



La maggior parte degli algoritmi di classificazione possiedono diversi parametri che ne regolano il funzionamento ed in particolare l'addestramento.

Esempio:

- Idc , qdc : la stima della matrice di covarianza è regolata dai parametri R ed S .
- $K\text{-NN}$: la classificazione dipende dal valore di K , oltre che dalla funzione 'distanza' utilizzata.

Come nel caso della scelta dell'algoritmo di classificazione, anche la scelta dei parametri ottimali può influire pesantemente sulle prestazioni del nostro sistema.



La scelta dei parametri del classificatore risulta quindi di primaria importanza durante lo sviluppo del proprio sistema di classificazione

La stima di queste grandezze deve essere eseguita in modo sistematico e affidabile sulla base dei dati e delle informazioni a nostra disposizione.

NB. Spesso è necessario individuare anche altre grandezze che influiscono sul processo di classificazione come l'algoritmo di classificazione, il punto di lavoro, il numero di feature, etc.

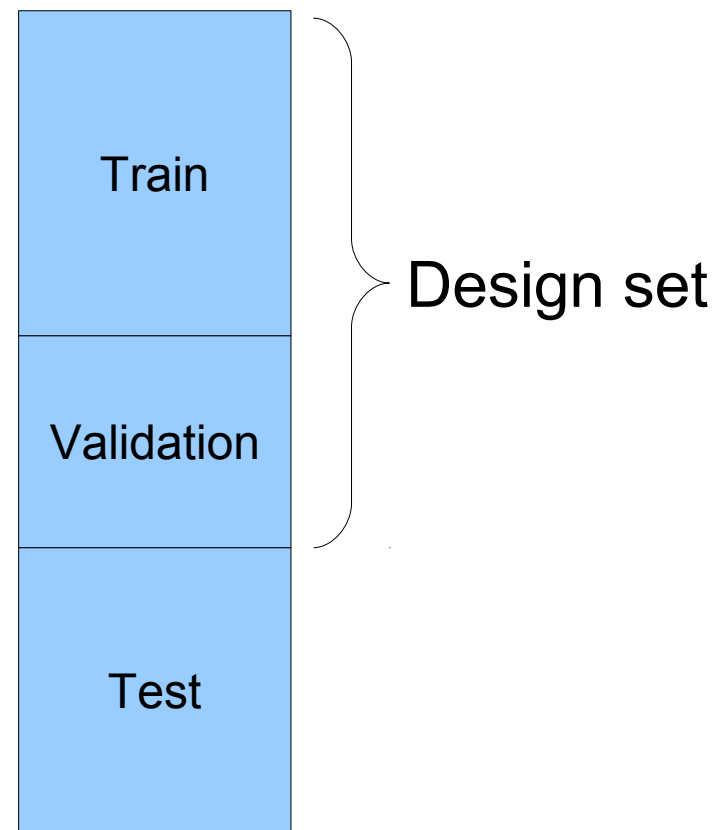


Stima dei parametri



Il metodo più semplice per la stima dei parametri ottimali avviene tipicamente calcolando le prestazioni del sistema su un insieme di dati, chiamato **VALIDATION set**, **NON** utilizzati per l'addestramento (train-set) e per la valutazione delle prestazioni (test-set).

Le prestazioni vengono stimate per diversi insiemi di parametri, successivamente fra tutti gli insiemi verrà selezionato quello che fornisce le prestazioni migliori sul **VALIDATION set**.





Come visto in precedenza la stima delle prestazioni può essere più affidabile calcolando l'errore medio su più ripetizioni e/o tramite cross validation.

Ad esempio nel caso del classificatore K-NN la funzione `PRTTools knnc` è in grado di stimare il valore ottimale di `K` mediante il metodo del “leave-one-out”.

```
[W k] = knnc(ds) ;
```



Esercizio 1



Scelto un dataset, dividetelo in un set di design e un set di test. Dal set di design estraete un training set ed un validation set. Costruite un set di classificatori k-NN con diversi valori di K.

Valutate l'errore sul train (senza leave-one-out), test e validation.

- Tabellate i dati e fate un grafico dell'errore al variare di K
- Quale sarebbe stato l'errore sul test se avessimo scelto il K migliore basandoci sui risultati di errore sul training set?
- Quale sarebbe stato l'errore sul test se avessimo scelto il K migliore basandoci sui risultati sul VALIDATION set?



Esercizio 1 - Soluzione



```
%Genero train, validation e test set (distinti)
ds_full = gendats([600 400],2, 1.5);
[ds_design, ds_tst] = gendat(ds_full, 0.5);
[ds_trn, ds_val]= gendat(ds_design, 0.75);

%Stimo il valore ottimale di k
k_max = 20;
for k=1:k_max,
    trn_err(k) = testk(ds_trn, k, ds_trn);
    val_err(k) = testk(ds_trn, k, ds_val);
    tst_err(k) = testk(ds_trn, k, ds_tst);
end

% Calcolo valore ottimale di K
[tmp k_worst] = max(val_err);
[tmp k_best] = min(val_err);
disp(sprintf('K-Best= %d',k_best));
```



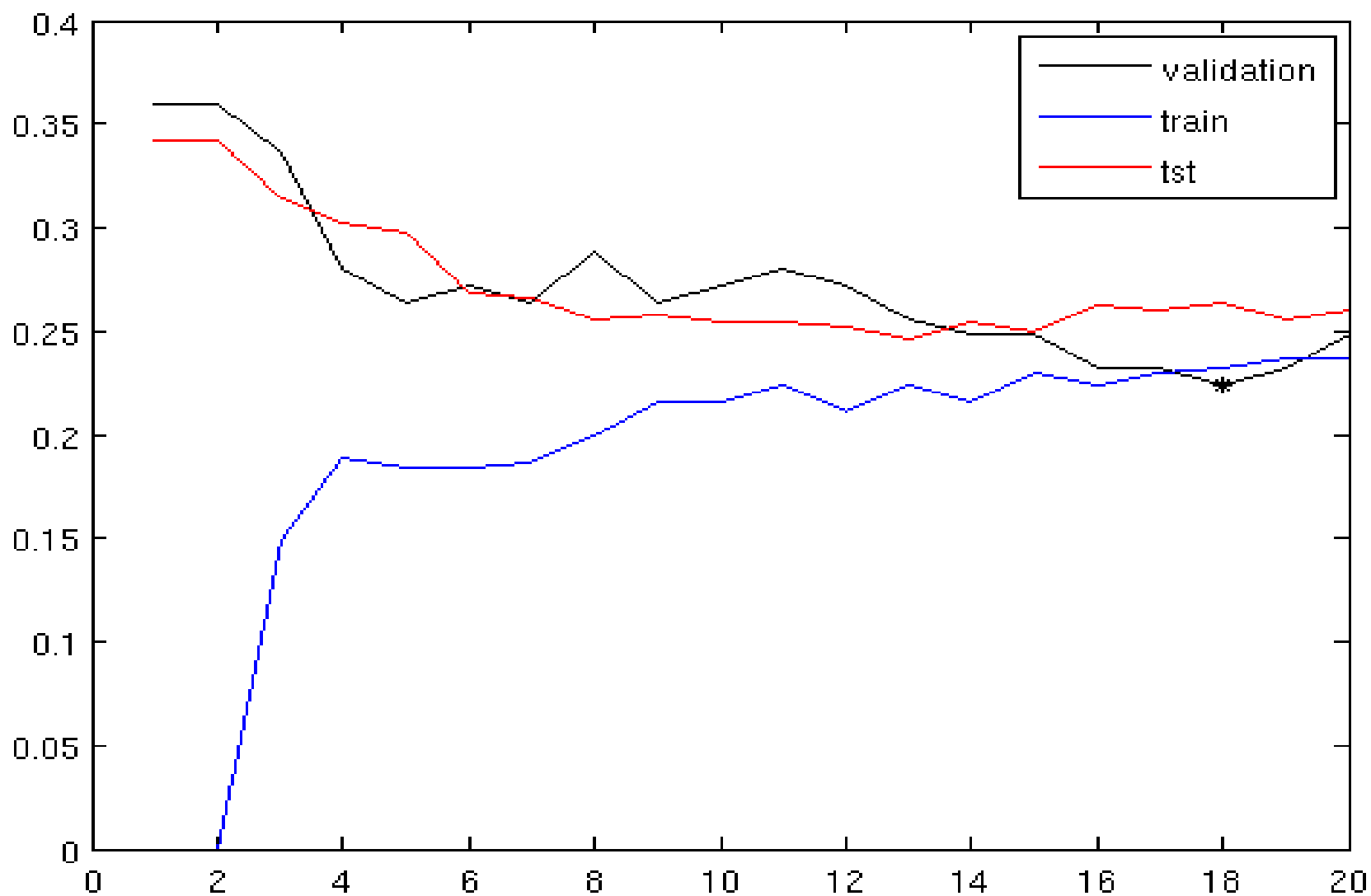
Esercizio 1 - Soluzione



```
%Visulizzo andamento errore  
figure(1); hold off;  
plot(val_err, 'k');  
hold on;  
plot(trn_err, 'b');  
plot(tst_err, 'r');  
legend('validation', 'train', 'tst')  
plot(k_best, tmp, 'k*');
```



Esercizio 1 - Soluzione





Esercizio 2



Ripetere l'esercizio precedente stimando il valore ottimale di K tramite 10-Fold Cross Validation sul training set.

NB. In questo caso il dataset scelto deve essere diviso solo in train e test set.



Esercizio 2 - Soluzione



```
%Utilizzo il dataset dell'esercizio precedente
%(il train set è l'unione di train e validation)
ds_trn = ds_design;

% Stimo il valore ottimale di k
k_max = 20;
for k=1:k_max,
    W = knnc([], k);
    cv_err(k) = crossval(ds_trn, W, 10,3);
    tst_err(k) = testk(ds_trn, k, ds_tst);
end

% Calcolo valore ottimale di K
[tmp k_worst] = max(cv_err);
[tmp k_best] = min(cv_err);
disp(sprintf('K-Best= %d',k_best));
```



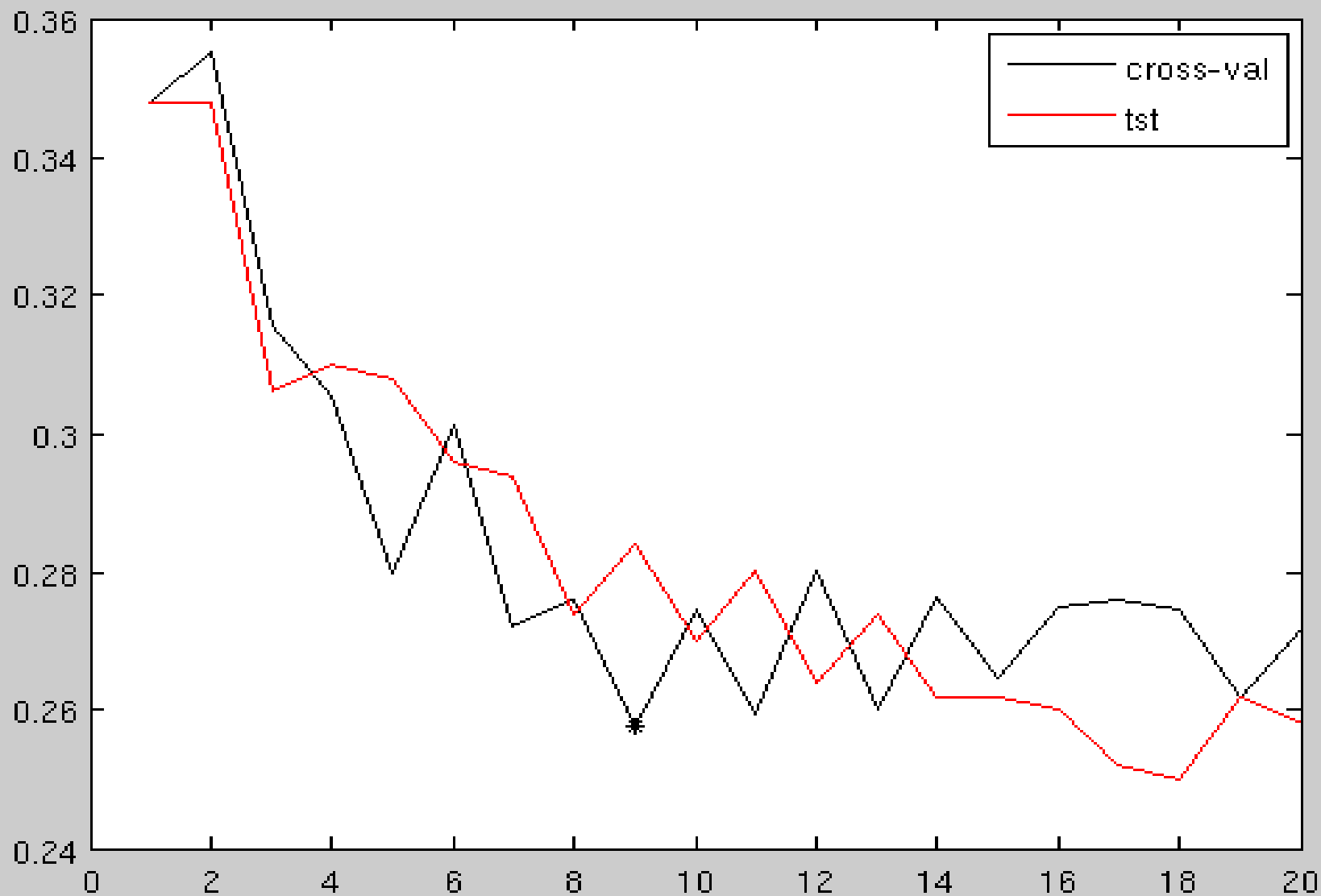
Esercizio 2 - Soluzione



```
%Visulizzo andamento errore  
figure(2); hold off;  
plot(cv_err, 'k');  
hold on;  
plot(tst_err, 'r');  
legend('cross-val', 'tst')  
plot(k_best, tmp, 'k*');
```



Esercizio 2 - Soluzione





Le procedure viste sino ad ora per la stima dei parametri ottimali possono essere applicate in modo molto simile anche alla stima di altri elementi del nostro sistema di classificazione come:

- feature da estrarre
- metodo di normalizzazione
- algoritmo di classificazione
- criterio di assegnamento delle categorie (non MAP)
- etc.