לאחר הפעלה של אלגוריתם [Gradient Descent], יתקבלו ערכי משקלים שעבורם תתקבל [התוצאה הנמוכה ביותר] מפונקציית השגיאה (לכאורה [מינימום גלובלי]). התוצאה תלוייה בגודל הצעד. ככל שהצעד [יותר קטן], כך ה[מינימום] שיתקבל יהיה יותר קרוב ל[מינימום גלובלי] ופונקציית השגיאה תיתן ערך [יותר קטן]) אך הזמן לחישוב יהיה ארוך יותר מכיוון שיהיו יותר (epochs). פתרון ה [Normal Equation] -ימצא משקלים שנותנים שגיאה שערכה שווה בדיוק ל[מינימום גלובלי].

התשובה הנכונה היא: ה[מינימום] שמתקבל כתוצאה מאלגוריתם Gradient Descent הוא בקירוב ה[מינימום] ה[גלובלי], מכיוון שאנו משתמשים ברגרסיה ליניארית (או לוגיסטית) ויש לנו [בדיוק] נקודת [מינימום] אחת (פונקציית השגיאה שלנו הינה [ריבועית]).
[אם נריץ את המודל על רשת נוירונים שבה יש נקודות [מינימום] רבות, אנו עלולים לקבל [מינימום] [מקומי] שרירותי כתלות בערכי המשקלים הרנדומליים ההתחלתיים.]
גודל הצעד משפיע על יכולת האלגוריתם להתקרב לנקודת ה[מינימום]; ככל שהצעד יהיה גדול יותר, ההתכנסות תהיה יותר [מהירה] אך השיפור בשגיאה יפסיק רחוק יותר מנקודת ה[מינימום].

מהי סיבוכיות הזמן של אלגוריתם Normal Equation ?

השתמשו באחד או יותר מהמשתנים הבאים:
- D - אוסף דוגמאות האימון.
- d - מספר המשתנים (הפיצ'רים) המקסימלי בדוגמת אימון אחת
- H - אוסף ההיפותזות (המשוואות שחוזות את תווית דוגמת האימון).

הכפלה של X^T ב- X: $O(d^2|D|)$

הכפלה של X^T ב- t: $O(d|D|)$

מציאת מטריצה הופכית ל- X^TX: $O(d^3)$

הכפלה של $(X^TX)^{-1}$ ב- X^Tt: $O(d^2)$

סה"כ: $O(d^2|D|+d|D|+d^3+d^2)$

לאחר צמצום: $O(d^2|D|+d^3)$

מהי סיבוכיות הזמן של אלגוריתםGradient Descent?

השתמשו באחד או יותר מהמשתנים הבאים:
- D - אוסף דוגמאות האימון.
- d - מספר המשתנים (הפיצ'רים) המקסימלי בדוגמת אימון אחת.
- H - אוסף ההיפותזות (המשוואות שחוזות את תווית דוגמת האימון).
- e - מספר ה epochs -המקסימלי.

כמה פתרונות קיימים לבעיית רגרסיה לינארית כאשר ב D -יש 2 דוגמאות (בלתי תלויות ליניארית) ממימד 2?
שימו לב שמספר המשקלים גדול באחד מהממימד, מכיוון שצריך להוסיף גם את ה- intercept.
תשובה: בהסתכלות גרפית מדובר על גרף של 3D (כאשר יש 2 דוגמאות ממימד 2, אז יש 2 פיצ'רים; צירים X ו Y-ויש את ערך ה-target; ציר ה-Z).
כאשר נשים 2 נקודות (2 דוגמאות) על הגרף, יש אין סוף משטחים שניתן להעביר דרך 2 הנקודות הללו ושיתנו ערך שגיאה מינימלי.
התשובה הנכונה: אין סוף פתרונות

כמה פתרונות קיימים לבעיית רגרסיה לינארית כאשר ב-D יש 3 דוגמאות (בלתי תלויות ליניארית) ממימד 2?
שימו לב שמספר המשקלים גדול באחד מהמימד, מכיוון שצריך להוסיף גם את ה-intercept.

בהסתכלות גרפית מדובר על גרף של 3D (כאשר יש 3 דוגמאות ממימד 2, אז יש 2 פיצ'רים; צירים X ו Y-ויש את ערך ה-target ציר ה-Z).
כאשר נשים 3 נקודות (3 דוגמאות) על הגרף, יש משטח יחיד שניתן להעביר דרך 3 הנקודות הללו ושיתן ערך שגיאה מינימלי.
התשובה הנכונה: פתרון יחיד

כמה פתרונות קיימים לבעיית רגרסיה לינארית כאשר ב-D יש 4 דוגמאות (בלתי תלויות ליניארית) ממימד 2?
שימו לב שמספר המשקלים גדול באחד מהמימד, מכיוון שצריך להוסיף גם את ה-intercept.

בהסתכלות גרפית מדובר על גרף של 3D (כאשר יש 4 דוגמאות ממימד 2, אז יש 2 פיצ'רים; צירים X ו Y-ויש את ערך ה-target, ציר ה-Z.
כאשר נשים 4 נקודות (4 דוגמאות) על הגרף, יש משטח יחיד שניתן להעביר ושיתן ערך שגיאה מינימלי.
התשובה הנכונה: פתרון יחיד

ברצוננו לסווג תמונה ל- 3 קטגוריות: A, B ו- C. הקטגוריות הן Mutual exclusive.
עושים זאת בעזרת רגרסיה לוגיסטית One vs Rest ו- Cross Entropy.
איזה וקטור t (Targets) יהיה שימושי עבור התמונות המסווגות?
התשובה הנכונה: t הוא וקטור בוליאני shape=[3] בקידודOne-Hot

נסתכל על מרחב קלט 2D:
פולינום מדרגה 1: 2 קלטים (ועוד ביאס)
x_1, x_2
פולינום מדרגה 2: 5 קלטים (ועוד ביאס)
$x_1, x_2, x_1x_2, x_1^2, x_2^2$
פולינום מדרגה 3: 9 קלטים
$x_1, x_2, x_1x_2, x_1^2, x_2^2, x_1^2x_2, x_1x_2^2, x_1^3, x_2^3$

מה היה קורה אם מרחב הקלט המקורי הוא 3D והפולינום הדרוש הוא מדרגה 2?

$x_1, x_2, x_3, \ x_1x_2, x_1x_3, x_2x_3, x_1^2, x_2^2, x_3^2,$

[8] בהינתן מסווג של naive Bayes בעל 3 קלטים בוליאנים X1,X2,X3 ופלט בוליאני Y.
a. מהו מספר הפרמטרים (המינימלי) שהמסווג צריך לשערך לצורך למידה? רשום אותם.
p(+), p(x1=1|+), p(X2=1,+), p(X3=1|+)
p(x1=1|-),p(x2=1|-), p(x3=1|-)
היתר ניתנים לחישוב למשל p(xi=0|+) =1-p(xi=1|+) , p(-)=1-p(+) ,
סה"כ 7

שערוך ההסתברות:

$P(cancer) =$ 0.008

$P(\neg cancer) =$ 0.992

$P(+|cancer) =$ 0.98

$P(-|cancer) =$ 0.02, False neg 2%

$P(+|\neg cancer) =$ 0.03 False pos 3%

$P(-|\neg cancer) =$ 0.97

$P(cancer|+) =$ P(+|cancer)P(cancer)/P(+)=0.98*0.008/p(+)=**0.00784 / P(+)**

P(¬cancer|+) = P(+|¬cancer)P(¬cancer)/P(+)=0.03*0.992/p(+)=**0.0298 / P(+)**

Joint probabilities: P(+,cancer)=p(+|cancer)p(cancer) ≈ 0.98*0.008=0.00784
P(+, ¬cancer))=p(+|¬cancer)p(¬cancer) ≈ 0.03*0.992=0.0298

- מה ההסתברות של P(cancer|+)?

- P(+)=P(cancer,+) + P(¬cancer,+) ≈ 0.00784+0.0298=0.03764
- P(cancer|+) ≈ 0.00784/0.03764≈**0.21**
- P(¬cancer|+) ≈ 0.0298/0.03764≈**0.79**

Naïve Bayes Classification Algorithm:

Naïve_Bayes_Classifier Training (D)

שלב ראשון: הערכת כל ההסתברויות הנחוצות:
- ה priors של הקלסים וגם ההסתברויות של
- הקטגוריות של הfeatures בהנתן קטגוריה

For each category k=1…K
- $P(k)$=estimate $p(y=k) \approx \frac{n_k}{n}$
- For each value v_j of each attribute x_j
  - $P(v_j|k)$= estimate $p(x_j = v_j \ \ |y=k) \approx \frac{n_{v_{j,k}}}{n_k}$

Naïve_Bayes_Classifier Inference(x)

שלב שני: קבלת קלט: x=(v_1,v_2, … .v_n)
וחישוב $argmax_k$ של p(x,k)

$$predict = argmax_k \left\{P(k)\prod_{j=1}^{n} P(v_j|k)\right\}$$

את ה Priors ניתן לשערך מ D אם הוא מדגם טוב של כל האוכלוסיה
אם המדגם איננו של כל האוכלוסייה, אולי ניתן לשערך ממקורות אחרים: ספרי לימוד, ידע של מומחים
במיקרים אשר בהם לא ניתן לשערך את ה priors, מניחים איזון בין הקטגוריות ומשתמשים ב Max-liklihihood

כן, exhausted feature selection. סיבוכיות - 2^n

תשובה: CV

$$minimize_w\left\{\sum_{i=1}^{m}\max[0, 1-y_i h(x_i)] + \gamma\sum_{j=1}^{n} w_j^2\right\}$$

Example: Classification using Bayes Rule

$$p(k\,|x) = \frac{p(k)\,p(x|k)}{P(x)} = \frac{p(k)\,p(x|k)}{\sum_i^K p(x|i)p(i)}$$

דוגמא: X=‘fever’ הינו סימפטום, וצריך לסווג: k=1 (notCorona), k=2 (Corona)
בהרבה מיקרים קל יותר להשיג סטטיסטיקה על
p(fever|corona) מאשר על P(corona|fever)
מספיק לדעת את p(fever|corona), את ה
prior של Corona ואת ה prior של fever
כדי לחשב p(Corona|fever)
נניח p(fever|corona)=0.7 , P(corona)=0.001 , p(fever)=0.0011
p(corona|fever)=0.001*0.7/0.0011
לחילופין , אם אין את הסטטיסיקה על p(feaver) אולי
נוכל להשיג את p(feaver|notcorona)=0.0004 ,
P(corona|fever)=0.001*0.7/ (p(fever|corona)p(corona)+p(fever|notcorona)p(notcorona))=
0.001*0.7/(0.7*0.001+0.0004*0.999)

תרגיל ביאזיאני:

איבחון מחלה על פי סימפטום/בדיקה

MAP vs. ML :

מטופל נבדק במעבדה והבדיקה חוזרת חיובית למחלת הסרטן.
האם תבחר לעשות לו ניתוח מורכב?

- **בדיקת מעבדה חיובית תתקבל ב 98% מהמיקרים בהם ישנה המחלה** (רגישות 0.98):
  2% של false negative rate – חולים שקיבלו בדיקה שלילית
  98% של true positives rate - חולים שקיבלו בדיקה חיובית

P(-|cancer) ≈ 0.02
P(+|cancer) ≈ 0.98

- **תוצאה שלילית נכונה תתקבל ב 97% מהמיקרים בהם המחלה איננה** (ספציפיות 0.97):
  3% של false positives rate - בריאים עם בדיקה חיובית
  97% של true negative rate- בריאים עם בדיקה שלילית

P(+|¬cancer) ≈ 0.03
P(-|¬cancer) ≈ 0.97

- **ידע מוקדם: 0.8% מהאוכלוסיה חולים במחלה**

P(cancer) ≈ 0.008
P(¬cancer) ≈ 0.992

מכיוון שהמספרים מבוססי מדגם, הם רק שערוכים של ההסתברויות

האם הסיכוי של המטופל להיות חולה גדול מהסיכוי שלו להיות בריא?

ל- y שתי קטגוריות: cancer, ¬cancer
x הוא תוצאת הבדיקה : מאפיין בוליאני + או –
סווג חולה או לא חולה על פי MAP וגם על פי ML.

מה מספר הפרמטרים המינימלי שהמסווג צריך לשערך מתוך D אם לא נשתמש בהנחה הנאיבית?

p(x|+) ו – p(x|-) לכל אחד מ 7 קלטים שיכולים להינתן (את השמיני ניתן לחשב). סה"כ 7+7=14 . את ה Priors ניתן לחשב.