Analisi dei test effettuati

Test di Stefano Biddau su F-MNIST con K-Nearest Neighbors

Cenni teorici K-Nearest Neighbors

DEFINIZIONE:

L'algoritmo k-nearest neighbors noto anche come KNN o k-NN è un classificatore di apprendimento supervisionato non parametrico che utilizza la prossimità per effettuare classificazioni o previsioni sul raggruppamento di un singolo punto di dati. Sebbene possa essere utilizzato per problemi di classificazione o regressione viene principalmente utilizzato per le problematiche di prima tipologia, basandosi sul presupposto che punti simili possono essere trovati l'uni vicino all'altro.

FUNZIONAMENTO

Per problemi di classificazione un' etichetta di classe viene assegnata sulla base di un voto a maggioranza, ad esempio viene utilizzata l'etichetta più frequente rappresentata attorno ad un determinato punto di dati. Ad esempio se ci sono solo due categorie la scelta ricadrà sull'etichetta che per maggioranza supererà il 50%. Quando ci sono più classi è logico pensare che la soglia si abbassi.

PRINCIPALI IPER-PARAMETRI DEL MODELLO

Vediamo quali sono (per la mia esperienza) i principali parametri o iper-parametri di un modello KNN, prendendo in considerazione l'implementazione della libreria scikit-learn in Python.

Per prima cosa si considera la **metrica di distanza** che aiutano a formare confini decisionali tra il punto di interrogazione e gli altri punti dati. Noi consideriamo diverse metriche per i nostri test:
- **euclidea**: misura una linea retta tra il punto di query e l'altro punto che si sta misurando;
- **manhattan**: misura il valore assoluto tra due punti;
- **minkowski**: questa misura della distanza è la forma generalizzata delle metriche di distanza euclidea e di manhattan;

Altro parametro molto importante che incide sulle prestazioni del classificatore è il **valore k** che è associato al numero di vicini che verranno controllati per determinare la classificazione di un punto specifico di query. Ad esempio se k = 1, l'istanza verrà assegnata alla stessa classe del suo singolo vicino più vicino. Definire k può essere un atto di bilanciamento in quanto valori diversi possono portare ad overfitting o underfitting. La determinazione di questo parametro dipende solo ed esclusivamente dai dati in input del dataset. Per convenzione si utilizzano principalmente dei valori dispari per evitare pareggi nella classificazione.

Infine l'ultimo parametro preso in considerazione è il **peso (weights)**, esso è associato alla funzione di peso utilizzata nella previsione dei possibili valori.
Può essere di due tipi:
- **uniforme**: tutti i punti in ogni vicinato sono ponderati allo stesso modo;
- **distanza**: in questo caso i vicini più vicini avranno un'influenza maggiore dei vicini più lontani;

PRO E CONTRO DELLA K-NEAREST NEIGHBORS

PRO

- Facile da implementare data la sua semplicità e allo stesso tempo forte accuratezza;
- Si adatta facilmente quando vengono aggiunti nuovi campioni
- Pochi iper parametri rispetto ad altri algoritmi di Machine Learning

CONTRO

- Non ha una buona scalabilità essendo un algoritmo pigro, occupa più memoria e spazio di Storage rispetto agli altri classificatori;
- Non funziona molto bene con input di dati ad alta dimensione;

TEST EFFETTUATI

F-MNIST

Eseguendo lo script per l'ottimizzazione degli iper parametri salta all'occhio che nelle prime 10 posizioni delle simulazioni effettuate, con i migliori responsi in termini di accuratezza troviamo principalmente il numero di vicini settato con questi valori: (3, 5, 7, 9, 11, 13)
Testiamo anche con il numero di vicini limite scelto che è 21 e poi testiamo con un numero di vicini che sfora il range prestabilito ossia 25.

1° Blocco di test (Aumento vicini, w: uniform, m: minkowski)

Indice test	Numero Vicini	Weights	Metric	Acc Train	Acc Test	Differenza MSE	Under-fit
1	3	Uniform	Minkowski	91,99%	85,38%	0,88	NO
2	**5**	**Uniform**	**Minkowski**	**89,98%**	**85,52%**	**0,59**	**NO**
3	7	Uniform	Minkowski	88,81%	85,41%	0,41	NO
4	9	Uniform	Minkowski	88,08%	85,18%	0,39	NO
5	11	Uniform	Minkowski	87,52%	84,93%	0,32	SI
6	13	Uniform	Minkowski	87,12%	84,68%	0,34	SI
7	21	Uniform	Minkowski	85,88%	84,08%	0,25	SI
8	25	Uniform	Minkowski	85,49%	83,76%	0,26	SI

Da questi primi 8 test si evince che all'aumentare dei vicini il modello perde in accuratezza sia sul set di test ma sopratutto sul set di train, il cui valore dopo il test 4 scende vertiginosamente. Infatti possiamo notare di come la differenza di MSE diventi sempre più piccola. Nonostante ciò non ci troviamo davanti ad un caso di overfitting perché come possa notare la differenza di MSE non sale con l'aumentare dei vicini tuttavia l'errore sul set di trai e sul set di test aumenta sempre più quindi è bene parlare di underfitting. I migliori risultati si riscontrano con il **test 2 col la percentuale di accuratezza sul set di test più alta.**

Adesso proviamo a cambiare metrica quindi passiamo da minkowski -> euclidean

2° Blocco di test (Aumento vicini, w: uniform, m: euclidean)

Indice test	Numero Vicini	Weights	Metric	Acc Train	Acc Test	Differenza MSE	Under-fit
1	3	Uniform	Euclidean	91,99%	85,38%	0,88	NO
2	**5**	**Uniform**	Euclidean	**89,98%**	**85,52%**	**0,59**	**NO**
3	7	Uniform	Euclidean	88,81%	85,41%	0,41	NO
4	9	Uniform	Euclidean	88,08%	85,18%	0,39	NO
5	11	Uniform	Euclidean	87,52%	84,93%	0,32	SI
6	13	Uniform	Euclidean	87,12%	84,68%	0,34	SI
7	21	Uniform	Euclidean	85,88%	84,08%	0,25	SI
8	25	Uniform	Euclidean	85,49%	83,76%	0,26	SI

Per questi altri 8 test otteniamo **risultati identici al Blocco 1.** Si evince che all'aumentare dei vicini il modello perde in accuratezza sia sul set di test ma sopratutto sul set di train, il cui valore dopo il test 4 scende vertiginosamente. Infatti possiamo notare di come la differenza di MSE diventi sempre più piccola. Nonostante ciò non ci troviamo davanti ad un caso di overfitting perché come possa notare la differenza di MSE non sale con l'aumentare dei vicini tuttavia l'errore sul set di trai e sul set di test aumenta sempre più quindi è bene parlare di underfitting. I migliori risultati si riscontrano con il **test 2 col la percentuale di accuratezza sul set di test più alta.**

Adesso proviamo a cambiare metrica quindi passiamo da euclidean -> manhattan e vediamo se otteniamo ancora gli stessi risultati

3° Blocco di test (Aumento vicini, w: uniform, m: manhattan)

Indice test	Numero Vicini	Weights	Metric	Acc Train	Acc Test	Differenza MSE	Under-fit
1	3	Uniform	Manhattan	92,26%	85,76%	0,92	NO
2	5	Uniform	Manhattan	90,40%	86,20%	0,59	NO
3	**7**	**Uniform**	**Manhattan**	**89,44%**	**86,30%**	**0,45**	**NO**
4	9	Uniform	Manhattan	88,69%	86,03%	0,39	NO
5	11	Uniform	Manhattan	88,16%	85,89%	0,33	SI
6	13	Uniform	Manhattan	87,75%	85,60%	0,28	SI
7	21	Uniform	Manhattan	86,54%	84,99%	0,20	SI
8	25	Uniform	Manhattan	86,22%	84,69%	0,21	SI

Per questi 8 test effettuati notiamo un comportamento analogo ai due blocchi precedenti, dopo aver raggiunto i migliori risultati con il **test 3** in termini di accuratezza sul set di test, all'aumentare dei vicini abbiamo una vertiginosa discesa in termini di punti percentuali sia sul set di train che sul set di test. Mentre all'aumentare dei vicini notiamo che la differenza di MSE si assottiglia sempre più. Anche in questo caso ci troviamo davanti ad un fenomeno di underfitting.

Diversamente dai due blocchi precedenti **la metrica manhattan fa lievitare considerevolmente il tempo di addestramento del modello** che passa da una manciata di minuti, a più di 30 minuti di addestramento, tuttavia la percentuale di accuratezza sia sul set di test che sul set di train è leggermente più alta rispetto agli altri due blocchi.

Adesso ripetiamo i test effettuati cambiando il peso da uniform -> distance

4° Blocco di test (Aumento vicini, w: distance, m: minkowski)

Indice test	Numero Vicini	Weights	Metric	Acc Train	Acc Test	Differenza MSE	Under-fit
1	3	Distance	Minkowski	100%	85,59%	1,94	-
2	5	Distance	Minkowski	100%	85,76%	1,94	-
3	7	Distance	Minkowski	100%	85,42%	1,95	-
4	9	Distance	Minkowski	100%	85,29%	2,00	-
5	11	Distance	Minkowski	100%	85,08%	1,99	-
6	13	Distance	Minkowski	100%	84,61%	2,00	-
7	21	Distance	Minkowski	100%	84,24%	2,11	-

Da questi altri 7 test si evince che il cambio di peso influisce nettamente sul modello rendendolo in utile in quanto l'accuratezza sul set di train raggiunge il 100% che è sintomo di un modello inefficiente che non predice ma impara meccanicamente.

5° Blocco di test (Aumento vicini, w: distance, m: euclidean)

Indice test	Numero Vicini	Weights	Metric	Acc Train	Acc Test	Differenza MSE	Under-fit
1	3	Distance	Euclidean	100%	85,59%	1,94	-
2	5	Distance	Euclidean	100%	85,76%	1,94	-
3	7	Distance	Euclidean	100%	85,42%	1,95	-
4	9	Distance	Euclidean	100%	85,29%	2,00	-
5	11	Distance	Euclidean	100%	85,08%	1,99	-
6	13	Distance	Euclidean	100%	84,61%	2,00	-
7	21	Distance	Euclidean	100%	84,24%	2,11	-

I risultati ottenuti da questo blocco di 7 test evidenziano il rapporto che abbiamo riscontrato con i risultati tra il **Blocco 1 e il Blocco 2** ossia che le metriche euclidean e minkowsky fanno ottenere gli stessi risultati in termini di percentuale sia sul set di test che sul set di train. Chiaramente se il Blocco 4 ci mostra un modello che non apprende ma impara meccanicamente, lo stesso risultato si ottiene con questo blocco di test.

6° Blocco di test (Aumento vicini, w: distance, m: manhattan)

Indice test	Numero Vicini	Weights	Metric	Acc Train	Acc Test	Differenza MSE	Under-fit
1	3	Distance	Manhattan	100%	85,97%	1,95	-
2	5	Distance	Manhattan	100%	86,13%	1,91	-
3	7	Distance	Manhattan	100%	86,14%	1,91	-
4	9	Distance	Manhattan	100%	85,98%	1,95	-
5	11	Distance	Manhattan	100%	85,92%	1,96	-
6	13	Distance	Manhattan	100%	85,71%	1,97	-
7	21	Distance	Manhattan	100%	85,10%	2,0	-

I risultati di questi 7 test mostrano l'andamento dei due blocchi precedenti con peso distanze, ossia un modello che non predice ma offre una classificazione in modo meccanico.

Possiamo stilare una lista di 3 test, quelli con i migliori risultati ottenuti in termini di percentuale di accuratezza sul set di test:

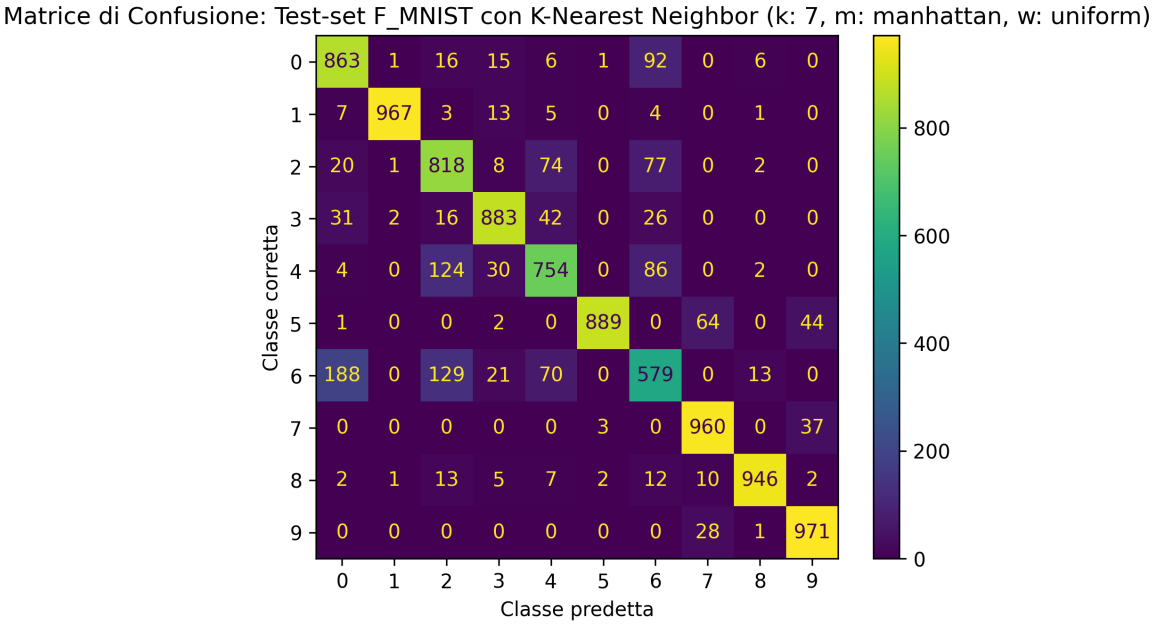
1) Blocco 3 - Test 3: Test del 2022-11-09 02:12:34
2) Blocco 1 - Test 2: Test del 2022-11-08 17:50:52
3) Blocco 2 - Test 2: Test del 2022-11-08 23:02:05

1) **Test del 2022-11-09 02:12:34 F_MNIST con KNN (k: 7, w: uniform, m: manhattan)**

Report di classificazione

	precision	recall	f1-score	support
0	0.77	0.86	0.82	1000
1	0.99	0.97	0.98	1000
2	0.73	0.82	0.77	1000
3	0.90	0.88	0.89	1000
4	0.79	0.75	0.77	1000
5	0.99	0.89	0.94	1000
6	0.66	0.58	0.62	1000
7	0.90	0.96	0.93	1000
8	0.97	0.95	0.96	1000
9	0.92	0.97	0.95	1000
accuracy			0.86	10000
macro avg	0.86	0.86	0.86	10000
weighted avg	0.86	0.86	0.86	10000

Dal report di classificazione notiamo che il nostro classificatore ottiene i migliori risultati classificando la **classe 1-> Pantaloni** e i peggiori risultati di classificazione con la **classe 6 -> Camicia**



La matrice di confusione conferma in parte ciò che viene mostrato nel report di classificazione. La **classe 1** che corrisponde ai **pantaloni** su **1000 campioni** disponibili ne classifica correttamente **967**, che tuttavia non è il miglior risultato ottenibile, visto che la **classe 9** che corrisponde allo

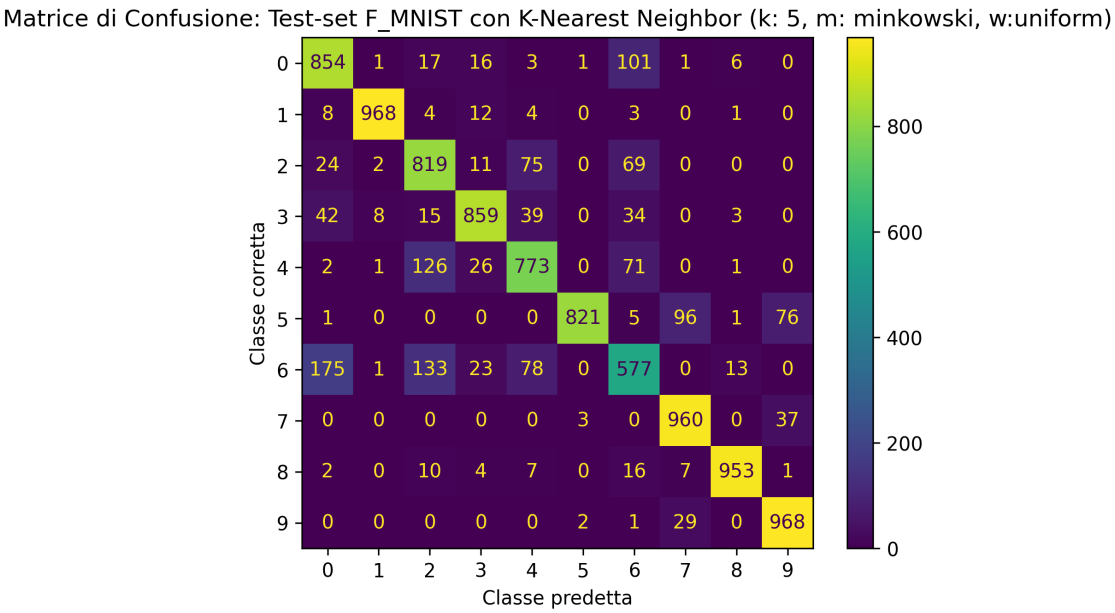
stivaletto sempre su 1000 campioni disponibili ne classifica correttamente **971**. Tuttavia la classe **9** a differenza della classe 6 distribuisce in malo modo gli errori di classificazione tra le altre classi, infatti tutti gli errori di classificazione (bel 28/29) corrispondono alla classe predetta **7** ossia **sneaker**. Al contrario la **classe 6** che corrisponde alla camicia su **1000** campioni disponibili ne classifica correttamente soltanto **579** poco più della metà, i maggiori errori di classificazione avvengono con **t-shirt/top, maglione e cappotto** (rispettivamente in ordine decrescente di campioni errati)

2) **Test del 2022-11-08 17:50:52 F_MNIST con KNN (k: 5, w: uniform, m: minkowski)**

Report di classificazione

	precision	recall	f1-score	support
0	0.77	0.85	0.81	1000
1	0.99	0.97	0.98	1000
2	0.73	0.82	0.77	1000
3	0.90	0.86	0.88	1000
4	0.79	0.77	0.78	1000
5	0.99	0.82	0.90	1000
6	0.66	0.58	0.61	1000
7	0.88	0.96	0.92	1000
8	0.97	0.95	0.96	1000
9	0.89	0.97	0.93	1000
accuracy			0.86	10000
macro avg	0.86	0.86	0.85	10000
weighted avg	0.86	0.86	0.85	10000

Dal report di classificazione notiamo che il nostro classificatore ottiene i migliori risultati classificando la **classe 1-> Pantaloni** e i peggiori risultati di classificazione con la **classe 6 -> Camicia**



La matrice di confusione conferma i dati mostrati dal report di classificazione con la classe 1 -> pantaloni su 1000 campioni disponibili ne classifica correttamente ben 968, valore più alto al pari ancora una volta della classe 9 -> stivaletto, con l'unica differenza che la distribuzione di errore è migliore per la classe 1. Al contrario la **classe 6** che corrisponde alla camicia su **1000** campioni

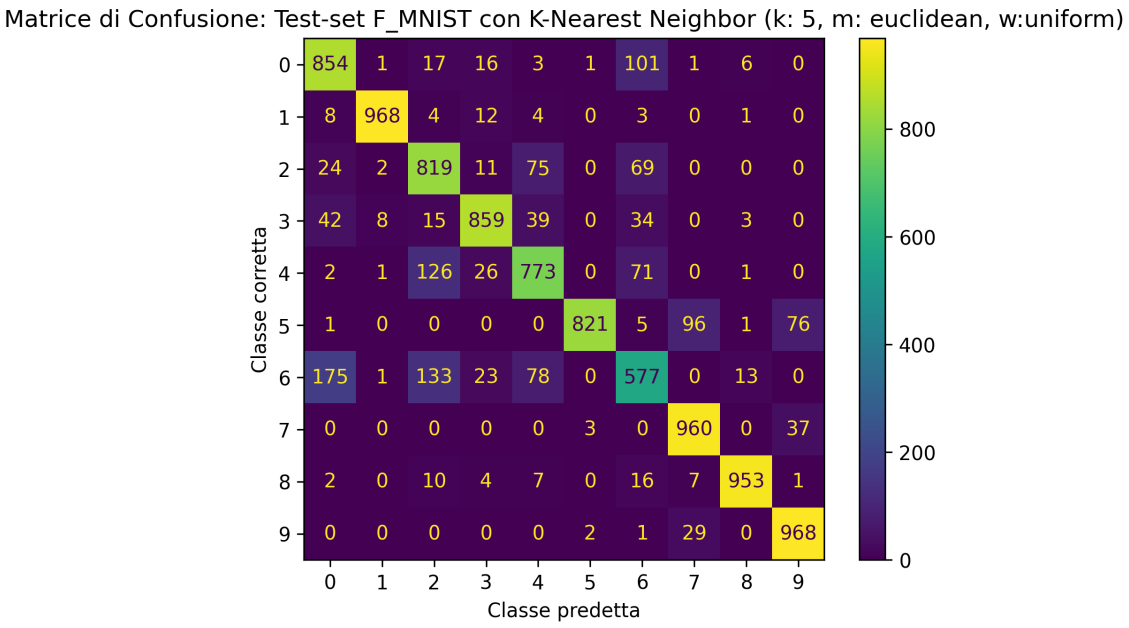
disponibili ne classifica correttamente soltanto **577** poco più della metà, i maggiori errori di classificazione avvengono con **t-shirt/top, maglione e cappotto** (rispettivamente in ordine decrescente di campioni errati)

3) **Test del 2022-11-08 23:02:05 F_MNIST con KNN (k: 5, w: uniform, m: euclidean)**

Report di classificazione

	precision	recall	f1-score	support
0	0.77	0.85	0.81	1000
1	0.99	0.97	0.98	1000
2	0.73	0.82	0.77	1000
3	0.90	0.86	0.88	1000
4	0.79	0.77	0.78	1000
5	0.99	0.82	0.90	1000
6	0.66	0.58	0.61	1000
7	0.88	0.96	0.92	1000
8	0.97	0.95	0.96	1000
9	0.89	0.97	0.93	1000
accuracy			0.86	10000
macro avg	0.86	0.86	0.85	10000
weighted avg	0.86	0.86	0.85	10000

Dal report di classificazione notiamo che il nostro classificatore ottiene i migliori risultati classificando la **classe 1-> Pantaloni** e i peggiori risultati di classificazione con la **classe 6 -> Camicia**



La matrice di confusione conferma i dati mostrati dal report di classificazione con la classe 1 -> pantaloni su 1000 campioni disponibili ne classifica correttamente ben 968, valore più alto al pari

ancora una volta della classe 9 -> stivaletto, con l'unica differenza che la distribuzione di errore è migliore per la classe 1. Al contrario la **classe 6** che corrisponde alla camicia su **1000** campioni disponibili ne classifica correttamente soltanto **577** poco più della metà, i maggiori errori di classificazione avvengono con **t-shirt/top, maglione e cappotto** (rispettivamente in ordine decrescente di campioni errati).