אלגוריתם כלל השכן הקרוב Nearest Neighbour

- משפחה של מסווגים חסרי זיכרון
- אין צורך בהתאמת מודל.

כאשר נתונה נקודת מבחן $x^{(0)}$

מוצאים את k השכנים עם המרחק הקרוב ביותר

$$x^{(i)}, \quad i = 1, 2, ..., k$$

ומסווגים בהתאם להחלטת רוב בקרב שכנים אלה, כלומר המחלקה שרוב השכנים שייכים אליה.

- במקרה של מספר שווה של שכנים לשתי מחלקות, בוחרים ביניהן באופן אקראי, או לפי המרחק הקטן ביותר.

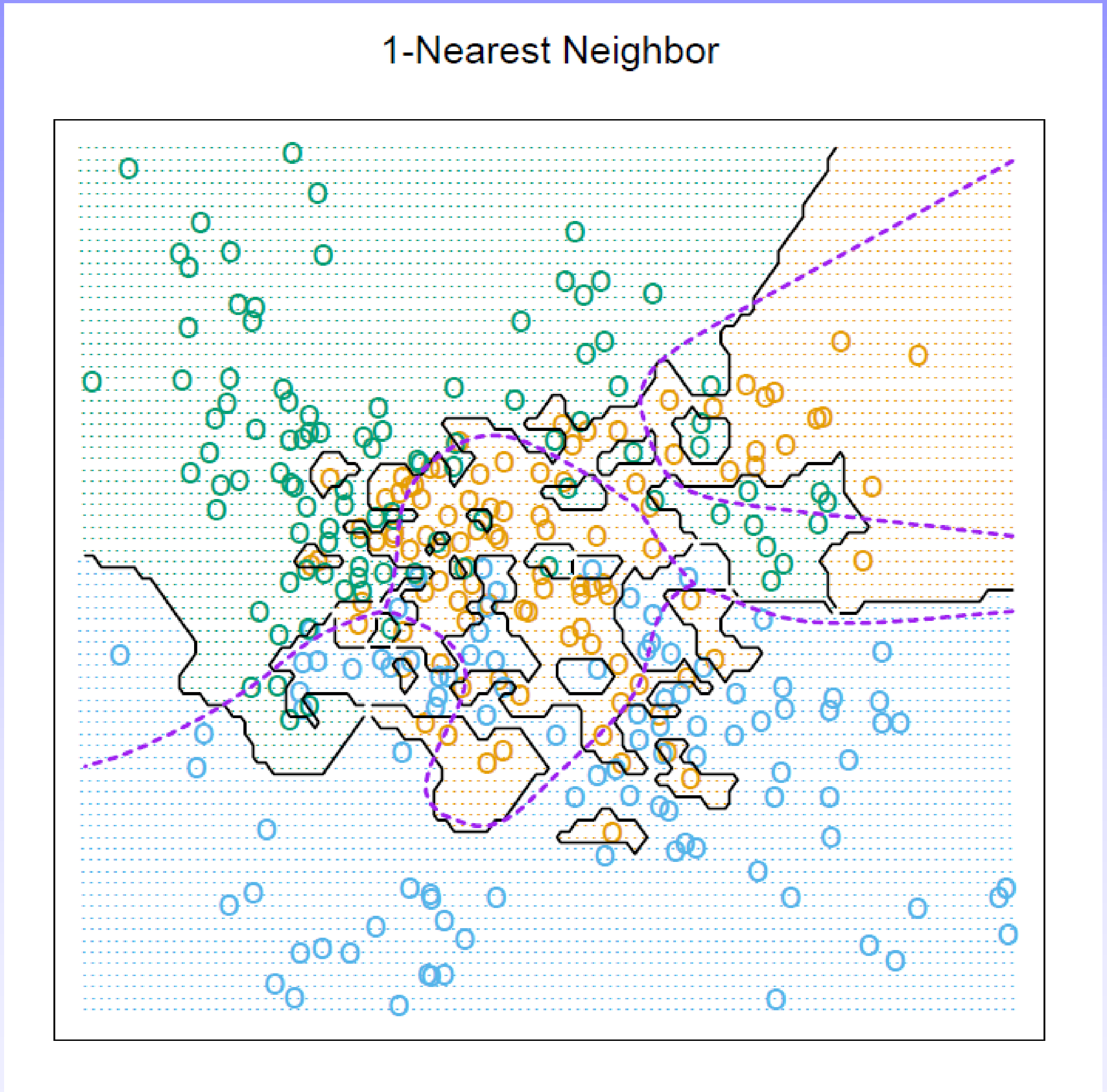
האלגוריתם

1. נניח כי $x^{(i)}$ נקודת קלט חדשה.
2. מוצאים את k השכנים הקרובים ביותר לנקודה
נסמן את השכנים הנ״ל:

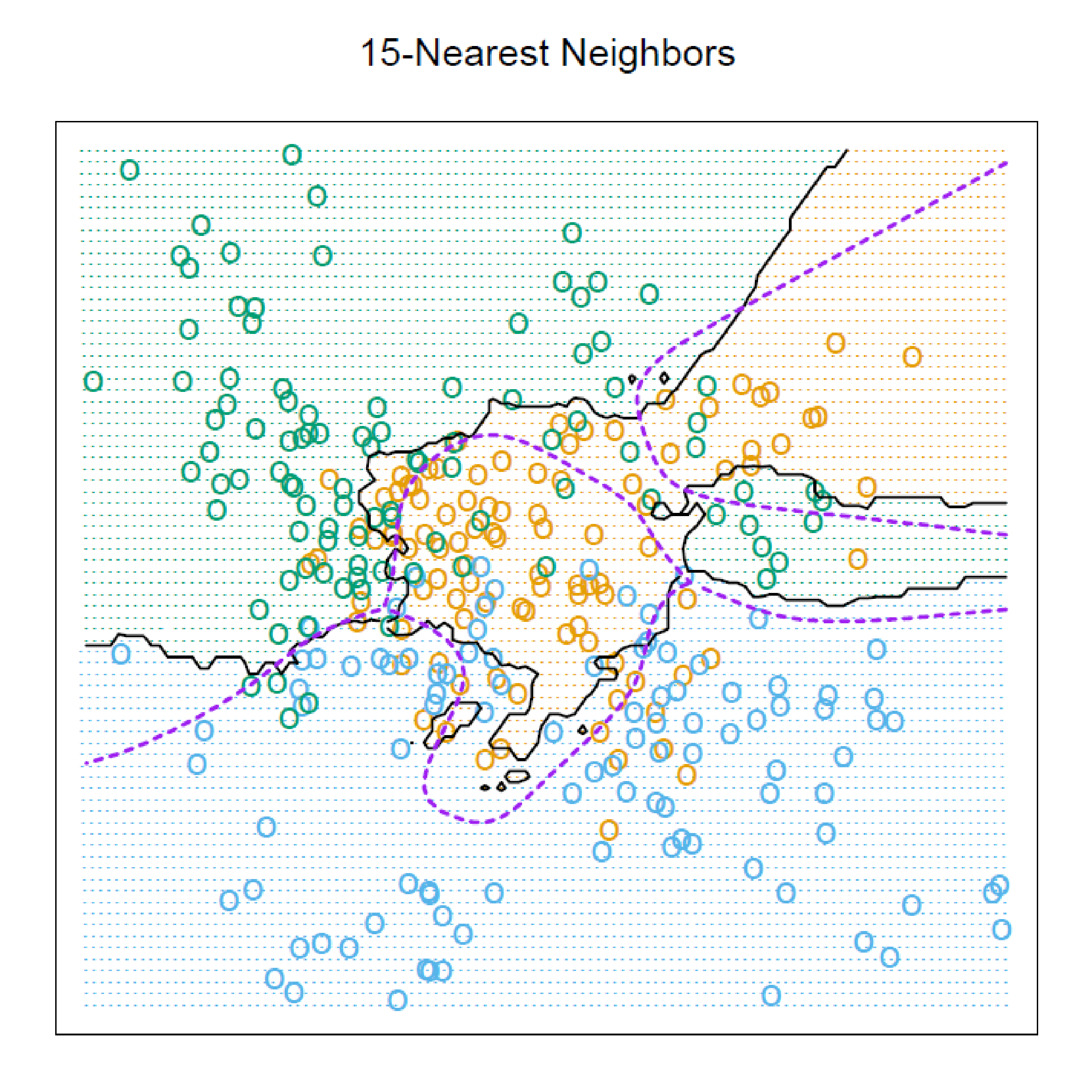
$$x^{(i)}, \quad i = 1, 2, \ldots, k$$

3. מוצאים את קבוצת הרוב, כלומר סופרים עבור כל מחלקה אפשרית מה מספר השכנים השייכים אליה.
4. אם יש שוויון בין שתי מחלקות בוחרים באופן אקראי או משווים את סכום המרחקים של השכנים מנקודת המבחן, עבור כל אחת מהמחלקות.
5. אם גם כאן יש שוויון, בוחרים באופן אקראי.

כלל השכן הקרוב



כלל K השכנים הקרובים



חסרונות ויתרונות

1. צריך לשמור בזיכרון את סדרת האימון

$$\left\{x^{(i)},\omega^{(i)}\right\},\quad i=1,2,\ldots,m$$

- כאשר מספר הדוגמאות של סדרת האימון גדול עומס חישובי רב (computational load).
  זמן החישוב ארוך.
  - עבור מספר דוגמאות קטן – ביצועים תת-אופטימליים.

2. מסווגים כאלה הם פשוטים למימוש ולתכנון.

הסתברות שגיאת הסווג

הגירסה הפשוטה ביותר של האלגוריתם היא כאשר k=1 , הידוע ככלל השכן הקרוב.
בהינתן שמספר דוגמאות האימון הוא מספיק גדול, כלל פשוט זה משיג תוצאות טובות.
ניתן להוכיח כי כאשר $N \to \infty$ הסתברות שגיאת הסווג חסומה על-ידי פעמיים שגיאת הסווג הבייסיאנית האופטימלית:

$$P_B \le P_{NN} \le P_B \cdot \left(2 - \frac{M}{M-1} P_B\right) \le 2P_B$$

ההתנהגות האסימפטוטית של מסווג ה- KNN היא אף טובה יותר :

$$P_B \le P_{KNN} \le P_B + \sqrt{\frac{2P_{NN}}{k}}$$

ועבור ערכים נמוכים של שגיאה בייאסינית אופטימלית הסתברות שגיאת ה- KNN היא מאותו סדר גודל של השגיאה האופטימלית.