

Gépi tanulás a gyakorlatban, előadás vázlat

Vincze Nándor

January 4, 2026

1 Bevezető, alapfogalmak

Definíció (Gépi tanulás). *Az intelligens viselkedés azon része, amely a tanulás képességén alapul. Minden olyan megoldás, ahol a rendszer teljesítménye javul a tapasztaltok gyűjtése által.*

Megjegyzés. *Gépi tanulás \subset Mesterséges intelligencia*

- Képfelismerés
- Önvezető autó
- Ajánlórendszerek

Definíció (Gépi tanulás 2). *Ha adott egy feladat T , és egy P teljesítménymetrika, akkor gépi tanulásról beszélünk, ha a rendszer egyre több E tapasztalat hatására a P teljesítménye a T feladaton javul.*

Definíció (Data science). *Üzleti problémák megoldása, statisztikai elemzések, gépi tanulás alkalmazásával.*

Megjegyzés. *Gépi tanulás esetén egy konkrét feladatot akarunk megoldani minél jobban.*

Definíció (Felügyelt tanulás). *Megfigyelés és célérték is áll rendelkezésre, a rendszer célja, hogy nem látott példákra is a lehető legjobb célértéket jósolja meg.*

Definíció (Felügyelet nélküli tanulás). *Csak a megfigyelés áll rendelkezésre, a rendszer célja a mintázatok és összefüggések felismerése.*

Definíció (Megerősítési tanulás). *A rendszer egy környezettel lép kölcsönhatásba, és a visszajelzések alapján tanul.*

Definíció (Osztályozási feladat). *Egyedek előre meghatározott osztályokba való besorolása.*

Definíció (Regressziós feladat). *Az egyedekhez tartozó folytonos célértékek előrejelzése.*

Definíció (Jellemzők). *Az egyedek leíró tulajdonságok. Lehet folytonos vagy diszkrét.*

Definíció (Gépi tanulás loop). *gépi tanulási feladat megoldásának lépései:*

1. *adatgyűjtés*
2. *előfeldolgozás*
3. *jellemző kinyerés*
4. *modell kiválasztás*
5. *tanítás*
6. *kiértékelés*

2 Egyszerű statisztikai döntések

Az adathalmaz statisztikáiból egyszerű döntési szabályokat hozunk létre. **Átlag, szórás, medián, min, max stb...**

Hisztogramok, scatter plot.

Definíció (Konstans döntés). *A jellemzőket nem vesszük figyelembe, és csak a célváltozó alapján predikálunk. Diszkrét esetben módusz, folytonos esetben átlag.*

Definíció (Egyetlen jellemző alapján történő döntés). *Triviális.*

Definíció (Korreláció (pongyola)). *Két változó közötti **lineáris** kapcsolat mértéke. $\rho \in [-1, 1]$*

A jellemzők kiválasztására több módszer is létezik, pl χ^2 teszt.

3 Osztályozási feladatok, döntési fák

Osztályozási feladat példák, train/test split, metrikák.

Definíció (k -fold cross validation). Az adathalmazt k részre bontjuk, és k iterációt/kísérletet futtatunk. Minden iterációban pontosan 1 részhalmazt használunk tesztelésre, a maradék $k - 1$ részhalmazt tanításra. Ekkor minden egyed pontosan egyszer kerül a teszhalmazba.

	Igaz pozitív	Igaz negatív
Definíció (Confusion matrix). Predikált pozitív	TP	FP
Predikált negatív	FN	TN

Definíció (Accuracy). $acc = \frac{TP+TN}{TP+TN+FP+FN}$

Definíció (Precision). $prec = \frac{TP}{TP+FP}$

Definíció (Recall). $recall = \frac{TP}{TP+FN}$

Definíció (F1 score). $F1 = 2 \cdot \frac{prec \cdot recall}{prec + recall}$

Definíció (Döntési fa). Egy olyan tanuló algoritmus, ami egy fa struktúra alapján osztályozási feladatot lát el. A belső csúcsok jellemzőkkel címkézettek, melyek szerint a bemeneti egyedeket szétválasztjuk. A levelek az adott szétválasztások szerint predikált osztálycímkek.

Folytonos és diszkrét jellemzők esetén is alkalmazható.

Néhány jellemző:

- nem minden jellemző kerül felhasználásra
- a levelek mélysége nem azonos
- jól vizualizálható
- explicit szabályokat tanul meg a jellemzők közötti

DE

- greedy algoritmus
- sok jellemző esetén sok példa kell

Definíció (Erdő osztályozó). Több, kisebb döntési fa együttes szavazata alapján jön létre a végső predikció.

4 Text mining és lineáris gépek

Strukturálatlan szöveges adatok automatikus feldolgozása. Több nehézség, mint a szinonímaszavak, szleng, nyelvjárások, hibák stb. . .

Néhány alkalmazása:

- chatbot
- fordítás
- információ kinyerés
- dokumentum osztályozás
- spam szűrés

Módszer (Szövegek előfeldolgozása). 1. *Tokenizálás: szavakra bontás*

2. *Normalizálás: kis/nagybetű, ékezetek stb. . .*

3. *Lemmatizálás: szótókinyerés, vagy Stemming: ragok egyszerű levágása*

4. *írásjelek eltávolítása*

5. *Stopszó szűrés: kötőszavak, névelők, jelentés nélküli*

6. *POS tagging: szófajok megjelölése*

Módszer (Szózsák modell). *Dokumentumok osztályozása során egy szótárt készítünk, az összes előforduló szóból, majd az egyes dokumentumokat a szótárból megjelenő szavak előfordulási gyakoriságával jellemezzük.*

A szavak sorrendje elveszik. Ez megoldható ha n -grammokat is használunk, kizárólag unigrammo helyett.

Definíció (Term frequency - inverse document frequency). *Szeretnénk, hogy azon szavak amelyek több dokumentumban is előfordulnak, kisebb súlyt kapjanak. Legyen adott w szó, d dokumentum, D az összes dokumentum halmaza, $tf(w, d)$ a w szó előfordulási gyakorisága a D dokumentumban, $df(w, D)$ pedig azon dokumentumok száma az D halmazban, amelyek tartalmazzák a w szót.*

Ekkor a TF-IDF súlyozás:

$$TFIDF(w, d) = tf(w, d) * \log \left(\frac{1}{df(w, D)} \right)$$

Definíció (Lineáris osztályozó). *Egy adott n dimenziós jellemző térben egy hipersík segítségével választja szét az osztályokat. A hipersík egyenlete:*

$$w_1x_1 + w_2x_2 + \dots + w_nx_n + b = 0$$

ahol w_i a súlyok, b az eltolás, megtanulandó paraméterek a predikcióhoz. Az osztályozás a következőképpen történik:

$$f : \mathbb{R}^n \rightarrow \{-1, 1\}, \quad f(x) = \text{sign}(w \cdot x + b)$$

ahol $f(x) = 1$ az egyik osztály, $f(x) = -1$ a másik osztály.

Definíció (Diszkriminatív vs. generatív módszerek). *A diszkriminatív modellek (mint a lineáris gépek) célja, hogy minél pontosabban elválasszák az osztályokat, míg a generatív modellek (mint a döntési fák) a bemeneti adatok eloszlását próbálják modellezni az egyes osztályokra külön-külön.*

Megjegyzés. *A lineáris gépek sok jellemzőt tudnak kezelni, ezek mind hozzájárulnak a döntéshez, viszont ha nem lineáris a változók közötti kapcsolat, akkor gyenge lesz a modell.*

5 Deep learning

Egy olyan gépi tanulási módszer, mely során több rejtett rétegből álló mesterséges neurális hálókat alkalmazunk.

Definíció (Neuron).

$$y = f \left(\sum_{i=1}^n w_i x_i + b \right)$$

ahol f a nem-lineáris aktivációs függvény, n a bemenetek száma, w_i a súlyok, b az eltolás, x_i a bemenetek, y a kimenet.

Egy neurális háló tanítása alatt a w_i és b paraméterek olyan beállítását értjük, amikor a **hibafüggvény** értéke minimális.

Definíció (Hibafüggvény). Számszerűsíti a modell predikciójának pontatlanságát. Ez lehet MSE, Cross entropy stb. . .

Definíció (Epoch). Az adathalmazon történő egy teljes tanulási ciklus.

Intuíciót erőtelve a rétegek értelmezésére, az első rétegek végzik az "automatikus" jellemzőkinyerést, míg a későbbi rétegek már ezekkel a jellemzőkkel dolgoznak.

Definíció (Konvolúciós neurális háló (CNN)). Olyan neurális háló, amely konvolúciós rétegeket is tartalmaz. Ezek a rétegek kis szűrőket alkalmaznak a bemeneti adatokra, hogy helyi jellemzőket nyerjenek ki. Gyakran használják képfeldolgozási feladatokban.

A mély gépi tanulás előnyei a klasszikus módszerekkel szemben: képesek nyers adatokkal dolgozni és változatos feladatokat meg tudnak oldani. Cserébe sok adat szükséges a tanításhoz, ami meglehetősen számításigényes, valamint black-box.

6 Reprezentáció tanulás, autodecoder

Olyan módszerek amelyek célja, hogy a bemeneti adatoknak egy hatékony elkódolását tanulják meg. Jellemzően a cél, hogy egy adott egyedet egy beágyazási vektorral jellemezzünk, amely magában hordozza az adott egyed fontos tulajdonságait.

Definíció (Önfelügyelt tanulás). *Olyan tanulási paradigma, ahol a modell a bemeneti adatok egy részét használja fel a bemenetként, míg a bemeneti adatok másik részét célértékként. Így a modell képes megtanulni a bemeneti adatok belső struktúráját anélkül, hogy külső címkékre lenne szükség.*

Módszer (Szóbeágyazások). *word2vec*

Cél: van egy szótárunk, amely minden eleméhez egy megfelelő vektort rendelünk. Szeretnénk, hogy a hasonló jelentésű szavakhoz közeli vektorok tartozzanak.

- *CBOW (continuous bag of words): adott egy adott szó környezete, a cél, hogy a környezetből megjósoljuk a középső szót (jegyzetben 5 szó)*
- *Skip-gram: adott egy szó, a cél, hogy megjósoljuk a környezetében található szavakat*

Módszer (LLM-ek). *A word2vec megközelítések a szavak jelentését statikusan tárolják. A nagy nyelvi modellek ezzel ellentétben teljes mondatok vagy bekezdések reprezentációját tanulják meg. A szóbeágyazások itt kontextusfüggőek.*

Definíció (BERT modellek). *Amikor egy szöveget akarunk elemezni, ismerjük a teljes szöveget, így a tanítás során egy adott szót letakarunk (maszkolunk), és a cél, hogy a modell a környező szavak alapján megjósolja a letakart szót.*

Definíció (GPT modellek). *Amikor szövegek generálása cél, akkor egy szónak csakis a baloldali környezete ismert, ez alapján kell megjósolni a következő szót. Itt használjuk a transzformer architektúrát. (???)*

Definíció (autodencoder). *Olyan neurális háló, amely egy adott bemenetet kódol (jellemzően alacsonyabb dimenzióba), majd visszakódolja. A cél, hogy a kimenet az eredeti bemenet legyen.*

7 Fine-tuning és generatív mesterséges intelligencia

Rendelkezésre állnak előre betanított modellek, amelyek szöveg-, vagy képfeldolgozási beágyazásokat végeznek. A kódolás eredményét felhasználhatjuk egy adott feladathoz tartozó modell betanításához. Ekkor az alsó rétegekben a betanított súlyokat használjuk, míg a felső rétegek súlyait véletlenszerűen inicializáljuk. A tanulás során természetesen az alsó rétegek súlyait is módosítjuk, vagyis **finomhangoljuk** az eredeti modellt.

7.1 Szövegenerálás

A GPT típusú modellek az interneten előtanított beágyazások segítségével képesek adott szószorozathoz megjósolni a következő szót a lehetséges angol szavak szótárából. Ezt egy osztályozási feladatként kezelik, vagyis a modell megadja a legvalószínűbb következő szót.

A ChatGPT egy finomhangolt GPT modellt, ahol emberi visszajelzések alapján tovább tanították a modellt, így képes a felhasználóval folytatott párbeszédre.

7.2 Képgenerálás

Ennek az alapjai is egy generálásra finomhangolt beágyazás. Lényegében a tanítópéldák képek, amelyek bizonyos részeit letakarjuk, és a modell feladata a hiányzó pixelek kiegészítése.

7.3 Képgenerálás instruckió alapján

Az előző két módszer ötvözte: egy szöveges leírást (instruckiót) kap a modell, és egy kép generálását kell elvégeznie. Itt a szöveges leírást egy előre betanított szöveg beágyazó modell segítségével kódolják.

A diffúziós modellek visszavezetik a generálási feladatot egy zajmentesítési feladatra. Az alapgondolat az, hogy van egy zajos képünk, de tudjuk annak tartalmát (szövegesen megadva), akkor a szöveges leírás alapján még a nagyon zajos képekből is helyre tudunk állítani egy képet, ami megfelel a leírásnak. Ez szintén egy önfelügyelt tanulási megoldás, hiszen a tanítópéldák képeiből mesterségesen zajosítunk képeket, és a modell feladata a zajmentesítés.

8 Képfeldolgozás és lokális osztályozók

Definíció (Képosztályozás). *Egy adott képhez tartozó címke (pl.: kutya, macska, autó stb...) predikálása.*

Definíció (Objektum detektálás). *Egy adott képen található objektumok helyének és címkéjének meghatározása.*

Definíció (Augmented reality). *Egy technológia, amely a valóságot kiegészíti digitális információkkal.*

Megjegyzés. *Az előfeldolgozás kiemelten fontos képes esetén, és eltérő lehet a felhasználási területtől függően.*

egységes méret, színcsatorna egységesítés

Néhány példa a képek jellemzésére:

- *két kép között a távolságot az egyes pixelek páronkénti hasonlósága alapján adjuk meg (pixelvektorok, hasonló struktúrájú képek esetén működik)*
- *a kép színeloszlását histogramokkal jellemezzük*
- *képfeldolgozási eszközökkel kinyerhetünk alacsony szintű jellemzőket (élek, sarkok, textúrák stb. . .) majd a szózsák modellhez hasonlóan jellemezhetjük a képeket*
- *mély neurális hálók segítségével, az utolsó előtti réteg kimenetét használhatjuk jellemzőként*

Definíció (KNN). *Ha adott egy távolság metrika, akkor a jellemzőtérben, valós időben megkeressük a legközelebbi k tanító példát, és ezek címkéinek többségi szavazata alapján predikálunk.*

Érzékeny a jellemzők skálázására, így érdemes normalizálni az adatokat, illetve dimenzionalitás problémája itt is fenáll.

9 Regressziós feladat

Olyna gépi tanulási probléma, ahol a célváltozó folytonos értékű.

Definíció (MSE).

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Ahol y_i a predikált érték, \hat{y}_i a valós érték, n a példák száma.

Definíció (lineáris regresszió). Egy olyan regressziós modell, ahol a célváltozó lineáris függvénye a bemeneti változóknak.

Definíció (regressziós döntési fa). Hasonlóan a döntési fához, de a levelekben folytonos értékek találhatók, a köztes csúcsokban pedig egy-egy regressziós modell, ami egy jellemző mentén osztja szét az adatokat.

Definíció (Regressziós KNN). Hasonlóan az osztályozási KNN-hez, de a legközelebbi k példa alapján prediktálja az értékeket, lényegében a k példa célértékeinek átlagát veszi.

10 Túltanulás

Definíció (Torzítás). *A modell tanítóadatbázison számított hibája. Magas torzítás esetén a modell nem képes jól megtanulni a tanító adatokat.*

Definíció (Variancia). *A modell tesztadatbázison számított hibája és a tanító adatbázison számított hiba különbsége.*

Definíció (Torzítás-variancia dilemma). *A modell torzítása és varianciája közötti kiegyensúlyozás.*

Néhány gyakorlati példa:

- Döntési fa: a fa magassága
- Lináris gép: a hibafüggvényhez hozzávesszük a súlyok valamilyen függvényét (L1, L2 regularizáció)
- KNN: a k értéke
- Neurális háló: rétegek száma, neuronok száma, dropout, korai leállítás

Definíció (Validációs halmaz). *Meta-paraméterek finomhangolás során használatos. Az adathalmazt három részre bontjuk: tanító, validációs, teszt. A paraméterek a tanító és validációs halmazon kerülnek beállításra, majd a teszt halmazon kerül kiértékelésre a végső modell.*

11 Reinforcement learning

Mestint I. alapú elvek: Egy ágens interaktál a környezetével, és az állapot megváltozása után visszajelzést kap. Az ágens célja, hogy maximalizálja a hosszútávú jutalmat.

Definíció (Exploration vs. exploitation). *Exploration: új, még nem ismert lehetőségek keresése. Exploitation: már ismert, jól működő lehetőségek kihasználása.*

Definíció (ϵ -greedy stratégia). *Az ágens ϵ valószínűséggel véletlenszerűen választ egy lehetőséget (exploration), és $1 - \epsilon$ valószínűséggel választja ki a legjobbnak tűnő lehetőséget (exploitation).*

Definíció (Value-based módszerek). *Az érték-alapú módszerek egy adott állapotból kiinduló összes akcióhoz egy értéket tanulnak, ami azt próbálja megbecsülni, hogy az adott állapotból az adott akció hatására mennyi jutalmat lehet összeszedni a jövőben.*

Definíció (Policy-based módszerek). *A stratégia-alapú módszerek közvetlenül az egyes döntések valószínűségét próbálják megbecsülni a jövőben összeszedhető jutalom ismerete nélkül.*

Megjegyzés. *Amennyiben már rendelkezésre áll egy a saját feladatunkban használt adattípushoz (kép, szöveg, stb...) előre betanított modell, akkor elhagyva az utolsó réteget, egy megfelelő beágyazást kaphatunk, amit felhasználhatunk a saját feladatunkhoz.*

Mindenképp figyelni kell arra, hogy habár az adattípus azonos, nem biztos, a saját célunknak megfelelő beágyazást kapunk.