מכון טכנולוגי חולון
Holon Institute of Technology

50058 מדע נתונים- תאוריה ומעשה | סמסטר קיץ תשפ״א

עבודה מספר 2- חיזוי סוג כיסוי יער

במסגרת עבודה זו, יש לפתח מודל לחיזוי סוג כיסוי יער, מתוך משתנים קרטוגרפיים בלבד (ללא נתוני חישה מרחוק).

סוג כיסוי היער בפועל לתצפית נתונה (30 x 30 מטר תא) נקבע מנתוני מערכת מידע המשאבים (RIS) של שירות היער האמריקאי (USFS). המאפיינים שבסט הנתונים נגזרו מנתונים שהתקבלו במקור מהסקר הגיאולוגי האמריקאי (USGS) ונתוני USFS. אזור המחקר ממנו נגזרו הנתונים, כולל ארבעה אזורי שממה הממוקמים ביער הלאומי רוזוולט שבצפון קולורדו בארה״ב. אזורים אלה מייצגים יערות עם הפרעות מינימליות הנגרמות על ידי בני אדם, כך שסוגי כיסוי היער הקיימים הם תוצאה של תהליכים אקולוגיים. בסט נתונים זה שבעה סוגי כיסוי יער, ומטרתנו לחזות מה סוג כיסוי היער הנכון.

תאור הבעיה ובסיס הנתונים מצויים בקישור הבא (ראה גם מטה):
https://archive.ics.uci.edu/ml/datasets/covertype
מחברת ג׳ופיטר לקריאת הנתונים:
https://drive.google.com/file/d/1T-03NBNnec5-acdGrWxBfwW14enWPobv/view?usp=sharing

סט הנתונים המוגדר לאימון כולל 15,120 אובזרבציות, הכוללות 54 מאפיינים ואת סוג הכיסוי. לצורך עבודה זו נשתמש רק בסט אימון זה, באופן הבא: 70% מסט זה ישמש אותנו לאמן את המודלים, 15% ישמשו לצורך ולידציה ו- 15% נוספים ישמשו לצורך הערכת ביצועים.

המשימות:

א. בצע אקספלורציה של הבעיה.
ב. בנה את מטריצת המאפיינים שתשמש אותנו לפיתוח המודל.
ג. פתח מודל לחיזוי כיסוי היער. בחירת המודל נתונה לשיקולכם, אך מומלץ לבחון מספר מודלים, למשל XGBoost ו- TabNet (https://github.com/dreamquark-ai/tabnet).
ד. השתמש באחת השיטות של feature selection (ראה למשל כאן: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SequentialFeatureSelector.html#sklearn.feature_selection.SequentialFeatureSelector) בכדי למקסם את ביצועי המודל. שים לב: את חיפוש סט המאפיינים יש לבצע באמצעות סט הולידציה, כאשר הערכת הביצועים הסופית תינתן על סט הבחינה (test set).

מכון טכנולוגי חולון
Holon Institute of Technology

הערות:

- ההגשה יכולה להתבצע באחת מבין שתי אפשרויות:

א. הגשת דו"ח הכולל פירוט (תרשים + הסבר תמציתי) של הנעשה, תוצאות, ניתוח תוצאות ומסקנות. יש להגיש את הדו"ח ב**קובץ PDF אחד בלבד** בהתאם להנחיות.

ב. הגשה במתכונת של מחברת Jupyter, ובלבד שהמחברת כוללים הסברים של מה שנעשה, ניתוח התוצאות והמסקנות. במקרה כזה, יש לעדכן בקישור שיפורסם לצורך כך, את הקישור למחברת שלכם ב- Google colab או GitHub ולהעלות את המחברת עצמה למודול.

- עיקר העבודה הוא ההסברים שלכם- הסבר תמציתי על הנעשה, ניתוח תוצאות ומסקנות.
- התרגיל יוגש ביחידים או בזוגות, באמצעות המודל בלבד. בהגשה בזוגות, מצופה מכל אחד מבני הזוג לשלוט בכל נדבכי העבודה והדו"ח. רק אחד מבני הזוג יגיש את העבודה במודל. יש לרשום שמות + מס' תעודות זהות בראש העבודה.
- עבודות דומות תיפסלנה ויינקטו צעדיים משמעתיים.
- מומלץ מאוד לבצע את העבודה באמצעות שפת פייתון.
- לוח הזמנים להגשה- בהתאם למוגדר במודול.

בהצלחה!

צוות הקורס

מכון טכנולוגי חולון
Holon Institute of Technology

Data Description

The study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. Each observation is a 30m x 30m patch. You are asked to predict an integer classification for the forest cover type. The seven types are:

1 - Spruce/Fir
2 - Lodgepole Pine
3 - Ponderosa Pine
4 - Cottonwood/Willow
5 - Aspen
6 - Douglas-fir
7 - Krummholz

The training set (15120 observations) contains both features and the Cover_Type. Please use only the training set for this work.

Data Fields

Elevation - Elevation in meters
Aspect - Aspect in degrees azimuth
Slope - Slope in degrees
Horizontal_Distance_To_Hydrology - Horz Dist to nearest surface water features
Vertical_Distance_To_Hydrology - Vert Dist to nearest surface water features
Horizontal_Distance_To_Roadways - Horz Dist to nearest roadway
Hillshade_9am (0 to 255 index) - Hillshade index at 9am, summer solstice
Hillshade_Noon (0 to 255 index) - Hillshade index at noon, summer solstice
Hillshade_3pm (0 to 255 index) - Hillshade index at 3pm, summer solstice
Horizontal_Distance_To_Fire_Points - Horz Dist to nearest wildfire ignition points
Wilderness_Area (4 binary columns, 0 = absence or 1 = presence) - Wilderness area designation
Soil_Type (40 binary columns, 0 = absence or 1 = presence) - Soil Type designation
Cover_Type (7 types, integers 1 to 7) - Forest Cover Type designation

The wilderness areas are:

1 - Rawah Wilderness Area
2 - Neota Wilderness Area
3 - Comanche Peak Wilderness Area
4 - Cache la Poudre Wilderness Area

מכון טכנולוגי חולון
Holon Institute of Technology

The soil types are:

1 Cathedral family - Rock outcrop complex, extremely stony.
2 Vanet - Ratake families complex, very stony.
3 Haploborolis - Rock outcrop complex, rubbly.
4 Ratake family - Rock outcrop complex, rubbly.
5 Vanet family - Rock outcrop complex complex, rubbly.
6 Vanet - Wetmore families - Rock outcrop complex, stony.
7 Gothic family.
8 Supervisor - Limber families complex.
9 Troutville family, very stony.
10 Bullwark - Catamount families - Rock outcrop complex, rubbly.
11 Bullwark - Catamount families - Rock land complex, rubbly.
12 Legault family - Rock land complex, stony.
13 Catamount family - Rock land - Bullwark family complex, rubbly.
14 Pachic Argiborolis - Aquolis complex.
15 unspecified in the USFS Soil and ELU Survey.
16 Cryaquolis - Cryoborolis complex.
17 Gateview family - Cryaquolis complex.
18 Rogert family, very stony.
19 Typic Cryaquolis - Borohemists complex.
20 Typic Cryaquepts - Typic Cryaquolls complex.
21 Typic Cryaquolls - Leighcan family, till substratum complex.
22 Leighcan family, till substratum, extremely bouldery.
23 Leighcan family, till substratum - Typic Cryaquolls complex.
24 Leighcan family, extremely stony.
25 Leighcan family, warm, extremely stony.
26 Granile - Catamount families complex, very stony.
27 Leighcan family, warm - Rock outcrop complex, extremely stony.
28 Leighcan family - Rock outcrop complex, extremely stony.
29 Como - Legault families complex, extremely stony.
30 Como family - Rock land - Legault family complex, extremely stony.
31 Leighcan - Catamount families complex, extremely stony.
32 Catamount family - Rock outcrop - Leighcan family complex, extremely stony.
33 Leighcan - Catamount families - Rock outcrop complex, extremely stony.
34 Cryorthents - Rock land complex, extremely stony.
35 Cryumbrepts - Rock outcrop - Cryaquepts complex.
36 Bross family - Rock land - Cryumbrepts complex, extremely stony.
37 Rock outcrop - Cryumbrepts - Cryorthents complex, extremely stony.
38 Leighcan - Moran families - Cryaquolls complex, extremely stony.

מכון טכנולוגי חולון
Holon Institute of Technology

39 Moran family - Cryorthents - Leighcan family complex, extremely stony.

40 Moran family - Cryorthents - Rock land complex, extremely stony.