DATA SCIENCE AZ IT BIZTONSÁGBAN

1. forduló

SOPHOS

A kategória támogatója: SOPHOS

Ismertető a feladathoz

Kérjük, hogy a feladatlap indítása előtt mindenképp olvasd el az alábbi útmutatót:

- MINDEN kérdésre **van helyes válasz**.
- Olyan kérdés **NINCS**, amire az összes válasz helyes, ha mégis az összes választ bejelölöd, arra a feladatra automatikusan 0 pont jár.
- A **radio button-os kérdésekre egy helyes válasz van**.
- **Ha lejár a feladatlap ideje, a rendszer AUTOMATIKUSAN** beküldi azt az addig megjelölt válaszokkal.
- Azokat a feladatlapokat, amelyekhez **csatolmány** tartozik, javasoljuk **NEM mobilon** elindítani, erre az érintett feladatlapok előtt külön felhívjuk a figyelmet.
- Az **adatbekérős feladatokra NEM jár részpontszám**, csak a feleletválasztósakra.
- **Helyezéseket a 4. forduló után mutatunk**, százalékos formában: adott kategóriában a TOP 20-40-60%-hoz tartozol.
- **Badge-ket** szintén a 4.forduló után kapsz majd először.
- Ha egyszerre több böngészőből, több ablakban vagy több eszközről megnyitod ugyanazt a feladatlapot, **nem tudjuk vállalni** az adatmentéssel kapcsolatban esetlegesen felmerülő anomáliákért a felelősséget!
- A hét forduló során az egyes kategóriákban (de nem feltétlenül mindegyikben) **könnyű-közepes-nehéz kérdésekkel** egyaránt találkozhatsz majd.

Jó versenyzést kívánunk!

1.forduló

Az első fordulóban egy gyors emléteti kérdéssorron megyünk át, hogy megalapozzuk a gyakorlatiasabb jellegű második fordulót.

Felhasznált idő: 00:00/30:00 **Elért pontszám:** 0/12

1. feladat 0/1 pont

A felsoroltak közül melyik "felügyelt tanulási" probléma (supervised learning)?

Válasz

- (●) Bináris fájlok osztályozása XGBoost segítségével malware/benignware-ként
  Ez a válasz helyes, de nem jelölted meg.
- ( ) Anomália detektálás tűzfal logokban Isolation Forest segítségével
- ( ) Near duplicate detekció MinHash segítségével

Magyarázat

Felügyelt tanítás során címkézett adatokat használunk fel.

Isolation Forest és Minhash esetén nincs szükség címkékre.

2. feladat 0/1 pont

Az alábbiak közül mi igaz a machine learning-re (ML) az IT biztonságban?

Válaszok

- [x] Az IT biztonsági rendszerek rengeteg adatot generálnak, ezek kiértékelésében az ML segítséget nyújthat
  Ez a válasz helyes, de nem jelölted meg.
- [ ] Az ML képes teljesen kiváltani a hagyományos antivírus (AV) megoldásokat
- [x] A mesterséges intelligenciára való túlzott hagyatkozás a biztonság hamis érzetét keltheti
  Ez a válasz helyes, de nem jelölted meg.
- [x] Egy ML tanítása múltbeli adatokon lehetővé teheti még nem ismert malware-ek felismerését a jövőben
  Ez a válasz helyes, de nem jelölted meg.

Magyarázat

1. Az IT biztonsági rendszerek jellemzően hatalmas mennyiségű adaton dolgoznak, hiszen egy eszközön rengetek process fut egyidejűleg, és mindnek lehet jelzés értéke. Az ML egyik fő erénye hogy hagyományos szabályalapú rendszekkel nehezen megfogható tudást tud felskálázni nagy mennyiségű adatra.
2. A tanító adatok előbb utóbb elavulttá válnak az IT securityban, ezért az ML nagy valószínűséggel soha nem fogja tudni kiváltani a hagyományos megoldásokat. Ezenfelül az ML sosem ad tökéletes biztonságot, csak egy a sok komponens közül.
3. Jellemzően egy IT security rendszer több komponensből épül fel. Ezek együttes védelme ad megbízható biztonságot, a komponensek önmagukban könnyebbne megkerülehetőek. Az ML modellek hírhedten könnyűen becsaphatóak.
4. Amennyiben ez nem teljesül, az ML-nek nincs hozzáadott értéke, a szabály alapú rendszerekhez képest.

3. feladat 0/1 pont

A feladat egy egyszerű bináris klasszifikációs modell építése. A cél megjósolni, hogy az adott minta kártékony, vagy ártalmatlan-e.

Melyik algoritmust lenne érdemes használni ebben az esetben, ha a **legegyszerűbb** megoldást keressük?

Válaszok

- [x] Linear regression
  Ez a válasz helyes, de nem jelölted meg.
- [x] Logistic regression
  Ez a válasz helyes, de nem jelölted meg.
- [ ] Random forest regression

Magyarázat

Kategórikus (bináris) kimenetet várunk el a modellünktől. Ezt legegyszerűbben Logistic regression segítségével tudjuk megtenni. A többi algoritmus folytonos érték becslésére alkalmazható.

4. feladat 0/1 pont

Mi mondható el a Sigmoid függvényről Logistic Regression esetén?

Válasz

- (●) Valószínűséget modellezünk vele
  Ez a válasz helyes, de nem jelölted meg.
- ( ) Kimenete -végtelen és +végtelen közé esik

Magyarázat

Sigmoid függvény tetszőleges bemenetet a [0, 1] intervallumra képez le, igy tekinthetünk rá valószinűség modellezéseként.

5. feladat 0/1 pont

Mely állitások igazak a One-Hot Encoding (OHE) működésére?

Válasz

- (●) Az OHE új feature-öket hoz létre a bemenetben szereplő egyedi érték és számossága alapján
  Ez a válasz helyes, de nem jelölted meg.
- ( ) Az OHE dimenzionalitást csökkent legkisebb négyzetek módszerrel
- ( ) Az OHE egy rendezést biztosít az általa kódolt értékek között, ezzel egyszerűsitve a ráépitett modellek optimalizációját

Magyarázat

https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

6. feladat 0/1 pont

Mely állítások igazak az alábbi logloss függvényre, ahol y' a becsült érték, y pedig a minta címkéje, x pedig a feature vektor?

$$\text{Log Loss} = \sum_{(x,y)\in D} \underbrace{-y\log(y')}_{\text{Első tag}} \underbrace{-(1-y)\log(1-y')}_{\text{Második tag}}$$

Válasz

- ( ) y = 0 és y' = 0 esetén a loss a második tag miatt jelentősen nőni fog
- ( ) y = 0 és y' = 0 esetén a loss az első tag jelentősen miatt nőni fog

- [ ] y = 0 és y' = 0.99 esetén a loss az első tag miatt jelentősen nőni fog
- [x] y = 0 és y' = 0.99 esetén a loss a második tag miatt jelentősen miatt nőni fog
  Ez a válasz helyes, de nem jelölted meg.

Magyarázat

ML tanítás során jellemzően a loss függvény minimalizáljuk. Intuitívan azt várjuk el a logloss függvénytől, hogyha a becsült érték megegyezik a minta címkéjével, akkor kicsi legyen (lehetőleg 0), illetve minél nagyobb az eltérés a címke és a becsült érték között, a loss annál nagyobb legyen.

Az 1. és 2. válasznál megegyezik a becsült érték a címkével, így egyik tag sem növeli a loss értékét.

3. válasz; y=0 miatt az első tag nulla lesz, az nem járul hozzá a losshoz

4 válasz; a losshoz ez a minta -1*log(0.01)-t ad hozzá.

7. feladat 0/2 pont

Az alábbi állitások közül melyek igazak?

Válasz

- [x] A kutatási környezetben használt adatok általában statikusak, míg az ipari környezetben használt adatok dinamikusak és folyamatosan változnak
  Ez a válasz helyes, de nem jelölted meg.
- [ ] Ipari környezetben általában egyetlen adott metrika maximlizálása, például nagyfokú pontosság a cél, míg kutatási környezetben nagyobb hangsúly kerül a skálázhatóságra
- [ ] Az ipari környezetben használt tanitó adatok publikusak, mig kutatási területen általában titkositják őket

Magyarázat

Tisztán kutatási területen statikus publikus adatokon jellemző fejleszteni modelleket egyetlen metrika optimalizálása érdekében.

Ipari alkalmazásban jellemzően az adatok változnak, a cég tulajdonát képezik és érzékeny információkat tartalmaznak. Bár az algoritmusok pontossága elvárás ipari környeyetben is, a költséghatékony deployment jellemzően szintén egy fontos aspektus.

8. feladat 0/2 pont

Klasszifikáció esetén a likelihood maximalizálása megegyezik a cross-entropy minimalizálásával.

Válasz

- [x] Igaz
  Ez a válasz helyes, de nem jelölted meg.
- [ ] Hamis

Magyarázat

https://stats.stackexchange.com/questions/364216/the-relationship-between-maximizing-the-likelihood-and-minimizing-the-cross-entr

9. feladat 0/2 pont

Az alábbiak közül melyek igazak a Logistic regression-re?

Válasz

- (●) Szükséges előfeltétele a lineáris függőség a független változók és a log-odds között
  Ez a válasz helyes, de nem jelölted meg.
- ( ) Szükséges előfeltétele a multikollinearitás a független változók között
- ( ) Robosztus a kiugró adatokkal szemben

Magyarázat

https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/