<u>הגיאומטריה של המידע: על מימדים גבוהים ובינה מלאכותית - ד"ר רונן אלדן, 17.12.2019</u>

ד"ר רונן אלדן: אהלן. שמי רונן, אני מהמחלקה למתמטיקה פה במכון ויצמן, ומה שאני אנסה – אני אדבר על גיאומטריה במימדים גבוהים, שזו תיאוריה מתמטית שבשנות ה-70 ה-80 התחילו לחקור אותה; 20-10 שנה אחר כך הבינו שיש לה קשרים ובעיות רלוונטיות ל-data science ו-machine learning. עכשיו, לא קל לתת הרצאת מדע פופולרי שהנושא שלה מתמטיקה. אבל אני אשתפשף עליכם אנסה כמיטב יכולתי.

קריאה: אתגר.

ד"ר רונן אלדן: ננסה להבין אם אפשר לעשות את זה. אחת המטרות שלי היא לשכנע אתכם שהבנה גיאומטרית והסתכלות גאומטרית על בעיות שאין להם קשר אפריורי לגיאומטריה, יכולה לתת לנו תובנות ולעזור בפתרון בעיות שבהן גיאומטריה מופיעה בצורה לא צפויה. חוץ מלגרום סיוט לתלמידי תיכון, גיאומטריה יכולה לעזור בדברים כמו עיבוד תמונה, פענוח שפה או חיזוי בבורסה.

בואו נתחיל. אנחנו מדברים היום על מרחבים במימדים גבוהים, או גיאומטריה במימד גבוה. נבין קודם כל מה זה מימד, מה אני מתכוון כשאני אומר יש לי מרחב במימד 2, 3 או 15. מרחב במימד 2 – אני יכול לחשוב על הלוח כמרחב שבו כל נקודה אני יכול לייצג על ידי שני מספרים. כל נקודה על הלוח אני יכול לייצג על ידי שאגיד את המיקום בציר ימינה שמאלה ובציר למעלה למטה, כולנו ראינו צירים דו-מימדים. נקודה בתלת-מימד אפשר לייצר ב-3 נקודות – XYZ, למעלה למטה, קדימה אחורה, ימינה שמאלה. אם אני מדבר על פנים כדור הארץ אז נקודה על כדור הארץ היא קואורדינטה GPS; הפנים של כדור הארץ הם יריעה דו-מימדית, זה מין משטח דו-מימדי שהוא בתוך מרחב תלת-מימדי. קואורדינטות GPS בכדור הארץ מיוצגות על ידי קו רוחב וקו גובה, 2 מספרים, אפשר לתאר בצורה חד ערכית מיקום על כדור הארץ. היום נראה הרבה דוגמאות למרחבים ממימדים שונים. הפואנטה היא שֶמַה שאנחנו מתכוונים כשאנחנו אומרים מימד, זה פשוט מספר הקואורדינטות שצריך לתת כדי לייצג נקודה. יש כל המרחבים שאנחנו מכירים, בואו ננסה לחשוב איזה עוד מרחבי קואורדינטות יכולים להיות ב-data science ו- machine learning, איזה עוד מרחבי מידע. בשקף יש כמה דוגמאות למרחבים שנקודה עליהם מיוצגת על ידי אוסף קואורדינטות. אפשר לחשוב על כל התמונות האפשריות. תמונה, כפי שמחשב רואה אותה, היא אוסף פיקסלים. תמונה פה היא חלק מהציור Starry night- של ואן גוך, אפשר לתאר אותה על ידי שאקריא את הבהירות של כל פיקסל משמאל לימין, מלמעלה למטה. את התמונה הזאת אני יכול לתרגם לאוסף של מספרים, שמזהים את התמונה בצורה חד חד-ערכית. המרחב של כל התמונות זה מרחב שהמימד שלו הוא כמספר הפיקסלים של התמונה. למשל, תמונות שיש בהן 100x100 פיקסלים, זה מרחב עם מימד 10,000, כל נקודה במרחב מייצגת תמונה. עוד מרחב הוא מרחב כל הרצפים הגנטיים האפשריים. רצף גנטי של אדם מכיל בערך כמה מאות מיליונים של סימנים, אני יכול לחשוב על מרחב של כל הרצפים הגנטיים האפשריים; DNA של אדם מסוים הוא נקודה במרחב הזה.

מרחב נוסף הוא מידע רפואי. כולנו עשינו בדיקות דם, התוצאה של בדיקת דם זה כל מיני שמות של דברים שאת רובם אנחנו לא מכירים – סוגים של תאי דם, לבנים, אדומים, ברזל, כבד וכו', ליד כל אחד מהם יש מספר. מרחב התוצאות של בדיקות הדם האפשריות הוא מרחב במימד 50, שכל נקודה בו היא תוצאה ספציפית של בדיקת דם שעשינו, וכל תוצאה כזאת יש לה נקודה במרחב שמיוצגת על ידי מספרים מסוימים.

מרחבים נוספים – מידע פיננסי, סטטיסטיקות של משתמשים ברשת חברתית. כל משתמש הוא סדרה של דברים מייצגים - לְמַה הוא עשה לייק, מי החברים שלו, איפה הוא גר. כל משתמש הוא נקודה במרחב של כל המשתמשים בפייסבוק. גם טקסט/מסמכים, העדפות ב-Netflix, כל אלה הם מרחבי קואורדינטות במימד שהוא בדרך כלל גבוה מ-3. גם במרחבים האלה אני יכול לחשוב על קבוצות שיש להן גיאומטריה, או צורות שיש להם גיאומטריה. הבנת הגיאומטריה של הצורות אלה הולכת לתת לנו תובנות שיעזרו לנו לפתור בעיות של חיזוי, סיווג, עוד מעט נראה איזה סוג בעיות בדיוק.

בואו נחשוב רגע – אתחיל מדוגמא שממחישה מהי בעיה סיווג. נסתכל על דוגמה פשוטה של מרחב דו-מימדי שבו יש כלבים וחתולים (בשקף). הצירים במרחב הזה יהיו – ציר אחד מייצג את הצורה של האוזן, יותר שפיצית לעומת יותר עגלגלה; הציר השני מייצג את גודל האף, אף גדול לעומת קטן. ובמרחב הזה אני רוצה לפתור את בעיית הסיווג. אני מקבל נקודה במרחב, לכל נקודה יש שתי קואורדינטות. אני רוצה להחליט אם הנקודה הזאת היא כלב או חתול. זו כנראה לא בעיה מסובכת, כי לכלבים יש אף גדול יותר, ולחתולים נוטה להיות אף קטן יותר אבל אוזניים מחודדות יותר. מהי גיאומטריה של קבוצת כלבים וחתולים במרחב דו-מימדי? אני לא מדבר על הגיאומטריה של הכלב או החתול עצמו, אלא גיאומטריה של קבוצת נקודות במרחב דו-מימדי שמייצגות חתול או כלב, האם הגיאומטריה של קבוצת חתולים או כלבים נראית כך או אחרת. זה ייתן לי תובנות איך לתכנן אלגוריתמים שיסווגו אחר כך כלבים לעומת חתולים. פה הבעיה פשוטה. אם תיתנו לי נקודה בלי התמונה הנלווית, אוכל להגיד בסבירות גבוהה אם זה כלב או חתול. אם ניקח את האינטואיציה הזאת ונשליך אותה, נסתכל על דוגמא של מידע רפואי. יש לי את מרחב כל התוצאות האפשריות של בדיקות דם, שהוא מרחב במימד 50 בערך. אני רוצה לשאול מה הגיאומטריה של קבוצת התוצאות של בדיקות דם שמהווה סיכון גדול לחלות במחלת הסוכרת, למשל. שאלה רלוונטיות. אני לא רוצה לאפיין ממש איך נראית הקבוצה, אני לא בהכרח רוצה להגיד על כל תוצאה של בדיקת דם האם היא מהווה סיכון גבוה או נמוך. אני רוצה איזושהי תובנה יותר בסיסית לגבי הגיאומטריה של הקבוצה. האם זו קבוצה שתלויה רק במספר קטן של קואורדינטות, האם אני יכול לאפיין סיכון גבוה לסוכרת רק על ידי הסתכלות במספר קטן של מדדים, האם הקבוצה נראית יחסית עגלגלה ונראית כמו גוש אחד שנמצא במקום אחד במרחב, או כמו הרבה איים קטנים. יש שאלות שאני יכול לשאול על גיאומטריה של קבוצה, קבוצה במרחב 50, שאולי יעזרו לי אחר כך לפתור את בעיית הסיווג ולהשתמש במידע סטטיסטי כדי להפעיל אלגוריתם; ואותו אלגוריתם יֵדע אחר כך להפיק תובנות ולהגיד לי בסופו של דבר, כשאני רואה תוצאת בדיקת דם עם בדיקות רפואיות נלוות, האם יש פה סיכון גבוה לפתח מחלה כלשהי.

עוד שאלה שאני יכול לשאול; בקבוצת כל התמונות האפשריות – ונניח שזה מרחב ממימד 10,000 – במרחב התמונות האפשריות, איך נראית תת הקבוצה של תמונות שבהן מופיע כלב או חתול או תמרור עצור. כשאנחנו בונים מכונית אוטונומית, זה רלוונטי לענות על השאלה אם מופיע פה תמרור עצור. איך אדע לזהות את הקבוצה זאת? מהי הגיאומטריה של הקבוצה? תנו לי לנסות לחדד את העניין כדי שתבינו על מה אני מדבר בגיאומטריה.

בואו נסתכל פה על עוד דוגמא מדומיינת. נניח שחייזרים פלשו לכדור הארץ, חלק ידידותיים וחלק עוינים, חלק נחמדים וחלק פחות; ואני לא יודע מי זה מי. יש לי שני מדדים על החייזרים האלה, ציר אחד מייצג את אורך מחושים שלהם, הציר הזה מייצג נגיד את אורך הזרועות שלהם. יש לי מחושים ויש זרועות. אז אני נתקל בכל מיני חייזרים ואני מסמן נקודה כחולה עבור כל חייזר ידידותי, ונקודה אדומה עבור חייזר

לא ידידותי. זו התמונה שיוצאת (בשקף). מגיע חייזר חדש, אני רוצה לחזות אם הוא יהיה ידידותי או לא נחמד. מסתכלים על התמונה הזאת, רואים מה שיצא; אם אגיד לכם שפגשתי חייזר שהוא בנקודה הזאת – בערך פה בנקודות הכחולות - והוא יהיה ידידותי, די קל להבין את הגיאומטריה של קבוצת הנקודות הכחולות. גם הגיאומטריה של קבוצת הנקודות האדומות זו גיאומטריה ברורה. אלו שתי קבוצות שיחסית הן מִקשה אחת שנמצאות פחות או יותר באותו אזור במרחב דו-מימדי, לכן קל לאפיין מה הנקודות שֶׁיִפלו לכחולות, מה הנקודות שיפלו לקבוצה של האדומות. כשיש יותר משני מימדים, זה קשה יותר. לפעמים יש תמונות – אפשר לדמיין מערכת צירים תלת-מימדית עם נקודות כחולות ואדומות, בדרך כלל יהיה יותר קשה למצוא חוקיות. מה שכולכם עשיתם כאשר הסתכלתם על התמונה, המוח שלכם הסתכל על איזושהי תמונה דו-מימדית וניסה למצוא בה תבנית, ניסה למצוא חוקיות של מה מאפיין את הנתונים של נקודות כחולות לעומת אדומות. כשנסתכל על תמונה תלת-מימדית, יהיה יותר קשה למצוא תבנית. ובתמונה 50 מימדית - אף אחד לא יכול ממש להסתכל על תמונה ממימד 50, גם אם הייתם יכולים, אני מבטיח לכם שהיה קשה למצוא חוקיות. גם במחשב זה יותר קשה, אנסה להסביר למה.

מה הגורמים המתמטיים מאחורי זה שיותר קשה למצוא חוקיות במימדים גבוהים? אחר כך אנסה להסביר איך מתגברים על קשיים אלה. מה שאני רוצה להתחיל לנסות לשכנע, זה שכל אחד מאיתנו בחיי היומיום מנסה למצוא חוקיות במרחבים ממימד יחסית גבוה. דוגמא מחיי: לפני שנתיים נולדה לי בת, בחודשים הראשונים היא היתה מצוברחת הרבה, ואני כאדם אובססיבי לסטטיסטיקה ניסיתי למצוא את החוקיות, מה גורם לה להיות מצוברחת; באיזה ימים, מה גורם לה להיות שמחה לעומת עצובה. יש פרמטרים רבים, זה שיגע אותי שלא הצלחתי להבין, היא לא יכולה לתַקשר מה היא אוהבת ומה פחות. עשיתי לי טבלת אקסל, מימד הדגימות בטבלה שלי היה מספר עמודות, בכל טבלת אקסל אפשר לחשוב על השורות והנקודות במרחב ממימד כלשהי. המימד שלי היה 1 2 3 4 5 6 7 8 9, היו נקודות במרחב ממימד 9, וניסיתי להבין מה ההשפעה של כל הקואורדינטות האלה על הקואורדינטה שממנה אכפת לי, שהיא מצב הרוח. בסופו של דבר אני מוכרח לומר שלא הגעתי למסקנה חותכת. אבל היו לי כל מיני הנחות שעשיתי, כשניסיתי להבין מה הגורמים שמשפיעים על מצב רוח. ההנחות הראשונות היו – כבר כשעשיתי את הטבלה, לא שמתי טור שהכותרת שלו היא צבע החולצה. עשיתי הנחה שחלק מהפרמטרים שיש לנו בעולם כנראה לא משפיעים על מצב הרוח שלה. אולי בגילאים מבוגרים יותר, הייתי כן שָׂם צבע חולצה. אבל זה סיפור אחר.

אז לפני שאנחנו עוברים לדבר על גיאומטריה ממימד גבוה, אני רוצה לסכם את החלק הזה בצורה הבאה. יש לנו נקודות במרחב ממימד גבוה שמייצגות נתונים סטטיסטיים. אנחנו מנסים להבין מה הגיאומטריה של הקבוצה שֶׁבָּה אנחנו מתעניינים. קודם היתה קבוצת חתולים לעומת כלבים; אם יש לי רק אף ואוזניים, זו היתה גיאומטריה נחמדה. במקרים אחרים יכולה להיות גיאומטריה פחות נחמדה. אולי הקבוצה שאני מסתכל על מרחב הנקודות שלה, היא לא מִקשה אחת אלא אוסף של איים מנותקים שמאוד קשה לי לתאר. שוב, יכולה להיות גיאומטריה נחמדה של מקשה אחת עם גבול ברור בין קבוצה אחת לשנייה. או שהיא יכולה להיראות כמו קבוצת איים מנותקים שאין קשר ביניהם, ואז כמובן יהיה לי קשה יותר לתאר במילים את הגיאומטריה של הקבוצה הזאת. גם למחשב יהיה קשה יותר ללמוד את קבוצה הזאת. אתם מזהים מה התמונה פה בשקף? זו מפה של הפיליפינים...

על מה אנחנו מדברים כשאומרים גיאומטריה? מהי גיאומטריה במימד גבוה. אנחנו מכירים גיאומטריה דו ותלת-מימדית. האם יש מחקר של אובייקטים במימד גבוה יותר ממימד 2 ו-3? מרחבים ממימד נמוך הם 2 ו-3, במרחב ממימד 50 – שיכנעתי אתכם שאפשר להסתכל על נקודות במרחב.

מהי גיאומטריה של המידע שלנו? יש לנו מושגים, אובייקטים ותכונות שאנחנו מנסים לאפיין ולחקור. מושג בסיסי הוא מושג של מרחב, הוא מרחב דו-מימדי ותלת-מימדי. אפשר למדוד מרחק בין שתי נקודות, במרחבים דו ותלת-מימדיים יש צורות שונות, במתמטיקה קוראים להם קבוצות. יש צורות קָנוֹנִיוֹת – קובייה, כדור, מנסרה, חרוט וכו'. יש תכונות של קבוצות – לכדור יש תכונה שהוא עגלגל, לחרוט יש פינה. יש כל מיני מאפיינים גיאומטריים של צורות שאפשר לשאול עליהם שאלות. והנה שאלה אחת שאפשר לשאול את עצמנו על קובייה במימד 3 – אפשר לחתוך אותה בכל מיני צורות, להעביר דרכה מישור דו-מימדי. אפשר לשאול מה המישור שייתן לי את שטח החתך הכי גבוה עבור קובייה תלת-מימדית. זו שאלה גיאומטרית לגיטימית שאפשר לשאול על גיאומטריה של צורה במימד 3.

דבר ראשון שאשכנע אתכם זה שגם במימדים גבוהים יותר אפשר להגדיר דברים כמו מרחק, יש הכללות של הצורות האלה למימד גבוה יותר. נתחיל מקובייה. מהי קובייה תלת-מימדית? בתור התחלה, שימו לב – האם הקובייה שרואים פה, היא דו או תלת-מימדית? אני טוען שֶמה שכולכם רואים זו צורה דו-מימדית ולא תלת-מימדית. הלוח הזה – אפשר לשכנע אתכם שיש לו מימד 2 ולא 3. מה שאתם רואים פה זה צל של קובייה תלת-מימדית על מישור דו-מימדי. נכון?

ואיך הצל הזה נראה? למעשה קובייה תלת מימדית, אם תחשבו רגע, אני יכול לחשוב עליה בתור הכללה של צורה דו-מימדית. מהי צורה דו-מימדית? ריבוע. נכון? קובייה תלת-מימדית זה ריבוע, ועוד ריבוע, כאשר כל קודקוד על הריבוע שנמצא למעלה, אני מחבר אותו בקו לקודקוד המתאים של הריבוע שנמצא למטה. אני משכפל את הריבוע פעמיים ומחבר קווים, כך אני מקבל צורה תלת-מימדית שהיא הכללה של צורה דו-מימדית, צורת ריבוע במקרה הזה. אם נסתכל פה, בעצם נראה את הצל, אני יכול לחשוב על הריבוע הזה שהוא דו-מימדי, והריבוע הזה הוא דו-מימדי, שניהם הולכים בצירים X ו-Y, אבל יש עוד ציר שהולך לתוך הלוח ומחוץ ללוח. אני לא יכול ממש להקרין אותו פה, כי המקרן יוצר תמונות דו-מימדיות, אז מסתכלים על הצל של הדבר הזה, ואת הציר השלישי אנחנו מתרגמים לאיזשהו כיוון אלכסוני. אם הייתי רוצה לקחת את היצורים בעולם דו-מימדי ולהסביר מהי קובייה תלת-מימדית – יצורים מעולם דו-מימדי יודעים מה זה ריבוע בלבד. אז אגיד להם: קחו ריבוע, תשכפלו אותו ותחברו קו בין כל זוג קווים. אנחנו יצורים מעולם תלת-מימדי; אם יצור ארבע-מימדי ירצה להסביר את העולם שלו, הוא יגיד: קחו קובייה ועוד קובייה, תחברו כל קודקוד לכל קודקוד בקו, תקבלו קובייה ארבע-מימדית.

עכשיו אולי אראה לכם סרטון קצר, רק כדי להמחיש את זה יותר טוב. פה יש קובייה תלת-מימדית, אז הכל פה הוא בעצם דו-מימדי, אבל מה שרואים פה זה את הצל שלה; הדבר הזה הוא צל של קובייה ארבע-מימדית, מוטל על מישור דו-מימדי.

אני מקווה ששיכנעתי אתכם שלפחות לקונספט של קובייה יש הכללה במימד גבוה יותר, ולמעשה אני יכול להכליל אותו למימדים עוד יותר גבוהים.

פה רואים צללים דו-מימדים של קוביות שהמימדים שלהם הולכים ועולים. לפי אותה לוגיקה, קחו את המימד הקיים, תשכפלו אותו פעמיים ותקבלו מימד גבוה יותר.

מה שאתם רואים פה, זה צל דו-מימדי של סֶפֶרה ארבע-מימדית. זו הכללה של מה שאנחנו מכירים כפני כדור, סֶפֶרה למימד גבוה יותר שלקחנו את הצל שלו על מימד 2.

אז המוח שלנו יחסית יודע לתפוס צורות תלת-מימדיות, כך אנחנו בנויים. אני יכול לומר שריבוע אפשר לחלק לשני משולשים, וכנראה רובכם די בקלות יכולים לראות את זה. אם אשאל אם אפשר לרצף קובייה תלת-מימדית על ידי פירמידות, אולי לא תדעו לענות על זה מיד, אבל זו שאלה שאתם תוכלו איכשהו לדמיין אותה בראש. אם אשאל אתכם אם אפשר לרצף קובייה ארבע-מימדית על פירמידות ארבע-מימדיות, או על גרסה מתאימה ארבע-מימדית לפירמידות, יהיה לכם קשה לדמיין את זה. אבל זו שאלה מתמטית לגיטימית. סוג שאלות כזה, למרות שהוא נראה מנותק מהחלק הראשון של ההרצאה שלי – שבו דיברתי על דוגמאות מלמידת מכונה ומדע הנתונים – איכשהו שאלות כאלה במימדים גבוהים, יש להן השפעות מפתיעות על מה שאנחנו יודעים לעשות בפתרון בעיות בתחומים של מדע המדידה.

השקף הזה מראה גיאומטריה במימד גבוה, אולי על זה אדלג. הדוגמא בשקף הבא היא נחמדה. מה שאנחנו רואים פה – לקחתי קובייה במימד 100, ומה שרואים פה זה חתכים שלה בכל מיני אופנים. בשקף שאחרי כן חותכים את הקומקום עם מישור, ואת הקובייה בשקף הבא חתכנו עם מישור תלת-מימדי באיזשהו כיוון, וזה מה שיצא. זה חתך של קובייה ממימד 100, ואחת התובנות המפתיעות – שאדבר עליה גם אחר כך – היא שבאיזשהו מובן חתכים טיפוסיים של צורות ממימד גבוה, בדרך כלל יש להם התנהגות יפה ופשוטה והתנהגות נחמדה. אם הייתי לוקח קובייה ממימד עוד יותר גבוה, הייתם רואים שהחתך נהיה יותר כמו כדור. כדור זה צורה שיש לה תפקיד מיוחד, אנחנו רואים אותה הרבה כשלוקחים צורות ממימד גבוה, אני מקווה שאספיק לדבר על זה אחר כך.

בשקף הבא, שכותרתו קללת המימדיות, רואים קצת על מה בכלל מדברים כשמדברים על גיאומטריה ממימד גבוה. אחזור רגע ל-data science, כדי לנסות להסביר מושג שנקרא קללת המימדיות. המושג הזה מנסה להסביר לנו מה הקושי לבצע ניתוחים ולמצוא תבניות במרחבים ממימד גבוה. בואו נחזור לדוגמא של בדיקות דם. יש לי 50 מדדים, אני רוצה למצוא בהם תבנית, להבין איך נראית הקבוצה של בדיקות הדם שמאפיינת סיכוי גבוה למחלה מסוימת. והנה דרך – אני יכול להציע את הדרך הבאה לעשות זאת. נגיד שהיו לי רק שני מדדים בבדיקות דם; אם היו לי רק שני צירים, X ו-Y, זה מודד גלוקוז וזה מודד תאי דם לבנים, הייתי יכול לרצף את המרחב על ידי תאים, לחלק כל ציר ל-5 או 10 תאים, גם את הציר הזה וגם את זה, ובכל אחד מהתאים האלה הייתי אומר: אני אקח את כל ההיסטוריה הסטטיסטית שלי, אקח את כל מי שביצע בדיקת דם ונפל לתוך התא הזה, אעשה ממוצע ואראה כמה מתוכם לקו באיזה שלב בחצבת. אעשה אותו דבר על התא הזה. ואז כשיבוא אליי אדם חדש עם תוצאת בדיקת דם, ננסה להבין לאיזה תא הוא נופל, נראה אלו אנשים נפלו לתא הזה קודם, ולפי מה שקרה להם אנסה להסיק עליו. נשמע לא רע, נכון? אז הנה הבעיה המרכזית במימד גבוה. אם היה לי רק מימד אחד והייתי מחלק כל ציר ל-5, היו לי רק 5 תאים. כשיש לי 2 מימדים, זה כבר 5 בריבוע, 25 תאים. כשיש לי 3 מימדים, כמה תבניות יש? 5 בשלישית, שזה 125. קללת המימדיות היא העובדה שמספר התאים האלה גדל באופן אקספוננציאלי עם המימד. במילים אחרות, אם אנסה להעביר את זה לשפה יותר פשוטה, מימדים גבוהים הם ענקיים. כדי לרצף אותם על ידי דברים קטנים יותר, אני צריך המון-המון מהדברים הקטנים הללו. לא קשה לדמיין שאם אני אמשיך את התמונה הזאת ל-4, 5 מימדים וכו', אני אצטרך המון-המון-המון תאים כאלה כדי לרצף את המרחב וכדי להסיק מסקנה לגבי תא מסוים, ואני צריך שיהיה שם מספר לא מבוטל של דגימות. לא קשה לראות שאם אני מרצף מרחב ממימד 50, מספר הדגימות שאצטרך כדי שבתוך כל תא תהיה אפילו דגימה אחת, הוא עצום. אין סיכוי שיהיה לי מספיק מידע היסטורי כדי להסיק משהו בצורה הזאת. זוהי בדיוק קללת המימדיות. במימדים גבוהים יש מספר הרבה יותר מדי גבוה של

אפשרויות כדי לעשות משהו נאיבי כזה שאומר: אני רוצה לעשות משהו על אדם אחד, בואו נסתכל על כל האנשים שהיו דומים לו, כל האנשים שהתוצאות שלהם היו דומות לתוצאות של האדם המסוים הזה. כמובן שכשאנחנו מדברים על עיבוד תמונה למשל, אם אני רוצה להסתכל על תמונה ולהבין אם היה בה חתול או כלב, אין שום סיכוי לקחת את כמות התמונות שהיו לי בהיסטוריה ולרצף את המרחב בצורה כזאת. לכן אני צריך תובנות על הגיאומטריה של הקבוצות שלי כדי לעשות משהו קצת יותר מתוחכם מהריצוף הזה שאנחנו עושים פה.

בואו נדבר על מה הפתרון, מה יכולים להיות פתרונות אפשריים לְמה שנקרא קללת המימדיות. הפתרונות הללו הם רבים, אני רוצה לדבר על משפחה של פתרונות שנקראת הורדת מימד. זה משהו שהמוח עושה בכל יום וכמעט כל רגע, ואני יכול די בקלות לנסח מתמטית מה זה אומר. מסתבר שזה קונספט מאוד שימושי במדעי המידע. בכותרת – הורדת מימד, זה לקחת מידע שנמצא במימד גבוה ולתאר אותו כנקודות במרחב ממימד נמוך יותר. אני יכול לקחת קבוצה של כל התמונות האפשריות שיש, ולנסות להוריד להן מימד על ידי זה שבמקום לתאר כל ערך של פיקסל, מה הבהירות שלו, אני לוקח את התמונה ומתרגם אותה לקואורדינטה אחת שאומרת האם בתמונה יש משהו עם זנב, האם בתמונה יש אוזניים גדולות, או קואורדינטה שאומרת האם בתמונה יש אובייקט פרוותי, וכו'. אני יכול לעשות ניסיון, כך שבמקום להגיד מה הערך של כל פיקסל ופיקסל, רק אסכם לכם את מה שאני רואה בתמונה על ידי מספר קטן יותר של ביטים של אינפורמציה. למעשה אם אתן לכם עכשיו להסתכל על תמונה, לתאר מה אתם רואים בה, לא תגידו שבצד שמאל למעלה יש צבע אדום כהה, קצת ימינה האדום הופך לאט-לאט לירוק, עוד ימינה הוא הופך לכחול, למטה הוא הופך לשחור וכו'. אם תנסו לתאר לי ככה את התמונה שאתם רואים, ייקח לכם המון זמן וכנראה שלא אבין מזה כלום. יותר סביר שתגידו לי: אני רואה בתמונה יצור פרוותי עם זנב, משמאלו יש בית, למעלה יש רקע כחול ולמטה רקע ירוק. כשאתם נותנים לי את התיאור הזה, מה שעשיתם זה הורדת מימד. ראיתם מידע ממימד גבוה, והמוח שלכם מסכם לי את המידע הזה, מכווץ אותו למשהו שאפשר לתאר אותו על ידי מימד נמוך יותר. באופן כללי הורדת מימד זה לקחת נקודות מידע שיש להן המון קואורדינטות ולתאר אותן על ידי מעט קואורדינטות, בתקווה שזה עדיין יכיל את המידע הרלוונטי לנו. יש דרכים פשוטות לעשות הורדת מימד, כולנו עושים את זה המון שנים.

אם נחזור לדוגמא של סוכרת, עד לא כל כך מזמן, לפחות ל-FDA היה מדד קָנוֹנִי שהם הריצו על אנשים שאמרו: יש המון בדיקות שאני עושה אבל אני מסתכל רק על תוצאות של 7 בדיקות, שהן הכי רלוונטיות לזיהוי המחלה הזאת. אמנם, על הבדיקות של כל אחד אפשר לחשוב כנקודה במימד 100, אבל אסתכל רק על 5 הקואורדינטות חשובות, וכך עשיתי הורדת מימד. הורדת מימד יכולה להיות במקרה שאולי יש מדדים רלוונטיים לבעיה שאני מנסה לפתור, אך לא תמיד אפשר להתמקד רק בסט קטן של מדדים מתוך כל המדדים שיש לנו. יש תורה שלמה איך אני מסתכל על נקודות במרחב ממימד גבוה ומנסה לחלץ מהן נקודות במימד נמוך יותר. גיאומטרית אפשר לחשוב על זה כסוג של צל של נקודות. אנחנו מסתכלים על נקודות שצפות במרחב התלת-מימדי שלנו, במתמטיקה זה נקרא הטלה – מטילים את הנקודות, לוקחים את הצל שלהן על המישור שכדי לתאר נקודה בתוכו אני צריך רק שתי קואורדינטות. זאת בתקווה שמתוך שתי הקואורדינטות האלה, אוכל כבר לחלץ את כל מה שחשוב לי מתוך הנקודות אלה.

אני חושב שהכי קל להבין את הרעיון של מהי הורדת מימד על ידי דוגמאות, יש כמה דוגמאות נחמדות מאוד. אם נסתכל על מרחב כל הפנים האפשריות של אנשים, אני יכול לתאר פנים של אדם; פנים זה משהו תלת-מימדי, נניח שאני מתאר על ידי תמונה דו-מימדית. מה שעשו פה (בשקף), ניסו לעשות הורדת

מימד לקבוצה של כל מיני פנים של אנשים, לתאר אותם על ידי שתי קואורדינטות. אני לא אומר מה הקואורדינטות אלה, אלגוריתם מצא מה הקואורדינטות ההגיוניות, אולי היום היה מוצא קואורדינטות אחרות. הזינו את התמונות לאלגוריתם, ביקשו שימצא שתי קואורדינטות שמבטאות הכי הרבה שונות, שאם אדע אותן אדע כמה שיותר מאפיינים על הפנים של הבן-אדם הזה.

מה שיצא – קיבלנו חלוקה של קבוצת כל הפרצופים האפשריים לשתי קואורדינטות. כשאני הולך לשמאל למטה רואה בעיקר גברים עם זקן ומשקפיים. אם אני אלך פה לימין למעלה, אז אין בכלל גברים עם משקפיים, כנראה שיש אנשים עם בייבי פֶּייס, פה למטה באזור הזה יש קצת יותר נשים. מה שהאלגוריתם עשה – הוא חילק את כל קבוצות הפנים, ניסה לתמצת כמה שיותר ממה שאני רואה בפנים של אדם, שצריך מימד גבוה לתאר אותם, רק על ידי שתי קואורדינטות.

דוגמא נוספת נחמדה – נכון שב-Netflix רואים סרטים ומדרגים אותם, ואז Netflix נותן לנו המלצה איזה סרטים לראות אחר כך? בדרך כלל הוא עושה עבודה די טובה. ב-Netflix אפשר לחשוב על כל משתמש כנקודה במרחב העדפות, כל קואורדינטה בו היא מה הדירוג שנתתי עבור כל סרט. כל אדם ב-Netflix הוא נקודה במימד גבוה, שזה כל הסרטים אפשריים. היתה תחרות ב-Netflix, שהיו ממנה כל מיני תובנות נחמדות ב-data science. מה שניסו לעשות זה הורדת מימד למרחב ההעדפות ב-Netflix, ניסיון לקטלג את המשתמשים ואת מרחב ההעדפות במימד נמוך יותר. לקחו את כל המשתמשים עם כל ההעדפות שלהם, עשו לזה הורדת מימד. מה שרואים פה בשקף זה הטלה על שתי הקואורדינטות הראשונות. אם אני עושה הורדה למימד 2 של מרחב ההעדפות האפשריות, יש לנו כל מיני סרטים פה ואנחנו רואים איך הם נופלים במרחב הדו-מימדי. אז לכל קואורדינטה פה, אני יכול לתת תיאור כלשהו. בואו נסתכל מה הפלט של האלגוריתם הזה. אם מסתכלים בצד שמאל, רואים כל מיני סרטי אימה, סרטי פעולה, בדרך כלל הכוכב הראשי הוא גבר; בצד ימין רואים יותר דרמות, כמו בחירתה של סופי, סרטים נשיים יותר, הכוכבים של הסרטים הם נשים. נסתכל על הציר למעלה למטה – למטה רואים יותר הפקות גדולות כאלה כמו צלילי המוזיקה, כל מיני קומדיות. ולמעלה רואים סרטים שנוטים להגיע לפסטיבלים באירופה, אני יכול לחשוב על הציר ימינה שמאלה כציר של יותר אלימות מול פחות אלימות, אם תרצו; ועל הציר למעלה למטה כאיכותי יותר מול המוני יותר. לא היה אף אדם שאמר: זה התקנון שאני רוצה לעשות לסרטים, אלא זה מה שמחשב עשה. הוא קיבל נקודות במרחב גבוה מאוד, אמרנו לו: תנסה להוריד את זה למימד 2. וזה הפלט שלו. זו הדרך הטובה שהאלגוריתם מצא כדי לקטלג סרט על ידי שתי קואורדינטות. הורדת המימד הזאת בסופו של דבר נותנת ל-Netflix דרך הרבה יותר קלה להגיד: במקום לחשוב, במקום לעשות את הניתוח שלי במרחב ממימד של כמה מאות או אלפים; אני צריך רק להסתכל על מרחב ממימד קטן יותר, כל משתמש הוא נקודה פה, כל סרט אני חושב עליו כנקודה, ואני יכול להגיד מה הסיכוי שמשתמש מסוים יאהב סרט מסוים. זו הדוגמא.

עכשיו אַראה עוד דוגמא נחמדה של הורדת מימד על תמונות. מרחב התמונות הוא כמובן ממימד גבוה מאוד, אבל היה פה ניסיון להוריד אותו למימד 2. אם ננסה להבין את הצירים שזה מצא לנו, נסתכל על הציר למעלה למטה – למעלה רואים יותר מכוניות, בתים מחשבים; אם אני יורד למטה, יש פה יותר דברים מהטבע. שוב, זה פלט של אלגוריתם. אין פה שום עבודת אדם. כל מה שהאלגוריתם קיבל זה רק דאטה של הפיקסלים של התמונות, בלי שום דבר. אמרו לו: תנסה להוריד למימד 2, וזה היה הפלט. המחשב הבין לבד שאפשר לקטלג את התמונות – כנראה שֶפֹּה למטה יצאו דברים שהם יותר מהטבע

והחי, כי אולי ככה אנחנו מתארים את זה. אולי פה למעלה יש יותר קווים ישרים, זוויות ישרות, ובדברים מהטבע יש יותר עגלגליות ופחות צורות גיאומטריות של קווים ישרים.
אם אני מסתכל על ציר ימין שמאל, אני לא יודע אם נצליח למצוא פה תבנית, אבל בצד ימין אני רואה יותר סוסים וכלבים, ופה – אתם מצליחים להבין?
קריאה: בצד שמאל יש דוממים.
ד"ר רונן אלדן: אולי בצד שמאל יש יותר עצמים דוממים. זה די יפתיע אותי, כי באופן כללי קשה להפריד פה בין דוממים לבין דברים שזזים.
מה שאנחנו מנסים לעשות באופן כללי – יש לנו סט אלגוריתמים שלוקח נקודות עם הרבה קואורדינטות ומנסה לחלץ מהן מספר קטן של קואורדינטות שמסכמות בצורה טובה את המאפיינים שלהן, שאחר כך על פי המאפיינים אלה אוכל לעשות לפתור בעיות במדעי המידע. ואם נחזור לגיאומטריה, אז התובנות שאנחנו מגיעים אליהן על הגיאומטריה של הקבוצות שאנחנו מנסים לאפיין, הולכות להיות רלוונטיות מאוד לאופן שבו נעשה הורדת מימד. בואו ננסה להסביר לכם – זה יהיה לי קצת קשה בכמה דקות שנשארו – אני רוצה לשכנע אתכם שיש תובנות מתמטיות שמגיעות ממש מהתיאוריה של חקר אובייקטים מתמטיים, שיכולות לעזור לעשות הורדת מימד לדאטה אמיתי שמגיע מהעולם.
יש דוגמאות פשוטות כאלה. למשל, אם נחזור לדוגמא של החייזרים, פה כמובן יכולים לראות בעין שיש צורה גיאומטרית פשוטה שמאפיינת כל קבוצה (קבוצת הנקודות הכחולות וקבוצות האדומות). גם אם נסתכל על המידע של מצב הרוח של הבת שלי, נוכל לחלץ מזה את קואורדינטות שלהן יש השפעה גבוהה ביותר על מצב הרוח. למשל, בדוגמא הזאת, כשהסתכלתי על הטבלה, הבנתי שכנראה כן יש פה קבוצה יחסית קטנה של קואורדינטות שרק אותה אני צריך לנתח. אם נסתכל על מרחבי התמונות, אז יש תובנה שהגיעו אליה עוד לפני די הרבה שנים, שאפשר להוריד את המימדיות של התמונות על ידי שאנחנו מסתכלים על איזה תדירויות מופיעות בתמונות הללו. בעיבוד תמונה ופיתוח קול – אפשר לנסות להתאים את התדירויות לכל תמונה או גלי קול, ולחלץ מהן תובנות על התמונות.
האמת היא שיש דוגמאות מאוד טובות שהן יחסית יותר מהשנים האחרונות. איזשהו אפיון גיאומטרי של תוצאות MRI, גרם לאנשים להבין שפעם עשו סריקות MRI, עשו מדידות מהרבה כיוונים, לקחו הרבה דגימות כדי לעשות סריקות MRI לאנשים. ואז היתה תובנה גיאומטרית על איך נראית קבוצת סריקות MRI שבדרך כלל רואים, מה שגרם לכך שאפשר לקחת הרבה פחות מדדים ממה שחשבו שצריך לקחת, כדי לעשות כמעט את אותו בדיקה. היתה תובנה גיאומטרית על קבוצת תוצאות MRI שאיפשרה לייעל מאוד את הבדיקה הזאת.
אין לי עוד הרבה זמן, התכוונתי אולי לנסות לספר עוד קצת על ממש התובנות עצמן, שאנחנו מגיעים אליהן בגיאומטריה במימד גבוה. אבל אני חושב שאסיים פה. אפשר לשאול שאלות.
מחיאות כפיים.

שאלה: מה קורה כשיש משהו מכוון, אתה יודע לא כל נקודות.
שאלה: מה קורה שיש קווים מנחים – ולמרות שאתה עושה סטטיסטיקה, יש משקל יתר לנקודות מסוימות? גם אז אלה דברים שמשתמשים בהם?
ד"ר רונן אלדן: השאלה מה קורה, לפעמים יש לנו ידע אפריורי על איזה מהקואורדינטות הן חשובות יותר; על איך כדאי לראות את המידע שלנו מראש, כדי להסיק מסקנות. כמובן שהאלגוריתמים שיש תומכים

בזה. אז רוב העבודה של מה שנקרא מדעני מידע, data scientists, זה לקחת את המידע שלנו ואת הידע האפריורי שיש לנו, שאומר שכנראה הקואורדינטות האלה חשובות יותר. למשל, כנראה שיש אינטראקציה בין קואורדינטה 3 ל-4 שהיא מעניינת, ואז יש לתרגם את זה לאיזושהי נוסחה שאני נותן לה אלגוריתם, שיֵדע להוציא מזה את תובנות בצורה הכי טובה. כן, כמובן, כשיש אדם מאחור שיודע משהו אפריורי על איך לנתח את המידע, יש המון דרכים להשתמש בזה.

שאלה: חשבתי בהרצאה על שתי דוגמאות, אחת זה תאונות דרכים. יש עשרות מימדים – תכונות הנהג, רכב, מצב הדרך וכו'. זה מאוד אקטואלי, אפשר להשתמש בזה למניעת תאונות דרכים?

ד"ר רונן אלדן: אני לא יודע שיש data science שנעשה על תאונות דרכים בהקשר הזה. זה רלוונטי ואין סיבה שלא ייעשה. בעיקר עכשיו כאשר בחצי מהרכבים יש מצלמה, בטח עוד מעט תהיה בכולם. טסלה בטוח שעושים את זה.

דובר: מה עם כל הנושא של חורים שחורים, לאתר איפה יש חורים שחורים,

ד"ר רונן אלדן: יש הרבה מדע של מידע שעושים באסטרונומיה, פיסיקה, אני פחות מבין בניתוח של זה. כל מה שאני מכיר זה שמשתמשים ב-data science של תמונה גם בשביל אסטרופיזיקה.

דובר: המערכת שלך לא יכולה להתמודד עם זה?

ד"ר רונן אלדן: זו לא מערכת שלי. אני מתעסק בצד המתמטי של זה, ושום דבר ממה שתיארתי פה הוא לא שלי. אבל לשיטות האלה של הורדת מימד יש הרבה שימושים לעיבוד תמונה, למשל גם פיענוח של תמונות בסיגנלים שמקבלים בטלסקופים. כן.

דובר ב': אתה יכול לעזור לנו להבין מהי הורדת מימד מוצלחת. היו דוגמאות שונות. מהי הורדת מימד מוצלחת עם תמונות של אדם אחד? עם סרטים קשה לראות את הפחתת המימד.

ד"ר רונן אלדן: הפחתת מימד מוצלחת – השאלה תלויה בבעיה שאתה מנסה לפתור. הבעיה שמנסים לפתור פה – מגיע משתמש חדש, יש קצת דאטה עליו, ראה את הסרטים האלה והאלה, אני רוצה להבין איזה סרטים הוא יאהב בהמשך. הפחתת מימד מוצלחת פה היא הפחתה למימד שאין בו הרבה קואורדינטות, כי אני רוצה להוריד את המימד למימד נמוך כדי שאוכל לעשות עליו עיבוד. מצד שני אני רוצה שזה עדיין יכיל לי את כל המידע הרלוונטי לפתרון הבעיה. במקרה הזה אני רוצה קבוצה קטנה של קואורדינטות שמתארות סרט, שיכיל לי כמה שיותר מידע רלוונטי על איזה משתמש הולך בסופו של דבר לאהוב אותו.

דובר ב': בעצם לקבוצת סרטים שאני אוהב, יש תכונות גיאומטריות מסוימות, למשל רדיוס קטן או משהו כזה?

ד"ר רונן אלדן: יש פה שתי שאלות. שאלה ראשונה שאתה שואל היא איך לעשות את הורדת המימד, מהי הורדת מימד טובה? שאלה שנייה – קבוצת סרטים שמשתמש אוהב היא תת-קבוצה במרחב הסרטים, מה אני יכול להגיד על גיאומטריה שלה; האם היא יחסית לוקלית במרחב, האם היא נראית כמו כדור סביב נקודה כשהנקודה היא הסרט האהוב עליי, או האם היא נראית כמו הרבה איים? אז שתי השאלות האלה הן בדרך כלל כרוכות זו בזו. בדרך כלל הורדת המימד שאני אחפש לעשות היא כזאת שֶבָּה הגיאומטריה של קבוצות שבהן אני מתעניין תהיה גיאומטריה נחמדה. אם אעשה הפחתת מידע כך שקבוצה של סרטים שרוב האנשים אוהבים תהיה כמו איים במרחב, כנראה אצטרך לעשות עיבוד יחסית מורכב. שואפים לייצג את המידע כך שקבוצות רלוונטיות תהיה להם גיאומטריה נחמדה. לא הגדרתי מה זה גיאומטריה נחמדה, אבל תדמיינו – במשהו דו-מימדי, גיאומטריה נחמדה זה משהו שרואים בעין: הנה הצורה הזאת (בשקף).

דוברת: אתה אומר הורדת מימד, זה למשל להוריד את הנקודה האם הסרט הוא ישן או חדש, האם הוא שחור לבן או צבעוני, האם הוא סרט מלחמה או לא. אלה המדדים שמורידים על מנת לאפיין את האהבה או החיבה לסרטים מסוימים?

ד"ר רונן אלדן: נסכם. הורדת מימד באופן כללי, במקום לתאר מה קורה בסרט בכל סצנה, הקואורדינטות שאת ציינת – אם הסרט ישן או חדש, שחור-לבן או צבעוני, קומדיה או אקשן וכו' – אלה קואורדינטות שהן כנראה רלוונטיות בהורדת מימד. בהחלט. הפואנטה היא למצוא את הקואורדינטות, ולקואורדינטות לאו דווקא יהיה תיאור ברור כמו שאת אומרת. את הקואורדינטות האלה האלגוריתם מצא, אי אפשר לתאר אותן בצורה כל כך פשוטה כמו שתיארת.

דוברת: גם אלגוריתם פועל על פי הנחיות.

ד"ר רונן אלדן: לאלגוריתם הזה לא היו שום הנחיות, רק מספרים. תחשבי על אקסל שכל שורה זה משתמש Netflix וכל דירוגים שלו לכל הסרטים. האלגוריתם יודע רק את הדירוגים ולא את הסרטים. לסרטים יש שָם מספר, אבל לאלגוריתם אין מושג על שום מידע רקע לגבי הסרטים האלה, ועדיין הוא מצא את הדבר הזה.

דוברת: האם מנסים הרבה הורדות מימד עד שמוצאים אחת מוצלחת?

ד"ר רונן אלדן: את שואלת איך מפעילים את האלגוריתם של הורדת מימד. על זה אני יכול לדבר בקלות עוד שעה, איך בכלל לעשות את זה.

דרך אחת לעשות זאת זה לנסות כל מיני דברים, עד שמוצאים דבר אחד מוצלח. אבל מהמידע שאלגוריתם רואה אני לא יכול לחלץ אם הסרט שחור לבן, אין לי המידע הזה. בהחלט דרך אחת לעשות זאת זה לנסות כל מיני דברים, זה בהחלט אחד הדברים הראשונים שכדאי לחשוב עליהם כשעושים הורדת מימד. למשל ברפואה זה מה שעשו כמעט בכל המקרים עד לא מזמן. כשרוצים לדעת אם יש סיכון למשהו, אנחנו מנסים לראות על מה צריך להסתכל, זה בעצם חיפוש תבניות.

דוברת: אם יתנו לכל סרט שחור מספר ולכל סרט צבעוני מספר, האלגוריתם יֵדע לזהות אם יהיה מספר?

ד"ר רונן אלדן: נכון. אבל נניח שתהיה לכל סרט קואורדינטה שאומרת אם הוא שחור או צבעוני. אז נכון, האלגוריתם יוכל לחלץ גם מזה משהו, אבל לא ברור שזה ייתן לאלגוריתם יכולת מֵעבר לְמה שהוא כבר עושה.

שאלה: איך מנתחים מיקרו-ביו?

ד"ר רונן אלדן: זה נושא שאני פחות מבין בו. בגלל שיש פה אנשים שמבינים בזה הרבה, אני מעדיף לא לענות.

שאלה: האם יש שילוב בין שיטות של data science וניתוחים סטטיסטיים?

ד"ר רונן אלדן: השאלה אם יש שימוש בכלים סטטיסטיים כמו cluster analysis כדי להבין את התוצאות של אפוסטריורי.

ספציפית בהורדת מימד או כל דבר אחר, כמובן שלקחנו מידע וניסינו לייצג אותו בצורה יותר נחמדה, אז יש הרבה שיטות להבין אם עשינו עבודה טובה. אחת, אם פותרים את הבעיה יותר טוב. אבל כמובן שיש כלים סטטיסטים כמו ניתוח אשכולות (cluster analysis) שמוודא שהגיאומטריה של המידע שלי היא נחמדה, שהיא נמצאת בכל מיני קְלַסטֶרִים. אז כן.

דובר: איפה אפשר למצוא לזה סימוכין?

ד"ר רונן אלדן: יש המון מקורות על הורדת מימד, כל ה-Data Science Tutorial, או גיאומטריה במימדים גבוהים.

אוקיי, תודה רבה.

מחיאות כפיים.