})(\mathbf{u}_1^T \mathbf{x}_n - \mathbf{u}_1^T \bar{\mathbf{x}})^T \\ &= \frac{1}{N} \sum_{n=1}^N \mathbf{u}_1^T (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T \mathbf{u}_1 \\ &= \mathbf{u}_1^T \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T \mathbf{u}_1 \\ &= \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 \end{aligned}$$

Il problema della PCA consiste quindi nel trovare la direzione \mathbf{u} che massimizza tale quantità, sotto il vincolo $\|\mathbf{u}\| = 1$.

16.2.5 Autovalori e autovettori

Il problema di massimizzazione della varianza può essere risolto tramite lo studio degli autovalori e autovettori della matrice di covarianza \mathbf{S} . Questo perché gli autovettori di \mathbf{S} rappresentano le direzioni principali del dataset, mentre gli autovalori associati indicano la quantità di varianza spiegata lungo ciascuna di queste direzioni. Si può dimostrare che il problema di ottimizzazione ha varianza massima quando:

$$\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 = \lambda_1$$

dove λ_1 è il massimo autovalore di \mathbf{S} e \mathbf{u}_1 è l'autovettore associato a λ_1 .

La direzione che massimizza la varianza dei dati proiettati è l'autovettore associato al massimo autovalore di \mathbf{S} . Analogamente, le direzioni successive sono date dagli autovettori associati agli autovalori successivi, ordinati in senso decrescente.

16.2.6 Costruzione e proiezione delle componenti principali

Si può dimostrare che le prime M componenti principali possono essere trovate scegliendo i primi M autovettori di \mathbf{S} della matrice di covarianza S ordinati in base ai loro autovalori decrescenti, definiti dalla matrice W :

$$W = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_M \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1D} \\ u_{21} & u_{22} & \cdots & u_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ u_{M1} & u_{M2} & \cdots & u_{MD} \end{bmatrix}$$

dove $u_i = [u_{i1}, u_{i2}, \dots, u_{iD}]$ è l'autovettore associato all'autovalore λ_i . Da questo proiettiamo i dati originali della matrice X nella nuova base delle componenti principali:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{pmatrix}$$

tramite la seguente operazione di proiezione:

$$Z = XW^T$$

dove Z è la matrice dei dati proiettati nello spazio delle componenti principali, di dimensione $N \times M$. Chiamiamo Z una *variabile latente* in quanto rappresenta una nuova rappresentazione dei dati in uno spazio a dimensionalità ridotta.

In termini geometrici, la PCA fa una rotazione del sistema di riferimento originale e proietta i dati sugli assi principali, riducendo la dimensionalità del dataset mantenendo la massima varianza possibile.

16.2.7 Data Whitening: decorrelazione e normalizzazione

Può essere dimostrato che la PCA trasforma i dati in modo tale che le nuove feature risultino non correlate tra loro, ovvero formano una matrice di covarianza diagonale. Questo perché ruotando i dati in modo tale che i nuovi assi sono allineati con le direzioni di massima varianza, si eliminano le correlazioni tra le feature originali. Questo processo si chiama **data whitening**.

16.2.8 Scelta del numero di componenti principali

La scelta del numero di componenti principali M da mantenere dipende dall'obiettivo dell'analisi e dalla quantità di varianza che si desidera preservare. Una strategia comune è quella di scegliere M in modo tale che la somma degli autovalori associati alle prime M componenti principali rappresenti una percentuale significativa (ad esempio, il 95%) della varianza totale del dataset.

16.2.9 Interpretazione delle prime componenti principali

Una volta trasformati i dati nello spazio delle componenti principali, si possono analizzare le prime componenti per comprendere quali caratteristiche originali contribuiscono maggiormente alla varianza dei dati. Per fare questo, si utilizza i **pesi delle componenti principali**, ovvero gli autovalori associati a ciascuna componente principale, per valutare l'importanza relativa di ciascuna feature originale nella formazione delle nuove componenti. Questi pesi sono anche chiamati **loadings**.

Analizzando i loadings, è possibile identificare quali feature originali hanno un impatto maggiore sulle prime componenti principali e, di conseguenza, sulla variabilità complessiva dei dati. Per farlo, si può utilizzare un grafico chiamato **load plot** (sezione 17.8), che mostra i pesi delle feature originali per ciascuna componente principale.

Capitolo 17

Grafici

I grafici sono strumenti fondamentali per visualizzare e comprendere i dati. Essi permettono di rappresentare informazioni in modo visivo, facilitando l'interpretazione e l'analisi dei dati stessi.

Quando si crea un grafico, la domanda principale da porsi è "Cosa deve rappresentare questo grafico?". La scelta del tipo di grafico dipende dal tipo di dati che si desidera visualizzare e dal messaggio che si vuole comunicare. Alcune domande comuni possono essere:

- Voglio mostrare la distribuzione di una singola variabile?
- Voglio confrontare più gruppi o categorie?
- Voglio analizzare la relazione tra due variabili?
- Voglio rappresentare dati temporali o sequenziali?

Dopo essersi posti queste domande, si sceglie il tipo di grafico più adatto. Di seguito sono elencati alcuni dei tipi di grafici più comuni e le loro caratteristiche principali.

17.1 Istogramma

Un Istogramma è un tipo di grafico che rappresenta la distribuzione di un insieme di dati suddividendoli in intervalli (o "bin") e mostrando la frequenza di dati in ciascun intervallo mediante barre verticali. Gli istogrammi sono utilizzati solo per variabili quantitative (numeriche, misurabili) come per esempio l'altezza delle persone, l'età; non avrebbe infatti senso fare un istogramma per una variabile puramente categoriale come il colore degli occhi o il genere poiché l'istogramma rappresenta frequenze lungo un asse numerico continuo.

Esempio. Consideriamo un insieme di dati che rappresentano le altezze (in cm) di un gruppo di persone. Creiamo un istogramma per visualizzare la distribuzione delle altezze, come mostrato nella figura 17.1. Dalla figura si può evincere facilmente che la maggior parte delle altezze si concentra intorno alla media (circa 170 cm), con una distribuzione che sembra approssimare una curva normale (a campana). Inoltre, l'istogramma mostra che ci sono poche persone con altezze estreme (molto basse o molto alte).

Dalla sotto-sezione 5.3.1 sappiamo che spesso gli istogrammi sono utilizzati per stimare la funzione di densità di probabilità (PDF) di una variabile casuale continua. In questo caso, l'istogramma fornisce una rappresentazione visiva della distribuzione dei dati, permettendo di

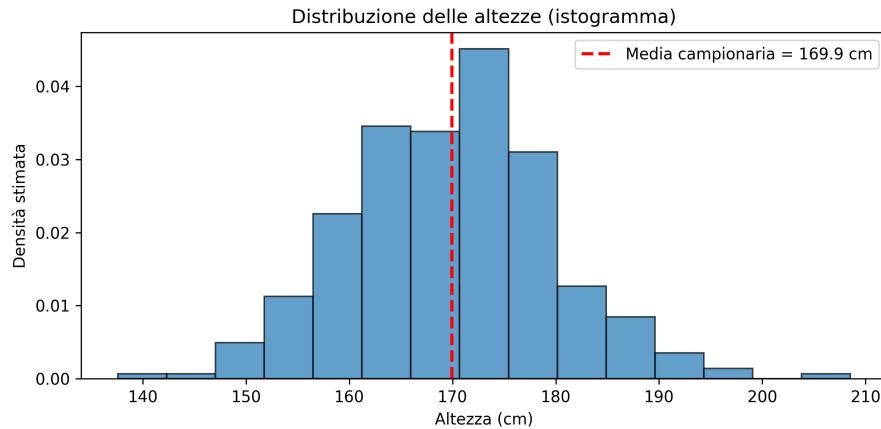


Figura 17.1: Distribuzione simulata delle altezze. Le barre mostrano la frequenza relativa (densità) dei valori osservati; la linea rossa tratteggiata indica la media campionaria.

osservare la forma della distribuzione, la presenza di picchi (modi), la simmetria o asimmetria, e altre caratteristiche importanti.

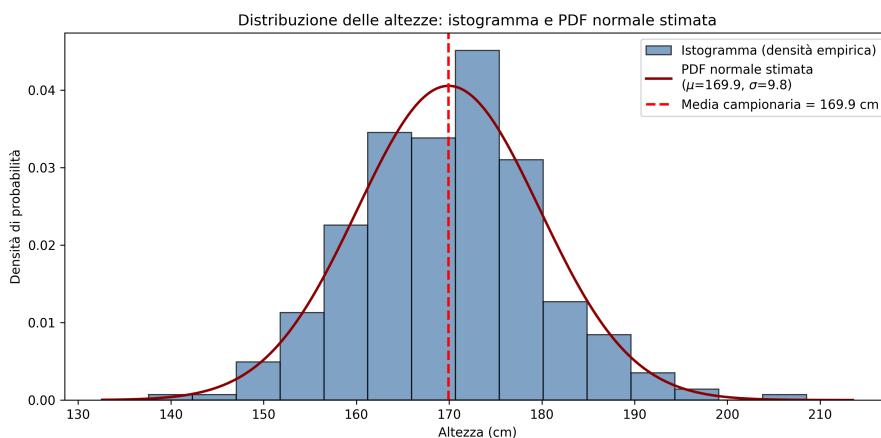


Figura 17.2: Istogramma delle altezze con sovrapposta la densità normale stimata dai dati. Le barre mostrano la distribuzione empirica, la curva rappresenta la PDF gaussiana stimata (media e deviazione standard campionarie) e la linea tratteggiata indica la media del campione.

Da questa figura si può osservare come grazie a un istogramma normalizzato sia possibile confrontare la distribuzione empirica dei dati con una distribuzione teorica (in questo caso, la distribuzione normale). Questo confronto può aiutare a valutare quanto bene i dati seguono una determinata distribuzione teorica, fornendo informazioni utili per l'analisi statistica e la modellizzazione.

17.2 Grafico a barre

Un grafico a barre è un tipo di grafico utilizzato per rappresentare dati categoriali mediante barre rettangolari. Ogni barra rappresenta una categoria e la lunghezza o l'altezza della barra è proporzionale alla frequenza o alla quantità associata a quella categoria.

I grafici a barre sono principalmente utilizzati per confrontare valori tra diverse categorie. Questi valori possono essere frequenze assolute, frequenze relative (percentuali) o altre misure quantitative.

Esempio. Consideriamo un insieme di dati che rappresentano il numero di studenti iscritti a diversi corsi universitari in un semestre. Creiamo un grafico a barre per visualizzare il numero di studenti in ciascun corso, come mostrato nella figura 17.3.

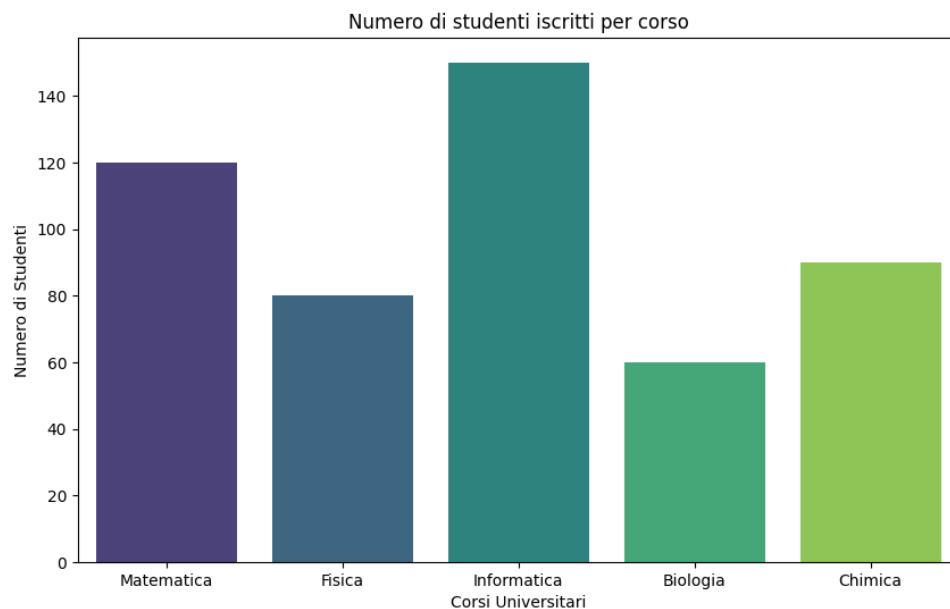


Figura 17.3: Numero di studenti iscritti a diversi corsi universitari in un semestre. Le barre rappresentano il numero di studenti per ciascun corso.

Dalla figura si può facilmente confrontare il numero di studenti iscritti a ciascun corso. Ad esempio, si nota che il corso di "Informatica" ha il maggior numero di iscritti, mentre il corso di "Biologia" ha il minor numero di iscritti. Questo tipo di grafico è utile per identificare rapidamente le differenze tra le categorie e per prendere decisioni basate sui dati.

17.3 Boxplot

Un boxplot è un tipo di grafico utilizzato per rappresentare la distribuzione di un insieme di dati quantitativi attraverso cinque statistiche riassuntive: il minimo, il primo quartile (Q1), la mediana (Q2), il terzo quartile (Q3) e il massimo. I boxplot sono utili per visualizzare la dispersione, la simmetria e la presenza di valori anomali (outlier) nei dati. È particolarmente utile perché consente di individuare facilmente, una volta imparato a leggerlo, caratteristiche importanti della distribuzione dei dati.

Esempio. Consideriamo un insieme di dati che rappresentano i punteggi ottenuti da studenti in un esame. Creiamo un boxplot per visualizzare la distribuzione dei punteggi, come mostrato nella figura 17.4.

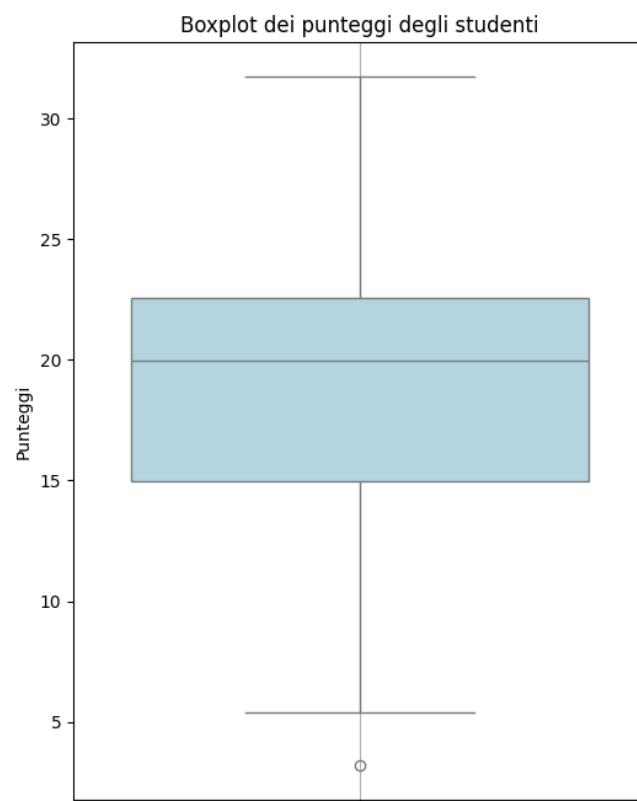


Figura 17.4: Boxplot dei punteggi ottenuti dagli studenti in un esame. Il boxplot mostra la mediana, i quartili, i valori minimi e massimi, e gli outlier.

Dalla figura si può osservare che la mediana dei punteggi è intorno a 20, quindi significa che il 50% degli studenti ha ottenuto un punteggio inferiore a 20 e il restante 50% ha ottenuto un punteggio superiore a 20. Inoltre, il boxplot mostra che la maggior parte dei punteggi si concentra tra il primo quartile (circa 15) e il terzo quartile (circa 23). Da questo si potrebbe calcolare l'intervallo interquartile $IQR = 23 - 15 = 8$. Si noti che il minimo è circa 6, il massimo è 31 (30L) mentre è presente un outlier sotto il valore di 5 (indicato con un cerchio). Questo outlier indica che c'è uno studente che ha ottenuto un punteggio significativamente più basso rispetto agli altri.

17.4 Grafici di dispersione

Un grafico di dispersione (scatter plot) è un tipo di grafico utilizzato per visualizzare la relazione tra due variabili quantitative. In un grafico di dispersione, ogni punto rappresenta un'osservazione del dataset, con la posizione orizzontale (asse x) che rappresenta il valore di una variabile e la posizione verticale (asse y) che rappresenta il valore dell'altra variabile. I grafici di dispersione sono utili per identificare correlazioni, tendenze e modelli nei dati durante l'analisi multivariata.

Esempio. Ipotizziamo di dover analizzare la relazione tra il numero di ore di studio settimanali e i punteggi ottenuti dagli studenti (200 in totale) in un esame (da 0 a 100). Creiamo un grafico di dispersione per visualizzare questa relazione, come mostrato nella figura 17.5.

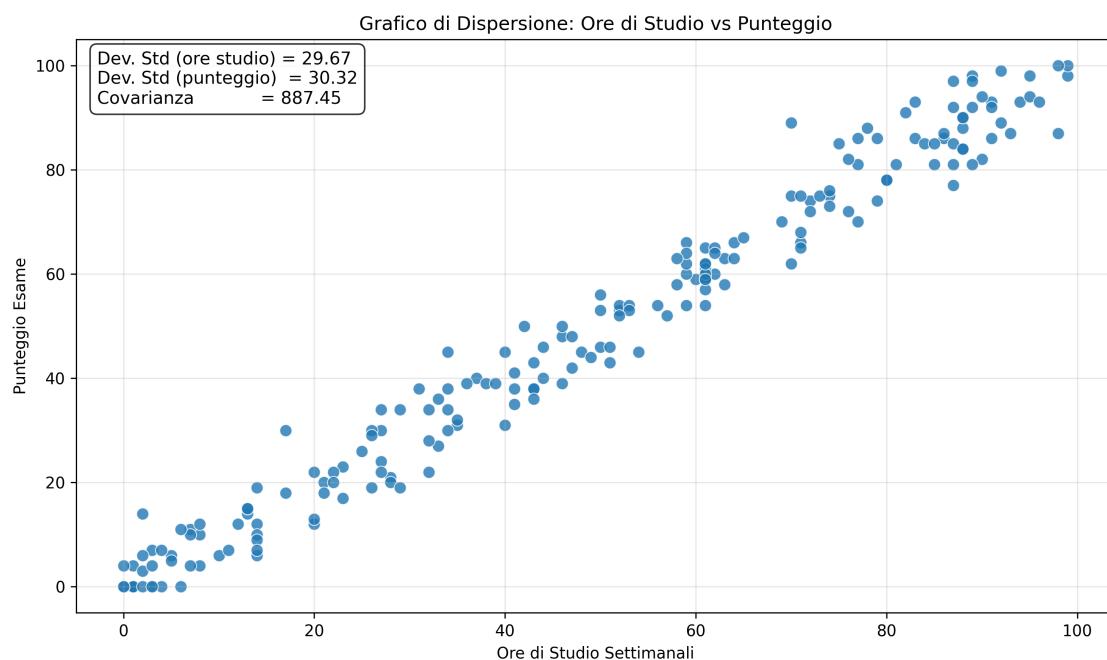


Figura 17.5: Grafico di dispersione che mostra la relazione tra il numero di ore di studio settimanali e i punteggi ottenuti dagli studenti in un esame.

Dalla figura si evince molto facilmente, che i punteggi tendono ad aumentare con l'aumentare delle ore di studio settimanali. Questo suggerisce una correlazione positiva tra le due variabili, indicando che gli studenti che dedicano più tempo allo studio tendono a ottenere punteggi più

alti negli esami. Inoltre, si possono notare alcune variazioni nei punteggi per un dato numero di ore di studio, il che indica che altri fattori potrebbero influenzare i risultati degli studenti.

Ci possiamo facilmente convincere di questa cosa calcolando il coefficiente di correlazione di Pearson tra le due variabili:

$$p = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \approx 0.99$$

Da questo calcolo confermiamo la forte correlazione positiva tra le due variabili, come suggerito dal grafico di dispersione.

17.4.1 Matrice di scatter plot

Una matrice di scatter plot (o scatter matrix) è una rappresentazione grafica che mostra i grafici di dispersione per tutte le coppie di variabili in un dataset multivariato. Ogni cella della matrice contiene un grafico di dispersione che rappresenta la relazione tra due variabili specifiche, mentre le diagonali possono contenere istogrammi o grafici a densità per ciascuna variabile.

Esempio. Ipotizziamo di continuare l'esempio precedente: abbiamo un dataset con più variabili relative agli studenti, come il numero di ore di studio settimanali, i punteggi ottenuti negli esami, il numero di assenze, il tempo dedicato ad attività extracurricolari e il livello di stress. Creiamo una matrice di scatter plot per visualizzare le relazioni tra tutte queste variabili, come mostrato nella figura 17.6.

Dalla figura si possono osservare diverse relazioni tra le variabili. Ad esempio, si nota una correlazione positiva tra il numero di ore di studio settimanali e i punteggi ottenuti negli esami, come già discusso in precedenza. Inoltre, si può osservare una correlazione negativa tra il numero di assenze e i punteggi degli esami, suggerendo che gli studenti che frequentano regolarmente le lezioni tendono a ottenere risultati migliori. La matrice di scatter plot consente di identificare rapidamente queste relazioni e di individuare eventuali pattern o tendenze nei dati.

17.4.2 Scatter plot con intervalli di confidenza

Uno scatter plot con intervalli di confidenza è un tipo di grafico di dispersione che include linee o bande che rappresentano gli intervalli di confidenza attorno a una stima della relazione tra le due variabili. Questi intervalli forniscono una misura della precisione della stima e aiutano a visualizzare l'incertezza associata alla relazione tra le variabili. Sono molto utili per valutare la significatività statistica della relazione osservata, in particolare quando si esegue una regressione lineare.

Esempio. Consideriamo nuovamente l'esempio del numero di ore di studio settimanali e dei punteggi ottenuti dagli studenti in un esame. Possiamo stimare una retta di regressione lineare che descriva, in media, come varia il punteggio al variare delle ore di studio e rappresentare graficamente sia tale retta sia l'incertezza sulla stima. In figura 17.7 è riportato uno scatter plot in cui ogni punto rappresenta uno studente, la linea centrale rappresenta la retta di regressione stimata e la banda ombreggiata attorno alla linea corrisponde all'intervalllo di confidenza (unicamente per fini didattici l'intervalllo sarà al 99% per mostrare in modo netto il grafico) della media condizionale del punteggio dato un certo numero di ore di studio.

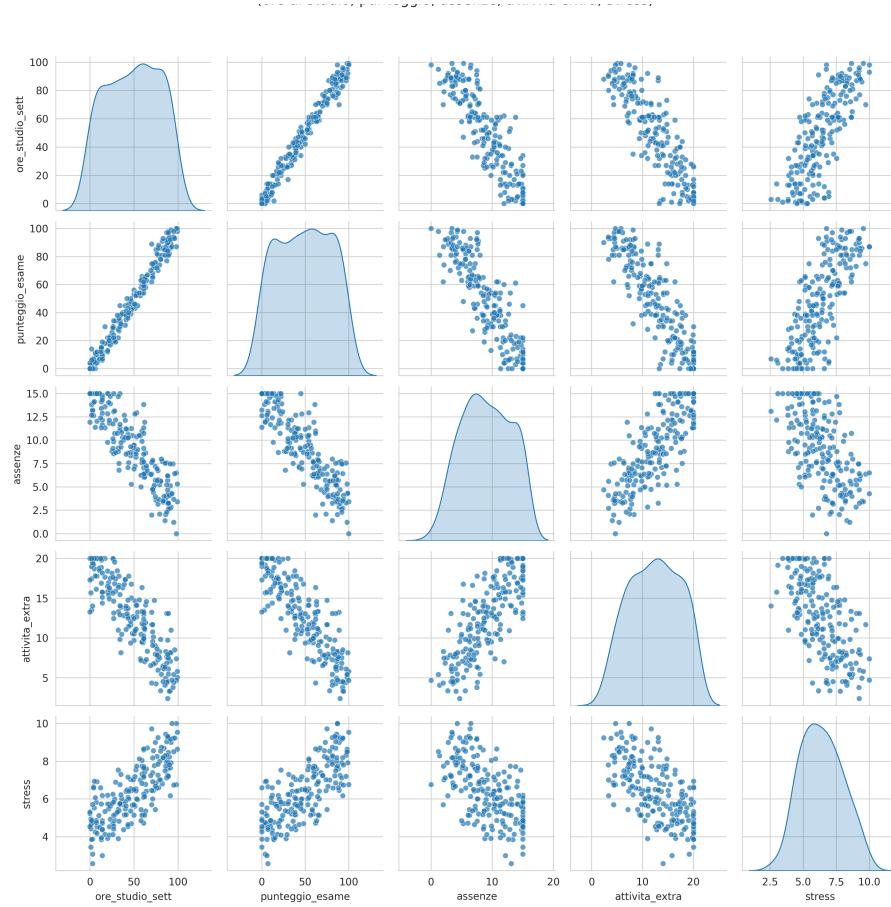


Figura 17.6: Matrice di scatter plot che mostra le relazioni tra diverse variabili relative agli studenti. Ogni cella contiene un grafico di dispersione per una coppia di variabili, mentre le diagonali mostrano istogrammi delle singole variabili.

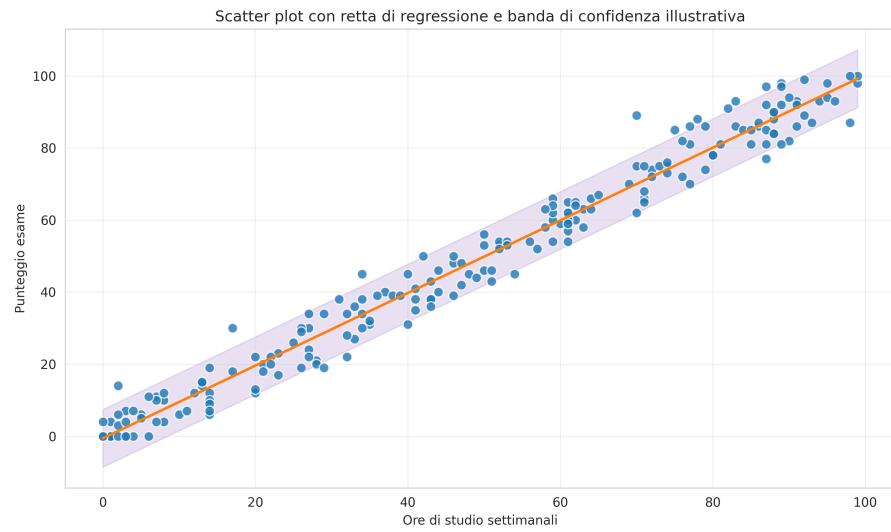


Figura 17.7: Scatter plot che mostra la relazione tra ore di studio settimanali e punteggio d'esame, con sovrapposta la retta di regressione lineare stimata e la banda di intervallo di confidenza al 99%.

Dalla figura si osserva che, in media, all'aumentare delle ore di studio il punteggio tende ad aumentare in modo quasi lineare. L'intervallo di confidenza è più stretto nella zona in cui si concentrano la maggior parte delle osservazioni (valori di ore di studio più frequenti) e tende ad allargarsi agli estremi, dove i dati sono più scarsi. Questo tipo di rappresentazione permette non solo di vedere la tendenza generale (la retta di regressione), ma anche di valutare l'incertezza associata alla stima di tale tendenza.

17.5 Hexbin plot

Un hexbin plot è un tipo di grafico utilizzato per visualizzare la densità di punti in un grafico di dispersione bidimensionale. Invece di rappresentare ogni punto individualmente, l'hexbin plot suddivide l'area del grafico in celle esagonali (hexagons) e conta il numero di punti che cadono in ciascuna cella. La densità dei punti in ogni cella viene quindi rappresentata mediante una scala di colori, con colori più scuri che indicano una maggiore densità di punti.

Viene considerato l'istogramma bidimensionale con celle esagonali, ed è particolarmente utile quando si lavora con grandi quantità di dati, poiché riduce il sovraffollamento e rende più facile identificare le aree di alta densità.

Esempio. Consideriamo un insieme di dati che rappresentano le altezze e i pesi di un gruppo di persone. Creiamo un hexbin plot per visualizzare la densità dei punti in relazione a queste due variabili, come mostrato nella figura 17.8.

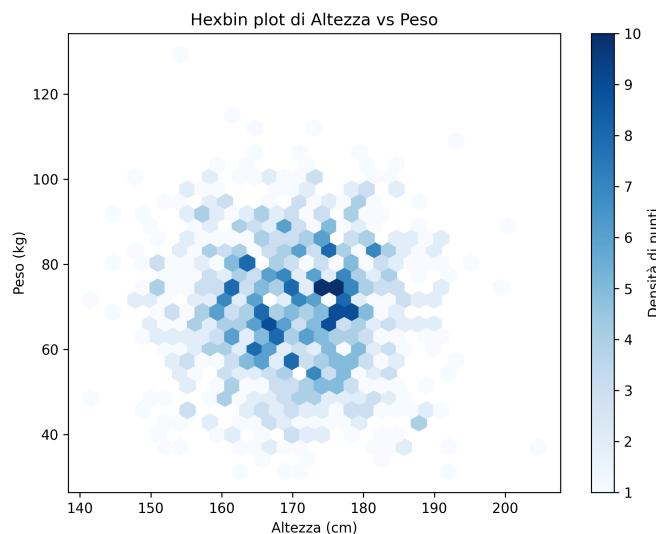


Figura 17.8: Hexbin plot che mostra la densità delle osservazioni in relazione all'altezza e al peso di un gruppo di persone. Le celle esagonali rappresentano la densità dei punti, con colori più scuri che indicano una maggiore densità.

Dalla figura si può osservare che la maggior parte delle persone si concentra in un'area specifica del grafico, indicando una relazione tra altezza e peso. Le celle esagonali più scure rappresentano le combinazioni di altezza e peso più comuni nel dataset, mentre le celle più chiare indicano

combinazioni meno frequenti. Anche questo grafico serve nell’analisi multivariata ed è utile per identificare pattern e tendenze nei dati.

17.6 Grafici di densità e di contorno

Un grafico di densità (density plot) è una rappresentazione grafica della distribuzione di una variabile continua. A differenza di un istogramma, che suddivide i dati in intervalli discreti, un grafico di densità utilizza una funzione di densità stimata per mostrare la distribuzione dei dati in modo più fluido e continuo. Questo tipo di grafico è utile per visualizzare la forma della distribuzione, identificare picchi (modi) e confrontare più distribuzioni.

Un grafico di contorno (contour plot) è una rappresentazione grafica che mostra le linee di livello (contorni) di una funzione di due variabili. In un grafico di contorno, le linee collegano punti con lo stesso valore della funzione, permettendo di visualizzare la topografia della funzione in uno spazio bidimensionale. I grafici di contorno sono utili per analizzare la relazione tra due variabili e per identificare aree di interesse, come massimi, minimi e punti di sella.

Esempio. Ipotizziamo di guardare la distribuzione delle altezze e del peso di un gruppo di persone. Creiamo un grafico di densità bidimensionale e un grafico di contorno per visualizzare la distribuzione congiunta di queste due variabili, come mostrato nella figura 17.9.

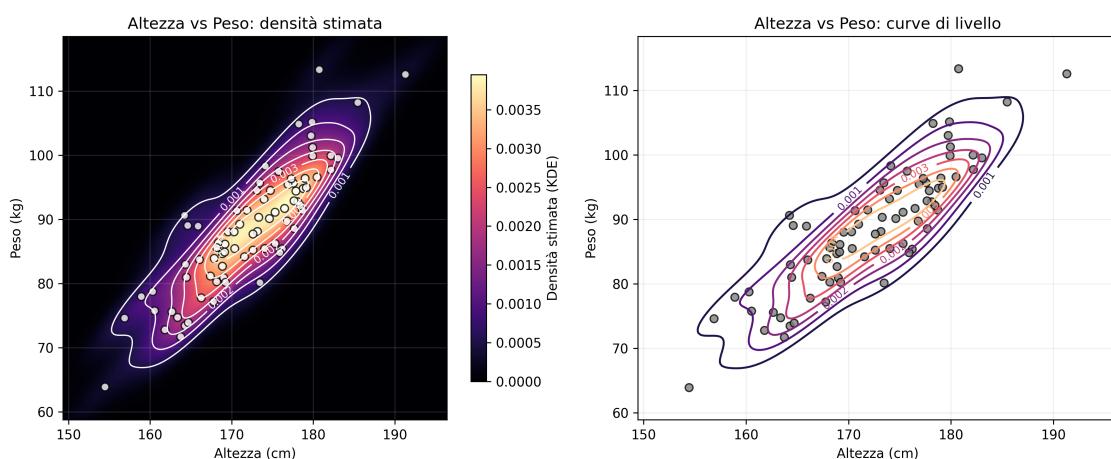


Figura 17.9: Grafico di densità bidimensionale e grafico di contorno che mostrano la distribuzione congiunta di altezza e peso in un gruppo di persone. Le aree più scure nel grafico di densità indicano una maggiore concentrazione di punti, mentre le linee nel grafico di contorno rappresentano i livelli di densità.

Come si evince dalla figura, il grafico di densità mostra che la maggior parte delle persone si concentra in un’area specifica del grafico, indicando una relazione tra altezza e peso. Le aree più scure rappresentano le combinazioni di altezza e peso più comuni nel dataset. Il grafico di contorno, invece, fornisce una rappresentazione visiva delle linee di livello della densità, permettendo di identificare facilmente le aree di alta e bassa densità.

Questi grafici sono ulteriori strumenti utili nell’analisi multivariata per comprendere meglio la distribuzione e la relazione tra due variabili quantitative.

17.7 Heatmaps

Dalle matrici di correlazione è possibile creare delle heatmaps (mappe di calore) per visualizzare le correlazioni tra più variabili in un dataset. Una heatmap utilizza una scala di colori per rappresentare i valori di una matrice, dove ogni cella della matrice corrisponde a una coppia di variabili e il colore della cella indica la forza e la direzione della correlazione tra quelle variabili.

Esempio. Riprendendo l'esempio della matrice di scatter plot nella sotto-sezione 17.4.1, possiamo calcolare la matrice di correlazione tra le variabili relative agli studenti e creare una heatmap per visualizzare queste correlazioni, come mostrato nella figura 17.10.

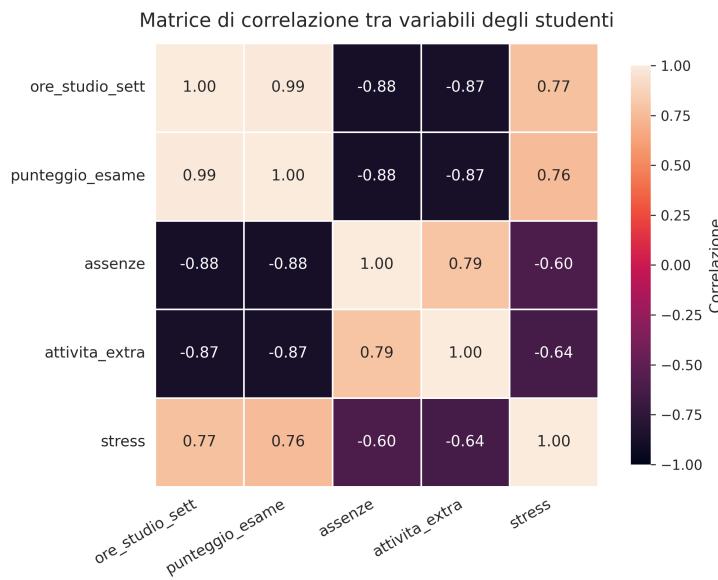


Figura 17.10: Heatmap che mostra la matrice di correlazione tra diverse variabili relative agli studenti. I colori indicano la forza e la direzione della correlazione, con colori più scuri che rappresentano correlazioni più forti.

Dalla figura si può osservare facilmente la forza e la direzione delle correlazioni tra le variabili. Ad esempio, si nota una forte correlazione positiva tra il numero di ore di studio settimanali e i punteggi ottenuti negli esami, come già discusso in precedenza. Inoltre, si può osservare una correlazione negativa tra il numero di assenze e i punteggi degli esami. La heatmap consente di identificare rapidamente queste relazioni e di individuare eventuali pattern o tendenze nei dati, facilitando l'analisi multivariata.

17.8 Load plots

Un load plot è un tipo di grafico utilizzato per visualizzare i carichi (loadings) delle variabili originali su componenti principali o fattori in un'analisi di componenti principali (PCA) o in un'analisi fattoriale. I loadings rappresentano l'importanza relativa di ciascuna variabile nella formazione delle componenti principali o dei fattori. Un load plot mostra questi carichi in modo visivo, permettendo di identificare quali variabili contribuiscono maggiormente a ciascuna componente o fattore.

In un loadplot:

- Le variabili che si "raggruppano" insieme in una direzione simile indicano una forte correlazione tra di esse.
- Le variabili che si trovano lontano dall'origine del grafico hanno un maggiore contributo alla componente principale o al fattore rappresentato.
- Le variabili che si trovano in direzioni opposte indicano una correlazione negativa tra di esse.

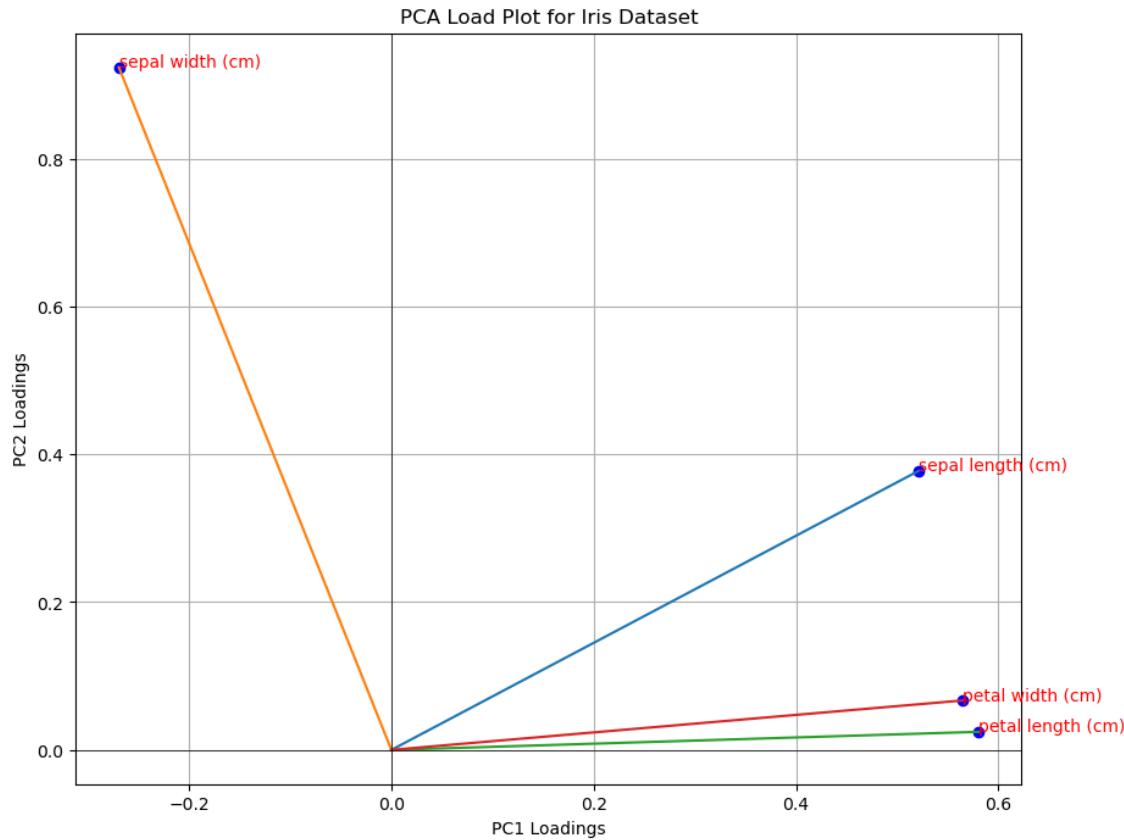


Figura 17.11: Load plot della PCA per il dataset Iris. Le frecce rappresentano i loadings delle feature originali sulle prime due componenti principali, indicando il contributo e l'orientamento di ciascuna variabile rispetto a PC1 e PC2.

