DIGITÁLIS BÖLCSÉSZETI KÖZPONT

Gépi tanulás 1.

Simon Eszter

2021. február 22.

PIM DBK

Tartalom

Racionalista és empirikus megközelítés

Racionalista és empirikus megközelítés

Racionalista:

- szabályalapú
- a nyelvész írja a szabályokat $\rightarrow$ a nyelvi információt expliciten adja át a számítógépnek

Empirikus:

- statisztikai alapú
- a számítógépes nyelvész az erőforrásokat adja oda a számítógépnek, ami azokat felhasználva magát tanítja

Kérdés: amikor számítógépes nyelvmodellt építünk, akkor korpuszadatokra vagy introspekcióra támaszkodunk?

A szabályalapú rendszerek előnyei és hátrányai

Előnyei:

- a fejlesztő nagyobb kontrollja a rendszer felett
- könnyebben értelmezhető visszacsatolás a rendszertől
- magas pontosság

Hátrányai:

- sok kézi munkát és nagy szakértelmet kíván
- nem hibatűrő
- bonyolult a fejlesztése, törékeny
- nehezen vihető át más nyelvre/doménre

A statisztikai alapú rendszerek előnyei és hátrányai

Előnyei:

- minden elemzéshez egy valószínűséget kapunk → rangsorolhatjuk őket → kiválaszthatjuk a leginkább odaillőt
- még akkor is adhat jó eredményt, ha a mögöttes nyelvmodell nem adekvát
- sokkal flexibilisebb megközelítés

Hátrányai:

- nagy mennyiségű annotált adatot igényelnek → a kézi munka nem tűnt el, csak átalakult
- a rendszer átvitele más nyelvre/doménre elég nagy teljesítménybeli visszaesést okoz

A statisztikai módszerek kategorizálása

- felügyelt (supervised)
- félig felügyelt (semi-supervised)
- felügyelet nélküli (unsupervised)

Felügyelet nélküli és félig felügyelt tanulás

Felügyelet nélküli tanulás

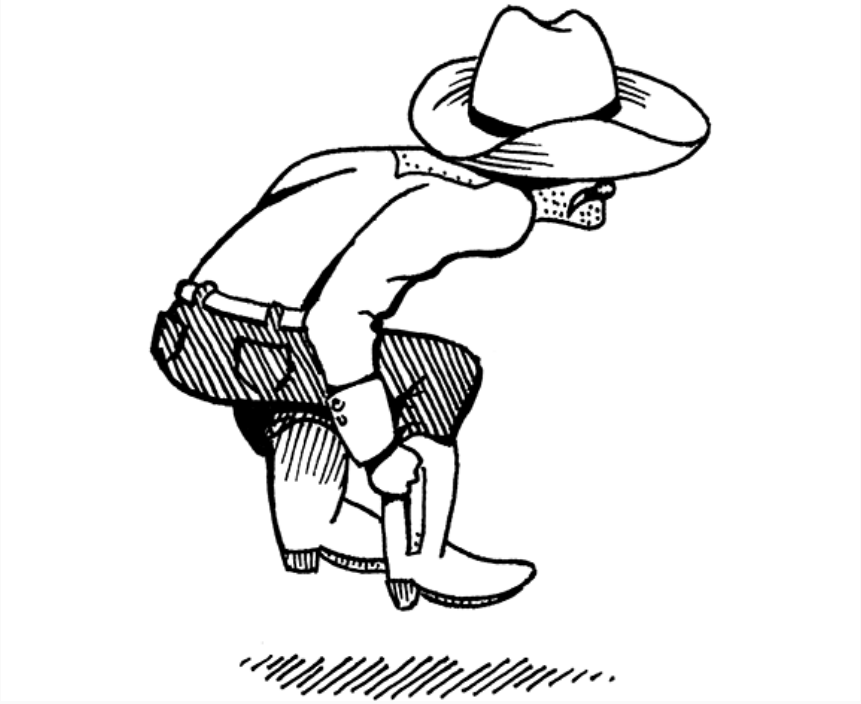
előfeldolgozott szöveg címkék nélkül → kérdés: mit lehet megtanulni a nyers szövegből?

Klaszterezés:
a hasonló grammatikai viselkedésű elemek fognak osztályokba csoportosulni

1. nincsenek előre definiált osztályok
   - ha új típusokat akarunk találni
2. előre megszabjuk az osztályok számát
   - ha az adott feladatban megszokott osztályok szerint akarjuk kiértékelni

Félig felügyelt tanulás

- címkézetlen szövegből tanul
- kézzel összeállított kiinduló halmaz (’seed’)
- bootstrapping
- az adatban előforduló természetes redundanciára építenek



A felügyelt gépi tanulás menete

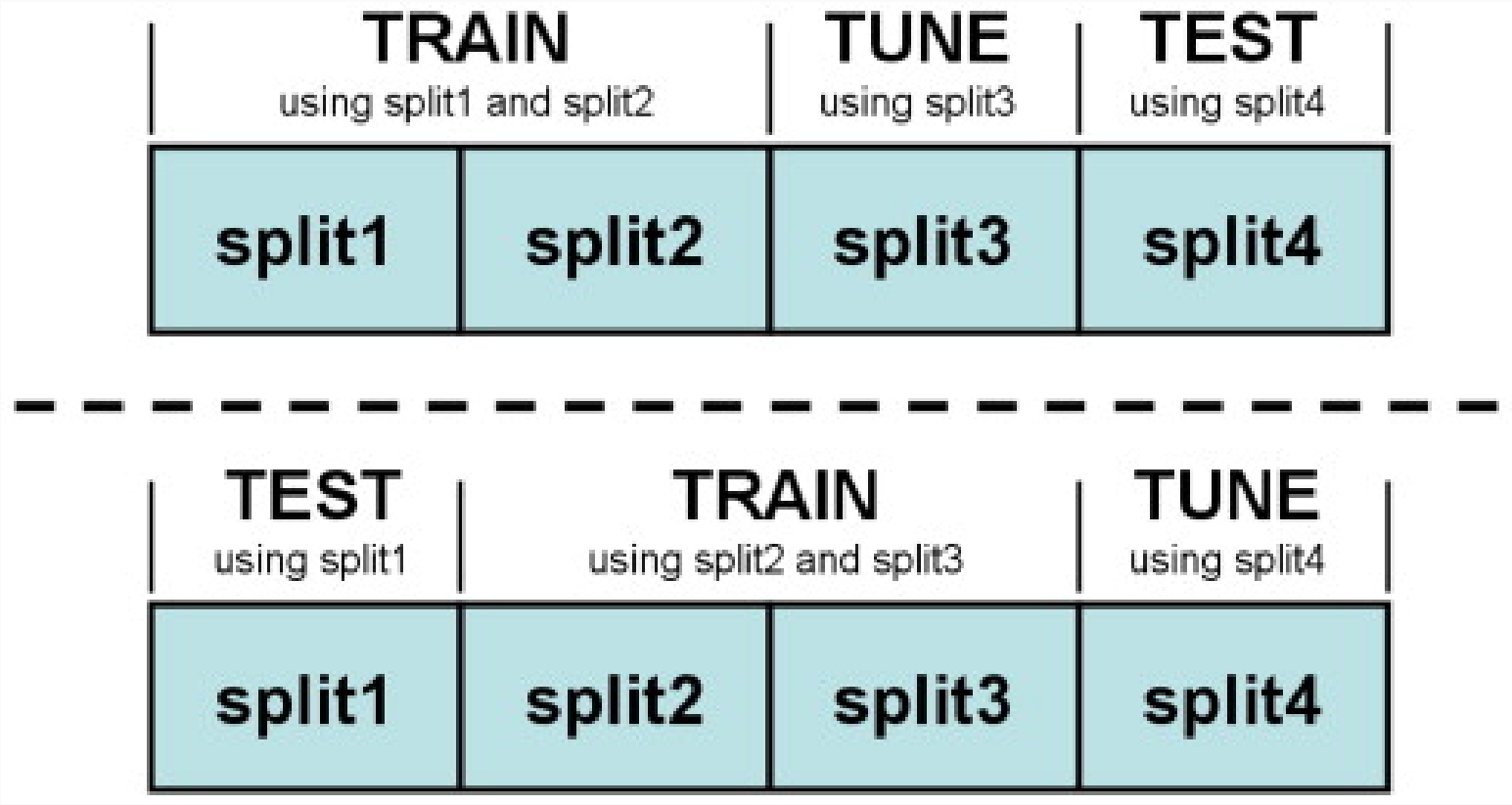
A felügyelt gépi tanulás alapvetése

- a felügyelt gépi tanulás azon a feltételezésen alapul, hogy az adatpontok egymástól független elemek, amelyeknek egyenletes az eloszlásuk
- feltesszük, hogy az eddig nem látott adatpontokra is igaz ez $\rightarrow$ így tudunk következtetni a már látott nyelvi elemekből a még nem látottakra
- nyelvi annotációval ellátott korpusz $\rightarrow$ ebből tanulja ki az adott adatpontokra jellemző jegyeket, és ez szolgál majd a kiértékelés alapjául is

Általános séma

1. gold standard korpusz
2. train–devel–test halmaz, keresztvalidáció
3. jegykinyerés
4. modellépítés
5. taggelés
6. kiértékelés

Train–devel–test & keresztvalidáció



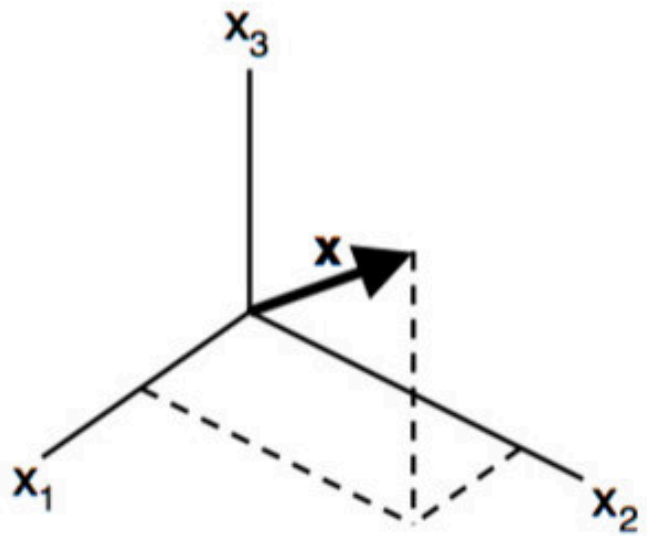
Túltanulás (overfitting)

- a teszthalmaz elemei nem szerepelhetnek a tanító halmazban
- fejlesztés közben mérni csak a develen szabad
- ha nem így teszünk, akkor az nem csak csalás, de a rendszerünk nem lesz képes általánosításokra $\rightarrow$ túlzottan rátanul az adott szövegre
- a rendszer a tanító halmazban levő random zajt reprezentálja, nem általánosít

Jegykinyerés (feature extraction)

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \\ x_d \end{bmatrix}$$

Feature vector



Feature space (3D)

További lépések

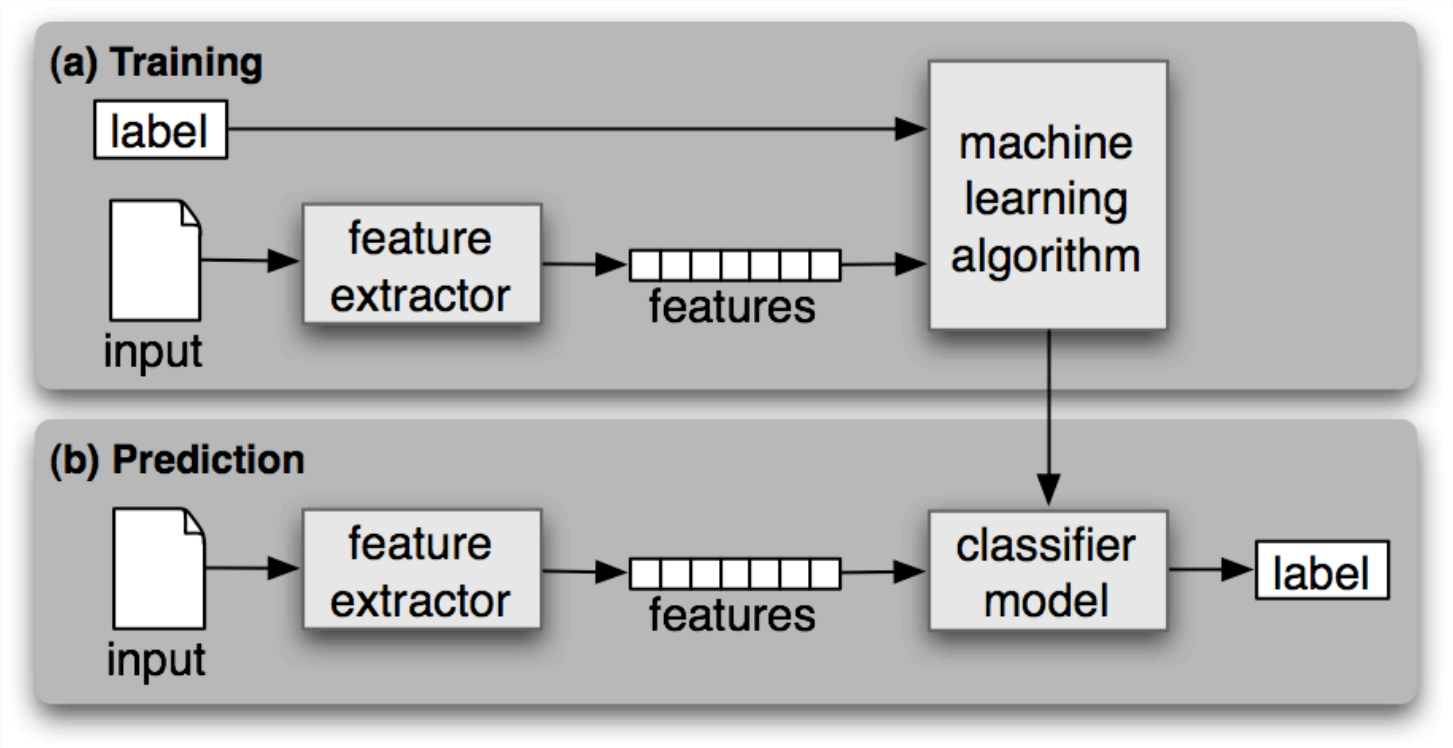
modellépítés:
jegy–címke párok, mindegyikhez egy súly hozzárendelve, ami azt mutatja meg, hogy az adott jegy mennyire van hatással arra, hogy az adott jeggyel rendelkező token az adott címkét kapja

taggelés:

1. a kiértékelő halmaz fícsörizálása
2. a kapott fícsörvektorok és a nyelvmodell alapján címkék kibocsátása

baseline:
a rendszertől minimálisan elvárt teljesítmény $\rightarrow$ jellemzően a leggyakoribb címke kiosztása minden adatpontra

Áttekintés


(a) Training
label
input
feature extractor
features
machine learning algorithm
(b) Prediction
input
feature extractor
features
classifier model
label

Címkézési feladat

Szekvenciális vs. tokenalapú címkézési feladat

szekvenciális
lánc-struktúra → időbeli vagy térbeli egymásutániság (pl. szavak egy mondatban) → a szekvencia minden egyes eleméhez címkét kell rendelni

tokenalapú
egy adatponthoz (egy tokenhez) egy címkét kell rendelni

Klasszifikáció vs. regresszió

klasszifikáció
kategóriacímkét bocsát ki (két vagy több diszkrét kategória)

regresszió
valós számot bocsát ki (folytonos)

logisztikus regresszió
valós számot bocsát ki, ami valószínűségként értelmezhető (két kategória)

multinomiális logisztikus regresszió
valós számot bocsát ki, ami valószínűségként értelmezhető (több kategória)

Maximum entrópia

- multinomiális logisztikus regresszió
- tokenalapú
- címkék fölötti teljes valószínűség-eloszlást bocsát ki

token	gold	C1	P1	C2	P2
tájékoztatta	O	O	1		
a	O	O	1		
Magyar	B-ORG	B-ORG	0.92	1-ORG	0.08
Nemzeti	I-ORG	I-ORG	0.96	O	0.04
Bank	E-ORG	E-ORG	1		
csütörtökön	O	O	1		
az	O	O	1		
MTI-t	1-ORG	B-ORG	0.35	1-ORG	0.65

Jegyek és paraméterek

n-gramok

George W. Bush

karakter *n*-gramok

unigramok: G, e, o, r, g, e, _, W, ., _, B, u, s, h

bigramok: Ge, eo, or, rg, ge, e_, _W, W., ._, _B, Bu, us, sh

trigramok: Geo, eor, org, rge, ge_, e_W, _W., W._, ._B, _Bu, Bus, ush

szó *n*-gramok

unigramok: George, W., Bush

bigramok: George W., W. Bush

trigramok: George W. Bush

címke *n*-gramok

unigramok: B-PER, I-PER, E-PER

bigramok: B-PER I-PER, I-PER E-PER

trigramok: B-PER I-PER E-PER

Rádiusz

- az aktuálisan vizsgált token nemcsak a saját magára vonatkozó fícsör–érték párokat kapja meg, hanem az előtte–utána levő *n* token fícsör–érték párjait is
- erre való a radius
- default NER: 3, NP: 5, de a config fájlban állítható, akár egyes fícsörökre külön-külön is
- ezzel a kontextus is figyelembe vehető

Mik a jegyek?

- a jegyek az adatpontok különféle tulajdonságait írják le
- a jegyeket a számítógépes nyelvész találja ki, definiálja és kódolja le (a “hagyományos” gépi tanulás során)
- a jegy hasznosságát az adat fogja meghatározni $\rightarrow$ a jegy megkülönböztető erejét ki kell mérni $\rightarrow$ utána lehet dönteni a jegy alkalmazásáról

A jegyek megkülönböztető ereje

- jegyek hozzáadása vagy paraméterek állítása egyesével $\rightarrow$ mérés $\rightarrow$ ha nem ront, akkor bennehagyjuk
- több jegyszelekciós módszer is létezik
  - a legegyszerűbb algoritmus: a fícsörtér minden lehetséges részhalmazát kimérni, és azt választani, amelyiknél a legnagyobb az F-mérték
  - a fícsörtér inkrementális bővítése (önkényesen vagy korábbi kísérletek eredményei alapján)