Fundamentals of Machine Learning (4341603) - Summer 2024 Solution

Milav Dabgar

June 15, 2024

પ્રશ્ન 1(a) [3 ગુણ]

Define Machine Learning using suitable example?

જવાબ

મશીન લર્નિંગ આર્ટિફિશિયલ ઇન્ટેલિજન્સનો એક ભાગ છે જે કમ્પ્યુટર્સને ડેટામાંથી શીખવા અને દરેક કાર્ય માટે સ્પષ્ટ રીતે પ્રોગ્રામ કર્યા વિના નિર્ણયો લેવા માટે સક્ષમ બનાવે છે.

કોષ્ટક 1. મશીન લર્નિંગના મુખ્ય ઘટકો

ઘટક	વર્ણન
Data	ટ્રેનિંગ માટે ઉપયોગમાં લેવાતી ઇનપુટ માહિતી
Algorithm	પેટર્ન શીખતા ગાણિતિક મોડેલ
Training	અલ્ગોરિધમને શીખવવાની પ્રક્રિયા
Prediction	શીખેલા પેટર્ન આધારિત આઉટપુટ

ઉદાહરણ: ઇમેઇલ સ્પામ ડિટેક્શન સિસ્ટમ હજારો ઇમેઇલોમાંથી "Spam" અથવા "Not Spam" તરીકે લેબલ કરેલા ઇમેઇલોમાંથી શીખે છે અને નવા ઇમેઇલોને આપોઆપ વર્ગીકૃત કરે છે.

મેમરી ટ્રીક

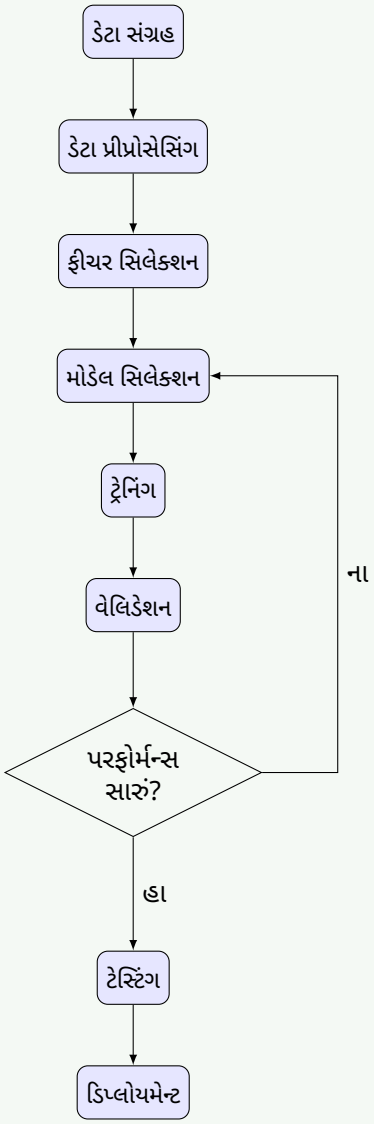
``Data Drives Decisions - ડેટા અલ્ગોરિધમને બુદ્ધિશાળી નિર્ણયો લેવા માટે પ્રશિક્ષિત કરે છે''

પ્રશ્ન 1(b) [4 ગુણ]

Explain the process of machine learning with the help of schematic representation

જવાબ

મશીન લર્નિંગ પ્રક્રિયામાં ડેટા સંગ્રહથી લઈને મોડેલ ડિપ્લોયમેન્ટ સુધીના વ્યવસ્થિત પગલાંઓનો સમાવેશ થાય છે.



આકૃતિ 1. મશીન લર્નિંગ પ્રક્રિયા

પ્રક્રિયાના પગલાં:

- **Data Collection**: સંબંધિત ડેટાસેટ એકત્રિત કરવું
- **Preprocessing**: ડેટાને સાફ અને તૈયાર કરવું
- **Training**: ટ્રેનિંગ ડેટાનો ઉપયોગ કરીને અલ્ગોરિધમને શીખવવું
- **Validation**: મોડેલની કામગીરીને ચકાસવી
- **Deployment**: વાસ્તવિક પ્રિડિક્શન માટે મોડેલનો ઉપયોગ

મેમરી ટ્રીક

``Computers Can Truly Think - Collect, Clean, Train, Test''

પ્રશ્ન 1(c) [7 ગુણ]

Explain different types of machine learning with suitable application.

જવાબ

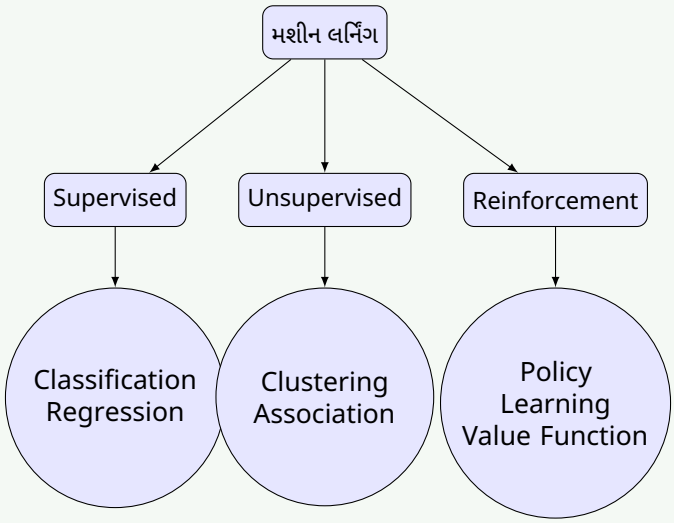
મશીન લર્નિંગ અલ્ગોરિધમ્સને લર્નિંગ એપ્રોચ અને ઉપલબ્ધ ડેટાના આધારે વર્ગીકૃત કરવામાં આવે છે.

કોષ્ટક 2. મશીન લર્નિંગના પ્રકારો

પ્રકાર	લર્નિંગ મેથડ	ડેટા આવશ્યકતા	ઉદાહરણ એપ્લિકેશન
Supervised	લેબલ્ડ ડેટાનો ઉપયોગ	ઇનપુટ-આઉટપુટ જોડીઓ	ઇમેઇલ ક્લાસિફિકેશન
Unsupervised	છુપાયેલા પેટર્ન શોધે	માત્ર ઇનપુટ ડેટા	કસ્ટમર સેગમેન્ટેશન
Reinforcement	રિવોર્ડ્સ દ્વારા શીખે	એન્વાયર્નમેન્ટ ફીડબેક	ગેમ પ્લેઇંગ AI

એપ્લિકેશન્સ:

- **Supervised Learning**: મેડિકલ ડાયગ્નોસિસ, ઇમેજ રેકોગ્નિશન, ફ્રોડ ડિટેક્શન
- **Unsupervised Learning**: માર્કેટ રિસર્ચ, એનોમેલી ડિટેક્શન, રેકમેન્ડેશન સિસ્ટમ્સ
- **Reinforcement Learning**: ઓટોનોમસ વેહિકલ્સ, રોબોટિક્સ, સ્ટ્રેટેજિક ગેમ્સ



આકૃતિ 2. મશીન લર્નિંગના પ્રકારો

મેમરી ટ્રીક

``Students Usually Remember - Supervised, Unsupervised, Reinforcement''

પ્રશ્ન 1(c) OR [7 ગુણ]

What are various issues with machine learning? List three problems that are not to be solved using machine learning.

જવાબ

કોષ્ટક 3. મશીન લર્નિંગની સમસ્યાઓ

સમસ્યા કેટેગરી	વર્ણન	અસર
Data Quality	અધૂરો, નોઇઝી, પક્ષપાતી ડેટા	નબળું મોડેલ પરફોર્મન્સ
Overfitting	મોડેલ ટ્રેનિંગ ડેટાને યાદ રાખે છે	નબળું જનરલાઇઝેશન
Computational	ઉચ્ચ પ્રોસેસિંગ આવશ્યકતાઓ	રિસોર્સ મર્યાદાઓ
Interpretability	બ્લેક બોક્સ મોડેલ્સ	પારદર્શિતાનો અભાવ

ML માટે અનુપયુક્ત સમસ્યાઓ:

1. **Simple rule-based tasks** - મૂળભૂત ગણતરીઓ, સિમ્પલ if-then લોજિક
2. **Ethical decisions** - માનવીય મૂલ્યોની આવશ્યકતા ધરાવતા નૈતિક નિર્ણયો
3. **Creative expression** - માનવીય લાગણીની આવશ્યકતા ધરાવતી મૂળ કલાત્મક સર્જના

અન્ય સમસ્યાઓ:
- **Privacy concerns**: સંવેદનશીલ ડેટા હેન્ડલિંગ
- **Bias propagation**: અન્યાયકારક અલ્ગોરિધમિક નિર્ણયો
- **Feature selection**: સંબંધિત ઇનપુટ વેરિએબલ્સ પસંદ કરવા

મેમરી ટ્રીક

``Data Drives Quality - ડેટા ક્વોલિટી સીધી રીતે મોડેલ ક્વોલિટીને અસર કરે છે''

પ્રશ્ન 2(a) [3 ગુણ]

Give a summarized view of different types of data in a typical machine learning problem.

જવાબ

કોષ્ટક 4. મશીન લર્નિંગમાં ડેટા પ્રકારો

ડેટા પ્રકાર	વર્ણન	ઉદાહરણ
Numerical	માત્રાત્મક મૂલ્યો	ઉંમર: 25, ઊંચાઈ: 170cm
Categorical	અસ્પષ્ટ કેટેગરીઓ	રંગ: લાલ, વાદળી, લીલો
Ordinal	ક્રમબદ્ધ કેટેગરીઓ	રેટિંગ: નબળું, સારું, ઉત્તમ
Binary	બે શક્ય મૂલ્યો	લિંગ: પુરુષ/સ્ત્રી

લક્ષણો:
- **Structured**: ટેબલોમાં વ્યવસ્થિત (ડેટાબેસેસ, સ્પ્રેડશીટ્સ)
- **Unstructured**: ઇમેજ, ટેક્સ્ટ, ઓડિયો ફાઇલો
- **Time-series**: સમય પર ડેટા પોઇન્ટ્સ

મેમરી ટ્રીક

``Numbers Count Better Than Words - Numerical, Categorical, Binary, Text''

પ્રશ્ન 2(b) [4 ગુણ]

Calculate variance for both attributes. Determine which attribute is spread out around mean.

જવાબ

આપેલ ડેટા:
- Attribute 1: 32, 37, 47, 50, 59
- Attribute 2: 48, 40, 41, 47, 49

ગણતરીઓ:

Attribute 1:
- Mean = $(32 + 37 + 47 + 50 + 59)/5 = 225/5 = 45$
- Variance = $[(32 - 45)^2 + (37 - 45)^2 + (47 - 45)^2 + (50 - 45)^2 + (59 - 45)^2]/5$
- Variance = $[169 + 64 + 4 + 25 + 196]/5 = 458/5 = 91.6$

Attribute 2:
- Mean = $(48 + 40 + 41 + 47 + 49)/5 = 225/5 = 45$

- Variance = $[(48-45)^2 + (40-45)^2 + (41-45)^2 + (47-45)^2 + (49-45)^2]/5$
- Variance = $[9 + 25 + 16 + 4 + 16]/5 = 70/5 = 14$

પરિણામ: Attribute 1 (variance = 91.6) એ Attribute 2 (variance = 14) કરતાં વધુ સ્પ્રેડ આઉટ છે.

મેમરી ટ્રીક

``Higher Variance Shows Spread - વધુ વેરિયન્સ વધુ વિખેરાઈને દર્શાવે છે''

પ્રશ્ન 2(c) [7 ગુણ]

List Factors that lead to data quality issue. How to handle outliers and missing values.

જવાબ

કોષ્ટક 5. ડેટા ગુણવત્તા સમસ્યાઓ

ફેક્ટર	કારણ	સોલ્યુશન
Incompleteness	મિસિંગ ડેટા કલેક્શન	ઇમ્પ્યુટેશન ટેકનિક્સ
Inconsistency	વિવિધ ડેટા ફોર્મેટ્સ	સ્ટેન્ડર્ડાઇઝેશન
Inaccuracy	હ્યુમન/સેન્સર એરર્સ	વેલિડેશન રૂલ્સ
Noise	રેન્ડમ વેરિએશન્સ	ફિલ્ટરિંગ મેથડ્સ

આઉટલાયર્સ હેન્ડલ કરવું:
- **Detection**: સ્ટેટિસ્ટિકલ મેથડ્સ (Z-score, IQR)
- **Treatment**: એક્સ્ટ્રીમ વેલ્યુઝને દૂર, ટ્રાન્સફોર્મ, અથવા કેપ કરવી
- **Visualization**: બોક્સ પ્લોટ્સ, સ્કેટર પ્લોટ્સ

મિસિંગ વેલ્યુઝ હેન્ડલ કરવું:
- **Deletion**: અપૂર્ણ રેકોર્ડ્સ રીમૂવ કરવા
- **Imputation**: મીન, મીડિયન, અથવા મોડ સાથે ભરવું
- **Prediction**: મિસિંગ વેલ્યુઝની આગાહી કરવા માટે ML નો ઉપયોગ

કોડ ઉદાહરણ:

```
# Handle missing values
df.fillna(df.mean()) # Mean imputation
df.dropna()          # Remove missing rows
```

મેમરી ટ્રીક

``Clean Data Makes Models - સાફ ડેટા બેહતર મોડેલ્સ બનાવે છે''

પ્રશ્ન 2(a) OR [3 ગુણ]

Give different machine learning activities.

જવાબ

કોષ્ટક 6. મશીન લર્નિંગ પ્રવૃત્તિઓ

પ્રવૃત્તિ	હેતુ	ઉદાહરણ
Data Collection	સંબંધિત માહિતી એકત્રિત કરવી	સર્વે, સેન્સર્સ, ડેટાબેસેસ
Data Preprocessing	ડેટાને સાફ અને તૈયાર કરવું	નોઇઝ રીમૂવ કરવું, મિસિંગ વેલ્યુઝ
Feature Engineering	અર્થપૂર્ણ વેરિએબલ્સ બનાવવા	રો ડેટામાંથી ફીચર્સ એક્સ્ટ્રેક્ટ કરવા
Model Training	અલ્ગોરિધમને પેટર્ન શીખવવા	ટ્રેનિંગ ડેટાસેટનો ઉપયોગ
Model Evaluation	પરફોર્મન્સ આકારણી	ટેસ્ટ એક્યુરસી, પ્રિસિઝન, રિકોલ
Model Deployment	મોડેલને પ્રોડક્શનમાં મૂકવું	વેબ સર્વિસેસ, મોબાઇલ એપ્સ

મુખ્ય પ્રવૃત્તિઓ:
- **Exploratory Data Analysis**: ડેટા પેટર્ન સમજવા
- **Hyperparameter Tuning**: મોડેલ સેટિંગ્સ ઓપ્ટિમાઇઝ કરવા
- **Cross-validation**: મજબૂત પરફોર્મન્સ આકારણી

મેમરી ટ્રીક

``Data Models Perform Excellently - Data preparation, Model building, Performance evaluation, Execution''

પ્રશ્ન 2(b) OR [4 ગુણ]

Calculate mean and median of the following numbers: 12,15,18,20,22,24,28,30

જવાબ

આપેલ સંખ્યાઓ: 12, 15, 18, 20, 22, 24, 28, 30
Mean ગણતરી: Mean = $(12 + 15 + 18 + 20 + 22 + 24 + 28 + 30)/8 = 169/8 = 21.125$
Median ગણતરી:
- સંખ્યાઓ પહેલેથી સૉર્ટ કરેલી છે: 12, 15, 18, 20, 22, 24, 28, 30
- સમ સંખ્યા (8 સંખ્યાઓ)
- Median = (4મી સંખ્યા + 5મી સંખ્યા)/2 = $(20 + 22)/2 = 21$

કોષ્ટક 7. સ્ટેટિસ્ટિકલ સમરી

માપદંડ	મૂલ્ય	વર્ણન
Mean	21.125	સરેરાશ મૂલ્ય
Median	21	મધ્યમ મૂલ્ય
Count	8	કુલ સંખ્યાઓ

મેમરી ટ્રીક

``Middle Makes Median - Middle value gives median''

પ્રશ્ન 2(c) OR [7 ગુણ]

Write a short note on dimensionality reduction and feature subset selection in context with data preprocessing.

જવાબ

Dimensionality Reduction અપ્રસ્તુત ફીચર્સને દૂર કરે છે અને કોમ્પ્યુટેશનલ જટિલતા ઘટાડે છે જ્યારે મહત્વપૂર્ણ માહિતી જાળવી રાખે છે.

કોષ્ટક 8. ડાયમેન્શનાલિટી રિડક્શન ટેકનિક્સ

ટેકનિક	મેથડ	વપરાશ
PCA	Principal Component Analysis	લીનિયર રિડક્શન
LDA	Linear Discriminant Analysis	ક્લાસિફિકેશન ટાસ્ક્સ
t-SNE	Non-linear embedding	વિઝ્યુઅલાઇઝેશન
Feature Selection	મહત્વપૂર્ણ ફીચર્સ પસંદ કરવા	ઓવરફિટિંગ ઘટાડવું

Feature Subset Selection Methods:
- **Filter Methods**: Statistical tests, correlation analysis
- **Wrapper Methods**: Forward/backward selection
- **Embedded Methods**: LASSO, Ridge regression

ફાયદાઓ:
- **Computational Efficiency**: ઝડપી ટ્રેનિંગ અને પ્રિડિક્શન
- **Storage Reduction**: ઓછી મેમરી આવશ્યકતાઓ
- **Noise Reduction**: અપ્રસ્તુત ફીચર્સ દૂર કરવા
- **Visualization**: 2D/3D પ્લોટિંગ સક્ષમ કરવું

```
from sklearn.decomposition import PCA
pca = PCA(n_components=2)
reduced_data = pca.fit_transform(data)
```

મેમરી ટ્રીક

``Reduce Features, Improve Performance - ઓછા ફીચર્સ ઘણીવાર બેહતર મોડેલ્સ તરફ દોરી જાય છે''

પ્રશ્ન 3(a) [3 ગુણ]

Does bias affect the performance of the ML model? Explain briefly.

જવાબ

હા, બાયસ પ્રિડિક્શન્સમાં સિસ્ટેમેટિક એરર્સ બનાવીને ML મોડેલના પરફોર્મન્સને નોંધપાત્ર રીતે અસર કરે છે.

કોષ્ટક 9. બાયસના પ્રકારો

બાયસ પ્રકાર	વર્ણન	અસર
Selection Bias	બિન-પ્રતિનિધિત્વકારી ડેટા	નબળું જનરલાઇઝેશન
Confirmation Bias	અપેક્ષિત પરિણામોની તરફેણ	ત્રાંસા નિષ્કર્ષો
Algorithmic Bias	મોડેલ ધારણાઓ	અન્યાયકારક પ્રિડિક્શન્સ

પરફોર્મન્સ પર અસરો:
- **Underfitting**: ઉચ્ચ બાયસ અતિ સરળ મોડેલ્સ તરફ દોરી જાય છે
- **Poor Accuracy**: સિસ્ટેમેટિક એરર્સ એકંદર પરફોર્મન્સ ઘટાડે છે
- **Unfair Decisions**: પક્ષપાતી મોડેલ્સ જૂથો સામે ભેદભાવ કરે છે

મેમરી ટ્રીક

``Bias Breaks Better Performance - બાયસ મોડેલની અસરકારકતા ઘટાડે છે''

પ્રશ્ન 3(b) [4 ગુણ]

Compare cross-validation and bootstrap sampling

જવાબ

કોષ્ટક 10. Cross-validation vs Bootstrap Sampling

પાસું	Cross-validation	Bootstrap Sampling
મેથડ	ડેટાને ફોલ્ડ્સમાં વિભાજિત કરવું	રિપ્લેસમેન્ટ સાથે સેમ્પલ કરવું
ડેટા ઉપયોગ	બધો ડેટા વાપરે છે	મલ્ટિપલ સેમ્પલ્સ બનાવે છે
હેતુ	મોડેલ ઇવેલ્યુએશન	અનિશ્ચિતતાનો અંદાજ
ઓવરલેપ	સેટ્સ વચ્ચે કોઈ ઓવરલેપ નથી	ડુપ્લિકેટ સેમ્પલ્સની મંજૂરી

મુખ્ય તફાવત:
- **Cross-validation**: ડેટાને k સમાન ભાગોમાં વહેંચે છે. k-1 ભાગોમાં ટ્રેન કરે છે, 1 ભાગમાં ટેસ્ટ કરે છે.
- **Bootstrap Sampling**: રિપ્લેસમેન્ટ સાથે રેન્ડમ સેમ્પલ્સ બનાવે છે. સમાન સાઇઝના મલ્ટિપલ ડેટાસેટ્સ જનરેટ કરે છે.

મેમરી ટ્રીક

``Cross Checks, Bootstrap Builds - Cross-validation checks performance, Bootstrap builds confidence''

પ્રશ્ન 3(c) [7 ગુણ]

Confusion Matrix Calculation and Metrics

જવાબ

આપેલ માહિતી:
- True Positive (TP): 83
- False Positive (FP): 7
- False Negative (FN): 5
- True Negative (TN): 5

	Predicted Buy	Predicted No Buy
Actually Buy	83 (TP)	5 (FN)
Actually No Buy	7 (FP)	5 (TN)

ગણતરીઓ:
a) Error Rate: Error Rate = $(FP + FN)/Total = (7 + 5)/100 = 0.12 = 12\%$
b) Precision: Precision = $TP/(TP + FP) = 83/(83 + 7) = 83/90 = 0.922 = 92.2\%$
c) Recall: Recall = $TP/(TP + FN) = 83/(83 + 5) = 83/88 = 0.943 = 94.3\%$
d) F-measure: F-measure = $2 \times (Precision \times Recall)/(Precision + Recall)$ F-measure = $2 \times (0.922 \times 0.943)/(0.922 + 0.943) = 0.932 = 93.2\%$

મેમરી ટ્રીક

``Perfect Recall Finds Everyone - Precision measures accuracy, Recall finds all positives''

પ્રશ્ન 3(a) OR [3 ગુણ]

Define in brief: a) Target function b) Cost function c) Loss Function

જવાબ

કોષ્ટક 11. ફંક્શન વ્યાખ્યાઓ

ફંક્શન	વ્યાખ્યા	હેતુ
Target Function	ઇનપુટથી આઉટપુટ સુધીની આદર્શ મેપિંગ	આપણે શું શીખવા માગીએ છીએ
Cost Function	એકંદર મોડેલ એરરને માપે છે	કુલ પરફોર્મન્સનું મૂલ્યાંકન
Loss Function	એક પ્રિડિક્શન માટે એરર માપે છે	વ્યક્તિગત પ્રિડિક્શન એરર

સંબંધ: Cost function સામાન્ય રીતે તમામ ટ્રેનિંગ ઉદાહરણોમાં લોસ ફંક્શન્સની સરેરાશ હોય છે.

મેમરી ટ્રીક

``Target Costs Less - Target function is ideal, Cost function measures overall error, Loss function measures individual error''

પ્રશ્ન 3(b) OR [4 ગુણ]

Explain balanced fit, underfit and overfit

જવાબ

કોષ્ટક 12. મોડેલ ફિટિંગ પ્રકારો

ફિટ પ્રકાર	Training Error	Validation Error	લક્ષણો
Underfit	ઊંચો	ઊંચો	ખૂબ સાદું મોડેલ
Balanced Fit	નીચો	નીચો	આદર્શ જટિલતા
Overfit	ખૂબ નીચો	ઊંચો	ખૂબ જટિલ મોડેલ

આકૃતિ 3. મોડેલ જટિલતા સ્પેક્ટ્રમ

સોલ્યુશન્સ:
- **Underfit**: મોડેલ જટિલતા વધારવી, ફીચર્સ ઉમેરવા
- **Overfit**: રેગ્યુલરાઇઝેશન, ક્રોસ-વેલિડેશન, વધુ ડેટા

મેમરી ટ્રીક

``Balance Brings Best Results - સંતુલિત મોડેલ્સ નવા ડેટા પર શ્રેષ્ઠ પરફોર્મ કરે છે''

પ્રશ્ન 4(a) [3 ગુણ]

Give classification learning steps.

જવાબ

કોષ્ટક 13. ક્લાસિફિકેશન લર્નિંગ સ્ટેપ્સ

સ્ટેપ	વર્ણન	હેતુ
Data Collection	લેબલ્ડ ઉદાહરણો એકત્રિત કરવા	ટ્રેનિંગ મટેરિયલ પ્રદાન કરવું
Preprocessing	ડેટાને સાફ અને તૈયાર કરવું	ડેટા ગુણવત્તા સુધારવી
Feature Selection	સંબંધિત એટ્રિબ્યુટ્સ પસંદ કરવા	જટિલતા ઘટાડવી
Model Training	ટ્રેનિંગ ડેટામાંથી શીખવું	ક્લાસિફાયર બનાવવું
Evaluation	મોડેલ પરફોર્મન્સ ટેસ્ટ કરવું	ચોકસાઈ આકારવી
Deployment	નવી આગાહીઓ માટે ઉપયોગ	પ્રેક્ટિકલ એપ્લિકેશન

મેમરી ટ્રીક

``Data Preparation Facilitates Model Excellence - Data prep, Feature selection, Model training, Evaluation''

પ્રશ્ન 4(b) [4 ગુણ]

Linear Relationship Calculation

જવાબ

આપેલ ડેટા: Hours (X) vs Exam Score (Y)
Linear Regression ગણતરી:
સ્ટેપ 1: Mean કેલ્ક્યુલેટ કરવા
- $\bar{X} = (2+3+4+5+6)/5 = 4$
- $\bar{Y} = (85+80+75+70+60)/5 = 74$

સ્ટેપ 2: Slope (b) કેલ્ક્યુલેટ કરવું
- ન્યુમેરેટર = $\sum(X-\bar{X})(Y-\bar{Y}) = -60$
- ડિનોમિનેટર = $\sum(X-\bar{X})^2 = 10$
- $b = -60/10 = -6$

સ્ટેપ 3: Intercept (a) કેલ્ક્યુલેટ કરવું
- $a = \bar{Y} - b \times \bar{X} = 74 - (-6) \times 4 = 74 + 24 = 98$

Linear Equation: $Y = 98 - 6X$
અર્થઘટન: સ્માર્ટફોન ઉપયોગના દરેક વધારાના કલાક માટે, પરીક્ષા સ્કોર 6 પોઇન્ટ ઘટે છે.

મેમરી ટ્રીક

``More Phone, Less Score - Negative correlation between phone use and grades''

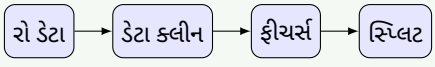
પ્રશ્ન 4(c) [7 ગુણ]

Explain classification steps in detail

જવાબ

Classification એ સુપરવાઇઝ્ડ લર્નિંગ પ્રક્રિયા છે જે ઇનપુટ ડેટાને પૂર્વનિર્ધારિત કેટેગરીઓમાં સોંપે છે.
વિગતવાર ક્લાસિફિકેશન સ્ટેપ્સ:
1. Problem Definition
- ક્લાસો અને ઉદ્દેશ્યો વ્યાખ્યાયિત કરવા
- ઇનપુટ ફીચર્સ અને ટાર્ગેટ વેરિએબલ ઓળખવા

2. Data Collection and Preparation



3. **Feature Engineering**
   - **Feature Selection**: સંબંધિત એટ્રિબ્યુટ્સ પસંદ કરવા
   - **Normalization**: ફીચર્સને સમાન રેન્જમાં સ્કેલ કરવા
4. **Model Selection and Training**

કોષ્ટક 14. સામાન્ય ક્લાસિફિકેશન અલ્ગોરિધમ્સ

અલ્ગોરિધમ	શ્રેષ્ઠ માટે	ફાયદાઓ
Decision Tree	ઇન્ટરપ્રિટેબલ રૂલ્સ	સમજવામાં સરળ
SVM	હાઇ-ડાયમેન્શનલ ડેટા	સારું જનરલાઇઝેશન
Neural Networks	જટિલ પેટર્ન્સ	ઉચ્ચ ચોકસાઈ

5. **Model Evaluation**
   - **Confusion Matrix**: વિગતવાર પરફોર્મન્સ એનાલિસિસ
   - **Metrics**: Accuracy, Precision, Recall, F1-score
6. **Final Evaluation and Deployment**
   - અદ્રશ્ય ડેટા પર ટેસ્ટ કરવું
   - પ્રોડક્શન ઉપયોગ માટે મોડેલ ડિપ્લોય કરવું

મેમરી ટ્રીક

``Proper Data Modeling Evaluates Performance Thoroughly - Problem definition, Data prep, Modeling, Evaluation, Performance testing, Tuning''

પ્રશ્ન 4(a) OR [3 ગુણ]

Does the choice of the k value influence the performance of the KNN algorithm? Explain briefly

જવાબ

હા, k વેલ્યુ KNN અલ્ગોરિધમના પરફોર્મન્સને નોંધપાત્ર રીતે પ્રભાવિત કરે છે.

કોષ્ટક 15. K વેલ્યુની અસર

K વેલ્યુ	અસર	પરફોર્મન્સ
Small K (k=1)	નોઇઝ પ્રત્યે સંવેદનશીલ	High variance, low bias
Medium K	સંતુલિત નિર્ણયો	આદર્શ પરફોર્મન્સ
Large K	સ્મૂથ બાઉન્ડરીઝ	Low variance, high bias

સિલેક્શન વ્યૂહરચના: આદર્શ k શોધવા માટે ક્રોસ-વેલિડેશનનો ઉપયોગ કરો, ઘણીવાર $k = \sqrt{n}$ થી શરૂઆત.

મેમરી ટ્રીક

``Small K Varies, Large K Smooths - Small k creates variance, large k creates smooth boundaries''

પ્રશ્ન 4(b) OR [4 ગુણ]

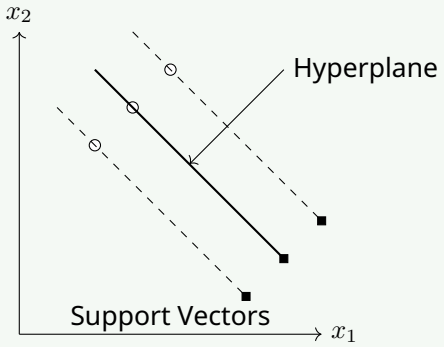
Define Support Vectors in the SVM model.

જવાબ

Support Vectors એ મહત્વપૂર્ણ ડેટા પોઇન્ટ્સ છે જે ડિસિઝન બાઉન્ડરી (hyperplane)ની સૌથી નજીક આવેલા હોય છે.

કોષ્ટક 16. Support Vector લક્ષણો

પાસું	વર્ણન	મહત્વ
Link	હાયપરપ્લેનની સૌથી નજીકના પોઇન્ટ્સ	ડિસિઝન બાઉન્ડરી વ્યાખ્યાયિત કરે
Distance	બાઉન્ડરીથી સમાન અંતર	Maximize margin
Role	હાયપરપ્લેનને સપોર્ટ કરે	આદર્શ વિભાજન નક્કી કરે



આકૃતિ 4. SVM હાયપરપ્લેન અને સપોર્ટ વેક્ટર્સ

મેમરી ટ્રીક

``Support Vectors Support Decisions - These vectors support the decision boundary''

પ્રશ્ન 4(c) OR [7 ગુણ]

Explain logistic regression in detail.

જવાબ

Logistic Regression એ બાઇનરી ક્લાસિફિકેશન માટે વપરાતી સ્ટેટિસ્ટિકલ મેથડ છે.
ગાણિતિક આધાર: Sigmoid Function: $\sigma(z) = 1/(1 + e^{-z})$ જ્યાં $z = \beta_0 + \beta_1 x_1 + ...$

કોષ્ટક 17. Linear vs Logistic Regression

પાસું	Linear Regression	Logistic Regression
Output	સતત મૂલ્યો	સંભાવનાઓ (0-1)
Function	Linear	Sigmoid (S-curve)
Purpose	આગાહી	ક્લાસિફિકેશન
Error	Mean Squared Error	Log-likelihood

મુખ્ય ઘટકો:
- **Logistic Function**: S-આકારનો કર્વ જે મૂલ્યોને [0,1] માં મેપ કરે છે.
- **Decision Rule**: જો $P(y = 1|x) > 0.5$, તો પોઝિટિવ તરીકે ક્લાસિફાય કરવું.
- **Training**: Maximum Likelihood Estimation નો ઉપયોગ કરે છે.

એપ્લિકેશન્સ: મેડિકલ ડાયગ્નોસિસ, ઇમેઇલ સ્પામ ડિટેક્શન, ક્રેડિટ એપ્રૂવલ.

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
model.fit(X_train, y_train)
```

```
probabilities = model.predict_proba(X_test)
```

મેમરી ટ્રીક

``Sigmoid Squashes Infinite Input - Sigmoid function converts any real number to probability''

પ્રશ્ન 5(a) [3 ગુણ]

Write a short note on Matplotlib python library.

જવાબ

Matplotlib એ વિઝ્યુઅલાઇઝેશન બનાવવા માટેની વ્યાપક Python લાઇબ્રેરી છે.

કોષ્ટક 18. Matplotlib મુખ્ય ફીચર્સ

ફીચર	હેતુ	ઉદાહરણ
Pyplot	MATLAB-જેવું ઇન્ટરફેસ	લાઇન પ્લોટ્સ
Formats	વિવિધ ફોર્મેટમાં સેવ કરવું	PNG, PDF, SVG
Subplots	એક ફિગરમાં મલ્ટિપલ પ્લોટ્સ	ગ્રિડ એરેન્જમેન્ટ્સ

મૂળભૂત ઉપયોગ:

```
import matplotlib.pyplot as plt
plt.plot(x, y)
plt.show()
```

મેમરી ટ્રીક

``Matplotlib Makes Pretty Plots - Essential tool for data visualization''

પ્રશ્ન 5(b) [4 ગુણ]

K-means clustering for two-dimensional data

જવાબ

આપેલ પોઇન્ટ્સ: બે જૂથો સ્પષ્ટ રીતે અલગ પડે છે. Group 1: (2,3) થી (8,3). Group 2: (25,20) થી (30,20).
અલ્ગોરિધમ સ્ટેપ્સ: Step 1: સેન્ટ્રોઇડ્સ ઇનિશિયલાઇઝ કરવા C1 = (4, 3), C2 = (27, 20)
Step 2: પોઇન્ટ્સ સોંપવા પોઇન્ટ્સ (2,3)...(8,3) C1 ની નજીક છે. પોઇન્ટ્સ (25,20)...(30,20) C2 ની નજીક છે.
Step 3: સેન્ટ્રોઇડ્સ અપડેટ કરવા નવું C1 = Group 1 નું Average = (5, 3) નવું C2 = Group 2 નું Average = (27.5, 20)
અંતિમ ક્લસ્ટર્સ: Cluster 1: ડાબા જૂથના પોઇન્ટ્સ. Cluster 2: જમણા જૂથના પોઇન્ટ્સ.

મેમરી ટ્રીક

``Centroids Attract Nearest Neighbors - Points join closest centroid''

પ્રશ્ન 5(c) [7 ગુણ]

Give functions and its use of Scikit-learn

જવાબ

a) Data Preprocessing:
- StandardScaler(): ફીચર્સને નોર્મલાઇઝ કરવા
- train_test_split(): ડેટાસેટ સ્પ્લિટ કરવું

b) Model Selection:
- GridSearchCV(): હાયપરપેરામીટર ટ્યુનિંગ
- cross_val_score(): ક્રોસ-વેલિડેશન

c) Model Evaluation:
- accuracy_score(): એકંદર ચોકસાઈ
- confusion_matrix(): એરર એનાલિસિસ
- classification_report(): વ્યાપક મેટ્રિક્સ

```
from sklearn.metrics import classification_report
print(classification_report(y_true, y_pred))
```

મેમરી ટ્રીક

``Preprocess, Select, Evaluate - Complete ML workflow in Scikit-learn''

પ્રશ્ન 5(a) OR [3 ગુણ]

List out the major features of Numpy.

જવાબ

NumPy વૈજ્ઞાનિક કોમ્પ્યુટિંગ માટેનું મૂળભૂત પેકેજ છે.
- **N-dimensional Arrays**: કાર્યક્ષમ એરે ઓબ્જેક્ટ્સ
- **Broadcasting**: વિવિધ સાઇઝના એરે પર ઓપરેશન્સ
- **Linear Algebra**: મેટ્રિક્સ ઓપરેશન્સ
- **Random Numbers**: સ્ટેટિસ્ટિકલ સિમ્યુલેશન્સ

મેમરી ટ્રીક

``Numbers Need Numpy's Power - Essential for numerical computations''

પ્રશ્ન 5(b) OR [4 ગુણ]

K-means clustering for one-dimensional data

જવાબ

Dataset: 1,2,4,5,7,8,10,11,12,14,15,17
K-means (k=3):
- **Cluster 1**: 1, 2, 4, 5 (Centroid ≈ 3)
- **Cluster 2**: 7, 8, 10, 11, 12 (Centroid ≈ 9.6)
- **Cluster 3**: 14, 15, 17 (Centroid ≈ 15.3)

મેમરી ટ્રીક

``Groups Gather by Distance - Similar points form natural clusters''

પ્રશ્ન 5(c) OR [7 ગુણ]

Give function and its use of Pandas library

જવાબ

a) Preprocessing:
- read_csv(): ડેટા લોડ કરવા
- head(), tail(): ડેટા જોવા

b) Inspection:
- info(): ડેટા ટાઇપ્સ, મેમરી
- describe(): સ્ટેટિસ્ટિકલ સમરી
- isnull(): મિસિંગ વેલ્યુઝ ચેક કરવા

c) Cleaning:
- dropna(), fillna(): મિસિંગ ડેટા હેન્ડલ કરવા
- groupby(): ડેટા એગ્રીગેટ કરવા

```
df = pd.read_csv('data.csv')
print(df.describe())
```

મેમરી ટ્રીક

``Pandas Processes Data Perfectly - Comprehensive data manipulation tool''