Machine Learning:
un nuovo approccio al data mining

Questionario autovalutativo

Vero o Falso?

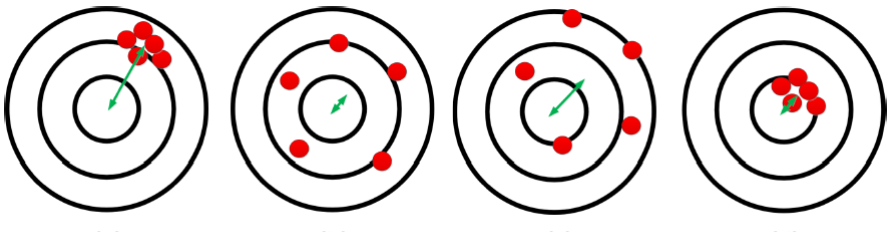
1. Il compito per cui è allenato il modello dipende dalla quantità di dati in possesso.
   - □ Vero
   - □ Falso
2. Nell'analisi univariata verifico la relazione tra una feature e il tempo.
   - □ Vero
   - □ Falso
3. Gli outlier possono essere eliminati dal dataset senza dover per forza fornire una previsione anche per questi punti.
   - □ Vero
   - □ Falso
4. Dopo aver corretto i dati anomali e mancanti posso permettermi di considerarli dati reali, senza dover condurvi analisi specifiche dopo aver allenato il modello.
   - □ Vero
   - □ Falso
5. Per costruire il test set, non sempre il partizionamento del dataset in modo casuale è una buona tecnica di validazione.
   - □ Vero
   - □ Falso
6. L'allenamento del modello consiste nella massimizzazione della risk function.
   - □ Vero
   - □ Falso
7. Se durante il training tramite KFold la deviazione standard è elevata allora è bene passare ad un partizionamento train/validation/test.
   - □ Vero
   - □ Falso
8. Nella cross validation ad ogni iterazione il modello è allenato tante volte quanti sono i fold, meno uno (perché di test).
   - □ Vero
   - □ Falso
9. L'accuratezza è sempre una buona misura delle capacità predittive del modello.
   - □ Vero
   - □ Falso
10. Applicare l'operatore gradiente ad una funzione implica il calcolo delle derivate parziali della funzione per ogni sua variabile.
    - □ Vero
    - □ Falso
11. L'analisi univariata dei residui permette di verificare la stabilità del modello.
    - □ Vero
    - □ Falso

12. L’analisi multivariata dei residui permette di verificare la coerenza delle previsioni del modello al variare del valore delle variabili esplicative.
    - □ Vero
    - □ Falso
13. La risk function, nel caso di apprendimento supervisionato, dipende in maniera funzionale anche dalle variabili esplicative e target, ma tale dipendenza è sempre sottointesa e quindi non specificata.
    - □ Vero
    - □ Falso
14. Se un fenomeno aleatorio è binario allora si può associare ad esso, con certezza, una variabile aleatoria bernoulliana.
    - □ Vero
    - □ Falso
15. Nell’empirical risk minimization, la metrica R^2 può essere usata come risk function ottenendo esattamente gli stessi risultati di un MSE.
    - □ Vero
    - □ Falso
16. La regressione logistica deve il suo nome al matematico A. W. Logistic, vissuto nel Diciannovesimo secolo.
    - □ Vero
    - □ Falso
17. Nela regressione logistica si modellizza con una “regressione lineare” (con una combinazione lineare tra parametri e covariate) il logaritmo delle odds.
    - □ Vero
    - □ Falso
18. Nella regressione logistica la risk function utilizzata è l’opposto della log-verosimiglianza.
    - □ Vero
    - □ Falso
19. La loss function utilizzata per allenare la regressione logistica è sempre utilizzata per problemi di classificazione binaria, anche se si applicano altri modelli di machine learning.
    - □ Vero
    - □ Falso
20. Nell’algoritmo k-Neares Neighbour, k è un parametro che l’algoritmo apprende allo scopo di minimizzare l’errore delle sue previsioni.
    - □ Vero
    - □ Falso
21. Nell’algoritmo k-Neares Neighbour, la risk function utilizzata è l’opposto della log-verosimiglianza.
    - □ Vero
    - □ Falso
22. Nella regressione lineare è complicato spiegare come, sulla base dei valori delle covariate, il modello produce la previsione.
    - □ Vero
    - □ Falso
23. Durante la fase di training gli alberi di decisione possono fornire previsioni con precisione arbitraria, anche fino ad errore nullo.
    - □ Vero
    - □ Falso
24. I CART tendono spesso all’overfitting del dataset, per questo si ricorre a tecniche di bagging.
    - □ Vero
    - □ Falso

25. Le random forest sono una particolare tecnica di bagging.
    - □ Vero
    - □ Falso
26. Il boosting consiste nell'utilizzo di una particolare risk function durante l'allenamento del meta-algoritmo.
    - □ Vero
    - □ Falso
27. Bagging e boosting possono essere applicati con molti modelli di machine learning come base learners, non solo con i CART.
    - □ Vero
    - □ Falso
28. Il coefficiente di variazione dei residui rapporta la deviazione standard dei residui al modulo del loro valore atteso: per questo motivo è una buona metrica per confrontare diversi modelli.
    - □ Vero
    - □ Falso
29. Gli algoritmi di machine learning sono abili nell'interpolazione.
    - □ Vero
    - □ Falso
30. Gli algoritmi di machine learning sono abili nell'estrapolazione, ad esempio nel caso di covariate shift.
    - □ Vero
    - □ Falso
31. La SVM è un classificatore lineare che per garantire buone performance richiede di trovare uno spazio delle covariate (eventualmente una sua trasformazione) che sia linearmente separabile.
    - □ Vero
    - □ Falso
32. Quando si ipotizza una dipendenza lineare tra le features e la variabile target conviene provare ad allenare in primis un decision tree.
    - □ Vero
    - □ Falso
33. Nell'equazione di seguito è un errore indicare che la risk function dipende dai parametri θ_1 e θ_2, infatti nell'espressione di destra tali quantità non sono presenti.

$$\mathcal{R}(\theta_1, \theta_2) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

    - □ Vero
    - □ Falso
34. Non è sempre necessario sviluppare modelli che, qualsiasi sia l'output, siano in grado di fornire una previsione.
    - □ Vero
    - □ Falso
35. Relativamente all'immagine riportata sotto, nel caso (b) siamo in presenza di un modello ad elevato bias, le previsioni infatti si discostano molto le une dalle altre.



    - □ Vero
    - □ Falso

36. Se questo questionario fosse usato come dataset, io che compilo il test con il massimo impegno mi aspetterei una ROC sotto la diagonale (SW – NE) e un AUC inferiore al 50%.
    - □ Vero
    - □ Falso
37. I CART sviluppano criteri decisionali sulla base di criteri randomici e minimizzazione della risk function.
    - □ Vero
    - □ Falso
38. La tecnica del pruning è utilizzata negli alberi decisionali per limitare l'overfitting.
    - □ Vero
    - □ Falso
39. I CART sono algoritmi così potenti da poter produrre un errore nullo su un qualsiasi training set.
    - □ Vero
    - □ Falso
40. Sviluppare una random forest con un base learner che tende ad avere un elevato bias e una bassa varianza permette di migliorare notevolmente la precisione delle previsioni.
    - □ Vero
    - □ Falso
41. Per un problema di classificazione è possibile costruire algoritmi di bagging o boosting usando la regressione logistica.
    - □ Vero
    - □ Falso
42. Lo stacking prevede l'allenamento di diversi modelli di ML sull'intero dataset.
    - □ Vero
    - □ Falso
43. Il boosting è una metodologia di costruzione di ensemble sequenziali.
    - □ Vero
    - □ Falso
44. Il termine "bagging" è l'acronimo di "bag aggregating".
    - □ Vero
    - □ Falso
45. L'algoritmo k-means è specifico per problemi di classificazione.
    - □ Vero
    - □ Falso
46. Gli algoritmi di dimensionality reduction permettono di trasformare i punti dallo spazio delle covariate ad un di proiezione conservando il più possibile una data metrica.
    - □ Vero
    - □ Falso
47. A causa dell'estrema complessità computazionale delle reti neurali, anche per piccoli problemi è impossibile implementare un processo di validazione tramite cross validation.
    - □ Vero
    - □ Falso
48. Il gradient descent è un algoritmo iterativo che ha lo scopo di trovare il minimo di una funzione usando come "bussola" il suo gradiente.
    - □ Vero
    - □ Falso
49. Quando il gradiente di una funzione si annulla in un punto allora tali coordinate identificano un punto di sella, massimo o minimo.
    - □ Vero
    - □ Falso

50. Quando l'algoritmo di gradient descent converge allora il punto trovato è di sella, massimo o minimo.
    - □ Vero
    - □ Falso
51. La funzione di attivazione sigmoidale è utilizzata anche nella regressione logistica.
    - □ Vero
    - □ Falso
52. Le reti neurali ricorrenti nella loro forma più semplice prevedono l'applicazione della stessa matrice dei pesi ad ogni input esterno.
    - □ Vero
    - □ Falso
53. Le reti neurali ricorrenti nella loro forma più semplice prevedono che lo stato successivo sia il risultato dell'applicazione di un percettrone sullo stato precedente e nuovo input.
    - □ Vero
    - □ Falso
54. Le reti neurali sono una generalizzazione dei modelli lineari come la regressione lineare o quella logistica.
    - □ Vero
    - □ Falso
55. Gli algoritmi di dimensionalty reduction mirano a ridurre il numero di variabili esplicative tentando di conservare una o più proprietà statistiche del dataset.
    - □ Vero
    - □ Falso
56. Il multi-percettrone prevede che tra l'input layer e quello di output siano presenti degli strati intermedi di percettroni detti shadow layer.
    - □ Vero
    - □ Falso
57. La sigla LSTM nelle omonime reti neurali ricorrenti significa "Large Small Timeseries Memory".
    - □ Vero
    - □ Falso
58. Le reti neurali ricorrenti LSTM introducono oltre allo stato del sistema anche uno stato della cella, che presumibilmente dovrebbe conservare nel tempo più informazione.
    - □ Vero
    - □ Falso
59. La più grande limitazione delle più semplici reti neurali ricorrenti è che è estremamente difficile allenarle con in input lunghe sequenze.
    - □ Vero
    - □ Falso
60. Quando il gradiente della risk function pressoché si annulla, ad esempio nei pressi di un minimo, si parla di fenomeno di vanishing gradient.
    - □ Vero
    - □ Falso

Risposte corrette: [ ] / 60