אוניברסיטת אריאל בשומרון

"דימות נתונים" – עבודת גמר, סמסטר ב תשפ״א

הקדמה:

במסגרת הקורס, מצ"ב פרטים על עבודת הגמר אשר יש להגיש. מטרת העבודה היא להכיר לסטודנט בצורה אישית ובלתי אמצעית את מגוון הנושאים אשר נלמדו בקורס. בין היתר, יש להבין את הנושאים שעברנו עליהם בקורס הקודם בצורה משמעותית וטובה. מה היו הנושאים המרכזים שלמדנו הסמסטר?

1. עבודה מתקדמת עם pandas
2. למידה ensemble
3. למידת unsupervised והדרך ליישם אותה בבעיות קלסיפיקציה
4. הורדת מימדים דרך PCA

מילה על ה-datasets:

אחד מה-datasets המפורסמים ביותר הינו Fashion-Mnist המכיל כ-70,000 תמונות שחור לבן בגודל של 28*28 או 784 פיקסלים. רשת למידה עמוקה קונבולוציה יכולה לקבל דיוק של מעל 99% ב-data זה אבל זה מאד מורכב לקבל דיוקים גבוהים בשיטות אחרות. data נוסף שנראה הינו cats-vs-dogs המכיל 25,000 תמונות של כלבים וחתולים.

משימות

1. עברו שוב על מחברת הקלסיפיקציה מהסמסטר הראשון. בדקו מה היה אחוז הדיוק שלכם ותראו האם אתם יכולים לשפר אותו בעקבות החומר שלמדתם בסמסטר הנוכחי. חובה להסביר במחברת בצורה מפורטת מה היה הדיוק שהגעתם אליו בסמסטר הקודם ומה הדיוק הנוכחי בעקבות השיפורים.
2. השתמשו בכל מה שלמדתם על מנת לייצר מודל המסווג טוב את FMNIST. שימו לב שמדובר בבעיית קלסיפיקציה multiclass. בשום שלב אסור לגעת ב10,000 הדוגמאות של הtesting. חשוב להסביר את המודל. שימו לב שניתן להשתמש בensemble, בטכניקות של clustering וכו׳ כדי להגיע למודל הטוב

ביותר. **אסור ליישם רשתות נוירונים בתרגיל זה.** יינתן ניקוד לא רק על מידת הדיוק אלא גם על מידת הדיוק תוך שימוש במינימום של מאפיינים. מודל קומפקטי יותר הינו מודל טוב יותר.

3. בעקבות היכולות שלכם בסעיף 2, נסו את כוחכם בdata השני (כלבים-חתולים). שימו לב שמדובר בתמונות צבעוניות בעלות גודל משתנה. גם בתרגיל זה אסור ליישם רשתות נוירונים.
4. קראו ועברו על המקורות הבאים.

- A Gentle Introduction to Imbalanced Classification
- Hands On Machine Learning Book - Chapters 7, 8, 9
- OpenCv Python Tutorials

5. מה מגישים? קישור לגיטהב בו אתם מציגים 4 מחברות. המחברת הראשונה זה שיפור הדיוק מהסמסטר הקודם. אין צורך לייצר מחברת חדשה אלא רק לשפר את הדיוק שהשגתם בסמסטר א. המחברת השנייה תראה את המודל של fMNIST, המחברת השלישית תראה את המודל של cat-vs-dogs והמחברת האחרונה תראה את העבודה בpandas המופיעה בנספח.

נספח א – פירוט עבודת סיווג בקורס "דימות נתונים"

מטרה

מטרת העבודה היא לסווג בין שלושה מצבים שונים באופן בו אנשים מתקשרים אחד עם השני. הראשון הינו מצב ספונטני (אוטונומי) בו שני אנשים מזיזים את הידיים שלהם בצורה חופשית אחד מול השני. השני הוא תנועה סינכרונית בה שני האנשים מזיזים את הידיים ביחד והשלישי הוא תנועה במצב לבד. בו רק הצד אחד מזיז את הידיים.

הרעיון הוא להסתכל על דפוסי הידיים ולנסות להסיק מהם האם מדובר במצב לבד, ספונטני או סינכרוני.

כפי שראיתם בתרגיל בית מספר 3, יש לכם כבר את כל המידע ב-DATAFRAME מסודר.

יש גם להציג את ההבדלים בצורה גרפית יפה.

בונוס יינתן ל-3 סטודנטים שיגיעו לרמות הדיוק הגבוהות ביותר. בנוסף, מצ"ב מאמר מדעי שפורסם על העבודה הראשונית הזו. כדאי מאד לקרוא אותו על מנת להבין מה נצרך מכם בעבודה זו.

שימו לב שבתיקיה המצורפת, יש תת-תיקייה שנקראת TRAINING ורק עליה יש לעבוד באימון הרשת. בנוסף, יש תת תיקיה הנקראת TESTING ורק עליה יש לעבוד בוולידציה.

הערות נוספות:

- במצב ALONE, קיימת רק הקלטה של יד אחת (יד שמאל). יש לבדוק את המצב הזה אל מול קובץ שנקרא HandRight ובו הקלטה של תנועות ביד ימין. שימו לב שאת כל מצבי ה-ALONE יש לבדוק אל מול הקובץ הזה.

- יש ספריות אשר יש שם 2 מקבצי הקלטות. במקרה כזה, יש לקחת את המקבץ האחרון.
- חשוב מאד לא לקחת את ה-7 שניות הראשונות של הקובץ.
- הרעיון פה הוא ליצצר פרמטר שבודק סינכרון בין שני אנשים. חשבו בעצמכם אילו פרמטרים חשובים ולמה.
- מטרת העבודה היא לסווג תנועות ידיים בין שני אנשים למצב (ALONE, AUTONMOUS ו—סינכרוני). ניתן לחשוב על כל SAMPLE כעל שנייה של הקלטה המורכבת מ-5 timestamps.
- חשוב להבין מה אחוז הדיוק ומהו הf-score של האלגוריתם שלכם. ניתן לייצר אלגוריתמים שונים אבל בשום מקרה אסור לאמן את המערכת על המידע מספריית הווילידציה.

בהצלחה לכולם.

רועי