

מדען נתונים, על נוסחת בייס כבר שמעת?

לפעמים בלמידת מכונה אנו מעוניינים לאמוד את ההסתברות לקבלת תוצאה מסוימת מתוך נתונים. התוצאה יכולה להיות לקוח שמגיע לחדלות פירעון על הלוואה מסוימת או עסקה שמתבררת כעסקת הונאה. על פי רוב, קיימת הסתברות א-פרוורית (התחלתית) מסוימת לקבלת התוצאה. כאשר מגיע מידע חדש, ההסתברות מתעדכנת להסתברות המותנית (Conditional) באמצעות נוסחת בייס (Bayes Theorem). נוסחת בייס משמשת לחישוב הסתברויות מותנות.

תומאס בייס הציג את נוסחת בייס בערך בשנת 1760. נסמן ב- $P(X)$ את ההסתברות לקרות מאורע X וב- $P(Y|X)$ את ההסתברות המותנית לקרות מאורע Y מותנה בכך שמאורע X קרה. נוסחת בייס גורסת כי-

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

ההוכחה של נוסחת בייס היא די אינטואיטיבית. מתוך המשמעות של הסתברויות מותנות:

$$P(Y|X) = \frac{P(X \text{ and } Y)}{P(X)}$$

וגם

$$P(X|Y) = \frac{P(X \text{ and } Y)}{P(Y)}$$

אם נבטא את שתי המשוואות לעיל באמצעות $P(X \text{ and } Y)$ ונשווה ביניהן, נקבל את נוסחת בייס.

נניח שבנק מסוים מנסה לזהות לקוחות שמנסים לבצע עסקאות הונאה בסניפיו. נניח ש- 90% מעסקאות ההונאה הן מעל ל- 100,000 ש"ח ומתבצעות בין 4 אחה"צ ל- 5 אחה"צ. עוד נניח כי רק 1% מהעסקאות הן עסקאות הונאה ו- 3% מכל העסקאות הן מעל ל- 100,000 ש"ח

ומתבצעות בין 4 אחה"צ ל- 5 אחה"צ.

במקרה שכזה נגדיר:

X : עסקאות מעל ל- 100,000 ש"ח שמתבצעות בין 4 אחה"צ ל- 5 אחה"צ.

Y : עסקאות הונאה.

אנחנו יודעים ש- $P(Y) = 0.01$, $P(X|Y) = 0.90$ ו- $P(X) = 0.03$. מתוך נוסחת בייס:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{0.90 \times 0.01}{0.03} = 0.3$$

ההסתברות שעסקה מקרית מסוימת היא עסקת הונאה היא רק 1%. אבל כאשר ידוע שאותה עסקה היא מעל ל- 100,000 ש"ח ושהיא התבצעה בין 4 אחה"צ ל- 5 אחה"צ, או אז נוסחת בייס מעדכנת את הסתברות מ- 1% ל- 30%. המשמעות של זה היא ברורה. אם לבנק יש מערכת לאישור עסקאות ב- Online, הרי שהיא אוטומטית לא צריכה לאשר עסקאות שהן מעל ל- 100,000 ש"ח ושמבצעות בין 4 אחה"צ ל- 5 אחה"צ.

נוסחת בייס מאפשרת לנו לחשב הסתברויות מותנות. לפעמים נוסחת בייס מייצרת תוצאות שנוגדות את האינטואיציה. נניח שמבחן לגילוי מחלה מסוימת הוא "מדויק ב- 99%". לאמור- כאשר לבנאדם מסוים יש את המחלה, אזי המבחן מספק תוצאה חיובית (קרי, הוא מנבא/חוזר שלבנאדם יש את המחלה) ב- 99% מהזמן. אנו מניחים גם שכאשר לבנאדם אין את המחלה, אז המבחן מספק תוצאה שלילית (קרי, הוא מנבא/חוזר שלבנאדם אין את המחלה) ב- 99% מהזמן. נניח שהמחלה היא כל כך נדירה כך שההסתברות (הלא מותנית) שלבנאדם יש את המחלה היא 1 מתוך 10,000 או 0.0001. נניח שעשית את המבחן והתוצאה יצאה חיובית, מהי ההסתברות שיש לך את המחלה?

התשובה הטבעית לשאלה זו היא 99%. (אחרי הכל, המבחן מדויק ב-99%). עם זאת, תוצאה זו לא עולה בקנה אחד עם נוסחת בייס. נניח ש- X מצביע על כך שתוצאת המבחן היא חיובית וש- Y מצביע על כך שלבנאדם יש את המחלה. אנחנו מעוניינים למעשה ב- $P(Y|X)$.

אנחנו יודעים ש- $P(X|Y) = 0.99$ וש- $P(Y) = 0.0001$. כעת נרחיב את הסימונים שלנו כך ש- \bar{X} מצביע על כך שתוצאת המבחן היא שלילית וש- \bar{Y} מצביע על כך שלבנאדם אין את המחלה.

$$P(\bar{X}|\bar{Y}) = 0.99 \text{ וש-} P(\bar{Y}) = 0.9999$$

$$\text{מאחר ו-} P(\bar{X}|\bar{Y}) + P(X|\bar{Y}) = 1 \text{ הרי ש-}$$

$$P(X|\bar{Y}) = 0.01$$

ואז אנחנו יכולים למעשה לחשב את ההסתברות שהמבחן יספק תוצאה חיובית כ-

$$P(X) = P(X|Y)P(Y) + P(X|\bar{Y})P(\bar{Y})$$

$$P(X) = 0.99 \times 0.0001 + 0.01 \times 0.9999 = 0.0101$$

באמצעות נוסחת בייס:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{0.99 \times 0.0001}{0.0101} = 0.0098$$

מכאן עולה שיש פחות מ-1% סיכוי שיש לך את המחלה אם תוצאת המבחן היא חיובית. תוצאת המבחן מגדילה הלכה למעשה את ההסתברות שיש לך את המחלה פי 98 (במקרה דגן שלפנינו כמובן), מהסתברות לא מותנית של 0.0001 להסתברות מותנית של 0.0098 (אבל עדיין הסתברות נמוכה). הנקודה העיקרית היא שרמת הדיוק מוגדרת באמצעות ההסתברות לקבל את התוצאה הנכונה בהינתן שלבנאדם מסוים יש את המחלה, אך לא להיפך.

פרטים אודות כותב המאמר: האקטואר רועי פולניצר, FRM

רועי בעל תואר שני במימון (התמחות בניהול סיכונים ואקטואריה) ותואר ראשון בכלכלה (התמחות במימון), שניהם מאוניברסיטת בן-גוריון בנגב, בעל דיפלומה בניהול סיכונים פיננסיים (FRM®) מאוניברסיטת אריאל בשומרון ולמד בתוכנית ללימודי תעודה באקטואריה באוניברסיטת חיפה. כמו כן, רועי אקטואר מלא



(Fellow) בלשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (F.I.L.A.V.F.A.), מוסמך כמעריך שווי מימון תאגידי (CFV) מטעם לשכת מעריכי השווי והאקטוארים הפיננסיים בישראל (IAVFA), מוסמך כמנהל סיכונים פיננסיים (FRM) מטעם האיגוד העולמי למומחי סיכונים (GARP) ומוסמך כמומחה לניהול סיכונים (CRM) מטעם האיגוד הישראלי למנהלי סיכונים (IARM).

לרועי ניסיון של מעל ל-15 שנה בביצוע ניתוחים כמותיים במכשירים פיננסיים, בהערכת שווי תאגידים ונכסים בלתי מוחשיים, באמידה וכימות סיכונים כמו תמותה, אריכות ימים, תחלואה, ביטולים והחלמה מנכות, ובמידול ומדידת סיכוני שוק, אשראי, תפעוליים, מודל, מזילות והשקעות לצורכי יישום הוראות רגולטוריות ותקינה חשבונאית, פיתוח, יישום ותיקוף מודלים בתחומים של הערכות שווי, ניהול סיכונים, אקטואריה והנדסה פיננסית, קביעת תעריפי ביטוח חיים, הערכת פרמיות סיכון והערכת עתודות ביטוח, קביעת עלות תנאי פנסיות (צוברות ותקציביות) והכנת מאזנים אקטואריים לקרנות פנסיה, ניתוח וחזוי מצבים פיננסיים מורכבים וכן העברת סמינרי הדרכה והשתלמויות בתחומי התמחותו: מימון, אקטואריה, הערכות שווי, בנקאות, ניהול סיכונים, אופציות והנדסה פיננסית.



ניסיונו של רועי בתחום ה-Data Analysis, כולל: עבודה עם מאגרי מידע גדולים Big Data תוך שימוש ב-Statistical Learning (כגון: סטטיסטיקה תיאורית, הסתברות, הסקה סטטיסטית, סטטיסטיקה א-פרמטרית, חלוקת נתונים, נרמול נתונים, Fitting ו- Bayes Theorem) ובאלגוריתמים מסוג Unsupervised Learning (כגון: Hierarchical Clustering, k-means Clustering, Density-based Clustering, Distribution-based Clustering ו- Principle Components Analysis) למציאת דפוסים וזיהוי מגמות ואנומליות בעולמות ניהול הסיכונים, ההשקעות, האקטואריה, הביטוח והפנסיה, פיתוח תשתית לצורך ניתוח נתונים, שילוב והטמעת כלים לצורך גישה ושליפה עצמאית של נתונים ממאגרי מידע, פיתוח דוחות, ממשקים ומסכים באמצעות כלי ויזואליזציה.

ניסיונו של רועי בתחום ה-Data Science, כולל: עבודה עם מסדי נתונים גדולים Big Data תוך שימוש באלגוריתמים מסוג Supervised Learning (כגון: Linear Regression, Ridge Regression, Lasso Regression, Elastic Net Regression, Logistic Regression, Maximum Likelihood Estimation, k-Nearest Neighbors, Decision Tree, Random Forest, Ensemble, Bagging, Boosting, Naïve Bayes Classifier, Linear Separation, Support Vector Machine, Non-Linear Separation, SVM Regression, Artificial Neural Network, Convolutional Neural Network ו- Recurrent Neural Network) לניבוי וסיווג בעולמות ניהול הסיכונים, ההשקעות, האקטואריה, הביטוח והפנסיה ובמודלים מסוג Reinforcement Learning (כגון: Q-learning, Monte Carlo, Simulation, Temporal Difference Learning ו- n-Step Bootstrapping) לקבלת החלטות מרובות שלבים בעולמות ניהול הסיכונים, ההשקעות, האקטואריה, הביטוח והפנסיה, זיהוי אתגרים עסקיים שבהם DATA יכול להוות גורם מכריע בשיפור קבלת החלטות, איתור ואיסוף מקורות מידע, הגדרה ואיפיון של שימושי המידע, בניית מסד המידע, אפיון והגדרת הצגת המידע



ותוצריו, פיתוח כלים, מודלים, תהליכים ומערכות בתחום האנליזה, תוך שימוש בכלי אנליזה מתקדמים (EXCEL, VBA ו-R).