יאיר צ'רטקוב - 206085532 גילה דורפמן 211702782

מדעי הנתונים ובינה עסקית - תרגיל בית 4 שפת Python

בקוד שלנו יש 2 מודולים GUI.py ו-naiveBayesModel.py:

קובץ GUI.py:

ב-GUI.py יש לנו את מחלקת naiveBayesGUI
פירט הפונקציות ב-**GUI.py:**
Init:
מתאחלת את כל האלמנטים הקשורים ב-GUI: כפתור ה-browse כפתור ה-build וכפתור ה-classify (עבור שלושת מוגדרים פונקציה מתאימות שמופעלות בעת לחיצה על הכפתור on_click_browse, ה-on_click_build וה-on_click_classify בהתאמה) שני ה-entries שלנו עבור ה-path ועבור מספר ה-binning ו-labels עבורם.

set_path: פונקציית set המעדכנת את המשתנה הלוקלי path המכיל את הנתיב לתקיה בה שמורים הקבצים הרלוונטים.

on_click_browse: פונקציה המופעלת בעת לחיצת הכפתור browse הפונקציה בודקת את תקינות הנתיב ואת התוכן של התקיה, בתקיה אמורים להיות הקבצים הבאים: train.csv, test.csv, Structure.txt במידה ונתיב איננו תקין או שהתוכן של התקיה לא מכיל את הקבצים הכרחיים מוצגת הודעה מתאימה למשתמש והכפתור של Build נשאר disabled (ערכו הדיפולטי) במידה והמשתמש לא הזין נתיב כלל גם תוצג הודעת שגיאה מתאימה.

set_num_of_bins: הפונקציה שנקראת בעת הזנת תו ב-entry של מספר ה-bin ים. הפונקציה בודקת את תקינות מספר הבינים (מספרים חיובים שלא מכילים אותיות) במידה ומספר ה-bin ים תקין אנחנו הופכים את הכפתור ל-normal ולכן הוא כבר איננו disabled מה שאומר שניתן ללחוץ עליו ולבצע את אימון המודל.
אחרת ה-entry לא מקבל את הערכים האסורים שמוזנים והכפתור נשאר disabled.

on_click_build: הלחיצה על הפתור מוודא מספר bin ים תקין (לא רק חיובי ומספרי כמו שנבדק בפונקציה Set_num_of_bins אלא גם גדול מ-2 כדי שיהיה משמעותי) ובמידה והכל תקין מבצעת את אימון המודל ומציגה את ההודעה המבוקשת. הפונקציה גם קובעת את הפתור classify ל-normal. אתה ניתן לבדוק את המודל על הקובץ train.csv.

on_click_classify: הפונקציה מופעלת בעת לחיצת הכפתור classify הפונקציה קוראת לפונקציות classify, import_data ו-save_predictions מהמודול naiveBayesModel.

מחלקת ה-GUI יוצרת אובייקט של tk והספריה tkinter ומפעילה את ה-mainloop() לשם יצירת חלון ה-GUI.

קובץ naiveBayesModel.py:

קובץ זה מכיל את כל הפונקציונאליות הנדרשת לעיבוד הנתונים, אימון המודל והחיזוי.

import_data:
מייבא את הנתונים לתוך $pd.DataFrame$ ובודק האם קובץ הנתונים ריק.

extract_feature:
פירוק טקסט קובץ מבנה הנתונים לטוקנים ובניית מילון המכיל את מבנה הנתונים. המפתחות הם הפיצ'רים והערכים הם רשימות של הערכים האפשריים של הפי'צר. ערכים נומריים יכילו ברשימה פריט יחיד $NUMERIC$.

read_structure:
קריאת קובץ מבנה הנתונים לפי נתיב נתון ובניית מילון מבנה הנתונים בעזרת פונקצית $extract - feature$.

fill_missing_values:
השלמת ערכים חסרים. משתנים נומריים ימולאו על ידי הממוצע וערכים קטגוריאלים על ידי השכיח.

discretize:
דיסקרטיזציה של משתנים נומריים לפי רוחב שווה על פי מספר תאים נתון.

to_numerical:
קידוד של כלל המשתנים לערכים נומריים באמצעות $LabelEncoder$ בפורמט מתאים עבור המודל של $sklearn$.

train_naive_bayes_model:
חלוקת הנתונים למשתני קלט ומשתנה מטרה ואימון המודל.

classify_with_naive_bayes:
חיזוי הסיווגים של משתנה המטרה $class$לפי מודל מאומן נתון.

refactor_prediction_labels:
המרה של הסיווגים מפורמט מקודד (0 -1) לפורמט מובן לבני אדם כנדרש בעבודה (no ו-yes).

save_predictions:
שמירה של תוצאות הסיווג בקובץ $output.txt$ לפי נתיב נתון לפי הפורמט המוגדר בעבודה.

build_model:
הפונקציה קוראת לפונקציה fill_missing_values, discretize ו-train_naive_bayes_model למעשה הפונקציה הזו נקראת מה-GUI בעת הלחיצה על כפתור Build ומהווה קישור לשאר הפונקציות הרלוונטיות במחלקה.

classify:
הפונקציה קוראת לפונקציה fill_missing_values, discretize ו-classify_with_naive_bayes למעשה הפונקציה הזו נקראת מה-GUI בעת הלחיצה על פתור classify ומהווה קישור לשאר הפונקציות הרלוונטיות במחלקה.