**Úvod do strojového učení**

Místo explicitních podmínek, dodáme vzorové data (trémovací data)

Pocitac se uci = zpresnovani spravnosti rozhodovani

Trenovaci data jsou od cloveka (anotace), popripade merenim [napr. pocasi]

Supervised Machine Learning (učení s učitelem)

Trénovací data spolu se správnými výsledky

Prediktor (funkce) – snaží se uhodnout spravnou hodnotu

Treninkova data = data na uceni a vytvareni modelů (prediktorů)

Testovaci data = data na rozpoznani lepsiho modelu (cim obsahlejsi a ruznorodejsi tim lepsi)

Ucici vs. Prediktorovy parametr

Development

Trenovaci data, pote testovaci dale development test data a overeni za pomoci unseen dat

Generalization error = ocekavana chyba (i pro budouci data)

Snaha aby byla co nejmensi

Nesnazit se o co nejnizsi training error, ale o obecnost funkce (modelu)

příznaky -skupiny

-přesnostni

-binarni (ano / ne)

🡨------------------11.10---------------------🡪

...

Entropie

je to cislo

velikost informace potrebna k popsani promene

zapisujeme H(Y|X) nebo H(C|A)

vlastnosti

normalita

konjuktivita

nezápornost a konečná

symetričnost

rekurzivnost

Hodnota entropie

-cim vyssi tim mensi rozdily hodnot (nejasnost rozhodovani) (mensi je lepsi)

Podminena entropie

Vzájemná informace je rozdil dvou entropií (jejich prunik)

Proces učení

schromazdeni dat

vytvoreni rozhodovaci stromu (alternativa “nahodne lesy”)

v kazdem uzlu je podminka (podle priznaku)

listy jsou vyresene priklady

faze: nejdrive roste (do hloubky), pote jeho ořezání a tvorba listů)

vyberu to štěpení které mi ho nejvic zmeni

Gini index

🡨------------------18.10---------------------🡪

Analyza dat a kvantitativni stavy

Pearson linear analyze

Koeficient stejnosti

Zahazovani nekvalitnich priznaků (neuzitecne)

Rizene vs nerizene uceni

Treba hledat strukturu

Grafové vzdálenosti:

Euklidovo - prime

Manhatanske – pravouhle

Hamming

Rozdeleni do clusteru (shluky)

Typy reseni clusteru:

Zeshora – podobnost + dalsi cinitele

Zezdola – podobnost dvou clusteru

Vazby clusteru

Vzdalenost dvou nejblizsich bodu - single

Vzdalenost dvou nejvzdalenejsich bodu - complete

Prumerna vzdalenost – average

🡨------------------25.10---------------------🡪

F jako funkce přesnosti

Testování hypotéz

Nulová hypotéza

Předpoklad výběru se štěstim

Statistické testování

Pravděpodobnost, že vyjde správně i když se vykonává random

Nulová hypotéza se **zamítne**, pokud je P-hodnota dostatecne mala

P-hodnota = (dosažená hladina testu)

α je míra rizika, že udělám chybu

🡨------------------15.11.---------------------🡪

Algoritmy strojového učení

Instant-based

Nepředpočítává při fázi učení

-hledání podobnosti v okolí (vzdálenost)

vzdálenosti: Euclideova, Manhattanská

-přidáváni penalizací

Naive Bayes classifier

Na bázi věruhodnosti (likehood)

Přiřazení četnosti a nerozhodnosti

Na bázi grafů

Princip maximální věruhodnosti

🡨------------------22.11.---------------------🡪

Kombinování různých typů učení

Bagging and bosting (algoritmus SU)