מודלים הסתברותיים

הרצאה 1 (26.10.2014)

מנהלה - אתר הקורס – cs.biu.ac.il/~89-919
דרישות הקורס – מבחן 70%, תרגילי בית 30%. התרגיל הראשון הוא תיאורטי – חזרה על מבוא להסתברות (הגשה ביחידים) ו-3 תרגילים תכנותיים (ב- Java או פייתון) בזוגות.

הקורס יעסוק בגישות אמפיריות – פיתוח מודלים ומערכות שמתמודדות עם תופעות אמפיריות. בניגוד לאלגוריתמים שאפשר להוכיח את הנכונות שלהם באופן פורמלי, יש מערכות שנותנות חיזוי (למשל חיזוי מזג האוויר, בורסה וכו'). בתוך התחום הזה יש הרבה גישות, והגישה ההסתברותית היא הדומיננטית ביותר, כלומר רוצים לחזות את ההסתברות באופן הטוב ביותר.

נושאים עיקריים:

1. מבוא להסתברות
2. מבוא לתורת האינפורמציה
3. מודלים בסיסיים (בינום, מולטינום)
4. אומדן הסתברויות (MLE)
5. שיטות החלקה (Smoothing)
6. מודלים מרקוביים
7. מודלים חבויים (סיווג, קלאסטרינג)
8. אלגוריתם שערוך EM

מבוא להסתברות

המטרה: למדל תופעות לא וודאיות.
נקודת מבט על התופעה שממדלים היא ביצוע ניסוי עם תוצאות אפשריות מסוימות.

מרחב המדגם (Sample Space) – קבוצת כל תוצאות הניסוי האפשריות (סימון: S).
לדוגמה, עבור הטלת מטבע, $S = \{H, T\}$ ועבור הטלת קובייה $S = \{1, \dots, 6\}$.
בהטלת שתי מטבעות: $S = \{< H, H >, < H, T >, < T, H >, < T, T >\}$. S יכולה להיות סופית או לא. למשל, מרחב המדגם של משקל של אדם: $S = [0, \infty)$. בסיווג מסמכים לכמה נושאים, S תהיה קבוצת החזקה של קבוצת הנושאים.

מאורע (Event) – תת קבוצה של S (סימון: E). לדוגמה, $E = \{H\}$ מסמנת את המאורע שתוצאת הניסוי של הטלת מטבע הייתה "עץ", ו- $E = \{1,3,5\}$ מסמנת את המאורע שתוצאת הטלת הקובייה הייתה אי-זוגית. $E = \{< H, T >, < H, H >\}$ היא המאורע שתוצאת הטלת המטבע הראשונה הייתה "עץ". כיוון שמאורע הוא קבוצה, ניתן לדבר על איחוד וחיתוך של מאורעות, למשל: תוצאת הטלת הקובייה היא זוגית ומתחלקת ב- 3: $E = \{2,4,6\} \cap \{3,6\} = \{6\}$, וניתן גם לדבר על הפרש מאורעות והמשלים של מאורע: $\bar{E} = S - E$.

הסתברות מאורעות (Probability) – פונקציה של המאורעות.
אינטואיטיבית, אם נחזור על הניסוי הרבה פעמים, ב- $P(E)$ אחוז מהניסויים התוצאה תהיה E.
למשל, במטבע מאוזן, נצפה ש- $P(H) = P(T) = 0.5$.
$P(E)$ היא פונקציית הסתברות אם היא מקיימת את שלושת התנאים הבאים:

1. $0 \leq P(E) \leq 1$
2. $P(S) = 1$
3. לכל סדרת מאורעות זרים $E_1, E_2, \dots$ מתקיים ש- $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$

תכונות שנובעות מההגדרה:

1. $P(\bar{E}) = 1 - P(E)$.
   הוכחה: מתכונה (2), $P(S) = 1$ ומהגדרת המשלים מתורת הקבוצות $E \cup \bar{E} = S$. מתכונה (3), $P(S) = P(E) + P(\bar{E})$. משני אלה: $1 = P(E) + P(\bar{E}) \Longrightarrow P(\bar{E}) = 1 - P(E)$. ■

2. עבור שני מאורעות (לאו דווקא זרים), E, F: $P(E \cup F) = P(E) + P(F) - P(EF)$.
   הוכחה: מתורת הקבוצות:
   $$E \cup F = E \cup (F \backslash EF)$$
   $$F = (F \backslash EF) \cup EF$$
   מתכונה (3):
   $$P(F) = P\big((F \backslash EF) \cup EF\big) = P(F \backslash EF) + P(EF)$$
   $$P(F \backslash EF) = P(F) - P(EF)$$
   $$P(E \cup F) = P\big(E \cup (F \backslash EF)\big) = P(E) + P(F \backslash EF) = P(E) + P(F) - P(EF)$$

■

הסתברות מותנית (Conditional Probability) - $P(E|F)$ הסתברות של קיום מאורע E בהינתן שידוע שמאורע אחר F קרה. נקודת המבט היא צמצום מרחב המדגם ל-F וביחס אליו רוצים לבטא את ההסתברות של המאורע EF.

$$P(E|F) = \frac{P(EF)}{P(F)}$$

כאשר $P(F) > 0$.

לדוגמה, נניח שרוצים לדעת מהי ההסתברות של המילה "to" באנגלית, כלומר, אם בוחרים מילה אקראית באנגלית, מה ההסתברות שזו המילה "to". מרחב המדגם יהיה כל המילים בקורפוס. לעומת זאת, ניתן לשאול "מה ההסתברות של המילה to בהינתן שהמילה want הופיעה מיד לפניה?", ואז נסתכל בקורפוס רק על המקומות שלאחר המילה "want". נצפה ש- $P(to|want) > P(want)$.

דוגמה נוספת: כד מכיל 5 כדורים לבנים ו-5 כדורים שחורים. מוציאים באקראי שני כדורים אחד אחרי השני. מה ההסתברות ששניהם שחורים?

נסמן – E – השני שחור, F – הראשון שחור. רוצים לחשב את $P(EF)$.
$P(F) = \frac{5}{10} = \frac{1}{2}$ ההסתברות שהכדור הראשון שחור, ו- $P(E|F) = \frac{4}{9}$, לכן $P(EF) = \frac{1}{2} \cdot \frac{4}{9} = \frac{2}{9}$.

כלל השרשרת (Chain Rule) – הכללה של $P(E_1E_2) = P(E_1)P(E_2|E_1)$:

$$P(E_1E_2 \dots E_n) = P(E_1)P(E_2|E_1) \dots P(E_n|E_1E_2 \dots E_{n-1}) = P(E_1)\prod_{i=2}^{n} P(E_i|E_1 \dots E_i)$$

מאורעות בלתי תלויים (ב"ת) (Independent Events) – מאורעות E, F נקראים בלתי-תלויים אם $P(EF) = P(E)P(F)$. מההגדרה נובע שאם E, F בלתי תלויים, אז מתקיים:

$P(E|F) = \frac{P(EF)}{P(F)} = P(E), P(F|E) = \frac{P(EF)}{P(E)} = P(F)$ כלומר, הסתברות E אינה מושפעת מקרות המאורע F ולהיפך. זה לא בהכרח אומר שהמאורעות אינם קשורים זה לזה, אלא רק שמבחינה הסתברותית הם לא משפיעים אחד על השני.

לדוגמה, בהטלת זוג קוביות, המאורע F הוא שיצא 4 בקובייה הראשונה, ו-E_1 שסכום הקוביות 6.

$$P(F) = \frac{1}{6}$$

$$P(E_1F) = P(<4,2>) = \frac{1}{36} \neq P(E_1)P(F) = \frac{1}{6}\cdot\frac{5}{36}$$

לכן המאורעות תלויים.

לעומת זאת, אם E_2 – סכום הקוביות הוא 7:

$$P(E_2F) = P(<4,3>) = \frac{1}{36} = P(E_2)P(F) = \frac{1}{6}\cdot\frac{6}{36}$$

לכן המאורעות בלתי-תלויים.

הכללה של הגדרת אי-התלות: קבוצת המאורעות $E_1, \dots, E_n$ אם לכל תת-קבוצה שלהם $E_{i_1}, \dots, E_{i_k}$ מתקיים ש- $P\left(\{E_{i_1}, \dots, E_{i_k}\}\right) = \prod_{j=1}^{k} P(E_{i_j})$.

<u>הרצאה 2 (02.11.2014)</u>

<u>נוסחת בייס (Bayes)</u> –

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

הוכחה – מהגדרת ההסתברות המותנית: $P(EF) = P(E|F)P(F) = P(F|E)P(E)$. ■

ניתן להשתמש בנוסחה כאשר צריך לחשב את $P(E|F)$ וקל יותר לחשב את $P(F|E)$ ו-$P(F|\bar{E})$.

נרצה לייצג את המכנה $P(F)$ לביטויים במונה, ע"י נוסחת ההסתברות השלמה:

$$P(F) = P(FE) + P(F\bar{E}) = P(F|E)\cdot P(E) + P(F|\bar{E})\cdot P(\bar{E}) \Longrightarrow$$

$$P(E|F) = \frac{P(F|E)P(E)}{P(F|E)P(E) + P(F|\bar{E})P(\bar{E})}$$

זו נוסחת בייס עבור שני מאורעות. המקרה הכללי מוכלל ל-n מאורעות, כלומר חישוב ההסתברות F כתלות במאורע E_i מתוך סדרת מאורעות זרים $E_1, \dots, E_n$ המכסים את S ($\cup_{i=1}^{n} E_i = S$). במקרה כזה:

$$P(F) = \sum_{i=1}^{n} P(FE_i) = \sum_{i=1}^{n} P(F|E_i)P(E_i) \Longrightarrow P(E_i|F) = \frac{P(F|E_i)P(E_i)}{\sum_{j=1}^{n} P\left(F\middle|E_j\right) P\left(E_j\right)}$$

<u>שאלה לדוגמה</u>: נתונה תיקייה עם שלוש מגירות. מכתב נמצא באחת משלוש המגירות בהסתברות שווה. בחיפוש מכתב במגירה בה הוא נמצא, ההסתברות למצוא אותו היא α. נניח שחיפשנו את המכתב במגירה הראשונה ולא מצאנו אותו, מהי ההסתברות שהוא בכל זאת שם?

<u>תשובה</u>: נסמן ב- F_1, F_2, F_3 את ההסתברות שהמכתב במגירה הראשונה, השנייה והשלישית בהתאמה. מתקיים $\cup_{i=1}^{3} F_i = S$. נסמן ב- E את המאורע הנתון – סרקנו את המגירה הראשונה ולא מצאנו. צריך לחשב $P(F_1|E)$. לפי נוסחת בייס:

$$P(F_1|E) = \frac{P(E|F_1)P(F_1)}{\sum_{j=1}^{3} P\left(E\middle|F_j\right) P\left(F_j\right)} = \frac{(1-\alpha)\cdot\frac{1}{3}}{(1-\alpha)\cdot\frac{1}{3} + 1\cdot\frac{1}{3} + 1\cdot\frac{1}{3}} = \frac{1-\alpha}{3-\alpha}$$

משתנים מקריים (מ"מ, Random Variables) – כאשר מתעניינים לא ישירות בתוצאת הניסוי, אלא בפונקציה שלה, לדוגמה סכום הטלה של זוג קוביות. לפונקציה כזו קוראים משתנה מקרי. משתנה מקרי X הוא פונקציה ממרחב המדגם לקבוצת ערכים כלשהי. בדר"כ נסמן $x \in X$ כערך אפשרי של הפונקציה. מאחר שערכו של המשתנה המקרי x נקבע עפ"י תוצאת הניסוי, אזי ניתן לייחס לו הסתברות $p(x)$ שהיא סכום ההסתברויות של תוצאות הניסוי שעבורן ערך X הוא x. במילים אחרות, זו הסתברות המאורע שערך המשתנה המקרי הוא x.

לדוגמה, $p(2) = P(X = 2) = P(\{< 1,1 >\}) = \frac{1}{36}, p(3) = P(\{< 1,2 >, < 2,1 >\}) = \frac{2}{36}$.
קל לראות שמתקיים $\sum_{i=2}^{12} p(i) = 1$ כיוון ש:

א. המאורעות $p(i)$ עבור $i = 2, \dots, 12$ הם זרים כיוון ש-X היא פונקציה.
ב. איחוד המאורעות מכסה את S ולכן $P(\bigcup_i x_{=i}) = 1$.

הערה: הסתברות של ערך של משתנה מקרי מיוצגת ע"י p, והסתברות של מאורע מיוצגת ע"י P.

מ"מ בדיד – הוא משתנה שהטווח שלו הוא סופי או בר-מניה (לדוגמה: צבע שיער).
עבור ערך אפשרי a של המשתנה, $\mathrm{p(a) = P(X = a)}$.
p נקראת פונקציית מסת ההסתברות (Probability Mass Function). מתקיים $\sum_{x \in X} p(x) = 1$.
לסט $\{p(x) | x \in X\}$ נקרא הפרמטרים של התפלגות המשתנה המקרי, לדוגמה בהטלת מטבע הוגן הפרמטרים הם $\left\{\frac{1}{2}, \frac{1}{2}\right\}$. נסמן $X \sim p(x)$, כלומר המשתנה המקרי X מתפלג לפי ההתפלגות $p(x)$.

דוגמאות בסיסיות למ"מ בדידים

- מ"מ ברנולי – עבור ניסוי עם שתי תוצאות אפשריות, המ"מ X ממפה לערכים 1 ("הצלחה") ו-0 ("כישלון"). $p(1) = P(\{x_{=1}\}) = p$. מתקיים: $p(0) = 1 - p(1)$.
- מ"מ בינומי - המ"מ X הוא מס' ההצלחות בסדרת n ניסוי ברנולי בלתי-תלויים.
  עבור $i = 0, \dots, n$ מתקיים: $p(i) = \binom{n}{i} p^i (1-p)^{n-i}$. מתקיים $\sum_{i=0}^{n} p(i) = 1$.

תוחלת (Expectation) – תוחלת של מ"מ בדיד X מוגדרת ע"י: $E[X] = \sum_{x \in X} x \cdot p(x)$. זהו ממוצע משוקלל של ערכי X כאשר כל ערך משוקלל בהסתברות שלו.
אינטואיטיבית, בהרבה דגימות אקראיות, הערך הממוצע של X יהיה $E[X]$.
לדוגמה, התוחלת של קובייה הוגנת היא: $E[X] = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$.
במשתנה ברנולי, התוחלת היא: $E[X] = 1 \cdot p + 0 \cdot (1 - p) = p$.
במשתנה בינומי, התוחלת היא: $E[X] = \sum_{i=0}^{n} i \cdot p(i) = \sum_{i=0}^{n} i \cdot \binom{n}{i} p^i (1-p)^{n-i} = np$.
אינטואיטיבית, אם נחזור על ניסוי בינומי הרבה פעמים, בממוצע נקבל np הצלחות בניסוי.
תוחלת של פונקציה $g(x)$ של מ"מ – נסתכל על $g(x)$ כעל מ"מ חדש שערכיו הם $g(x)$ וההסתברות לכל ערך $g(x)$ היא $p(x)$: $E[g(x)] = \sum_{x:p(x)>0} g(x) \cdot p(x)$.

התפלגות משותפת (Joint Probability) – כאשר יש תופעה שממדלים אותה ע"י מס' משתנים מקריים (לדוגמה במזג האוויר מסתכלים על משתנים של מהירות הרוח, הלחות וכו'). נתעניין בהסתברות של צירופי משתנים מקריים, למשל X, Y כאשר מניחים שהסתברות זו נקבעת לפי התפלגות משותפת $p(x, y)$, כלומר ההתפלגות קובעת ישירות לכל צירוף ערכים x, y:
$p(x, y) = P(\{X = x, Y = y\})$.

התפלגות שולית (Marginal) – התפלגות של ערכי משתנה אחד: $p_X(x) = \sum_y p(x, y)$ ו-$p_Y(y) = \sum_x p(x, y)$. ההתפלגות השולית נגזרת מההתפלגות המשותפת.

מ"מ רציף – הוא משתנה שהטווח של הוא רצף ערכים, כגון הממשיים (לדוגמה: גובה).
p נקראת פונקציית צפיפות ההסתברות (Probability Density Function).

הרצאה 3 (09.11.2014)

בהתפלגות משותפת, לא ניתן לדבר על תוחלת הערך של מס' משתנים. ניתן להגדיר משתנה מקרי חדש מעל התפלגות משותפת, שהוא פונקציה של המשתנים המקריים, ולחשב את התוחלת שלו.
לדוגמה, אם X, Y תוצאות הטלה של שתי קוביות, $x + y$ וניתן לחשב את התוחלת של $x + y$.

תוחלת של משתנה מקרי שהוא פונקציה של x, y: $g(x, y)$ מוגדרת ע"י:

$$E[g(x,y)] = \sum_{x,y} g(x,y) \cdot p(x,y)$$

משתנים בלתי-תלויים – X ו-Y נקראים בלתי-תלויים (ב"ת) אם לכל זוג (x, y) כך ש-$x \in X$ ו-$y \in Y$, מתקיים $p(x, y) = p_X(x) \cdot p_Y(y)$.

התפלגות מותנית של מ"מ – התפלגות מותנית $p(x|y)$ מגדירה התפלגות נפרדת על ערכי X בהינתן ערך מסוים של Y: $p(x|y) = P(X = x|Y = y) = \frac{P(X=x, Y=y)}{P(Y=y)}$.

מתכונות ההתפלגות, מתקיים לכל $y \in Y$: $\sum_{x \in X} p(x|y) = 1$.
התוחלת המותנית מוגדרת ע"י: $E[X|Y = y] = \sum_{x \in X} x \cdot p(x|y)$.
אם X, Y בלתי תלויים, אז מתקיים: $p(x|y) = p(x)$.

מודלים הסתברותיים חישוביים

מטרה – מודל הסתברותי שניתן יהיה לחשב הסתברות לפיו, שיהיו שימושיות לחיזוי, פענוח או הסבר. המודל מגדיר הסתברויות שתופעות מסוימות יקרו, כאשר המודל מתייחס לתופעות שנצפו, וגם יכול להתייחס לתופעות שאינן נצפות (חבויות) אבל מניחים שהן קיימות. נגדיר מרחב מדגם ומשתנים, שנניח שתופעות השונות (נצפות וחבויות) נוצרות לפיהן. נניח שהמודל קובע את ההתנהגות הגנרטיבית שלפיה מיוצרות התצפיות. נניח שהמודל מקרב בצורה סבירה את היווצרות התצפיות במציאות.

דוגמאות

- מודל פשוט שמייצר טקסט עברי, יכול להיות סדרת הגרלות אקראיות ממולטינום עם 23 ערכים (א'-ב' + רווח).
- מודל שיחזה את שכיחות המילה "אהבה" בטקסט חדשותי בעברית. ניתן למדל לפי מודל ברנולי כאשר p היא ההסתברות שהמילה תופיע במקום אקראי בטקסט. בהינתן טקסט חדשותי, על מנת לחזות את ההסתברות שהמילה "אהבה" תופיע k פעמים בקטע באורך n מילים, משתמשים בנוסחת הבינום. הרעיון הוא ללמוד את ההסתברות p מטקסטים מסוימים ולבחון אותם על טקסטים חדשים. אם ההסתברות על הטקסטים החדשים קרובה להסתברות שנלמדה במודל, המודל לא מספיק טוב, ויש לשפר אותו (אולי להוסיף משתנים).

סימון מקובל לאוסף הפרמטרים במודל (ההסתברויות של המשתנים המקריים במודל) הוא θ.

<u>שערוך המודל - אומדן ההסתברויות</u>
כדי לאפשר שימוש במודל שהגדרנו, נצטרך ראשית לאמוד את ההסתברויות שלו.

לדוגמה, שערוך ההסתברות הפרמטר p של הטלת מטבע (ההסתברות לעץ). נטיל את המטבע $n = 100$ פעמים, ונניח שקיבלנו 70 פעמים עץ. ההסתברות משוערכת ל-$p = \frac{70}{100} = 0.7$.

<u>אומדן נראות מקסימלי (Maximum Likelihood Estimate - MLE)</u> - האומדן הבסיסי ביותר.
<u>נראות (Likelihood)</u> – היא ההסתברות לקבלת/יצירת תצפית נתונה בהינתן הפרמטרים של המודל.
הנראות מוגדרת ע"י הפונקציה $L: \theta \to [0,1]$, כלומר פונקציה של הפרמטרים של המודל, כאשר התצפית נתונה. לדוגמה, הנראות של $k = 70, n = 100$ בהינתן $p = 0.5$:

$$L(< 70,100 >) = \binom{100}{70} 0.5^{70} 0.5^{100-70}$$

נבחר כאומדן ML את ערך הפרמטר הממקסם את ההסתברות לקבלת התצפית שראינו בפועל:

$$\hat{p}_{<70,100>} = argmax_p L(< 70,100 >; p)$$

באופן כללי, נסמן ב-O את התצפית (או אוסף תצפיות), ונרצה לשערך את סט הפרמטרים θ:

$$\hat{\theta}_{ML}(O) = argmax_\theta L(O; \theta)$$

נוכיח ש-MLE של ניסוי בינומי הוא $\frac{n}{k}$. עבור n, k נתונים, מתקיים:

$$L(< n, k >; p) = \binom{n}{k} p^k (1-p)^{n-k}$$

כאשר n, k קבועים. נגזור לפי p על מנת למצוא את p המקסימלי. יהיה נוח יותר לגזור את $\log L$ (כיוון שלוג היא פונקציה מונוטונית, המקסימום של L ושל $\log L$ שווה). לפעמים מסמנים ב-L את ה-log-likelihood.

$$L(p) = \log \binom{n}{k} p^k (1-p)^{n-k} = \log \binom{n}{k} + k \cdot \log p + (n-k) \cdot \log(1-p) \Longrightarrow$$

(תזכורת: $((\log x)' = \frac{1}{x}$): $L'(p) = 0 + \frac{k}{p} - \frac{n-k}{1-p} = \frac{k(1-p)-(n-k)p}{p(1-p)} = \frac{k-np}{p(1-p)}$.

נשווה ל-0 לקבלת מקסימום: $k - np = 0 \Longrightarrow p = \frac{k}{n}$.
נותר להראות שהפונקציה קעורה (הנגזרת השנייה שלילית בכל התחום) ולכן הנגזרת הראשונה מתאפסת במקסימום. ■

עבור תצפיות שנוצרות ע"י פרמטרים אמיתיים מסוימים שאותם מנסים לשערך, אומדן ML מתכנס לערך האמיתי של הפרמטר עם הגדלת המדגם. משערך המקיים תכונה זו נקרא עקבי אסימפטוטית.

הרצאה 4 (16.11.2014)

מודל מלטינומי
מ"מ X שמקבל אחד מ-m ערכים שונים: $x_1, \dots, x_m$, כל ערך בהסתברות p_i כך ש-$\sum_{i=1}^{m} p_i = 1$.
נוסחת המלטינום – ההסתברות שב-n דגימות, נקבל k_i פעמים את הערך x_i לכל $i = 1, \dots, m$, כאשר $\sum_{i=1}^{m} k_i = n$.

$$p(k_1, \dots, k_m, n) = \frac{n!}{k_1! \cdot \dots \cdot k_m!} \cdot p_1^{k_1} \cdot \dots \cdot p_m^{k_m}$$

הגורם הראשון הוא מס' האפשרויות לווקטור דגימות שבו k_1 דגימות של x_1 וכו'. הגורם השני הוא ההסתברויות לכל הערכים.

נשים לב שכאשר מתעניינים בהסתברות לערך מסוים x_i אז ניתן להסתכל על המלטינום כעל משתנה בינומי, כאשר x_i הוא הצלחה וכל ערך אחר הוא כישלון.

תוחלת של משתנה מלטינומי – תוחלת מס' הפעמים לקבלת ערך x_i ב-n דגימות היא $n \cdot p_i$ (כמו במשתנה בינומי).

אומדן MLE ל-p_i - $\hat{p}_{i_{ML}} = \frac{k_i}{n}$, כלומר מס' הדגימות של x_i לחלק לכל הדגימות.

יישום של מודל שפה – Language Model (LM)
המטרה: מידול רצף של סימנים בשפה כלשהי, למשל טקסט, דיבור או לוג של מכונת בקרה. סימן יכול להיות ברמה כלשהי: אות, מילה וכו'. מטרת מודל השפה לשערך הסתברות לכל רצף אפשרי בשפה. נקודת מבט מקובלת היא להסתכל על מודל שפה כמודל גנרטיבי שמייצר סדרות בשפה, כך שנאמוד את ההסתברות של סדרת סימנים בשפה כהסתברות שהמודל שלנו ייצר אותה.

שימוש לדוגמה: העדפת משפטים סבירים יותר בשפה בזיהוי דיבור.

נסמן: $S = w_1 \dots w_n$ רצף בשפה.

$$p(s) = p(w_1 \dots w_n) = p(length(s) = n) \cdot p(w_1 \dots w_n | length = n)$$

ההסתברות למשפט מסוים היא ההסתברות לאורך שלו כפול ההסתברות למילים בהינתן האורך שלו. כאשר האורך n נתון (קבוע), מתקיים לפי כלל השרשרת:

$$p(w_1 \dots w_n | length = n) = p(w_1) \cdot p(w_2|w_1) \cdot \dots \cdot p(w_n|w_1 \dots w_{n-1})$$

הגישה הכללית: כדי לפשט את המודל, נרצה להניח הנחות אי-תלות, כלומר למדל רק חלק מהתלויות במודל כדי לקבל מס' פרמטרים סביר ושניתן לאמוד אותם בצורה סבירה מסטטיסטיקות נגישות.

המודל הפשוט יותר: מודל יוניגרם (unigram) – ממדלים הסתברות כל מילה בנפרד ללא תלות במילים האחרות:

$$= p(w_1) \cdot \dots \cdot p(w_n) = \prod_{i=1}^{n} p(w_i)$$

המודל לא מתייחס לסדר המילים, אבל יש חזרות (multi-set). דרך נוספת לייצג את ההסתברות הזו היא כמכפלה של כל המילים בשפה (הערכים האפשריים במלטינום), כאשר לכל מילה מכפילים את ההסתברות שלה בשכיחות שלה במשפט: $\prod_{w \in W} p(w)^{f(w,s)}$.

מודל יוניגרם הוא מודל מלטינומי:
W – קבוצת הערכים.
$w \in W$ ערך אפשרי עם פרמטר $p(w)$ - הסתברות הופעת w בדגימה אקראית של מילה במשפט.
מודל זה מניח אי-תלות בין הופעת המילים ב-s. המודל ממדל את השכיחות היחסית של המילים.
אומדן MLE – בהינתן מדגם גדול של משפטים באורך כולל של n מילים, $\hat{p}_{ML}(w) = \frac{f(w)}{n}$.

מודל אלטרנטיבי למשפט/מסמך עם הנחת אי-תלות בין המילים - Multiple Bernoulli
במודל זה נמדל את S כקבוצת כל המילים שהופיעו במשפט, ללא חשיבות לשכיחות ההופעה (התעלמות מחזרות). כל משפט הוא תת-קבוצה של המילון: $S \subseteq W$. ניתן לייצג משפט ע"י $|W|$ משתני ברנולי בינאריים המקבלים 0 עבור מילים שלא הופיעו במשפט ו-1 עבור מילים שהופיעו במשפט (ווקטור בינארי באורך $|W|$). לכל משתנה ברנולי יש פרמטר $p(w)$ שמציין את ההסתברות ש-w תופיע במשפט S אקראי (ללא קשר למס' הפעמים שהופיעה).

התהליך הגנרטיבי ליצירת משפט: מגרילים לכל מילה w בהסתברות $p(w)$ האם היא תופיע במשפט או לא. ההסתברות למשפט בשפה היא מכפלת ההסתברויות של כל המילים במילון לפי הופעה או אי-הופעה של המילה במשפט:

$$\prod_{w \in S} p(w) \cdot \prod_{w \notin S} \big(1 - p(w)\big)$$

אומדן MLE – מס' המשפטים במדגם בהם w הופיעה לחלק למס' המשפטים הכולל במדגם.

שימוש במודל שפה
דוגמה – זיהוי דיבור: נכנס סיגנל דיבור אקוסטי (A) לקופסה (Noisy Channel) ויוצאת סדרת מילים (S). המטרה היא למקסם את ההסתברות $p(S|A)$. מניחים שמידול התופעה מערב את $P(A|S)$ – ההתאמה בין אקוסטיקה למילים, ואת $p(S)$ – ההסתברות לסדרת המילים (מודל השפה).

$$argmax_S P(S|A) = argmax_S \frac{P(A|S)P(S)}{P(A)} = argmax_S P(A|S)P(S)$$

הנוסחה הזו משמשת בתרגום מכונה סטטיסטי, כאשר A הוא המשפט בשפת המקור ו-S הוא המשפט בשפת היעד. מודל השפה הוא בשפת היעד ומודל התרגום נותן את ההסתברות המשותפת של המשפטים.

הרצאה 5 (23.11.2014)

מודל שפה n-gram

על מנת לשערך את ההסתברות לסדרת מילים (משפט), לפי כלל השרשרת:
$$p(w_1 \dots w_n) = p(w_1) \cdot p(w_2|w_1) \cdot \dots \cdot p(w_n|w_1 \dots w_{n-1})$$

במודל יוניגרם, הנחנו שההסתברות למילה מסוימת בלתי-תלויה במילים שלפניה, כלומר:
$p(w_i|w_1 \dots w_{i-1}) = p(w_i)$. באותו האופן, ניתן להגדיר מודל ביגרם (bigram):
$p(w_i|w_1 \dots w_{i-1}) = p(w_i|w_{i-1})$, ובאופן כללי במודל N-gram המילה ה-Nית מותנית ב-N-1 המילים שלפניה: $p(w_i|w_1 \dots w_{i-1}) = p(w_i|w_{i-N+1} \dots w_{i-1})$.

זה סוג של מודלים שנקראים מודלים מרקוביים, ובהם יש תלות בהיסטוריה באורך חסום.

הרעיון במודל הזה, הוא שההסתברות של מילה מסוימת תלויה במילים שלפניה. לדוגמה, יכול להיות ש-$p(\text{הלך}) = 10^{-4}$, $p(\text{ילד}|\text{הלך}) = 10^{-3}$ ו-$p(\text{אווירון}|\text{הלך}) = 10^{-20}$.

מספר הפרמטרים במודל:
אם $|X|$ הוא גודל המילון (מספר המילים = מס' הערכים האפשריים למשתנה X).
במודל יוניגרם, מס' הפרמטרים הוא $|X|$, ובמודל n-gram, מס' הפרמטרים הוא $|X|^n$ (הסתברות לכל n-יה של מילים). ככל ש-n גדול יותר, קשה יותר לשערך את ההסתברות. זאת הסיבה שבדר"כ עובדים עם n=4,5 ולא יותר מזה.

נחזיק מונים ואומדני הסתברות מתאימים רק ל-n-יות מילים שנצפו בקורפוס. דרישת הזיכרון קטנה מגודל המדגם. בנוסף, מקובל להתעלם ממאורעות שנצפו רק פעם אחת במדגם.

אומדן MLE: $p(w_i|w_{i-N+1} \dots w_{i-1}) = \frac{c(w_{i-N+1} \dots w_{i-1} w_i)}{c(w_{i-N+1} \dots w_{i-1})}$, כלומר השכיחות היחסית של מס' ההופעות של ה-nיה לחלק למס' ההופעות של n-1 המילים שלפני המילה ה-i.

בעיית דלילות המידע
עבור מאורע (n-gram) ששכיחותו 0 במדגם, נקבל $p = 0$ וזה יגרום לשערוך הסתברות 0 לכל סדרה שתכיל את המאורע הזה. הפתרון הוא להשתמש בהחלקה.

החלקה (Smoothing)
אומדן למלטינום: $p_{MLE}(x) = \frac{c(x)}{N}$ כאשר $c(x)$ הוא מס' ההופעות של המאורע במדגם ו-N הוא גודל המדגם. כדי למנוע מתן הסתברות 0, נרצה לתת הסתברות קטנה חיובית למאורעות שלא נצפו, ובהתאם (כדי לשמור על סכום הסתברויות 1), נפחית מסת הסתברות כוללת שווה מהמאורעות שנצפו. השיטה נקראת גם discounting.

החלקת Lidstone – הוספת קבוע λ למונה של כל מאורע ונרמול אומדן ההסתברות בהתאם. אם S הוא מדגם האימון ו-$|X|$ הוא מס' המאורעות, אז: $p_{Lid}(x) = \frac{c(x)+\lambda}{|S|+\lambda|X|}$.
עבור $\lambda = 1$ זה נקרא החלקת Laplace. באופן כללי, λ הוא פרמטר של שיטת ההחלקה וצריך לכייל אותו אמפירית. אם משתמשים ב-λ קטנה, נותנים חשיבות גדולה יותר לתצפיות. ההצדקה המתמטית לאומדן לידסטון היא אומדן בייסיאני שמניח הסתברות א-פריורית שווה לכל המאורעות. בהתאם, מתקיים שניתן לרשום את אומדן לידסטון באופן שקול כאינטרפולציה ליניארית של אומדן MLE (מהמדגם) עם התפלגות אחידה:
$$p_{Lid}(x) = \mu \cdot \frac{c(x)}{|S|} + (1-\mu) \cdot \frac{1}{|X|}$$

כאשר $\frac{1}{|X|}$ היא ההסתברות למאורע בהתפלגות אחידה, ו-$\frac{c(x)}{|S|}$ היא ההתפלגות לפי MLE.
עבור $\mu = \frac{|S|}{|S|+\lambda|X|}$, שזו הפרופורציה בין המדגם האמיתי לבין המדגם המנופח (כולל תצפיות הדמה), מתקבלת אותה הנוסחה.
פרקטית, זו שיטה בסיסית ופשוטה, אם כי אינה מוצדקת תיאורטית (מתמטית) כאשר ההתפלגות הא-פריורית אינה אחידה (למשל מילים בשפה). משתמשים בה הרבה כאשר המודל לא רגיש מדי לדיוק ההחלקה. המודל עשוי לתת תוצאות גרועות כשיש רגישות להחלקה.

<u>התנהגות לא רצויה של לידסטון</u>
במקום להפחית את האומדן לכל המאורעות שנצפו, ביחס לאומדן MLE, p_{Lid} מגדיל את האומדן למאורעות מסוימים. זה יקרה מכיוון ש-p_{Lid} הוא אינטרפולציה של p_{MLE} ו-p_{Uni}. אז יתקיים ש-$p_{Lid} > p_{MLE}$ כאשר $p_{Uni} > p_{MLE}$. בהתפלגות אחידה, השכיחות המצופה לכל מאורע היא $\frac{|S|}{|X|} = |S| \cdot \frac{1}{|X|}$. בהתאם, למאורע ששכיחותו $\frac{|S|}{|X|}$ נקבל $p_{Uni} = \frac{1}{|X|} = \frac{\frac{|S|}{|X|}}{|S|} = p_{MLE}$.
כאשר $c(x) < \frac{|S|}{|X|}$ נקבל $p_{MLE} < p_{Uni}$. כלומר, לכל מאורע שהשכיחות שלו קטנה מהשכיחות הממוצעת במדגם, יתקיים $p_{MLE} < p_{Uni}$ ולכן $p_{Lid} > p_{MLE}$.

<u>נושאים לטיפול</u>

1. כיצד לכייל את λ?

- <u>אמפירית</u> – כאשר המודל משמש לאפליקציה ספציפית שעבורה יש נתונים לכיול (Development Set), ננסה טווח ערכים סביר של הפרמטר ונבחר את הערך שהביא לתוצאות הטובות ביותר באפליקציה.

- <u>גישה כללית שאינה תלוית יישום</u> – נשתמש במודלים אלטרנטיביים כגון אלו שנוצרו מערכי λ שונים כדי לשערך הסתברות של מדגם Development חדש ונבחר את המודל שנותן לו את ההסתברות הגבוהה ביותר.

הרצאה 6 (07.12.2014)

שיטה להערכת איכות מודל שפה בעזרת סט נתוני בדיקה
הרעיון הכללי: בהינתן שני מודלי שפה שונים (או מודל שפה מאותו הסוג עם פרמטרים שונים), נרצה לבדוק איזה מהם טוב יותר. בודקים אותם על סט נתוני בדיקה נפרד ונותנים לכל אחד מהם ציון.
נחשב את ההסתברות $p_m(T)$ שהמודל נותן לסדרת בדיקה חדשה - T (שונה מאשר האימון) ונעדיף מודל שעבורו $p_m(T)$ הגבוה ביותר.

לדוגמה, במודל יוניגרם:

$$T = w_1^n = w_1 \dots w_n \Longrightarrow p(T) = p(w_1^n) = \prod_{i=1}^{n} p(w_i)$$

כדי שגודל סט הבדיקה לא ישפיע על התוצאה, מנרמלים את ההסתברות, ומקבלים הסתברות ממוצעת (ממוצע גיאומטרי) שהמודל חוזה למילה אקראית אחת בסדרה: $\sqrt[n]{p(w_1^n)}$.

בדר"כ נעדיף לעבוד ב-log-space, אחרת במכפלה של הרבה הסתברויות מתקבל Underflow:

$$\log\left(p(w_1^n)\right) = \frac{1}{n}\sum_{i=1}^{n} \log\left(p(w_i)\right)$$

מכיוון שלוג היא פונקציה מונוטונית, עדיין ניתן להשוות בין הציונים של המודלים השונים.

Perplexity - מקובל להסתכל על הגודל - $\frac{1}{\sqrt[n]{p(w_1^n)}}$. נשים לב שמס' זה שקול לגודל המילון בהתפלגות אחידה שעבורה חיזוי ההסתברות הממוצעת יהיה זהה לזו של המודל p_m. המס' הזה נקרא Perplexity, והוא מסמל את אי-הוודאות של המודל (ככל שהוא נמוך יותר, כך המודל טוב יותר).
חישוב ה-Perplexity: $2^{\log\left(\frac{1}{\sqrt[n]{p(w_1^n)}}\right)} = 2^{-\frac{1}{n}\sum_{i=1}^{n}\log(p(w_i))}$.2

שיטת החלקה Held-out
זו שיטה אמפירית (שלומדת את ההתפלגות מהנתונים), ולא שיטה אנליטית (שמניחה מראש התפלגות). ההחלטות שצריך לקבוע בשיטת החלקה של discounting (שמפחיתה מהמאורעות שנצפו ומוסיפה לכאלה שלא נצפו):

1. כמה מסת הסתברות להקצות למאורעות שלא נצפו?
   למשל, בלידסטון: $N_0 \cdot \frac{\lambda}{|S|+\lambda|X|}$ כאשר N_0 הוא מס' המאורעות שלא נצפו במדגם האימון.
2. כמה מסת הסתברות להפחית מהאומדן של המאורעות בכל אחת מהשכיחויות שנצפו $c(x)$? (תחת נקודת מבט של מאורעות "אטומיים" ללא מבנה פנימי, ולכן תחת ההנחה שלכל המאורעות עם אותה השכיחות נרצה לבצע את אותה הפחתת מסה).

נשתמש בשני מדגמים: מדגם אימון – T ומדגם חדש – H. לכל קבוצת מאורעות בשכיחות מסוימת r ב-T, נבדוק מה השכיחות הכוללת שלה ב-H, ובהתאם נקצה את מסת ההסתברות.

סימון:
r – שכיחות ב-T

N_r – מס' המאורעות שהופיעו r פעמים ב-T
$c_T(\cdot), c_H(\cdot)$ - שכיחות מאורע ב-T, H בהתאמה

נחשב את t_r (total) – השכיחות הכוללת ב-H של כל המאורעות שהופיעו r פעמים ב-T:

$$t_r = \sum_{x:c_{T(x)}=r} c_H(x)$$

השכיחות האמפירית הממוצעת ב-H של המאורעות שהופיעו r פעמים ב-T היא $f_{emp} = \frac{t_r}{N_r}$.

ולבסוף, לקבלת אומדן להסתברות ל-x שמופיע r פעמים ב-T, נחלק בגודל H:

$$p_{Ho}(x: c_T(x) = r) = \frac{t_r/N_r}{|H|} = \frac{\sum_{x:c_{T(x)}=r} c_H(x)}{N_r|H|}$$

מבצעים לכל r.

1. כמה מסת הסתברות להקצות למאורעות שלא נצפו?
$\frac{\sum_{x:c_{T(x)}=0} c_H(x)}{N_0|H|}$ - ההסתברות שמילה שאינה מופיעה ב-T מופיעה ב-H. אם המדגמים מספיק גדולים, בטוח תהיה לפחות מילה אחת ב-H שאינה מופיעה ב-T ולכן אף מאורע לא יקבל הסתברות 0.
2. כמה מסת הסתברות להפחית מהאומדן של המאורעות בכל אחת מהשכיחויות שנצפו $c(x)$?

נבדוק שההסתברויות האלה מסתכמות ב-1:

$$\sum_x p_{Ho}(x) = \sum_{r=0}^{\infty} \frac{t_r/N_r}{|H|} \cdot N_r = \sum_{r=0}^{\infty} \frac{t_r}{|H|} = \frac{1}{|H|} \sum_{r=0}^{\infty} \sum_{x:c_{T(x)}=r} c_H(x) = \frac{1}{|H|} \cdot |H| = 1$$

מבחינת נוסחת האומדן, אין הכרח שגדלי T ו-H יהיו שווים.

<u>הערה</u> - איך מגדירים את N_0? אם גודל המילון $|V|$ ידוע, N_0 הוא ההפרש בין גודל המילון לבין מס' המילים השונות שהופיעו בסט האימון. במשימות שבהן לא יודעים מראש את גודל המילון, צריך לשערך את גודל המילון, וזה הופך לפרמטר נוסף של שיטת ההחלקה.

ראינו דוגמה להחלקת Held-out וראינו שככל ש-r גדול יותר, ההבדל בין f_{emp} ל-r קטן.

<u>וריאציה טכנית על שיטות החלקה – ביצוע החלקה רק למונים נמוכים יחסית:</u>
לעיתים נעדיף להשתמש באומדן MLE למאורעות שכיחים (כי יש מעט מהם, כי ההחלקה עלולה לעוות את ההסתברות שלהם וכי MLE מספיק אמין עבורם), ואז:
לכל המאורעות בשכיחות $0 < r \leq R$ נבצע החלקה - $p_d(x)$.
לכל המאורעות בשכיחות $r > R$ נשתמש ב-$p_{MLE}(x)$.
נדאג לסכום 1 ע"י הקצאה מתאימה עבור $r = 0$:

$$p(x: c(x) = 0) = \frac{1}{n_0}\left[1 - \sum_{t=1}^{R} n_r \cdot p_d(r) - \sum_{r=R}^{\infty} n_r \cdot p_{MLE}(r)\right]$$

כך ש-$p(r)$ הוא אומדן ההסתברות למאורע בשכיחות r. ההנחה בשיטה היא ש-p_d מבצע הפחתה ביחס ל-MLE (אחרת הסכום היה יוצא 0).

הרצאה 7 (14.12.2014)

שיטת החלקה Backoff – החלקה לאומדן הסתברות ל-ngrams
המטרה: לאמוד הסתברויות מוחלקות ל-bigrams מהצורה: $p(w'|w)$. באופן כללי, ב-bigram, ישנה התפלגות נפרדת לכל w (מילה מתנה), ומבצעים החלקה לכל התפלגות בנפרד.

אם נבצע החלקה כגון לידסטון, אז לכל זוג מילים w', w'' כך ש-$c(ww') = c(ww'') = 0$ נקבל אותו אומדן להסתברויות המותנות $p(w'|w) = p(w''|w)$. אבל אם יש בקורפוס מידע על $p(w'), p(w'')$, למשל כך ש-$p(w') > p(w'')$, אזי נרצה אומדן שיקיים $p(w'|w) > p(w''|w)$.

לדוגמה, אם לא ראינו את ה-bigrams "ילד שר" ו"ילד צלל", לפי שיטת ההחלקה הרגילה (לדוגמה לידסטון), שניהם יקבלו הסתברות שווה. נרצה ש"ילד שר" יקבל הסתברות גבוהה יותר מאשר "ילד צלל" כי ההסתברות של המילה "שר" באופן כללי גבוהה יותר מההסתברות של המילה "צלל".

הרעיון:

1. למילים שנצפו אחרי w נשערך הסתברות מוחלקת כרגיל לפי הסתברות p_d.
2. את המסה ההסתברותית שהתפנתה עבור מילים שלא הופיעו אחרי w, נחלק ביניהן באופן פרופורציונלי להסתברות היוניגרם שלהן:

$$p_B(w'|w) = \begin{cases} p_d(w'|w) & \text{אם } c(ww') > 0 \\ \alpha(w)p_d(w') & \text{אחרת} \end{cases}$$

$\alpha(w)$ הוא גורם נרמול שדואג שסכום ההסתברויות המותנות ב-w יהיה 1: $\sum_{w'} p(w'|w) = 1$.
פיתוח $\alpha(w)$: נסמן את המסה שמתפנית ממקרה 1 ע"י $\beta(w)$:

$$\beta(w) = 1 - \sum_{w':c(ww')>0} p_d(w'|w)$$

צריך שיתקיים:

$$\beta(w) = \sum_{w':c(ww')=0} \alpha(w)p_d(w')$$

כלומר:

$$\alpha(w) = \frac{1 - \sum_{w':c(ww')>0} p_d(w'|w)}{\sum_{w':c(ww')=0} p_d(w')}$$

מבחינת יעילות חישובית, לא כדאי לסכום את כל ההסתברויות במכנה, כי יש הרבה יותר מילים שלא הופיעו אחרי w מאשר מילים שכן הופיעו אחריה: $|\{w': c(ww') = 0\}| \gg |\{w': c(ww') > 0\}|$.
במקום, ניתן להפחית מ-1 את הסתברויות היוניגרם של המילים שכן הופיעו אחרי w:

$$\alpha(w) = \frac{1 - \sum_{w':c(ww')>0} p_d(w'|w)}{1 - \sum_{w':c(ww')>0} p_d(w')}$$

אם $c(w') = 0$, $p_d(w')$ דואגת להחלקה. אם $c(w) = 0$, מתקיים:

$$\alpha(w) = \frac{1 - \sum_{w':c(ww')>0} p_d(w'|w)}{1 - \sum_{w':c(ww')>0} p_d(w')} = \frac{1-0}{1-0} = 1$$

כלומר לכל w': $p_B(w'|w) = p_d(w')$.

<u>שיטת Backoff ל-ngram כללי:</u>
לדוגמה trigram: $p(w_3|w_1w_2)$. אם $c(w_1w_2w_3)=0$, נעשה backoff ל-$p(w_3|w_2)$, ואם $c(w_2w_3)=0$, נעשה backoff ל-$p(w_3)=0$.

באופן כללי:

$$p_B(w_n|w_1^{n-1})=\begin{cases} p_d(w_n|w_1^{n-1}) & \text{אם } c(w_1^n)>0 \\ \alpha(w_1^{n-1})p_B(w_n|w_2^{n-1}) & \text{אחרת} \end{cases}$$

כלומר, אם אין מידע לגבי ngram, משתמשים במידע לגבי ה-(n-1)gram וכך רקורסיבית.

פיתוח $\alpha(w_1^{n-1})$:

$$\alpha(w_1^{n-1})=\frac{1-\sum_{w_n:c(w_1^n)>0}p_d(w_n|w_1^{n-1})}{1-\sum_{w_n:c(w_1^n)>0}p_B(w_n|w_2^{n-1})}$$

תנאי ה"עצירה" הוא במקרה של יוניגרם, כאשר $p_B(w'|w)=p_d(w'|w)$.
נשים לב ששיטת ה-backoff מתעלמת מהעובדה שאם w,w' שכיחים, אבל $c(ww')=0$, יש בכך מידע על העובדה שהצירוף ww' אינו סביר.

<u>גישות אינטרפולציה (סכימה)</u>
מודל ה-backoff שילב בין מודל ביגרם למודל יוניגרם ע"י נסיגה. גישה אחרת: ניתן לעשות שילוב ממושקל בין המודלים, כגון אינטרפולציה ליניארית.

$$p(w_3|w_1w_2)=\lambda_1p(w_3|w_1w_2)+\lambda_2p(w_3|w_2)+\lambda_3p(w_3)$$

כאשר $\sum_i\lambda_i=1$ (כיוון שמובטח שכל ההסתברויות מסתכמות ל-1).
בשיטת אינטרפולציה צריך שיטה לבחור ערכי λ_i אופטימליים. במקרה הכללי, ניתן לעשות אופטימיזציה לערך λ בנפרד לכל מאורע מתנה. קיימים אלגוריתמי אופטימיזציה, ובפרט גרסה של אלגוריתם EM, שמחפשת סט ערכי λ אופטימליים במטרה למקסם את ה-likelihood של המודל על מדגם held-out.

באופן כללי, גישות אינטרפולציה יכולות לשמש לשילוב של מודלים כלשהם שרלוונטיים לאומדן.
לדוגמה, אומדן יוניגרם לשפה בתחום מקצועי מסוים:

$$p(w)=\lambda p_{domain}(w)+(1-\lambda)p_{general}(w)$$

כאשר יש מודל שאומן על מדגם קטן של טקסטים בתחום ומודל שאומן על מדגם גדול יותר של השפה הכללית.

הרצאה 8 (28.12.2014)

מודלים של משתנים חבויים

סכימה כללית: נתונה תופעה נצפית y שהיא בדר"כ מורכבת: $y =< y^1 y^2 \dots y^m >$, לדוגמה מסמך (סדרת מילים), ייצוג תמונה ע"י מאפיינים וכו'. מניחים ש-y מתפלג לא באופן עצמאי, אלא לפי התפלגות משותפת עם משתנה נוסף X, חבוי, שבו בדר"כ נתעניין. לדוגמה, נושא המסמך, הסצנה בתמונה. נסמן: $p(X, Y; \theta)$, כאשר θ הפרמטרים של ההתפלגות, כאשר X ערך אטומי.

מבט גנרטיבי מקובל: לייצור זוג (x, y) מסוים (מסמך ונושא) שבו צופים ב-y, קודם נבחר $x \in X$ לפי התפלגות $p(x)$ (ההתפלגות השולית של x), ואח"כ נייצר y לפי ההתפלגות $p(y|x)$ (התפלגות המסמכים לפי נושא).

מטרות לשימוש במודל חבוי

1. סיווג – כאשר יודעים או מניחים את קיום ה-Xים ורוצים לשחזר אותם.
   לדוגמה, אנחנו יודעים שמסמכים מתחלקים לנושאים, ואנחנו רוצים לסווג מסמכים לנושאים.
2. כאשר מניחים שקיימים Xים כלשהם והתחשבות בהם תתן מודל מדויק יותר לתצפיות.
   לדוגמה, אם מודל שפה מתחשב בנושא (החבוי) של השיחה, הביצועים שלו יהיו טובים יותר (ואין צורך לשחזר את הנושא במפורש).

נק' מבט של סיווג – שתי סיטואציות:

- סיווג מבוקר (Supervised) – זהות ערכי X האפשריים וזהות ה-x_t של כל תצפית y_t נקבעת חיצונית ע"י מנחה. מטרת המערכת היא לשחזר את ערך ה-x המתאים לכל y חדש. הלמידה בדר"כ מתוך מדגם אימון שבו לכל y_t נתון x_t ולפיו צריך לשערך את הפרמטרים θ.

- סיווג לא מבוקר (Unsupervised) – קיבוץ / Clustering – זהות ערכי X לא ידועה (בדר"כ נניח $|X|$) והמטרה היא לחלק את Y לקבוצות כך שכל קבוצה שייכת ל-x_i מסוים. יש שני וריאנטים: hard – כל y משויך ל-x מסוים, ו-soft שכל y משויך בהתפלגות $p(x_i|y)$ ל-x_i כך ש-$\sum_{i=1}^{|X|} p(x_i|y) = 1$. המטרה למצוא את החלוקה הסבירה ביותר של ה-Yים.

דוגמה:

מידול של מסמכים עם הנושאים שלהם שמיוצגים כעירוב היסטוגרמות (Mixture of histograms)

ניתן להסתכל על כל מסמך כעל היסטוגרמה של כל המילים בשפה (כמה שכיחה כל מילה במסמך).

נוטציה\ייצוג – כמו במודל המולטינומי, בצירוף שיוך לנושאים:

ייצוג מסמך y: היסטוגרמה של מילים – התפלגות מולטינומית מול מילון בגודל V מילים.

סט המילים האפשריות: $w_1, \dots, w_V$.

נתון מדגם של N מסמכים: $y_1, \dots, y_N$.

כל y_t הוא דגימה מהמשתנה הנצפה Y (לכל $t = 1, \dots, n$).

סימון היסטוגרמה של מסמך: n_{t_k} – שכיחות w_k בתוך y_t.

$$y_t =< n_{t_1}, \dots, n_{t_V} >$$

לפשטות, נניח שכל המסמכים באורך שווה m, כלומר $\sum_{k=1}^{V} n_{t_k} = m$.

מודל עם משתנים חבויים:

נניח קיום של משתנה נוסף X עם ערכים $x_i \in X$ $(i = 1, \dots, |X|)$ המייצג קטגוריה\נושא של המסמך.

מניחים התפלגות משותפת $P(X, Y; \theta)$.

<u>מודל גנרטיבי – עירוב היסטוגרמות:</u>
ליצירת זוג (x, y):

- שלב א' – קביעת נושא המסמך: $x_i \in X$ לפי התפלגות $p(x_i) = P(X = x_i)$.
- שלב ב' – קביעת המילים ב-y לפי התפלגות $p(y|X = x_i)$. בפרט נניח שמסמך y נוצר עפ"י התפלגות מולטינומית של המילים תחת הנושא x_i: לפי פרמטרים $p(w_k|x_i)$. בהתאם, ההסתברות ליצירת מסמך לפי נושא היא הסתברות יוניגרם על כל המילים במסמך: $p(y_t|x_i) = \prod_{k=1}^{V} p(w_k|x_i)^{n_{t_k}}$.

לסיכום, הפרמטרים במודל:
התפלגות הנושאים - $\theta = \{p(x_1), \dots, p(x_{|X|})\}$.
התפלגויות המילים לפי נושא –

$$p(w_1|x_1), \dots, p(w_V|x_1)$$
$$p(w_1|x_2), \dots, p(w_V|x_2)$$
$$\dots$$
$$p(w_1|x_{|X|}), \dots, p(w_V|x_{|X|})$$

<u>השאלות שנתעניין במודל חבוי:</u>

1. <u>הסתברות יצירת תצפיות על פי θ: $p(y; \theta)$</u>
לפי ההסתברות השולית, כלומר ההסתברות ליצירת מסמך עבור כל נושא אפשרי:
$$p(y; \theta) = \sum_{i=1}^{|X|} p(x_i, y; \theta) = \sum_{i=1}^{|X|} p(x_i) p(y|x_i; \theta) = \sum_{i=1}^{|X|} p(x_i) \prod_{k=1}^{V} p(w_k|x_i)^{n_{t_k}}$$
2. <u>הסתברות סמויה (קביעת הסתברות x_i עבור y נתון): $p(x_i|y; \theta)$</u>
נשתמש בנוסחת בייס:
$$p(X = x_i|y_t; \theta) = \frac{p(y|x_i; \theta) p(x_i; \theta)}{p(y_t; \theta)} = \frac{p(x_i) \prod_{k=1}^{V} p(w_k|x_i)^{n_{t_k}}}{\sum_{i=1}^{|X|} p(x_i) \prod_{k=1}^{V} p(w_k|x_i)^{n_{t_k}}}$$
3. <u>אומדן מודל נראות מקסימלי: θ_{ML}</u>
במקרה המבוקר (Supervised), נתון מדגם מסמכים: $y_1, \dots, y_N$ ולכל y_t נתון ה-x_t המתאים.

$$p(x_i) = \frac{\lambda + \text{מספר מסמכי } x_i}{\lambda|X| + N}$$

$$p(w_k|x_i) = \frac{\lambda' + \text{מספר מופעי } w_k \text{ במסמכי } x_i}{\lambda' \cdot V + \text{אורך כולל של מסמכי } x_i}$$

כלומר, מחשבים בשכיחות יחסית עם החלקת לידסטון.

לחישוב הסיווג:

1. בסיווג soft, מפעילים את נוסחה 2 לכל $x_i \in X$.
2. בסיווג hard, מפעילים את נוסחה 2 לכל $x_i \in X$ ובוחרים את הקטגוריה הסבירה ביותר (מסווג Naïve Bayes). מבחינה חישובית, אין טעם לחשב מחדש את המכנה הקבוע, ובנוסף, כדאי לעבור ל-log space:

$$x_{i^*} = \arg\max_i p(x_i|y_t;\theta) = \arg\max_i p(x_i)\prod_{k=1}^{V} p(w_k|x_i)^{n_{t_k}} =$$

$$= \arg\max_i \left[logp(x_i) + \sum_{k=1}^{V} n_{t_k} logp(w_k|x_i) \right]$$

<u>הרצאה 9 (04.01.2015)</u>

שיעור השלמה יום ב' 10:00-12:00 ב-26.01.15 השבוע האחרון של הסמסטר.

<u>סיווג מסמכים לנושאים</u>
נשים לב שהנוסחאות אינן בהכרח תלויות במודל הספציפי שלמדנו וניתן להשתמש בהן גם עבור מודל אחר. לדוגמה, אם נבחר מודל Multiple Bernoulli ליצירת מסמך $p(y|x_i)$:

מסמך מיוצג ע"י סט המילים המרכיבות אותו: $y =< y^1, \dots, y^{|V|} >$ - ווקטור בינארי באורך המילון, כלומר $y^i \in \{0,1\}$ מציין אם w_i במסמך y. כדי לייצר מסמך, עוברים על כל המילים במילון ומגרילים האם היא נמצאת במסמך או לא. לכל מילה יש הסתברות להופיע בכל נושא.

$$p(y|x_i) = \prod_{w\in y} p(w|x_i;\theta) \prod_{w\notin y} \big(1 - p(w|x_i;\theta)\big)$$

<u>אומדן הפרמטרים במקרה ה-Supervised:</u>
התפלגות הנושאים - $\theta = \{p(x_1), \dots, p(x_{|X|})\}$. האומדן הוא: $p(x_i) = \frac{\text{מספר מסמכי } \lambda + x_i}{\lambda|X| + N}$.
התפלגויות המילים לפי נושא –

$$p(w_1|x_1), \dots, p(w_V|x_1)$$
$$p(w_1|x_2), \dots, p(w_V|x_2)$$

...

$$p(w_1|x_{|X|}), \dots, p(w_V|x_{|X|})$$

האומדן הוא: $p(w|x_i) = \frac{\lambda' + w \text{ מספר מסמכי } x_i \text{ שמופיעה בהם}}{\lambda' \cdot 2 + x_i \text{ מספר מסמכי}}$ – ההסתברות שהמילה w תופיע במסמך בנושא x_i. הסיבה שמכפילים את λ' ב-2 היא שיש ל-w, משתנה ברנולי, 2 ערכים (המילה מופיעה או לא).

<u>סיווג לא מבוקר (Unsupervised)</u>
נתון מדגם $Y = y_1, \dots, y_N$ מיוצגים ע"י עירוב היסטוגרמות (לצורך הדגמה). זהות ערכי X לא ידועה, אבל מחליטים על |X|. לגישה הכללית יש שתי מטרות אפשריות:

1. <u>Clustering (קיבוץ)</u> – מציאת |X| התפלגויות שונות שמייצרות את המדגם ולפיהן נוכל לשייך כל $y_t \in Y$ ל-x_i המתאים. יש שני וריאנטים: hard – כל y משויך ל-x מסוים, ו-soft שכל y משויך בהתפלגות $p(x_i|y)$ ל-x_i כך ש-$\sum_{i=1}^{|X|} p(x_i|y) = 1$. המטרה למצוא את החלוקה הסבירה ביותר של ה-Yים.

2. <u>מידול יותר מדויק של הנתונים הנצפים במדגם</u> – למשל, מודל שפה מדויק יותר (לפי נושא).

<u>הגישה</u>: ננסה למצוא מודל אחר θ_{ML} שעבורו מתקבלת הסתברות מקסימלית ליצירת המדגם:

$$\theta_{ML} = argmax_\theta p(y;\theta) = argmax_\theta \prod_{t=1}^{n} p(y_t;\theta)$$

ננסה להבין למה החלוקה הנכונה של מסמכים לנושאים מייצרת likelihood יותר גבוה מאשר אם היינו מייצרים התפלגות מלטינומית אחת לכל הקורפוס.

<u>דוגמה</u>: שימוש במשתנים חבויים מאפשר ליצור מודל שייתן נראות גבוהה יותר למדגם.
נניח שיש לנו 10 מילים שונות $|V| = 10$: $w_1, \dots, w_{10}$, 10 נושאים $|X| = 10$: $x_1, \dots, x_{10}$ ו-10 מסמכים $|Y| = 10$: $y_1, \dots, y_{10}$. בכל y_i מופיעה רק 10 פעמים w_i.

במודל יוניגרם מולטינומי, לכל i: $p(w_i) = \frac{10}{100} = 0.1$ (לכל מילה מס' ההופעות שלה הוא 10 ומס' המילים בקורפוס הוא 100). ההסתברות לכל מסמך y_i היא $p(w_i)^{10} = \left(\frac{1}{10}\right)^{10}$.

במודל חבוי, θ_{ML} ייתן הסתברות אחידה לכל הנושאים: $p(x_i) = \frac{1}{10}$. כל מסמך יהיה בנושא אחר. הסתברות המילים לפי נושא: $p(w_i|x_i) = 1$ ו-$p(w_j|x_i) = 0$ לכל $j \neq i$.
ההסתברות ליצירת מסמך y_i: $p(y_i|x_i) = 1^{10} = 1$, $p(y_i|x_j) = 0$.

$$p(y_i) = \sum_{j=1}^{10} p(x_j)\, p(y_i|x_j) = \frac{1}{10} \cdot 1 + \frac{9}{10} \cdot 0 = \frac{1}{10}$$

במקרה הכללי: $|X| \ll N$. מנסים ללמוד |X| התפלגויות שממדלות |X| נושאים/התפלגויות במדגם המסמכים. אין פתרון אנליטי למציאת θ_{ML} (לא ניתן לגזור ולמצוא מקסימום בגלל התלות בפרמטרים). קיים אלגוריתם קירוב איטרטיבי EM שמתחיל מערכי $\hat{\theta}$ התחלתיים ומשנה אותם כך שמובטח שבכל איטרציה $p(y;\hat{\theta})$ עולה. כתוצאה מתקבל אלגוריתם שמוצא מקסימום לוקאלי לפונקציות הנראות, אך לא בהכרח מקסימום גלובלי.

<u>נוסחת EM למודל עירוב היסטוגרמות:</u>
<u>אתחול:</u> נאתחל את הפרמטרים לערכים כלשהם (רנדומי, ניחוש מושכל).
בהינתן הפרמטרים, נחשב את הנראות – likelihood: $p(y;\theta)$.

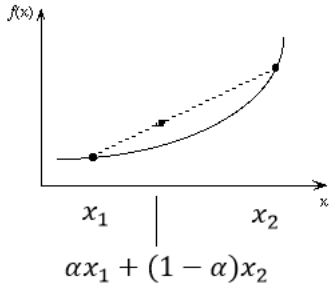
<u>בכל איטרציה:</u>

1. <u>שלב ה-E (Expectation)</u> – לכל מסמך y_t נחשב את הסתברות הסיווג לכל נושא $p(X_t = x_i|y_t)$ (כאשר X_t הוא ערך המשתנה החבוי - הנושא של y_t), לפי הנוסחה:
$$p(X = x_i|y_t;\theta) = \frac{p(y|x_i;\theta)p(x_i;\theta)}{p(y_t;\theta)} \triangleq w_{ti}$$
2. <u>שלב ה-M (Maximization)</u> – מחשבים ערכים חדשים לפרמטרים:
$$p(x_i) = \frac{\sum_{t=1}^{N} w_{ti}}{\sum_{j=1}^{|X|}\sum_{t=1}^{N} w_{tj}} = \frac{\sum_{t=1}^{N} w_{ti}}{\sum_{t=1}^{N}\sum_{j=1}^{|X|} w_{tj}} = \frac{\sum_{t=1}^{N} w_{ti}}{\sum_{t=1}^{N} 1} = \frac{1}{N}\sum_{t=1}^{N} w_{ti}$$
תוחלת מניית x_i במדגם (בשברי מופעים): במקום לתת הסתברות נצפית 1 לכל נושא שמסמך כלשהו שייך אליו, נותנים לו הסתברות להיות שייך אליו. ההסתברות למסמך מסוים היא סכום ההסתברויות של הנושא, מנורמלת בהסתברויות של כל הנושאים.
$$p(w_k|x_i) = \frac{\sum_{t=1}^{N} n_{t_k} \cdot w_{ti}}{\sum_{l=1}^{|V|}\sum_{t=1}^{N} n_{t_l} \cdot w_{ti}} = \frac{\sum_{t=1}^{N} n_{t_k} \cdot w_{ti}}{\sum_{t=1}^{N} n_t \cdot w_{ti}}$$
תוחלת מניית ההופעות של w_k בנושא x_i.

בסוף כל איטרציה מחשבים $p(y;\theta)$ עבור הפרמטרים החדשים ועוצרים כשמזהים התכנסות.
הערה: בדיבאג לוודא ש-$p(y;\theta)$ עולה בכל איטרציה.

הרצאה 10 (11.01.2015)

<u>אי-שוויון Jensen</u>
<u>הגדרה</u> – פונקציה $f(x)$ היא קמורה (Convex) אם המיתר נמצא מעל לפונקציה:
$$f(\alpha x_1 + (1-\alpha)x_2) \le \alpha f(x_1) + (1-\alpha)f(x_2)$$



ההגדרה עבור פונקציה קעורה הפוכה: $f(\alpha x_1 + (1-\alpha)x_2) \ge \alpha f(x_1) + (1-\alpha)f(x_2)$.
ניתן להרחיב את ההגדרה ל-n משתנים: לכל $x_1, \dots, x_n$ ולכל $\alpha_1, \dots, \alpha_n \ge 0$ כך ש-$\sum_{i=1}^{n} \alpha_i = 1$ מתקיים: $f(\sum_{i=1}^{n} \alpha_i\, x_i) \le \sum_{i=1}^{n} \alpha_i f(x_i)$.

אינטרפרטציה הסתברותית: $\alpha_i = p(x_i)$:

$$f\left(\sum_{i=1}^{n} p(x_i)\, x_i\right) \le \sum_{i=1}^{n} p(x_i) f(x_i) \Longrightarrow f(E[x]) \le E[f(x)]$$

כלומר, לכל פונקציה קמורה, הפעלת הפונקציה על התוחלת של x קטנה מהתוחלת של הפעלת הפונקציה על x.

<u>אנטרופיה יחסית (Relative Entropy) – (KL) Kullback-Leibler</u>
<u>הגדרה</u> - נתונות שתי התפלגויות דיסקרטיות מעל n ערכים: $x_1, \dots, x_n$. נסמן p_i ו-q_i. נגדיר:

$$D_{KL}(p||q) \triangleq \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right)$$

זהו המדד המקובל בתורת האינפורמציה להשוואת ה-Relative Entropy. המדד סופר כמה מידע מפסידים כאשר משתמשים בהתפלגות q כאומדן להתפלגות p. אם ההתפלגויות זהות (p=q), $D_{KL}(p||p) = \sum_{i=1}^{n} p_i \log(1) = 0$. ככל שההתפלגויות שונות, נקבל מס' קרוב יותר ל-1. המדד אינו סימטרי. אם קיים $q_i = 0$ אזי המרחק אינו מוגדר.

<u>טענה</u> – לכל p, q, $D_{KL}(p||q) \ge 0$.
<u>הוכחה</u> – נסתכל על $-D_{KL}(p||q) = -\sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right) = \sum_{i=1}^{n} p_i \log\left(\frac{q_i}{p_i}\right)$.
נסמן $f(x) = \log\big(g(x)\big)$, $g(p_i) = \frac{q_i}{p_i}$. זוהי פונקציה קמורה ולכן עפ"י אי-שוויון Jensen מתקיים:

$$\sum_{i=1}^{n} p_i \log\left(\frac{q_i}{p_i}\right) \leq \log\left(\sum_{i=1}^{n} p_i \frac{q_i}{p_i}\right) = \log\left(\sum_{i=1}^{n} q_i\right) = 0$$

■

<u>הצורה הכללית של אלגוריתם EM</u>

<u>מאמר מקורי</u>:

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." Journal of the Royal Statistical Society. Series B (Methodological) (1977): 1-38.

אנחנו נלמד גרסה מאוחרת יותר שמבוססת על המאמר:

Neal, Radford M., and Geoffrey E. Hinton. "A view of the EM algorithm that justifies incremental, sparse, and other variants." Learning in graphical models. Springer Netherlands, 1998. 355-368.

<u>המטרה</u>: מציאת θ_{ML} כשיש משתנים חבויים ואין פתרון אנליטי. ע"י אלגוריתם איטרטיבי, נמצא $\hat{\theta}$ שמגדיל נראות בכל איטרציה.

מניחים התפלגות משותפת $p(X,Y)$. לכל תצפית y יש ערך חבוי x. נראות הנתונים הנצפים (likelihood) מוגדרת ע"י $p(y;\theta) = \sum_x p(x,y;\theta)$.

$$\hat{\theta}_{ML} = argmax_\theta \log\big(p(y;\theta)\big) = argmax_\theta \log\left(\sum_x p(x,y;\theta)\right)$$

בדר"כ לא קיים פתרון אנליטי ל-$\hat{\theta}_{ML}$. לכן נפתח אלגוריתם איטרטיבי שיעלה את הנראות בכל איטרציה עד להתכנסות למקסימום.

<u>סכימת Alternate Maximization</u>

נניח שיש פונקציה $g(\psi)$, כאשר ψ קבוצת המשתנים (הפרמטרים) של הפונקציה. נרצה למצוא את $argmax_\psi g(\psi)$. נניח מצב שבו בעיה זו קשה מדי לפתרון, אבל ניתן לחלק את הפרמטרים לשתי קבוצות: θ, ϕ, ונרשום $g(\theta,\phi)$. מתקיים שכאשר מקבעים קבוצת פרמטרים אחת, אזי אפשרי (קל) למצוא מקסימום ל-g עבור קבוצת הפרמטרים השנייה. כלומר, עבור ϕ_0 מקובע ניתן למצוא את $argmx_\theta g(\theta,\phi_0)$ ועבור θ_0 מקובע ניתן למצוא את $argmx_\phi g(\theta_0,\phi)$.

ל-g כזו ניתן להגדיר אלגוריתם איטרטיבי שבכל איטרציה ימצא ערכים חדשים ל-ϕ,θ כך שבכל איטרציה ערך g יעלה.

<u>אתחול</u> - נקבע θ_0

(1) עבור θ_0 מקובע, נמצא $\phi_0 = argmx_\phi g(\theta_0,\phi)$.
(2) עבור ϕ_0 מקובע, נמצא $\theta = argmx_\theta g(\theta,\phi_0)$.
(3) $\theta \to \theta_0$ וחזור לשלב (1).

<u>הכללה של הסכימה</u>
נניח שרוצים למקסם פונקציה מהצורה $f(\theta)$ שלא ניתן להפעיל עליה ישירות את סכימת Alt-Max.
<u>המטרה המקורית</u>: $argmax_\theta f(\theta)$.
<u>הרעיון</u>: נשתמש בפונקציית עזר g שהיא פונקציה של θ ובנוסף קבוצת משתני עזר נוספים ϕ, כך שמתקיים: לכל θ, $f(\theta) = \max_\phi g(\theta, \phi)$. כלומר, יתקיים:

$$argmax_\theta f(\theta) = argmax_\theta \left(\max_\phi g(\theta, \phi) \right)$$

כלומר: עם מציאת ערכי ϕ, θ אופטימליים שממקסם את g בפרט ימקסמו גם את f. לכן, ננסה למצוא g כזו שעבורה ניתן להפעיל Alt-Max ונפעיל סכימה זו.

<u>אתחול</u> - נקבע θ_0

(1) עבור θ_0 מקובע, נמצא $f(\theta_0) = \max_\phi g(\theta_0, \phi)$.
(2) עבור ϕ_0 מקובע, נמצא $\theta = argmax_\theta g(\theta, \phi_0)$.
(3) $\theta \to \theta_0$ וחזור לשלב (1).

לכן בשלב (1) בכל איטרציה נקבל ערך $f(\theta)$ גדול יותר.

<u>הרצאה 11 (18.01.2015)</u>

<u>פיתוח סכימת EM – המשך:</u>
נחפש פונקציה F מהצורה $F(\theta, \phi)$ שחסומה מלמעלה ע"י פונקציית הנראות ובמקסימום שלה מתלכדת עם פונקציית הנראות. נפתח את פונקציית הנראות:

$$\log\big(p(y;\theta)\big) = \log\left(\sum_x p(x,y;\theta)\right) = \log \sum_x q(x) \cdot \left(\frac{p(x,y;\theta)}{q(x)}\right)$$

כלומר, הוספנו פרמטרים ϕ ע"י מכפלה וחילוק ב**התפלגות** $q(x)$ כלשהי.
לפי אי-שוויון Jensen: $f(\sum_{i=1}^n q(x_i)\, x_i) \geq \sum_{i=1}^n q(x_i) f(x_i)$ (אי-השוויון הפוך כי הפונקציה קעורה) ולכן:

$$\log \sum_x q(x) \cdot \left(\frac{p(x,y;\theta)}{q(x)}\right) \geq \sum_x q(x) \cdot log\left(\frac{p(x,y;\theta)}{q(x)}\right) = F(\theta, q)$$

$F(\theta, q)$ נקראת "האנרגיה החופשית" של q ו-θ. קיבלנו: $\log\big(p(y;\theta)\big) \geq F(\theta, q)$ לכל התפלגות $q(x)$. אם שוויון אפשרי, אזי עבור q שממקסם את $F(\theta, q)$ לכל θ, נקבל:
$\log\big(p(y;\theta)\big) = \max_q F(\theta, q)$. נוכיח ששוויון מתקיים. נסתכל על ההפרש:

$$\log\big(p(y;\theta)\big) - F(\theta, q) =$$

בגלל ש-$\sum_x q(x) = 1$, נכפיל בו ונקבל:

$$= \sum_x q(x) \cdot \log\big(p(y;\theta)\big) - \sum_x q(x) \cdot log\left(\frac{p(x,y;\theta)}{q(x)}\right) =$$

$$= \sum_x q(x) \left[\log\big(p(y;\theta)\big) - log\left(\frac{p(x,y;\theta)}{q(x)}\right)\right] =$$

$$= \sum_x q(x) \cdot \log\left(\frac{q(x)}{\frac{p(x,y;\theta)}{p(y;\theta)}}\right) = \sum_x q(x) \cdot \log\left(\frac{q(x)}{p(x|y;\theta)}\right) = D_{KL}(q(x)||p(x|y;\theta))$$

שוויון בין $\log\big(p(y;\theta)\big)$ ל-$F(\theta,q)$ מתקבל כאשר $q(x) = p(x|y;\theta)$. נשים לב שזו פונקציית הסיווג למשתנה החבוי. קיבלנו ייצוג של פונקציית הנראות כמקסימום מעל F.

$$\log\big(p(y;\theta)\big) = \max_q F(\theta,q) = F\big(\theta, p(x|y;\theta)\big)$$

$$\theta_{ML} = argmax_\theta \log\big(p(y;\theta)\big) = argmax_\theta \max_q F(\theta,q)$$

רוצים למצוא זוג של θ, q שממקסמים את F ובפרט ערך θ יהיה זה שממקסם את הנראות (θ_{ML}).
כעת נפעיל את סכימת Alt-Max על F כדי למצוא בכל איטרציה ערכי θ שמגדילים את ערך F ובמקביל את ערך הנראות.

<u>שלבי אלגוריתם EM (סכימת Alt-Max עבור F):</u>

<u>אתחול</u> - נקבע θ_0 כלשהו, למשל רנדומי או ניחוש מושכל

(1) צעד ה-E: נחשב את ערך q הממקסם את F כאשר מקבעים $\theta = \theta_0$:
$$q(\theta_0) = \operatorname{argmax}_q F(\theta,q) = p(x|y;\theta)$$
זו פונקציית הסיווג, מחשבים את ערכיה לכל ערך x עבור y (תצפית) נתון וערכי θ_0.

(2) צעד ה-M: עבור q מקובע שחישבנו בצעד ה-E, נמצא θ שממקסם את $F(\theta,q)$:

$$\theta = \operatorname{argmax}_\theta F(\theta,q) = \operatorname{argmax}_\theta \sum_x q(x) \cdot log\left(\frac{p(x,y;\theta)}{q(x)}\right) =$$
$$= \operatorname{argmax}_\theta \sum_x q(x)\log\big(p(x,y;\theta)\big) - \sum_x q(x)\log\big(q(x)\big)$$

מכיוון ש-$\sum_x q(x)\log\big(q(x)\big)$ קבוע בהינתן q:

$$= \operatorname{argmax}_\theta \sum_x q(x)\log\big(p(x,y;\theta)\big)$$

נציב את $q(x)$:

$$= \operatorname{argmax}_\theta \sum_x p(x|y;\theta)\log\big(p(x,y;\theta)\big) \triangleq \operatorname{argmax}_\theta Q(\theta,\theta_0)$$

$Q(\theta,\theta_0)$ נקראת Auxiliary Function. קיבלנו שבצעד ה-M צריך למצוא θ שממקסם את $Q(\theta,\theta_0)$, כאשר ל-Q צורה שונה בכל מודל בהתאם לנוסחאות ההסתברות באותו המודל. פעמים רבות ניתן למצוא פתרון אנליטי למקסום $Q(\theta,\theta_0)$, אחרת משתמשים בפתרון איטרטיבי. כל הנוסחאות המתקבלות לחישוב θ הממקסם את Q הן נוסחאות עדכון הפרמטרים בשלב ה-M (משוואות ה-M Step).

(3) $\theta \to \theta_0$ וחזור לשלב (1).

<u>נוסחת EM למודל עירוב היסטוגרמות:</u>

נפתח משוואות M-Step לעירוב היסטוגרמות. מקרה נפוץ: y הוא סדרת תצפיות. לדוגמה, במסמכים: $y =< y_1, \dots, y_N >$ סדרת מסמכים בלתי-תלויים, $x = (x_1, \dots, x_N)$ סדרת הנושאים שלהם.

$$p(x, y; \theta) = \prod_{t=1}^{N} p(x_t, y_t; \theta)$$

$$p(x|y; \theta) = \prod_{t=1}^{N} p(x_t|y_t; \theta)$$

אם נציב בנוסחת Q, אפשר להראות (באמצעות פיתוח אלגברי נוסף שלא נכנסנו אליו) שמתקבל:

$$Q(\theta, \theta_0) = \sum_{t=1}^{N} \sum_{i=1}^{|X|} p(x_i|y_t; \theta_0) \log\big(p(x_i|y_t; \theta)\big)$$

<u>אתחול:</u> נאתחל את הפרמטרים לערכים כלשהם (רנדומי, ניחוש מושכל).
בהינתן הפרמטרים, נחשב את הנראות – likelihood: $p(y; \theta)$.

(1) <u>שלב ה-E (Expectation)</u> – לכל מסמך y_t נחשב את הסתברות הסיווג לכל נושא i: $p(x_i|y_t; \theta_0)$, לפי הנוסחה:

$$p(X = x_i|y_t; \theta) = \frac{p(y|x_i; \theta)p(x_i; \theta)}{p(y_t; \theta)} \triangleq w_{ti}$$

(2) <u>שלב ה-M (Maximization)</u> – צריך למצוא את ערכי $p(x_i), p(w_k|x_i)$ שימקסמו את Q.
נציב את הערכים ב-Q ונמצא את הערכים שימקסמו אותה.

$$Q(\theta, \theta_0) = \sum_{t=1}^{N} \sum_{i=1}^{|X|} p(x_i|y_t; \theta_0) \log\big(p(x_i|y_t; \theta)\big)$$

<u>הרצאה 12 (25.01.2015)</u>

<u>נוסחת EM למודל עירוב היסטוגרמות – המשך:</u>

<u>שלב ה-M (Maximization)</u> – צריך למצוא את ערכי $p(x_i), p(w_k|x_i)$ שימקסמו את $Q(\theta, \theta_0)$ עבור נוסחאות המודל שלנו.

$$Q(\theta, \theta_0) = \sum_{t=1}^{N} \sum_{i=1}^{|X|} p(X_t = x_i|y_t; \theta_0) \cdot \log p(X_t = x_i|y_t; \theta)$$

כשלב עזר, נוכיח שאומדן ML למלטינום הוא: $p_i = \frac{n_i}{N}$ כאשר p_i הוא ההסתברות ל-x_i במלטינום עם m ערכים: $X = \{x_1, \dots, x_m\}$, n_i הוא מספר הפעמים ש-x_i נצפה ו-$N = \sum_{i=1}^{m} n_i$ הוא מספר הדגימות. צריך למצוא $\theta_{ML} = \prod_{i=1}^{m} p_i^{n_i}$. נמקסם את ה-log-likelihood במקום:

$$\log \theta_{ML} = \sum_{i=1}^{m} n_i \log p_i$$

ניתן לחלק ב-M (לנוחות), ונמקסם את: $\frac{1}{N}\sum_{i=1}^{m} n_i \log p_i$.

$$\frac{1}{N}\sum_{i=1}^{m} n_i \log p_i = \sum_{i=1}^{m} \frac{n_i}{N} \log p_i =$$

נוסיף ונפחית ביטוי:

$$= \sum_{i=1}^{m} \frac{n_i}{N} \log p_i + \sum_{i=1}^{m} \frac{n_i}{N} \log \frac{n_i}{N} - \sum_{i=1}^{m} \frac{n_i}{N} \log \frac{n_i}{N} = -\sum_{i=1}^{m} \frac{n_i}{N} \log \left(\frac{\frac{n_i}{N}}{p_i} \right) + \sum_{i=1}^{m} \frac{n_i}{N} \log \frac{n_i}{N} =$$

$$= -D_{KL}(\frac{n_i}{N} || p_i) + \sum_{i=1}^{m} \frac{n_i}{N} \log \frac{n_i}{N}$$

המחובר $\sum_{i=1}^{m} \frac{n_i}{N} \log \frac{n_i}{N}$ הוא קבוע ולכן ניתן להתעלם ממנו ולמקסם את $-D_{KL}(\frac{n_i}{N} || p_i)$. מתכונות ה-$D_{KL}$: הוא תמיד חיובי, מתקיים ש-$-D_{KL}(\frac{n_i}{N} || p_i) < 0$ והמקסימום מתקבל ב-0 כאשר $p_i = \frac{n_i}{N}$.

בהוכחה זו ראינו שכאשר צריך למקסם ביטוי מהצורה $\sum_{i=1}^{m} n_i \log p(x_i)$ כאשר $p(x_i)$ התפלגות מעל m ערכים, אזי המקסימום מתקבל כאשר $p(x_i) = \frac{n_i}{\sum_{i=1}^{m} n_i}$.

כעת נמקסם את $Q(\theta, \theta_0)$ למודל עירוב היסטוגרמות:

$$Q(\theta, \theta_0) = \sum_{t=1}^{N} \sum_{i=1}^{|X|} p(X_t = x_i|y_t; \theta_0) \cdot \log p(X_t = x_i|y_t; \theta)$$

נשים לב שלפי הפרמטרים של EM: $w_{ti} = p(X_t = x_i|y_t; \theta_0)$, לכן:

$$= \sum_{t=1}^{N} \sum_{i=1}^{|X|} w_{ti} \cdot \log[p(y_t|X_t = x_i; \theta) \cdot p(X_t = x_i)] =$$

ניזכר בפרמטרים: $p(x_i)$ – הסתברות כל נושא, ו-$p(w_k|x_i)$ – הסתברות w_k במלטינום של x_i.

$$= \sum_{i=1}^{|X|}\sum_{t=1}^{N} w_{ti}\log p(X_t = x_i) + w_{ti}\log p(y_t|X_t = x_i;\theta) =$$

ניזכר ש-$p(y_t|X_t = x_i;\theta) = \prod_{k=1}^{|V|} p(w_k|x_i)^{n_{tk}}$, לכן:

$$= \sum_{i=1}^{|X|}\sum_{t=1}^{N}\left[w_{ti}\log p(X_t = x_i) + w_{ti}\sum_{k=1}^{|V|} n_{tk}\log p(w_k|x_i)\right] =$$

$$= \sum_{i=1}^{|X|}\sum_{t=1}^{N} w_{ti}\log p(x_i) + \sum_{i=1}^{|X|}\sum_{k=1}^{|V|}\sum_{t=1}^{N} w_{ti}n_{tk}\log p(w_k|x_i)$$

צריך למצוא את ערכי $p(x_i), p(w_k|x_i)$ שימקסמו את הביטוי הזה, בהינתן ש-w_{ti} קבוע. $p(x_i)$ מופיע רק בביטוי השמאלי ו-$p(w_k|x_i)$ מופיע רק בביטוי הימני, לכן ניתן למקסם בנפרד.

מקסום $p(x_i)$:

$$\sum_{i=1}^{|X|}\sum_{t=1}^{N} w_{ti}\log p(x_i) = \sum_{i=1}^{|X|}\left(\sum_{t=1}^{N} w_{ti}\right)\log p(x_i)$$

כאשר ניתן להסתכל על $\sum_{t=1}^{N} w_{ti} = n_i$ – מס' המופעים של מסמך בנושא i (זהו מונים בדומה לשכיחות יחסית, אבל מונים של מופעים חלקיים - בתוחלת). לפי משפט העזר, ערכי $p(x_i)$ שימקסמו את הביטוי הם:

$$p(x_i) = \frac{n_i}{\sum_{i=1}^{|X|} n_i} = \frac{\sum_{t=1}^{N} w_{ti}}{\sum_{i=1}^{|X|}\sum_{t=1}^{N} w_{ti}}$$

מקסום $p(w_k|x_i)$:

לכל x_i אין תלות בין המחוברים ולכן נמקסם כל אחד בנפרד. עבור x_i כלשהו, צריך למקסם את:

$$\sum_{k=1}^{|V|}\sum_{t=1}^{N} w_{ti}n_{tk}\log p(w_k|x_i) = \sum_{k=1}^{|V|}\left(\sum_{t=1}^{N} w_{ti}n_{tk}\right)\log p(w_k|x_i)$$

לפי משפט העזר, ערכי $p(w_k|x_i)$ שימקסמו את הביטוי הם:

$$p(w_k|x_i) = \frac{n_k}{\sum_{k=1}^{V} n_k} = \frac{\sum_{t=1}^{N} w_{ti}n_{tk}}{\sum_{k=1}^{|V|}\sum_{t=1}^{N} w_{ti}n_{tk}}$$

<u>מושגים בסיסיים בתורת האינפורמציה</u>

נקודת המבט הבסיסית: מידול של כמות המידע בתוצאת דגימה ממשתנה מקרי. לדוגמה, בתוצאת הטלת קובייה יש יותר מידע מאשר בתוצאת הטלת מטבע (יותר ערכים = יותר אי ודאות), ובתוצאת הטלת מטבע מאוזן יש יותר מידע מהטלת מטבע לא מאוזן (התפלגות יותר מאוזנת = יותר אי ודאות). באופן שקול, ניתן לחשוב על כך כרמת אי הוודאות בהתפלגות.

דוגמה: התפלגות א' - $\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}$, התפלגות ב' - $\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}$

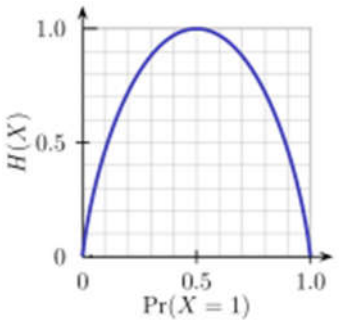
נסתכל על תוחלת מס' הביטים הדרושים לשדר תוצאת ניסוי בקידוד אופטימלי (נסמן ב-H).

$H\left(<\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}>\right)=2$ – נקודד כל תוצאה בשני ביטים: 00,01,10,11, וממוצע נקבל 2.

$H\left(<\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}>\right)=2$ – באופן אינטואיטיבי כדאי להשקיע כמה שפחות ביטים בתוצאה שההסתברות שלה הכי גבוהה, לכן נשקיע ביט 1 (0) בתוצאה הראשונה. אחרת נשדר מילה שמתחילה ב-1. באופן דומה, בחצי מהמקרים שנשדר 1 נשדר את התוצאה השנייה לכן נשקיע בה ביט נוסף (10), ואחרת נשדר עוד ביט לתוצאה השלישית (110) והרביעית (111).
התוחלת היא: $\frac{1}{2}\cdot 1+\frac{1}{4}\cdot 2+2\cdot\frac{1}{8}\cdot 3=\frac{7}{4}<2$. אי הודאות נמוכה יותר מאשר בהתפלגות המאוזנת.
בקוד הופמן, מס' הביטים האופטימלי לקידוד תוצאה המתקבלת בהסתברות p היא $-\log p$.

<u>אנטרופיה (Entropy)</u> – $H[X]=-\sum_x p(x)\log p(x)=E_X[-\log p(x)]$.

<u>דוגמה</u>: אנטרופיה של משתנה ברנולי (מטבע) כפונקציה של p:



אם ההסתברות היא 1 או 0, אין שום אי וודאות ולכן האנטרופיה היא 0. האנטרופיה המקסימלית מתקבלת בהתפלגות המאוזנת $p=0.5$.

באופן כללי, המקסימום תמיד מתקבל בהתפלגות אחידה, וטווח האנטרופיה היא:

$$0\le H[X]\le -|X|\cdot\frac{1}{|X|}\log\left(\frac{1}{|X|}\right)=\log|X|$$

<u>Perplexity</u> – של משתנה X מוגדרת להיות $2^{H[X]}$, והמשמעות היא שרמת אי-הוודאות של ההתפלגות שקולה לרמת אי-הוודאות של התפלגות אחידה עם $2^{H[X]}$ ערכים.

הרצאה 13 (26.01.2015)

מושגים בסיסיים בתורת האינפורמציה - המשך

Cross Entropy - $H(p||q) = -\sum_x p(x)\log q(x)$ מוגדר להיות מס' הביטים שנדרש בתוחלת לשדר דגימה מהתפלגות $p(x)$ כאשר הקוד האופטימלי משתמש בקירוב $q(x)$.
נשים לב שהתופעה מתפלגת לפי $p(x)$ (בטבע) אבל מס' הביטים הנדרש הוא מס' הביטים של המאורע בהתפלגות של המודל $q(x)$.

Relative Entropy – KL Divergence – מדד הממדל את הסטייה בין שתי התפלגויות p, q.
המשמעות של המדד במונחי תורת האינפורמציה היא כמה ביטים אנחנו מפסידים (הגידול באנטרופיה) ממידול של התפלגות $p(x)$ ע"י התפלגות $q(x)$.

$$D_{KL}(p||q) = H(p||q) - H(p) = -\sum_x p(x)\log q(x) + \sum_x p(x)\log p(x) = -\sum_x p(x)\log\frac{p(x)}{q(x)}$$

לא מוגדר אם קיים x שעבורו $p(x) > 0$ ו-$q(x) = 0$. המדד אינו סימטרי.

טענה - $H(p||q) \geq H(p)$
אינטואיטיבית, זה יהיה גבוה יותר מאשר האנטרופיה האמיתית כי $\log p(x)$ הוא מס' הביטים האופטימלי לקידוד מאורע מהתפלגות $p(x)$.
הוכחה – הוכחנו ש-$D_{KL}(p||q) \geq 0$ ולכן: $H(p||q) - H(p) \geq 0 \Longleftarrow H(p||q) \geq H(p)$. ■

מדד סימטרי לסטייה בין התפלגויות (Jensen-Shannon Divergence)

$$J(p||q) = \frac{1}{2}\left[D_{KL}\left(p||\frac{p+q}{2}\right) + D_{KL}\left(q||\frac{p+q}{2}\right)\right]$$

זהו המרחק הממוצע של כל אחת מההתפלגויות מההתפלגות הממוצעת. המדד תמיד מוגדר כי ההתפלגות הממוצעת לא נותנת 0 לאף x.

קיים גם מדד Skew Divergence (מדד הסטייה) שהוא ממוצע ממושקל בין שני המרחקים לפי פרמטר α.

מדידה / קירוב אמפירי לאנטרופיה של סדרות
נסתכל על משתנה שמייצר סדרת ערכים: $w =< w_1, \dots, w_n > \triangleq w_1^n$.
ערכי המשתנה: כל הסדרות האפשריות בשפה L שמיוצרת ע"י המשתנה. נסתכל על H:

$$H(L) = \sum_{w_1^n \in L} p(w_1^n)\log p(w_1^n)$$

בסדרות מקובל להסתכל על קצב האנטרופיה (Entropy Rate): $\frac{1}{n}H(L)$ (כיוון שלסדרה באורך ארוך יותר תהיה אנטרופיה גבוהה יותר). הגדרת אנטרופיה לשפה כתהליך סטוכסטי אינסופי:

$$H(L) = \lim_{n\to\infty}\frac{1}{n}H(w_1^n)$$

הבעיה עם המדד הזה הוא שקשה לחשב אותו כי צריך לרוץ על כל הסדרות של n מילים מהשפה.

משפט Shannon-McMillan-Brieman

$$H(L) = \lim_{n\to\infty} -\frac{1}{n}\log p\,(w_1^n)$$

כלומר $H(L)$ מוגדר ע"י הסתכלות על $\log p$ של סדרה אחת. זה מתקיים תחת הנחות מסוימות על

השפה, בפרט מניחים שהשפה היא Stationary (אין רגישות לנק' הזמן) ו-Ergodic – אין תלות ב"עבר" הרחוק (כמו במודל מרקובי). כשהסדרות הופכות מספיק ארוכות הן מתחילות להראות דומות אחת לשנייה מבחינת ההסתברות, לכן מספיק להסתכל על סדרה אחת מאשר על ערך משוקלל לפי ההסתברות לכל סדרה.

בדר"כ ההתפלגות האמיתית p (ההסתברות בטבע של השפה) אינה ידועה ומשתמשים במודל q עבורו. המטרות שנרצה:

א. להשתמש ב-q כדי למצוא חסם עליון לאנטרופיה של p: $H(p||q)$.
ב. להעריך איכות של מודלים q,q': $H(p||q) > H(p||q')$ – ככל שקטן יותר, המודל קרוב יותר ל-p.

איך מחשבים את ה-Cross Entropy Rate במקרה של סדרה אינסופית?

$$H(p||q) = \lim_{n\to\infty} -\frac{1}{n}\sum_{w_1^n \in L} p(w_1^n)\log q(w_1^n)$$

לפי המשפט ביחס ל-$H(p||q)$:

$$H(p||q) = \lim_{n\to\infty} -\frac{1}{n}\log q(w_1^n) = \lim_{n\to\infty} \log \frac{1}{\sqrt[n]{q(w_1^n)}}$$

לפי המשפט הזה אין שימוש ב-p ולכן ניתן להשתמש בו. לוקחים סדרת מדגם מספיק ארוכה (כך שאם נכפיל את אורכה עדיין נקבל את אותו ה-Entropy Rate) ואת המודל q (מודל שפה כלשהו) ומחשבים את הביטוי.

בהתאם אפשר לחשב גם את ה-Perplexity: $2^{H(p||q)} = \frac{1}{\sqrt[n]{q(w_1^n)}}$.

אלה אומדנים ל-Cross Entropy ול-Perplexity שלה בין ההתפלגות p של השפה להתפלגות המודל q.

<u>עבור זוג משתנים x, y עם התפלגות משותפת $p(x,y)$:</u>

<u>אנטרופיה משותפת (Joint Entropy)</u> - $H(x,y) = -\sum_x \sum_y p(x,y)\log p(x,y)$

<u>אינפורמציה הדדית (Mutual Information)</u>
כמה אינפורמציה משתנה אחד מספק על המשתנה השני.

$$I(x;y) = D_{KL}\big(p(x,y)||p(x)p(y)\big) = \sum_x \sum_y p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

מתקיים $I(x;y) \geq 0$ ו-$I(x;y) = 0$ כאשר x,y בלתי-תלויים.
ניתן להשתמש במדד הזה כדי לבדוק קשר בין שני משתנים. מודדים התפלגות משותפת ומחשבים כמה אינפורמציה הדדית יש בין המשתנים.

<u>דוגמה לשימוש חשוב ב-I: Feature Selection במסווגים – Information Gain</u>
<u>Feature Selection</u> - בחירה להתחשב רק בחלק מהמאפיינים של האובייקט לסיווג, ולסנן מהחישוב את כל היתר. לדוגמה, בנוסחת סיווג של Naïve Bayes (כמו במודל עירוב היסטוגרמות), נסנן קודם חלק מהמילים במסמך (רוצים להשאיר את המילים הכי אינדיקטיביות).

הגישה להחליט איזה מאפיינים לשמור: נמדוד מתאם/קורלציה כלשהו בין המאפיין לקטגוריה שאליה מסווגים. נסתכל על סיווג בינארי. נשתמש ב-I. נסמן ב-C את הקטגוריה (1 – המסמך בקטגוריה, 0 – לא בקטגוריה). המאפיין הוא F (1 – המאפיין מופיע, 0 – המאפיין לא מופיע). נחשב אינפורמציה הדדית בין שני המשתנים האלה:

$$I_G = I(C,F) = \sum_{c=\{0,1\}} \sum_{f=\{0,1\}} p(C=c, F=f) \log\left(\frac{p(C=c,F=f)}{p(C=c)p(F=f)}\right)$$

בפועל, משתמשים ב-top k המאפיינים הכי טובים.

מתקיים גם: $I(x;y) = H(x) - H(x|y) = H(y) - H(y|x)$.