Curs 2 - Învățare supervizată Clasificare

1: Prezentare generală

- **Titlu:** Învățare supervizată – clasificare
- **Structura cursului:**
  - Definiția problemei de clasificare
  - Naïve Bayes
  - Regressie logistică
  - Arbori de clasificare
  - Overfitting & regularizare (Lasso, Ridge)
  - Metrici de evaluare: accuracy, precision-recall, ROC-AUC
  - Aplicații în lumea reală

2: Introducere în problemele de clasificare

- **Definiţie:** Prezicerea unei etichete sau categorii discrete
- **Exemple:**
  - Spam sau non-spam
  - Pacient bolnav sau sănătos
- **De ce clasificare?**
  - Folosită frecvent în diagnostic medical, filtre de e-mail, detectarea fraudei

3: Concepte cheie pentru clasificare

- **Caracteristici & etichete:**
  - Caracteristici = variabile de intrare (predictori)
  - Etichete = clasele țintă
- **Antrenare vs. Testare:**
  - Antrenare pe date etichetate
  - Testare pentru validarea performanței
- **Exemplu: Prezintă un set de date mic cu caracteristici precum vârsta, venitul și o etichetă binară care indică „purchased” sau „not purchased.”**

4: Clasificatorul Naïve Bayes

- **Ideea de bază:**
  - Utilizează teorema lui Bayes cu presupunerea independenței caracteristicilor
- **Avantaje:**
  - Simplu, rapid, funcționează bine cu date zgomotoase
- **Dezavantaje:**
  - Presupunerea independenței poate fi nerealistă în anumite domenii
- **Exemplu: Clasificarea recenziilor ca pozitive sau negative pe baza frecvenței cuvintelor**

Articol

5: Regressie logistică

- **Concept:**
  - Modelează probabilitatea ca o anumită clasă să fie rezultatul
  - Utilizează funcția logistică (grafic sigmoid)
- **Interpretabilitate:**
  - Coeficienții pot fi analizați pentru a vedea impactul caracteristicilor
- **Când se utilizează:**
  - Clasificare binară cu o graniță de decizie liniară
- **Exemplu: Prezicerea probabilității ca un client să facă o achiziție pe baza datelor demografice**
- **Articol**

6: Arbori de clasificare

- **Structură:**
  - Serie de reguli decizionale bazate pe valorile caracteristicilor
- **Avantaje:**
  - Ușor de vizualizat și interpretat
- **Dezavantaje:**
  - Predispus la suprapotrivire dacă nu este tăiat (pruned)
- **Exemplu: Un arbore de decizie pentru aprobarea unui împrumut pe baza caracteristicilor precum venit, scor de credit și raportul datorii/venit**

Articol

7: Overfitting & regularizare

- **Suprapotrivire:**
  - Modelul se potrivește zgomotului sau detaliilor specifice din datele de antrenare
  - Simptome: acurateţe mare pe antrenare, generalizare slabă
- **Regularizare:**
  - Penalizează coeficienţii mari pentru a reduce overfitting
  - Metode comune: Lasso (L1), Ridge (L2)
- **Articol

8: Lasso (L1) vs. Ridge (L2)

- **Lasso (L1):**
  - Poate aduce coeficienții la zero → selecția caracteristicilor
- **Ridge (L2):**
  - Coeficienții sunt micșorați spre zero, dar rareori devin zero
- **Când se utilizează:**
  - Lasso: preferi modele mai simple, cu mai puține caracteristici
  - Ridge: vrei să păstrezi toate caracteristicile, dar să reduci variația mare
- **Exemplu: Aplicarea regresiei logistice cu regularizare L1 sau L2 pe un set de date cu multe caracteristici și observarea caracteristicilor semnificative**
- **Articol**

9: Metrici de evaluare

- **Accuracy:**
  - Proporţia etichetelor prezise corect
  - Utilă când clasele sunt echilibrate
- **Precision & recall:**
  - Precision = dintre predicţiile pozitive, câte sunt corecte
  - Recall = dintre cazurile pozitive reale, câte au fost identificate
- **ROC & AUC:**
  - Graficul ratei adevărat pozitive vs. ratei fals pozitive
  - ROC = Receiver-operating characteristic curve
  - AUC = Aria sub curba ROC (măsură a performanței generale)
- **Exemplu: Prezintă o matrice de confuzie pentru o problemă de clasificare binară și calculează precizia, recall-ul și acuratețea**
- **Articol**

10: Alegerea metricii corecte

- **Considerații cheie:**
  - Dezechilibru de clase: acuratețea poate fi înșelătoare
  - Costul fals pozitive vs. fals negative
  - Raportarea mai multor metrici pentru claritate
- **Exemplu: Set de date dezechilibrat în domeniul sănătății, unde fals negativele pot fi foarte costisitoare**

10.1: Tuning de hiperparametri cu GridSearchCV

- **Ce este este ajustarea hiperparametrilor?**
  - Procesul de a găsi cel mai bun set de parametri (ex. rata de învățare, puterea de regularizare) care nu sunt învățați direct în timpul antrenării modelului.
  - Alegerea unor hiperparametri corespunzători poate îmbunătăți semnificativ performanța și capacitatea de generalizare a modelului.
- **Prezentare Generală GridSearchCV:**
  - Căutare exhaustivă peste valorile specificate pentru anumiți parametri.
  - Utilizează **cross-validation** intern:
    1. Împarte datele în mai multe fold-uri.
    2. Antrenează și validează modelul pentru fiecare combinație de parametri.
    3. Alege combinația care oferă cea mai bună performanță medie.
- **Exemplu:**
  - Exemplu: Arată cum tuning-ul pentru `max_depth` într-un Decision Tree (între 2 și 10) poate găsi o adâncime optimă care să echilibreze underfitting și overfitting.

11: Aplicații reale

- **Clasificare de text:**
  - Detectarea spamului, analiza sentimentelor
- **Diagnostic medical:**
  - Clasificarea bolilor, evaluarea riscurilor
- **Domeniul financiar:**
  - Scor de credit, detectarea fraudei
- **Exemplu: Cum folosește o bancă regresia logistică pentru a identifica tranzacțiile frauduloase**

12: Concluzii

- **Rezumat:**
  - Clasificarea este centrală pentru multe sarcini predictive
  - Există mai multe algoritmi, fiecare cu puncte forte și slabe
  - Măsurile de evaluare ghidează selecția modelului
- **Următorii pași:**
  - Experimentează cu diferite tehnici de clasificare
  - Concentrează-te pe regularizare și selecția corectă a metricilor
  - Explorează modele avansate de clasificare (ex. Random Forest, SVM)
- **Întrebări / Discuții**