Machine learning

Feature Selection

Exercise V

פיתוח:
ד"ר יהונתן שלר
משה פרידמן

התפלגות במדגם – ממוצע, סטית תקן ושונות משותפת

מדגם (sample): מדגם הוא קבוצת פרטים, המהווה מודל לאוכלוסייה, שאליה היא שייכת. אצלינו – ה-train-set.

ממוצע במדגם: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

שונות במדגם: : $Variance(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}\ )\text{^}2}{n-1}$

סטיית התקן במדגם: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

השונות המשותפת: $Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}\ )\cdot(y_i - \bar{y})}{n-1}$

מקדם המתאם של פירסון

לצורך חישוב מקדם המתאם, נשתמש בחישובים, לפי המדגם (אע"פ שזה לא משנה לצורך המתאם):

Pearson Correlation Coefficient

ממוצע במדגם: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{S_X S_Y}$$

$$= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

סטיית התקן במדגם: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

השונות המשותפת: $Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$

תרגיל 1 - שונות משותפת ומקדם המתאם

ליאורה עובדת כיועצת השקעות.

כחלק מתפקידה, עליה לבנות תיק השקעות בעל סיכון מאוזן.

התיק של הלקוח שלה עוקב בעיקר אחר ביצועי מדד תל-אביב 100.

הלקוח שלה, רוצה להוסיף עוד מניות טבע, וליאורה רוצה לוודא שהוא אינו מוסיף עוד סיכון הדומה לסיכון הקיים.

לשם כך רוצה ליאורה להשתמש במדד Covariance.

תרגיל 1 - שונות משותפת ומקדם המתאם – פתרון – שלב א' – covariance – חלק i

	Tel Aviv 100	Teva
2015	1148	65
2016	1338	88
2017	1276	93
2018	1454	94
2019	1699	122

Covariance in the sample:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X}) \cdot (y_i - \bar{Y})}{n-1}$$

<u>הנתונים (המומצאים) מהשנים האחרונות:</u>

חלק i - חשבו קודם כל את הממוצעים

<u>ממוצע ת"א 100</u>

$$\frac{1148 + 1338 + 1276 + 1454 + 1699}{5} = 1383$$

<u>ממוצע טבע</u>

$$\frac{65 + 88 + 93 + 94 + 122}{5} = 92.4$$

תרגיל 1 - שונות משותפת ומקדם המתאם – פתרון – שלב א' – covariance – חלק ii

	Tel Aviv 100	Teva	TA 100 - Avg(TA 100)	Teva - Avg(Teva)
2015	1148	65	-235.0	-27.4
2016	1338	88	-45.0	-4.4
2017	1276	93	-107.0	0.6
2018	1454	94	71.0	1.6
2019	1699	122	316.0	29.6

חלק ii - חשבו את ההפרשים מהממוצעים

ממוצע ת"א 100 = 1383

ממוצע טבע = 92.4

Covariance in the sample:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X}) \cdot (y_i - \bar{Y})}{n-1}$$

תרגיל 1 - שונות משותפת ומקדם המתאם – פתרון – שלב א' – covariance – חלקים iii ו-iv

חלק iii - חשבו את מכפלות ההפרשים מהממוצעים

	Tel Aviv 100	Teva	TA 100 - Avg(TA 100)	Teva - Avg(Teva)
2015	1148	65	-235.0	-27.4
2016	1338	88	-45.0	-4.4
2017	1276	93	-107.0	0.6
2018	1454	94	71.0	1.6
2019	1699	122	316.0	29.6

	(TA100-avg(TA100))*(Teva-avg(Teva))
2015	6439.0
2016	198.0
2017	-64.2
2018	113.6
2019	9353.6

חלק iv – חשבו סכום וחלוקה

סכום ההפרשים= 16040

ולכן, השונות המשותפת =

16040 / (5 – 1) = 4010

כיוון המדד והמניה זהים ולכן לא מומלץ להשקיע במניה.

Covariance in the sample:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X}) \cdot (y_i - \bar{Y})}{n-1}$$

תרגיל 1 - שונות משותפת ומקדם המתאם – פתרון – שלב ב' – מקדם המתאם – חלק i

חלק i – נחשב ראשית את סטיות התקן - קודם, נחשב הפרשים בריבוע

	Tel Aviv 100	Teva	TA 100 - Avg(TA 100)	Teva - Avg(Teva)	[TA 100 - Avg(TA 100)]^2	[Teva - Avg(Teva)]^2
2015	1148	65	-235.0	-27.4	55225.0	750.76
2016	1338	88	-45.0	-4.4	2025.0	19.36
2017	1276	93	-107.0	0.6	11449.0	0.36
2018	1454	94	71.0	1.6	5041.0	2.56
2019	1699	122	316.0	29.6	99856.0	876.16

חלק ii – נחשב ראשית את סטיות התקן – כעת, נסכם, נחלק ונוציא שורש:

עבור ת"א 100: $var = 173596 / 4 = 43399$ ולכן $std \approx 208.32$

עבור טבע: $var = 1649.19 / 4 \approx 412.3$ ולכן $std \approx 20.3$

תרגיל 1 - שונות משותפת ומקדם המתאם – פתרון – שלב ב' – מקדם המתאם – חלק iii

חלק iii – נחשב את המתאם:

	Tel Aviv 100	Teva
2015	1148	65
2016	1338	88
2017	1276	93
2018	1454	94
2019	1699	122

s(TA100) ≈ 208.32 , s(Teva) ≈ 20.3

Cov(TA100,Teva) = 4010

Pearson Correlation Coefficient =

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{S_X S_Y}$$

$$= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

$$\text{Pearson Correlation Coefficient} = \frac{4010}{208.32 \cdot 20.3} \approx 0.948$$

גם חוזק הקשר גבוה

שימו לב: באותו אופן, גם שני מאפיינים עם מתאם גבוה, עלולים להוות בעיה