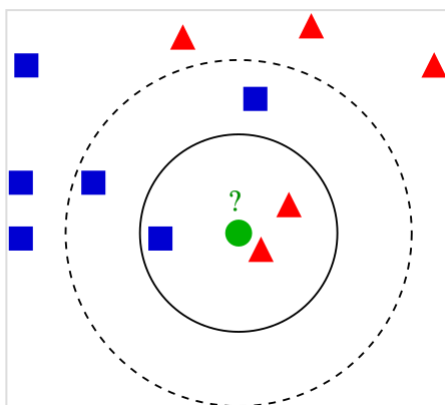




ויקיפדיה

אלגוריתם שכן קרוב



דוגמה לסיווג עבור אלגוריתם k-NN. המבחן לדוגמה (העיגול הירוק) צריך להיות מסווג או אל המחלקה הראשונה - קבוצת המרובעים הכחולים או לחלופין, אל המחלקה השנייה - קבוצת המשולשים האדומים. אם $k=3$ (המעגל הפנימי) הוא מוקצה לקבוצה השנייה כי ישנם 2 משולשים ורק מרובע אחד בתוך המעגל הפנימי. אם $k=5$ (עיגול מקווקו) הוא מסווג למחלקה הראשונה (שלושה ריבועים לעומת שני משולשים בתוך המעגל החיצוני).

אלגוריתם השכן הקרוב או **k-Nearest Neighbors algorithm** (או בקיצור **k-NN**) הוא אלגוריתם חסר פרמטרים לסיווג ולרגרסיה מקומית^[1] שפותח לראשונה על ידי אוולין פיקס וג'וזף הודג'ס ב-1951^[2]. בשני המקרים הקלט תלוי ב-k התצפיות הקרובות במרחב התכונות. k-NN יכול לשמש לסיווג או לרגרסיה:

- **k-NN לסיווג** – בהינתן קלט של דוגמה חדשה, האלגוריתם משייכה לקבוצה. הדוגמה משויכת למחלקה הנפוצה ביותר בקרב k השכנים הקרובים (כאשר k מוגדר כמספר חיובי שלם, בדרך כלל מספר קטן). אם $k=1$ האובייקט משויך למחלקה של השכן הבודד הקרוב ביותר.
- **k-NN לרגרסיה** – בהינתן דוגמה חדשה, האלגוריתם מחזיר ערך מאפיין לדוגמה. ערך זה הוא ממוצע ערכים של ערכי k השכנים הקרובים ביותר.

k-NN הוא אלגוריתם לימוד מבוסס מופעים, או למידה עצלה, שבו הפונקציה מקורבת באופן מקומי בלבד וכל החישובים נדחים עד סיווגה. אלגוריתם k-NN הוא מבין האלגוריתמים הפשוטים ביותר בתחום למידת המכונה.

שקלול תרומתם של השכנים יכול להיות שימושי גם במקרה של סיווג וגם במקרה של רגרסיה, כך שמשקל השכנים הקרובים תורם יותר לממוצע מהשכנים הרחוקים יותר. לדוגמה שיטת שקלול נפוצה מורכבת כך שנותנים לכל שכן משקל של $\frac{1}{d}$, כאשר d הוא המרחק לאותו שכן.^[3]

השכנים נלקחים מתוך סדרת אובייקטים של מחלקה (עבור k-NN לסיווג) או אפיון הערך (עבור k-NN לרגרסיה) ידועים. חיסרון בולט של האלגוריתם הוא רגישותו למבנה המקומי של הנתונים.

אלגוריתם

הקלט לשלב האימון של האלגוריתם הוא דוגמאות אימון, וקטורי תכונות במרחב רב ממדי כל אחד עם תווית סיווג (למשל $\{1, 2\} \times \mathbb{R}^d$ עבור וקטור תכונות של d ממדים ושתי מחלקות סיווג). שלב האימון מתבסס רק על אחסון תכונות הווקטור ותווית הסיווג של דוגמאות האימון במבנה נתונים שיאפשר בהמשך חיפוש מהיר בהם, כדוגמת עץ kd.

הקלט לשלב הסיווג הוא וקטור ללא תווית (למשל \mathbb{R}^d). בשלב הסיווג k מוגדר כקבוע, והמסווג קובע את תווית הסיווג על פי התווית השכיחה ביותר בקרב k דוגמאות האימון הקרובות לדוגמה הנבדקת.

מטריקת מרחק

מטריקה מקובלת למדידת מרחק בין משתנים רציפים היא מרחק אוקלידי. עבור משתנים בדידים, כגון בעיית סיווג טקסט, ניתן להשתמש במטריקה אחרת, כגון מרחק המינג. לעיתים קרובות, דיוק הסיווג של k -NN ניתן לשיפור באופן משמעותי אם המטריקה שבשימוש נלמדת באמצעות אלגוריתמים מיוחדים כמו "Large margin nearest neighbor" או "Neighbourhood components analysis".

מחלקות לא מאוזנות

חיסרון משמעותי של סיווג כזה, לפי השכיח המקומי, הוא כאשר התפלגות המחלקות מוטה, כלומר: מספר דוגמאות האימון באחת המחלקות גדול בהרבה ממספר דוגמאות האימון במחלקות האחרות. במקרה כזה, רוב הדוגמאות החדשות יסווגו למחלקה זו, כי בשל מספרן הגדול, דוגמאות האימון ממחלקה זו נוטות להיות נפוצות בקרב k השכנים הקרובים לדוגמה החדשה.^[4] דרך אחת להתמודדות עם בעיה זו היא הקצאת משקל לסיווג, שייקבע לפי מרחק הדוגמה מ־ k השכנים הקרובים. הסיווג (או הערך, במקרה של בעיית גרסיה) של כל אחת מ־ k הנקודות הקרובות יוכפל במשקל הפרופורציונלי להופכי של המרחק מנקודת האימון לנקודה הנוכחית. דרך אחרת כדי להתגבר על סטיות במדידה היא הפשטה בייצוג נתונים. לדוגמה, ברשת קוהונן, כל צומת מייצגת מרכז של קבוצת נקודות דומות, ללא קשר לצפיפותן בנתוני האימון המקוריים.

בחירת פרמטרים

הבחירה הטובה ביותר של k תלויה בנתונים; בדרך כלל, ערכים גבוהים יותר של k גורמים לצמצום ההשפעה של הרעש על סיווג,^[5] אבל גורמים לגבולות בין מחלקות להיות פחות מובהקים. k טוב יכול להיבחר באמצעות מספר שיטות. במקרה המיוחד בו נחזתה מראש סוג המחלקה כמחלקה הקרובה ביותר לנקודות האימון (כלומר כאשר $k=1$) נקרא אלגוריתם השכן הקרוב ביותר.

הדיוק של אלגוריתם k -NN יכול להיפגע קשות על ידי נוכחות של רעש או תכונות לא רלוונטיות, או אם סקלת התכונה אינה עקבית עם חשיבותה. מאמצי מחקר רבים הושקעו עבור בחירת תכונות או דירוג תכונות לשיפור סיווגם.

בסיווג בינארי (דו-ערכי; כאשר יש שתי מחלקות), כדאי לבחור את k להיות מספר אי-זוגי כדי להימנע ממצבי תיקון. דרך אחת פופולרית לבחירת k אופטימלי באופן אמפירי למצב זה היא באמצעות שיטת אתחול.^[6]

מאפיינים

k -NN הוא מקרה פרטי של הערכת צפיפות משתני קרנל, של הערכת רוחב פס משתנים ושל הערכת צפיפות קרנל "בלון" עם אחידות סטטיסטית בקרנל.^{[7][8]}

בצורתו הפשוטה של האלגוריתם הוא נדרש לחשב המרחקים בין הדוגמה לסיווג לכל דוגמאות האימון, אך גישה נאיבית כזו דורשת חישובים רבים כאשר יש דוגמאות אימון רבות. עם זאת שימוש באלגוריתם יעיל לחיפוש שכן קרוב מאפשר להשתמש ב-k-NN גם עבור דוגמאות רבות. במהלך השנים הוצעו מספר רב של אלגוריתמי חיפוש לשכן הקרוב; בכלל ניסו אלגוריתמים אלה לצמצם את מספר הערכות המרחק שמבוצעות בפועל.

כאשר מספר הדוגמאות שואף לאינסוף, לאלגוריתם מובטח שיעור שגיאה מרבי לא יותר מפעמיים שיעור השגיאה של בייס (השגיאה המינימלית הניתן להשגה בהתחשב בהתפלגות הנתונים).^[9]

הפחתת ממדים

עבור נתונים רב-ממדיים (לדוגמה, אם מספר ממדים גדול מ-10), הפחתת ממדים מבוצעת בדרך כלל לפני הפעלת k-NN על הנתונים, כדי למנוע את ההשפעות של קללת הממד (Curse of dimensionality).^[10]

משמעותה של קללת הממד, בהקשר של k-NN, היא שהמרחק האוקלידי אינו מדד יעיל למרחק במרחב מממד גבוה, כיוון שכל הווקטורים הם שווי מרחק ביחס לווקטור הנבדק (דמיינו מספר נקודות מונחות פחות או יותר על עיגול שלם עם נקודת שאילתה במרכז; המרחק בין נקודות השאילתה לכל נקודות המידע הוא כמעט זהה).

ניתן לשלב בפעולה אחת חילוף תכונות והורדת ממד באמצעות שיטות ניתוח גורמים ראשיים (PCA), ניתוח הבחנה ליניארי (LDA) או ניתוח מתאם קנוני (CCA) כשלב טרום עיבודי, ולאחר מכן יצירת אשכולות על ידי k-NN בשיכון במרחב בממד נמוך יותר (embedding).^[11] הורדת ממד יכולה להיעשות גם באמצעות הורדת ממד אקראית.

גבול ההחלטה

כללי השכן הקרוב ביותר למעשה מחשבים במרומז את גבול ההחלטה. כמו כן ניתן לחשב את גבול ההחלטה באופן מפורש, כדי לעשות זאת באופן יעיל כך שמורכבות החישוביות היא פונקציה של מורכבות הגבול.^[12]

צמצום נתונים

צמצום נתונים היא אחת הבעיות החשובות ביותר בבעיות עם מספר דוגמאות רב. בדרך כלל, לסיווג מדויק דרושות רק כמה נקודות, או "אבות טיפוס" וניתן למצוא אותן כך:

1. בחירת "Class-outliers", כלומר, דוגמאות שסווגו לא נכון על ידי k-NN עבור k נתון.

2. יש להפריד את שאר הנתונים לתוך שתי קבוצות: (א) אבות הטיפוס המשמשים לסיווג (ב) "הנקודות הנבלעות" - נקודות ש-k-NN יכול לתקן את סיווגן באמצעות אבות הטיפוס, ונקודות אלו ניתן להסיר ממדגם האימון.

בחירה של Class-outliers

דוגמאות האימון המוקפות בדוגמאות של מחלקות אחרות נקראות Class outlier. הגורמים של Class outliers כוללים:

- שגיאה אקראית
- לא מספיק דוגמאות אימון במחלקה זו (קיימת דוגמה בודדת במקום מקבץ דוגמאות)
- חוסר בתכונות חשובות (המחלקות מופרדות בממדים שונים לא ידועים)
- ריבוי דוגמאות ממחלקות אחרות (מחלקות לא מאוזנות) אשר יוצרות רקע עוין למחלקה הקטנה

בשילוב עם k-NN מייצר רעשים. הרעשים יכולים להיות מזוהים ומופרדים לצורך ניתוח עתידי. בהינתן שני מספרים טבעיים, $k > r > 0$, דוגמת אימון נקראת k,r-NN class-outlier אם k שכנים קרובים כוללים יותר מ-r דוגמאות של מחלקות אחרות.

CNN עבור צמצום כמות הנתונים

אלגוריתם ממוקד לשכן הקרוב ביותר (Condensed Nearest Neighbor, CNN, מכונה גם אלגוריתם הארט) הוא אלגוריתם שנועד לצמצם את מספר הדוגמאות לסיווג k-NN.^[13] אלגוריתם זה בוחר את אבות הטיפוס U מתוך דוגמאות האימון, כך ש-k-NN יכול לסווג את U כמעט במדויק, כפי ש-1NN מסווג את כל הנתונים.

CNN עובד באופן איטרטיבי, בהינתן מדגם אימון:

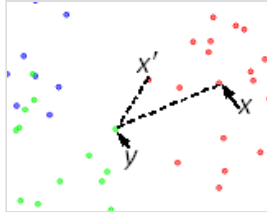
1. סרוק את כל האיברים ב-X, וחפש רכיב x שיש לו תווית שונה מאב הטיפוס הקרוב אליו (מקרב U).
2. הסר את x מ-X, והוסף אותו ל-U.
3. סרוק שוב, עד שלא יתווספו עוד אבות טיפוס ל-U.

השתמש ב-U במקום ב-X לשם הסיווג. הדוגמאות שאינן אבות טיפוס נקראות נקודות בלועות.

יחס הגבול

לשם שיפור יעילות הסריקה של CNN, נגדיר את יחס הגבול לדוגמת אימון x:

$$a(x) = \frac{\|x' - y\|}{\|x - y\|}$$

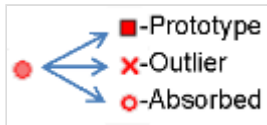


חישוב יחס הגבול.

כאשר $\|x-y\|$ הוא המרחק הקרוב ביותר לדוגמה y בעלת תווית שונה מאשר x , והמרחק $\|x'-y\|$ הוא המרחק בין y לדוגמה הקרובה ביותר אליה x' , המסווגת כמו x . יחס הגבול הוא בין 0 ל- 1 , היות שערכו של $\|x'-y\|$ לעולם אינו גדול יותר מערכו של $\|x-y\|$. סדר זה מעניק עדיפות לגבולות המחלקות לשם הכללתם בקבוצת אבות-הטיפוס U . נקודה בעלת תווית שונה מאשר x נקראת נקודה חיצונית ל- x . חישוב יחס הגבול מודגם באיור משמאל. הנתונים מסומנים בצבעים: נקודת הפתיחה היא x , והיא מסומנת באדום. הנקודות החיצוניות הן כחולות וירוקות. הנקודה החיצונית הקרובה ביותר ל- x היא הנקודה y . הנקודה האדומה הקרובה ביותר ל- y היא הנקודה x' . יחס הגבול $a(X)$ הוא תכונה של נקודת הפתיחה x .

בעזרת יחס הגבול, ניתן ליעל את סריקת דגימות האימון ב-CNN, אם הסריקה מבוצעת בסדר יורד של $a(x)$.^[14]

דוגמה



שלושה סוגים של

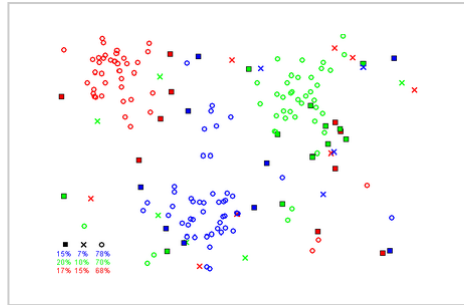
נקודות: Class-

outliers, אבות טיפוס

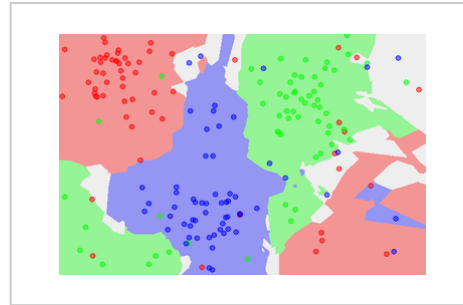
ונקודות בלועות.

להלן המחשה של CNN על קבוצת דוגמאות אימון המחולקות לשלוש מחלקות: אדום, ירוק וכחול. איור 1 מראה את המצב ההתחלתי, שבו יש 60 נקודות בכל מחלקה. איור 2 מציג את מפת הסיווג 1NN: כל פיקסל מסווג על ידי 1NN באמצעות כל הנתונים. איור 3 מציג את מפת הסיווג 5NN. אזורים לבנים תואמים את האזורים הלא מסווגים, שבהם היה תיקו בין מחלקות (למשל, אם יש שתי נקודות ירוקות, שתי נקודות אדומות ונקודה אחת כחולה מקרב חמשת השכנים הקרובים ביותר). איור 4 מציג את מקבץ הדוגמאות המצומצם. ה- x -ים הם Class-outliers שנבחרו על ידי כלל (3,2)NN (כל שלושת השכנים הקרובים ביותר של מקרים אלה שייכים למחלקות אחרות); הריבועים הם אבות-הטיפוס, והעיגולים הריקים הם הנקודות הבלועות. בצד שמאל למטה מוצגים מספרי ה-class-outlier, אבות-הטיפוס והנקודות הבלועות לכל אחת משלוש המחלקות. מספר אבות-הטיפוס נע בין 15% ל-20% עבור מחלקות שונות בדוגמה זו. איור 5 מראה שמפת הסיווג 1NN עם אבות-הטיפוס דומה מאוד למפה עם הנתונים הראשוניים. האיורים נוצרו בעזרת היישומון של מירקס.^[14]

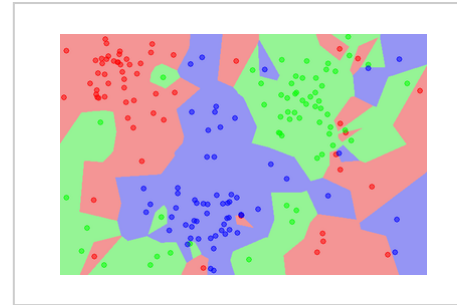
צמצום CNN עבור k-NN לסיווג



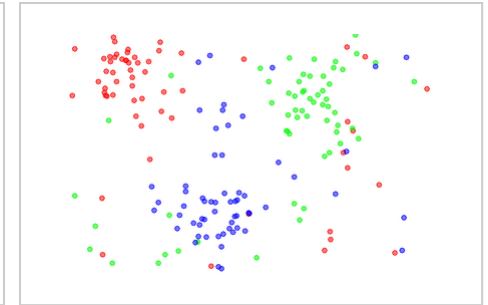
איור 4. דוגמאות האימון, לאחר צמצום CNN.



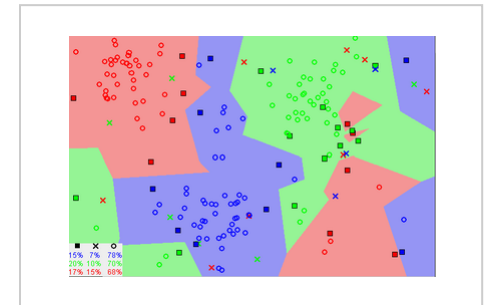
איור 3. מפת סיווג 5NN.



איור 2. מפת סיווג 1NN.



איור 1. דוגמאות האימון.



איור 5. מפת סיווג 1NN, על בסיס אבות הטיפוס שנבחרו על ידי CNN.

k-NN לרגרסיה

אלגוריתם K-NN משמש להערכת משתנים רציפים. אלגוריתם לדוגמה עשוי להשתמש במוצע משוקלל של K השכנים הקרובים ביותר, משוקללים לפי ההופכי של מרחקם. אלגוריתם זה פועל כדלהלן:

1. חשב את המרחק האוקלידי או מרחק מהלנוביס (Mahalanobis) מהנקודה הנבדקת לנקודות האימון המתוגות.
2. סדר את הנקודות המתוגות לפי מרחק עולה.
3. מצא k אופטימלי לאלגוריתם, על סמך שורש הטעות הריבועית הממוצעת. נעשה באמצעות אימות צולב.

4. חשב ממוצע משוקלל לפי המרחק ההופכי מ־k שכנים קרובים.

לקריאה נוספת

When Is "Nearest Neighbor" Meaningful? (<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.1422>)
 Belur V. Dasarathy, ed. (1991). *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. ISBN 0-8186-8930-7.
 Shakhnarovich, Darrell, and Indyk, ed. (2005). *Nearest-Neighbor Methods in Learning and Vision*. MIT Press. ISBN 0-262-19547-X.
 Mäkelä H Pekkarinen A (2004-07-26). "Estimation of forest stand volumes by Landsat TM imagery and stand-level field-inventory data". *Forest Ecology and Management*. **196** (2–3): 245–255. doi:10.1016/j.foreco.2004.02.049 (<https://doi.org/10.1016%2Fj.foreco.2004.02.049>).
 Fast k nearest neighbor search using GPU. In Proceedings of the CVPR Workshop on Computer Vision on GPU, Anchorage, Alaska, USA, June 2008. V. Garcia and E. Debreuve and M. Barlaud.
 Scholarpedia article on k-NN (http://www.scholarpedia.org/article/K-nearest_neighbor)
 google-all-pairs-similarity-search (<https://code.google.com/p/google-all-pairs-similarity-search/>)

קישורים חיצוניים

- אלגוריתם השכן הקרוב (<https://www.youtube.com/watch?v=zenh6DahgZU>), בקורס מבוא למערכות לומדות באוניברסיטה העברית

הערות שוליים

- Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". *The American Statistician*. **46** (3): 175–185. doi:10.1080/00031305.1992.10475879 (<https://doi.org/10.1080%2F00031305.1992.10475879>).
- Fix, Evelyn; Hodges, Joseph L., Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties (<https://apps.dtic.mil/dtic/tr/fulltext/u2/a800276.pdf>), 1951
- זוהי הכללה של אינטרפולציה ליניארית.
- D. Coomans; D.L. Massart (1982). "Alternative k-nearest neighbour rules in supervised pattern recognition : Part 1. k-Nearest neighbour classification by using alternative voting rules". *Analytica Chimica Acta*. **136**: 15–27. doi:10.1016/S0003-2670(01)95359-0 (<https://doi.org/10.1016%2FS0003-2670%2801%2995359-0>).
- Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011) *Miscellaneous Clustering Methods*, in *Cluster Analysis*, 5th Edition, John Wiley & Sons, Ltd, Chichester, UK.
- Hall P, Park BU, Samworth RJ (2008). "Choice of neighbor order in nearest-neighbor classification". *Annals of Statistics*. **36** (5): 2135–2152. doi:10.1214/07-AOS537 (<https://doi.org/10.1214%2F07-AOS537>).

7. D. G. Terrell; D. W. Scott (1992). "Variable kernel density estimation". *Annals of Statistics*. **20** (3): 1236–1265. doi:10.1214/aos/1176348768 (<https://doi.org/10.1214/aos/1176348768>).
8. Mills, Peter (2012-08-09). "Efficient statistical classification of satellite measurements" (<https://archive.org/details/arxiv-1202.2194>). *International Journal of Remote Sensing*.
9. Cover TM, Hart PE (1967). "Nearest neighbor pattern classification". *IEEE Transactions on Information Theory*. **13** (1): 21–27. doi:10.1109/TIT.1967.1053964 (<https://doi.org/10.1109/TIT.1967.1053964>).
10. Beyer, Kevin, et al.. "When is "nearest neighbor" meaningful? Database Theory—ICDT'99, 217-235|year 1999
11. Shaw, Blake, and Tony Jebara. 'Structure preserving embedding. Proceedings of the 26th Annual International Conference on Machine Learning. ACM,2009
12. Bremner D, Demaine E, Erickson J, Iacono J, Langerman S, Morin P, Toussaint G (2005). "Output-sensitive algorithms for computing nearest-neighbor decision boundaries". *Discrete and Computational Geometry*. **33** (4): 593–604. doi:10.1007/s00454-004-1152-0 (<https://doi.org/10.1007/s00454-004-1152-0>).
13. P. E. Hart, The Condensed Nearest Neighbor Rule. IEEE Transactions on Information Theory 18 (1968) 515–516. doi: 10.1109/TIT.1968.1054155
14. E. M. Mirkes, KNN and Potential Energy: applet. (<http://www.math.le.ac.uk/people/ag153/homepage/KNN/KNN3.html>) University of Leicester, 2011.