**Machine Learning - Esempi di domande**

1. Quali formule descrivono i **modelli a regressione lineare, logistica e polinomiale**?
2. Dare la definizione di **bias**.
3. Descrivere l 'algoritmo di **tree induction**.
4. Che cosa si intende per **distance-based clustering**?
5. Dare la **definizione di KDD** esistente in letteratura, spiegando il significato di tutti i termini in essa utilizzati. Descrivere quindi gli step che compongono il processo di KDD.
6. Scegliere un **dominio di propria invenzione** su cui si ipotizza di dover applicare machine learning di tipo **supervisionato predittivo**. Descrivere e formalizzare con la notazione adottata a lezione: il dataset, la funzione target, le feature, la classe, che tipo di previsione dovrà effettuare il modello.
   1. Scegliere un **dominio di propria invenzione** su cui si ipotizza di dover applicare **regressione** (lineare o polinomiale). Descrivere e formalizzare con la notazione adottata a lezione: il dataset, la funzione target, le feature, la classe, che tipo di previsione dovrà effettuare il modello.
   2. Scegliere un **dominio di propria invenzione** su cui si ipotizza di dover applicare learning **supervisionato**. Descrivere e formalizzare: il dataset, la funzione target, le feature, la classe, che tipo di previsione dovrà effettuare il modello
7. Dare la definizione di **varianza**.
8. Descrivere l’algoritmo di apprendimento di una **Random forest**.
   1. Come funziona l’algoritmo di **Bagging**? Per quali motivi viene impiegata questa tecnica?
9. Qual è lo scopo dell’algoritmo **k-NN**? Quali funzioni target utilizza nel caso base e nelle sue estensioni?
10. Qual è lo scopo di **K-means**? Qual è la formula della funzione obiettivo? Come si scelgono i centri iniziali?
11. Dare la definizione di **concept learning** (apprendimento di concetti) e fare un esempio in un dominio di propria invenzione.
12. Cosa si intende per **curse of dimensionality** relativamente all’algoritmo **k-NN**? Quali soluzioni vengono adottate per tale problema?
13. Dare le definizioni di ipotesi, funzione di loss, apprendimento, modello, inferenza.
14. Descrivere il **processo di KDD**.
15. Rappresentare sul **piano Errore vs Capacity** le curve relative alle **diverse tipologie di errore**, la capacità ottimale di un modello e le zone di underfitting e overfitting. Commentare il grafico.
16. Cos’è e a che cosa serve l’**AUC ROC**?
17. Dare la definizione di **regola associativa** e fornire un esempio reale. Descrivere il problema di scoperta di regole associative.

**Risposte**

**Domanda 6**:

Il dominio è la previsione del successo di una startup entro un anno dalla creazione. Si vuole quindi prevedere se una startup avrà successo o meno. Il dataset avrà dati raccolti da startup simili in un periodo, ad esempio gli ultimi 5 anni. Con xi tutte le feature per la startup i. E yi variabile che indica il successo o meno della stessa.

La funzione target è del tipo Rd → {0,1}, con d #feature. Il modello vorrà apprendere un’approssimazione f’ di f.

Le feature possono essere diverse: Fondi raccolti nel primo round di finanziamenti, Membri del team, Esperienza media dei membri, Presenza di un prodotto già sviluppato al momento del lancio, Percentuale di crescita degli utenti.

La classe è binaria. Se vale 0, allora non avrà successo. Se vale 1, ha successo, quindi riesce a raggiungere un obiettivo prefissato.

Si dovrà effettuare una classificazione binaria. È un problema supervisionato predittivo. Si deve assegnare un’etichetta (0,1 oppure P,N) a un nuovo caso non visto.

**Domanda 6a**:

Il dominio è la previsione del prezzo di una casa sulla base delle sue caratteristiche.

Il dataset conterrà informazioni su case vendute in una certa area negli ultimi anni. Con xi tutte le feature per la casa i. E yi prezzo di vendita della stessa. La funzione target è del tipo Rd → R e il modello vorrà apprendere un’approssimazione f’ che potrà essere una funzione lineare o polinomiale.

Le feature possono essere diverse: Superficie della casa, #Stanze, #Bagni, Età, Distanza dal centro.

La classe è una variabile continua, per il prezzo della casa.

Si dovrà effettuare una regressione a seconda se il prezzo delle case può essere approssimato da una relazione lineare o polinomiale con le feature. Si applicherà o regressione lineare o polinomiale.

**Domanda 6b**:

Il dominio è la previsione che un cliente abbandoni un servizio. Il dataset avrà dati sui clienti dell'azienda che offre un servizio in abbonamento. Con xi tutte le feature per il cliente i. E yi variabile che indica se il cliente abbandonerà (1) o meno (0).

La funzione target è del tipo Rd → {0,1}, con d #feature. Il modello vorrà apprendere un’approssimazione f’ per stimare la probabilità che y=1 dato x.

Le feature possono essere diverse: Durata dell’abbonamento, Frequenza d’uso del servizio nell’ultimo periodo, Importo medio dell’abbonamento, #Reclami in un periodo, Metodo di pagamento.

La classe è binaria. Se vale 0, allora continua a usare il servizio. Se vale 1, ha abbandonato.

Si dovrà effettuare una classificazione supervisionata. Restituire una probabilità P(y=1|x), con una soglia (½) per decidere se un cliente è a rischio.

**Domanda 15**:

Capacity - Asse x, la complessità di un modello. Prima si saranno ad esempio regressione lineare semplice, mentre a destra modelli più complessi, ad esempio polinomi ad alto grado.

Errore - Asse y, l’errore può essere

* di Training, un modello più complesso può adattarsi meglio ai dati di training.
* di Test, diminuisce con la capacità, ma aumenta dopo una soglia, quando il modello andrà incontro a overfitting. Andamento a U.

Il grafico mostra tre aree:

* Underfitting: Modello troppo semplice per catturare la complessità dei dati. Entrambi gli errori sono alti, e tipico per modelli su dati non lineari.
* Zona ottimale: Zona con errore di test minimo. Generalizza bene sui dati di test. È il compromesso ideale tra bias e varianza.
* Overfitting: Modello troppo complesso, si adatta ai dati di training, ma non riesce a generalizzare. Errore di test alto e di training basso.

Il grafico mostra l’importanza della giusta complessità del modello. Troppo poco → underfitting, troppa → overfitting. Per la zona ottimale si può utilizzare cross-validation.

**Domanda 13**:

Ipotesi è una funzione candidata appartenente allo spazio delle ipotesi. Questa mappa le feature x sui valori target y. Ciò che il modello cerca di apprendere

Loss Function misura quanto l’output predetto si discosta dal valore target reale. Per valutare la qualità delle previsioni. L’obiettivo è minimizzarla. (MSE)

Apprendimento è il processo con cui un modello ottimizza i suoi parametri per ridurre l’errore rispetto a un dataset di training. Si basa su algoritmi per minimizzare la loss function.

Modello è l’implementazione matematica di un’ipotesi, una relazione tra input x e output y.

Inferenza è il processo di uso di un modello addestrato per fare previsioni su dati non visti.

**Domanda 11**:

[Def di concept learning]

Un esempio è un sistema che vuole apprendere il concetto di frutta commestibile nel dominio di tutti i tipi di frutti. Saranno da fornire degli esempi positivi (frutta commestibile) e esempi negativi e chiedere al sistema di apprendere le caratteristiche distintive. Esempio: Positivi (Mela, banana, arancia) Negativi (Abete, Fungo velenoso) e le caratteristiche (colore, forma, consistenza). Il sistema apprenderà le caratteristiche e quando si presenterà un nuovo esempio, riuscirà a distinguerlo sulla base di ciò che ha appreso.

**Domanda 7**:

La varianza misura la dispersione di un insieme di dati rispetto alla loro media. Indica quanto i dati sono distribuiti intorno alla loro media. Più è alta, maggiore la dispersione, grandi differenze tra i singoli valori e la media. Specifica la variabilità di un dataset. .

**Domanda 2**:

Rappresenta la tendenza di un modello a fare previsioni sbagliate per alcuni esempi. Bias nel modello quando questo non è abbastanza complesso per rappresentare correttamente la relazione tra input e output. È l’incapacità di adattarsi ai dati. Bias nei dati quando non sono rappresentativi dell’intera varietà di casi che il modello incontrerà. È legato alla underfitting. Modello troppo semplice che non cattura la complessità dei dati, e quindi bias elevato.

È spesso dovuto a interpretazione errata dei dati o modelli troppo semplici.

**Domanda 1**:

La regressione lineare descrive la relazione tra una variabile dipendente continua y e una o più variabili indipendenti xI con una funzione lineare. Quindi y = B0 + B1x1 + BNxN + e

La regressione logistica quando l’output è una variabile categorica binaria (0,1). Probabilità di un evento y=0 in funzione delle variabili x. P(y=1|xi) = 1 / 1+ e-z, funzione sigmoide.

La regressione polinomiale, per relazioni non lineari tra x,y. La formula è un polinomio di grado d. Grado che determina la complessità.