מבוא לבינה מלאכותית – תרגול 9

נושא:

מדדי איכות להערכת פתרון

- **מה לא נכון לקחת כמדד איכות**
- **Training-test**
- **BIC, AIC**
- **הערכות מבוססות מרחק לאלגוריתמי clustering**

רקע:

בלמידה לא מפוקחת, אין לנו תיוגים. עד כה עסקנו באלגוריתמים שונים להתעסקות עם הנתונים שיש לנו, אבל לא הייתה שום הבטחה שהם יעבדו כמו שצריך.
שלא כמו בלמידה מפוקחת, שם אפשר בקלות לבדוק כמה טוב אנחנו מסוגלים לחזות תיוג של דוגמה, כאן מה שנוכל לעשות זה רק לדאוג שהתוצאה מספיק "יפה" באמצעות כלים כמו נראות.

מה לא לוקחים כמדד איכות:

נכון, כתוב שם למעלה "נראות". אבל הנראות לבדה לא משמשת כלי מספיק להערכת איכות המודל שלנו. כלומר, אם נבדוק מודל מסוים עם פרמטרים שונים בו, לא בהכרח נעדיף לקחת את הפרמטרים שנותנים למודל ערך נראות גבוה יותר.
הדוגמה הקלאסית היא אשכול – במצב של n דגימות, אם נבצע אשכול עם k רכיבים אז k הכי טוב לנו מבחינת הנראות יהיה $k = n$, הרי חלוקה של הדאטא כך שכל אשכול יכיל בדיוק דגימה אחת יהיה מושלם מבחינת צפיפות ומרחק משאר האשכולות. למרות זאת, כמובן שזה לא מה שאנחנו רוצים (השתמשנו ביותר מדי פרמטרים וקיבלנו overfitting).

אז מה כן יעזור לנו?

אפשרות אחת – חלוקה ל-training ו-test:

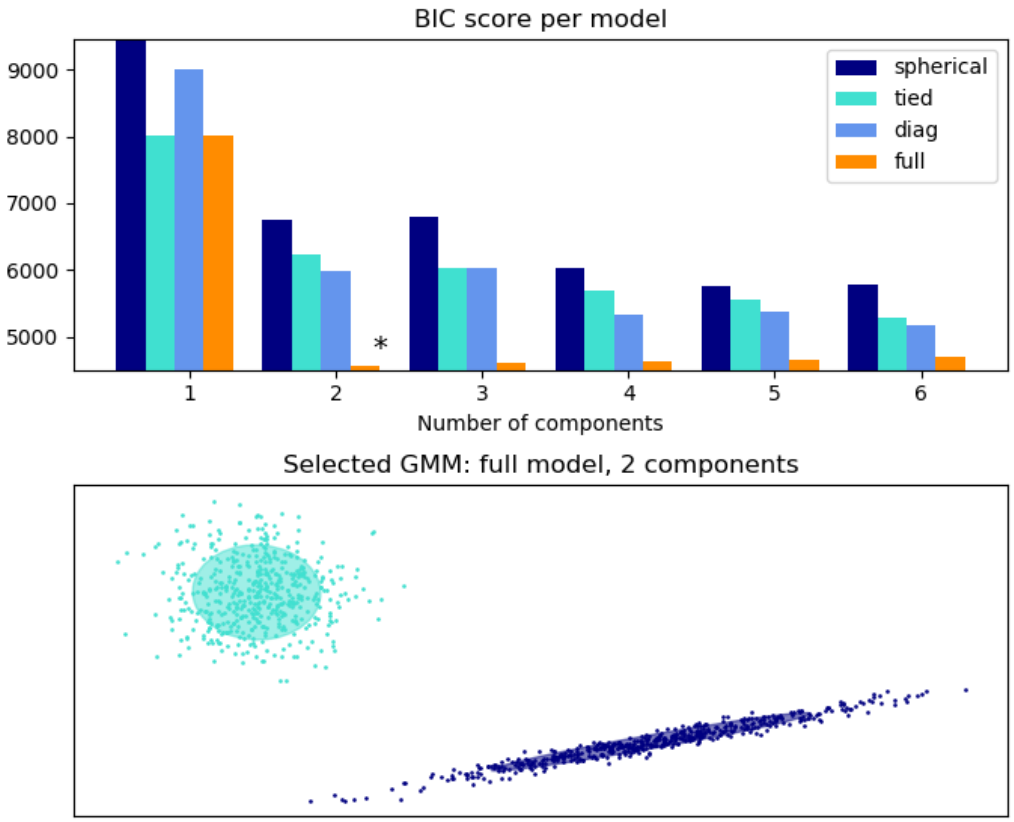
נחלק את הדאטא לקבוצת training שבעזרתה נבנה את המודל, ולקבוצה שנייה שבאמצעותה נעריך את איכותו, לפי פונקציית הנראות של הדוגמות.
שימו לב להבדל המשמעותי – כאן אנחנו מעריכים נראות לא על כל הנקודות ולא על הנקודות עליהן אימנו את המודל, אלא על נקודות אחרות. אם היינו מבצעים אשכול על n נקודות עם n אשכולות, הנראות של הנקודות החדשות הייתה גרועה.

אפשרות שנייה – BIC ו-AIC:

אלה שני מדדים שבהם מחשבים את לוג הנראות המקסימלית L ל-n הנקודות שיש לנו, אבל גם מתקנים אותה לפי מספר הפרמטרים החופשיים (כמו מספר האשכולות שבחרנו), k. ככל שנבחר יותר פרמטרים חופשיים, ככה נשלם יותר במדד שלנו.

מדד BIC (Bayesian Information Criterion) - $k\ln(n) - 2L$. ככל שיהיה נמוך יותר, כך טוב יותר.

הערה – המדד הזה טוב יותר כאשר $n \gg k$.



מדד AIC (Akaike Information Criterion) - $2k - 2L$. גם כאן נמוך יותר פירושו טוב יותר. אפשר לשים לב שהמדד לא תלוי ב-n, וזה טוב למצב שבו יש לנו תצפיות שתלויות זו בזו, שם אנחנו לא רוצים שכמות הדגימות תשפיע.

אפשרות שלישית – מדדים מבוססי מרחק עבור אשכול:

באשכולות, אלגוריתם איכותי ייתן מרחקים פנימיים (בתוך האשכול) קטנים לעומת מרחקים חיצוניים (בין נקודות מאשכולות שונים) גדולים.

Davies-Bouldin index – אם נסמן מרכזים של שני אשכולות ב-k, k', את המרחק הממוצע שבין הנקודות ששייכות לאשכול של k מ-k ב-$\sigma(k)$ ואת המרחק בין האשכולות ב-$d(k,k')$, אז לכל מרכז נשייך ערך

$$D_i = \max_{j \neq i} \frac{\sigma(i) - \sigma(j)}{d(i,j)}$$

ואז (נסמן את כמות האשכולות הכוללת שבחרנו ב-K):

$$DB = \frac{1}{K} \sum_{i=1}^{K} D_i$$

יש ורסיות דומות נוספות. הכי חשוב – הערך הטוב ביותר עבורנו הוא זה שייתן את הגודל הנמוך ביותר.

Dunn index – נסמן עבור שני אשכולות i, j את המרחק ביניהם (תלוי איך נגדיר אותו, אבל לא משנה לענייננו) ב-$d(i,j)$ ואת המרחק בתוך אשכול (שוב, כמה הגדרות ופחות מעניין כרגע) ב-$d(i)$.
אז:

$$DI = \frac{\min\limits_{i,j} d(i,j)}{\max\limits_{i} d(i)}$$

הפעם, ככל שהערך גדול יותר, כך טוב יותר.

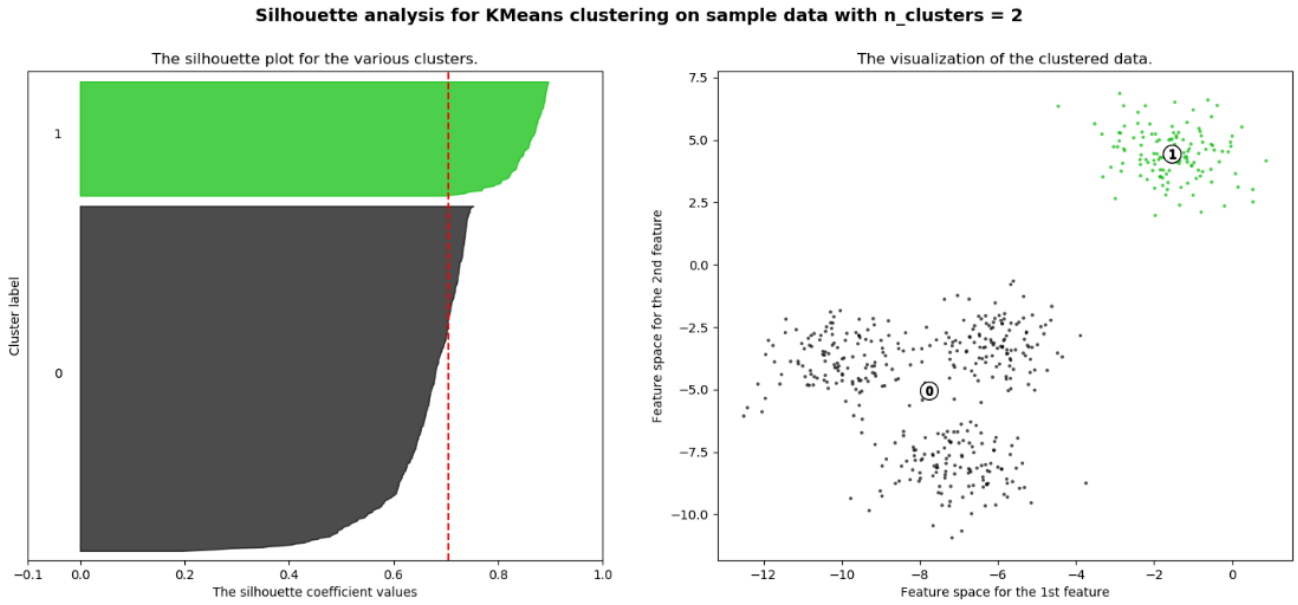
Silhouette score – זהו מדד יותר אינפורמטיבי מהשניים הקודמים, כי אפשר לחשב אותו לכל נקודה בדאטא בנפרד. הוא גם שימושי יותר כי הוא נוטה שלא לתת פתרונות טריוויאליים.
בהינתן קדקוד i ששייך לאשכול j, נגדיר: $a(i)$ – מרחק ממוצע משאר הנקודות ב-j, $b(i)$ – מרחק מינימלי של i מנקודה מחוץ ל-j.
אז:

$$S(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)}$$

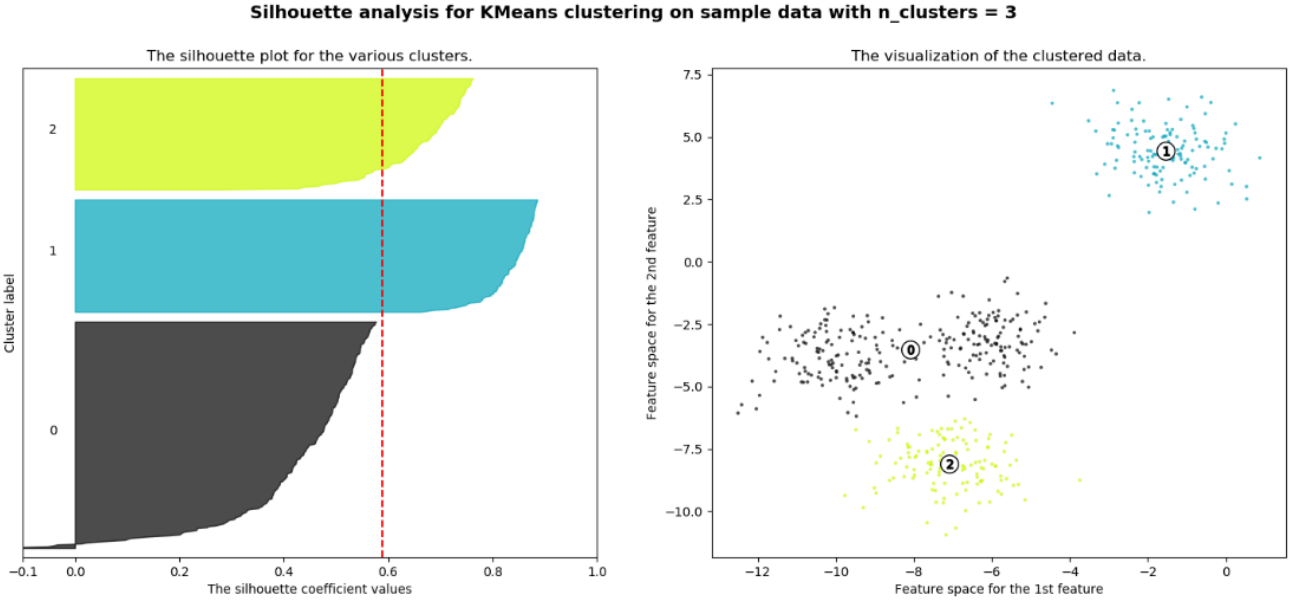
המדד עבור אשכול יהיה ממוצע של המדד לכל אחת מהנקודות באשכול, והמדד עבור כלל המערכת יהיה ממוצע על פני כל המדדים של האשכולות.

מדד זה נע בין 1- (הנקודה ממש לא שייכת לאשכול) לבין 1 (הנקודה מתאימה בצורה מושלמת לאשכול).

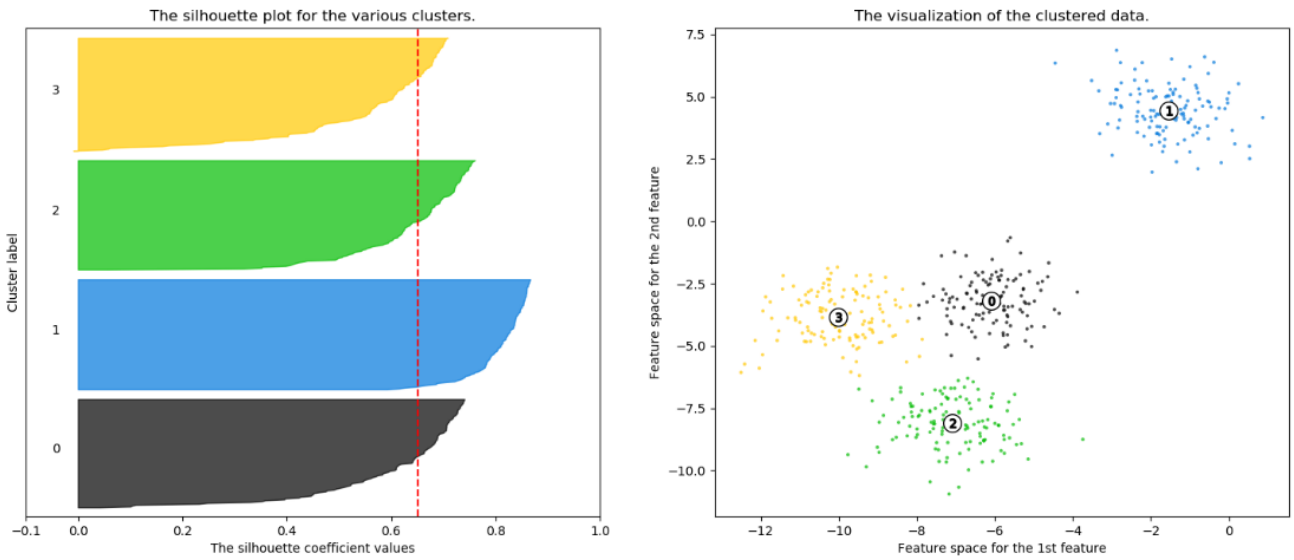
דוגמה לבחירת מספר אשכולות בעזרת silhouette score:



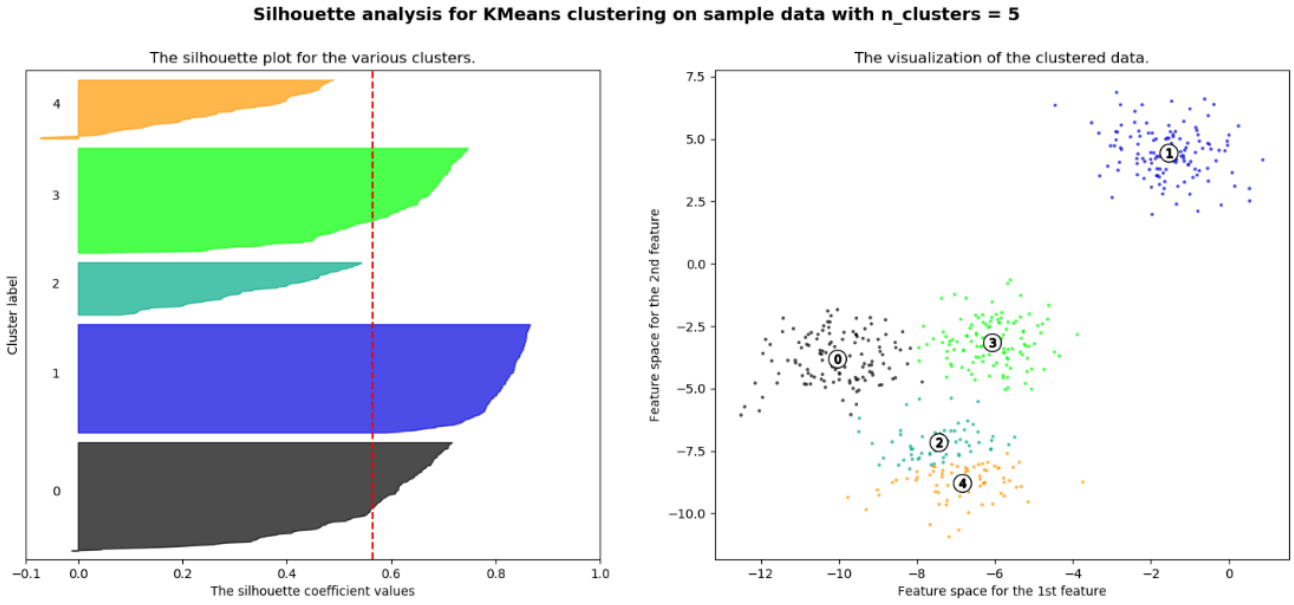
כאן ה-silhouette score היה 0.7049787496083262.



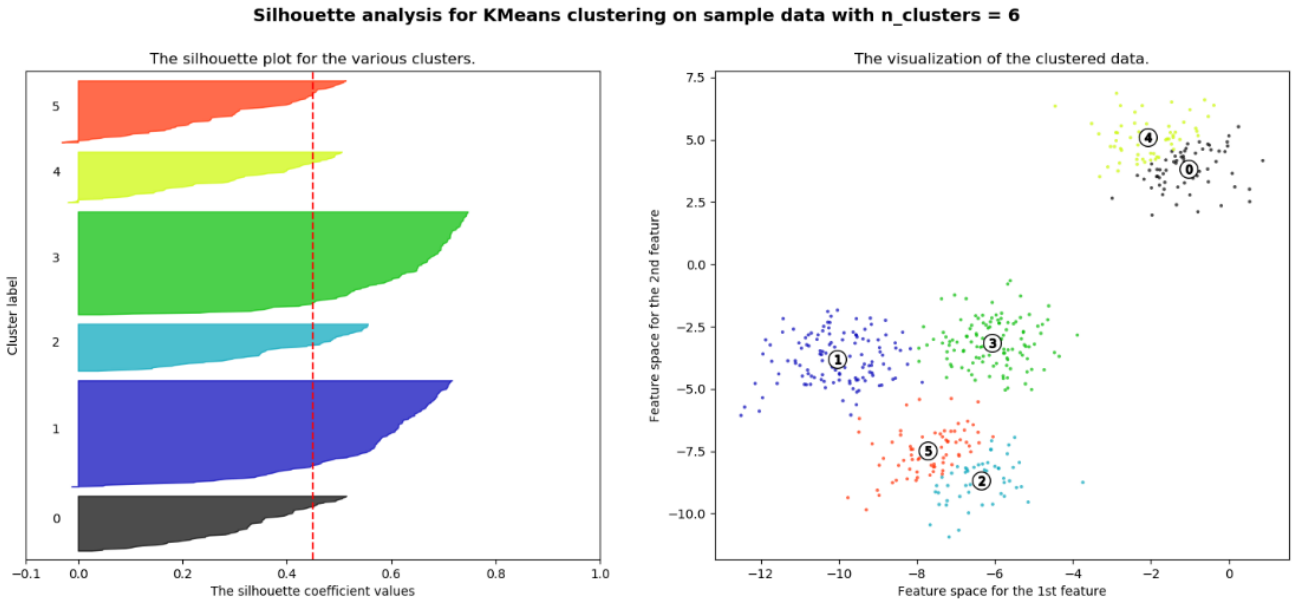
כאן ה-silhouette score היה 0.5882004012129721.



כאן ה-silhouette score היה 0.6505186632729437.



כאן ה-silhouette score היה 0.56376469026194.



כאן ה-silhouette score היה 0.4504666294372765.

בחירה מבוססת על גודל המדד לבדה משאירה אותנו לקחת 2 אשכולות. אולם, הסתכלות על כל אשכול לגופו מראה שאולי עדיף יותר לקחת 4 אשכולות.