עיבוד שפה טבעית ש4:
סיווג מסמכים (המשך)

פרקים: SLP 5, Eisenstein 2-4

מטריקת F - רענון

אמיתי /// חזוי	+	-
+	60 TP	20 FN
-	10 FP	10 TN

Precision = TP/(TP+FP) ← כמה המודל פוגע

Recall = TP/(TP+FN) ← כמה המודל תופס

- **ממוצע הרמוני** בין P ל-R
- פרמטר β קובע כמה כל חלק "חשוב"
  - תיקון שגיאות - תפיסה זה חשוב, אבל פגיעה הרבה יותר
  - (לא נל"פ) איתור וירוס קורונה בבדיקה מהירה - תפיסה יותר חשובה
  - כמעט תמיד נשתמש במטריקה המאוזנת בכל-זאת.
- הנוסחה:

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

- מה הערך המאזן?
- איך מחזקים את R?

אינטואיציה על F ועל β

- (format: $F_\beta(P,R)$)
- $F_1(0.5,0.5) = 0.5$
- $F_1(0.6,0.4) = 0.48$
- $F_1(0.8,0.2) = 0.32$
- $F_{0.5}(0.8,0.2) = 0.5$
- $F_{0.5}(0.2,0.8) = 0.24$
- $F_{0.5}(0.4,0.6) = 0.43$
- $F_2(0.2,0.8) = 0.5$

- לסיכום - אם אנחנו רוצים שה-precision ישפיע יותר, ניקח β [קטנה/גדולה] יותר

מה קורה כשיש יותר משני תגים?

/// חזוי אמיתי ///	**א**	**ב**	**ג**	**NIL**
א	120	20	5	5
ב	60	60	30	20
ג	5	0	5	20
NIL	10	10	10	10

הערכת סיווג מרובה-תגים

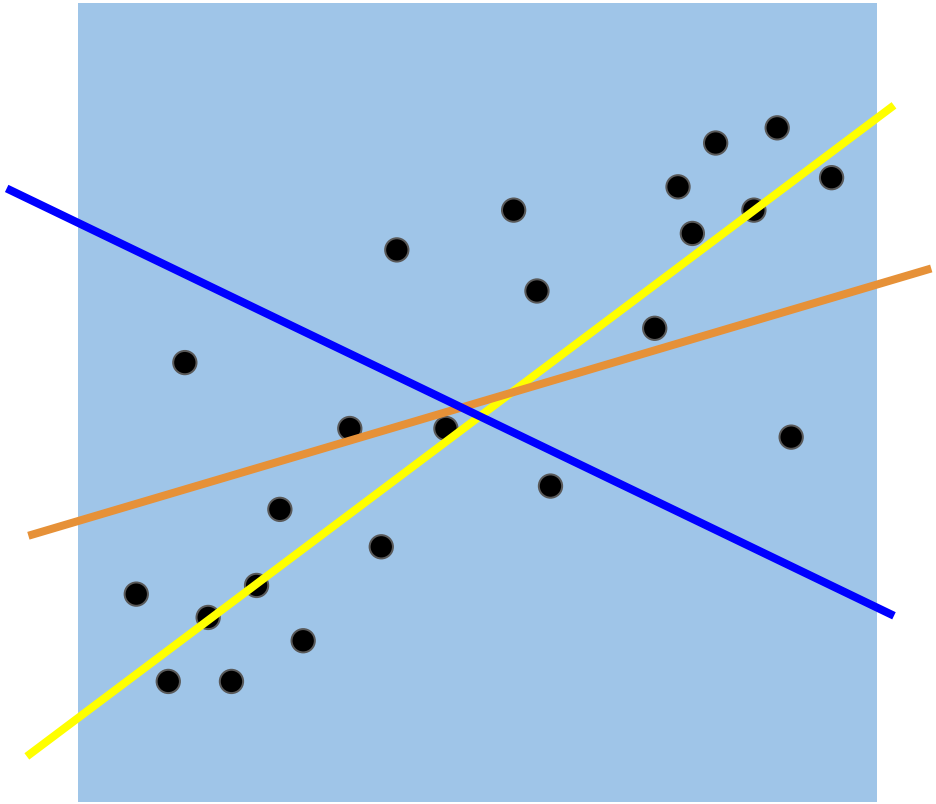
/// חזוי אמיתי ///	א	ב	ג	NIL
א	120	20	5	5
ב	60	60	30	20
ג	5	0	5	20
NIL	10	10	10	10

- עבור כל תג אפשר לחשב בנפרד P, R, F
- אנחנו רוצים לדווח **מספר אחד** למערכת שלנו
- אפשרות א': למצע F בין התגים
  - מיצוע מאקרו (macro-averaging) כי כל תג שווה חצי
  - למי הוא נותן יותר משקל?
- אפשרות ב': לחשב P, R על הכל ולחשב F אחד
  - מיצוע מיקרו (micro-averaging) כי חצי
  - מי מקבל כאן יותר משקל?

לימוד מכונה מפוקח: מרכיבי מסווג הסתברותי

- אופן ייצוג הקלט (**פיצ'רים**)
- **פונקציית סיווג** (לחישוב ההסתברות של כל קלאס בהינתן הדוגמא)
- **פונקציית מטרה** (objective function), לפעמים תוגדר בצורה הפוכה כפונקציית הפסד (loss), שנרצה להביא למקסימום/מינימום על-פני כל הנתונים
- **אלגוריתם למידה** (למעבר על דאטא האימון ושיפור פונקציית המטרה)

דוגמה ללימוד מכונה מפוקח: מציאת קו מגמה (linear regression)

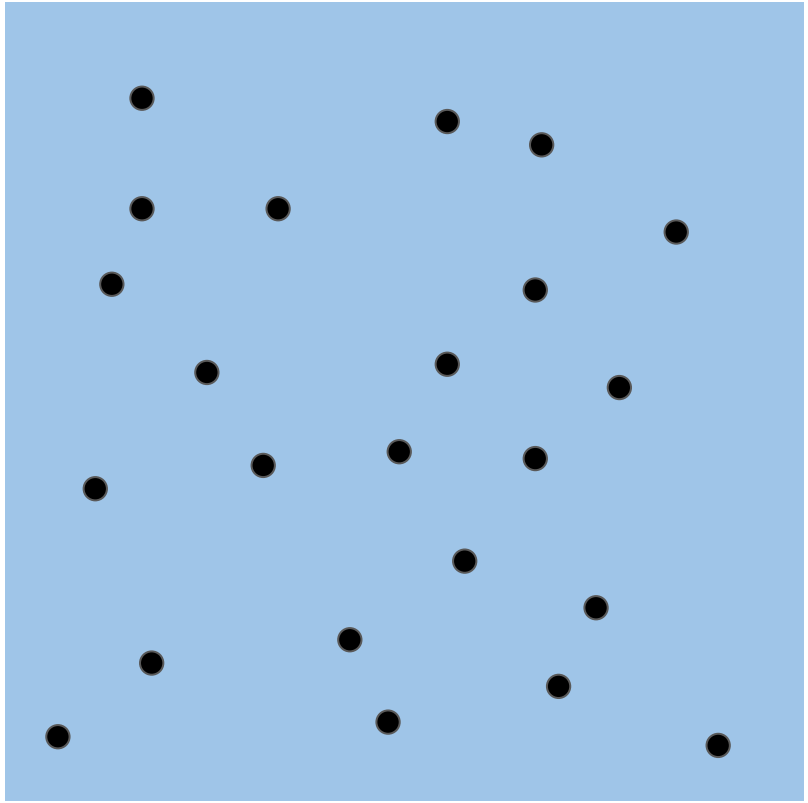


מה הפיצ'רים?

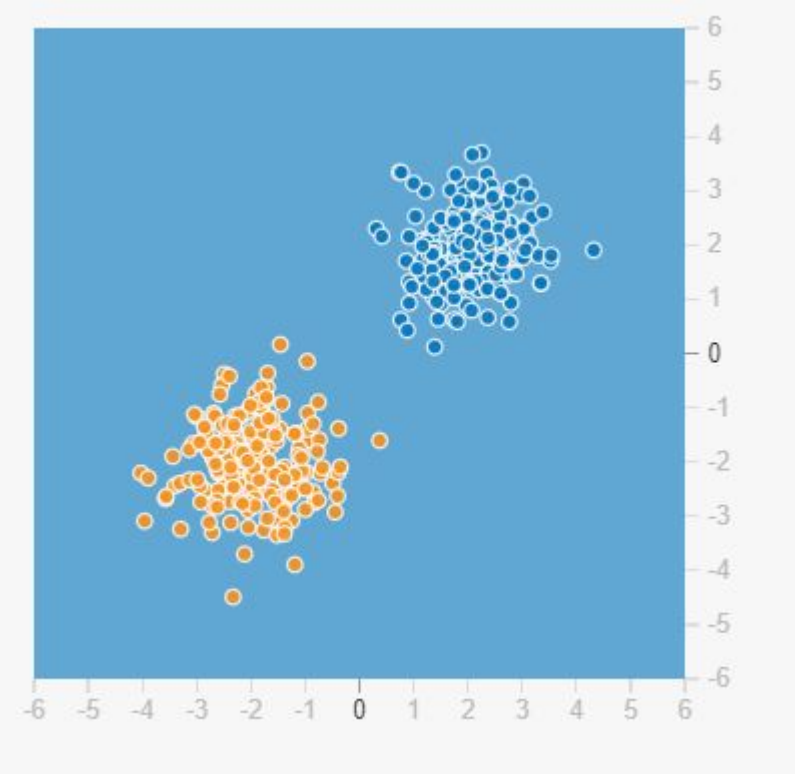
מה פונקציית הסיווג?

רעיון לפונקציית מטרה?

דוגמה ללימוד מכונה מפוקח: מציאת קו מגמה (linear regression)

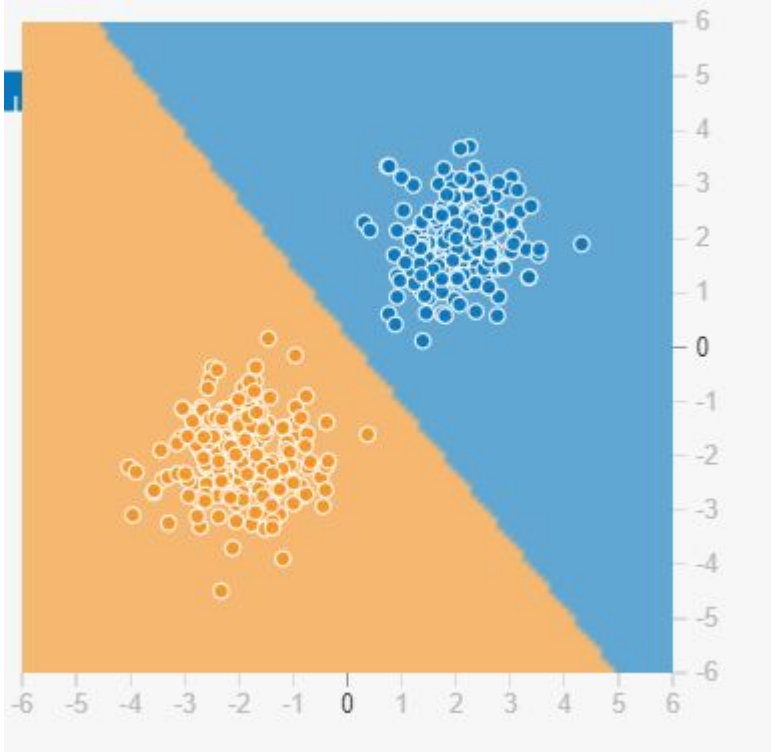


דוגמה שנייה: מציאת קו גבול (linear separation)

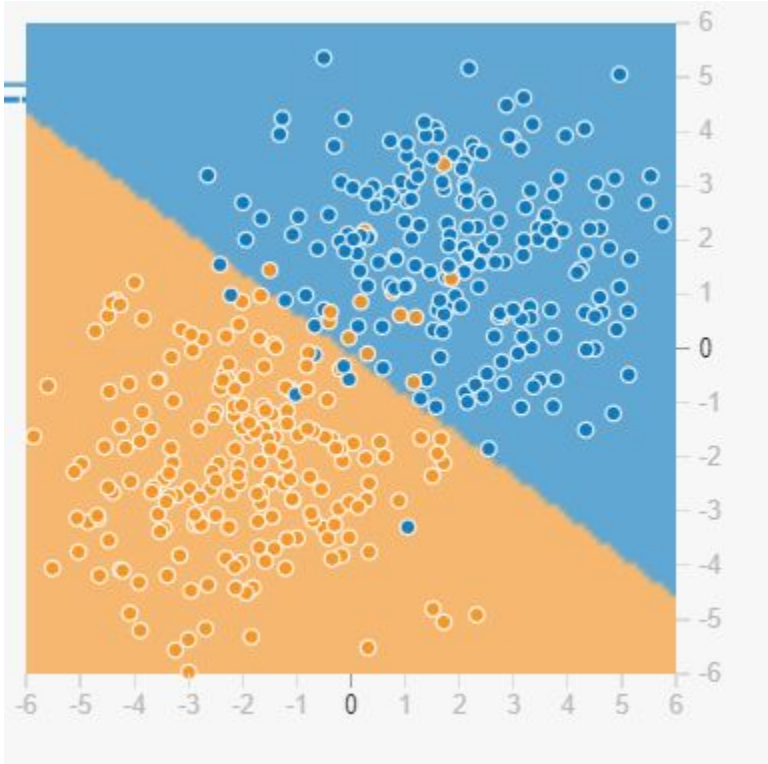


https://playground.tensorflow.org/

דוגמה שנייה: מציאת קו גבול (linear separation)

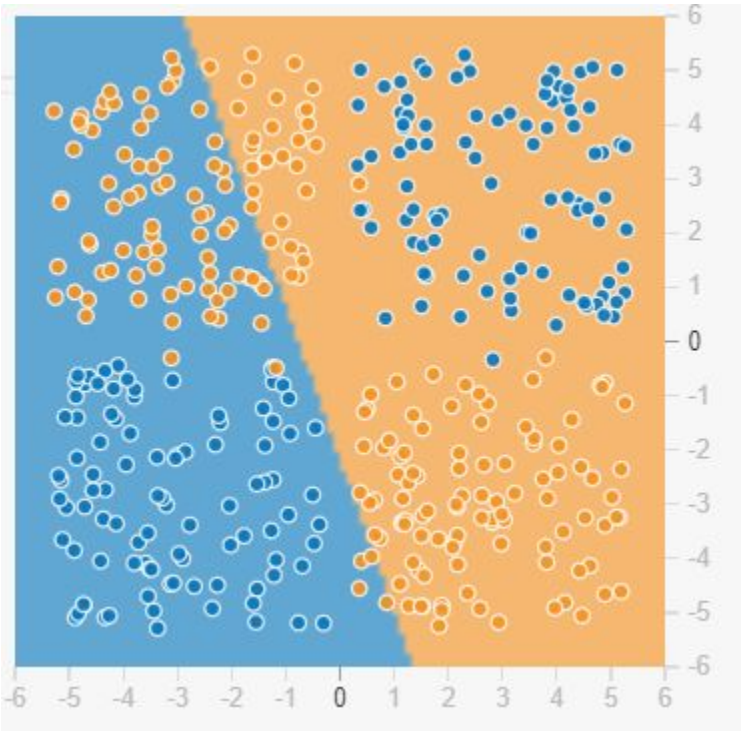


דוגמה שנייה: מציאת קו גבול (linear separation)



Training loss 0.079

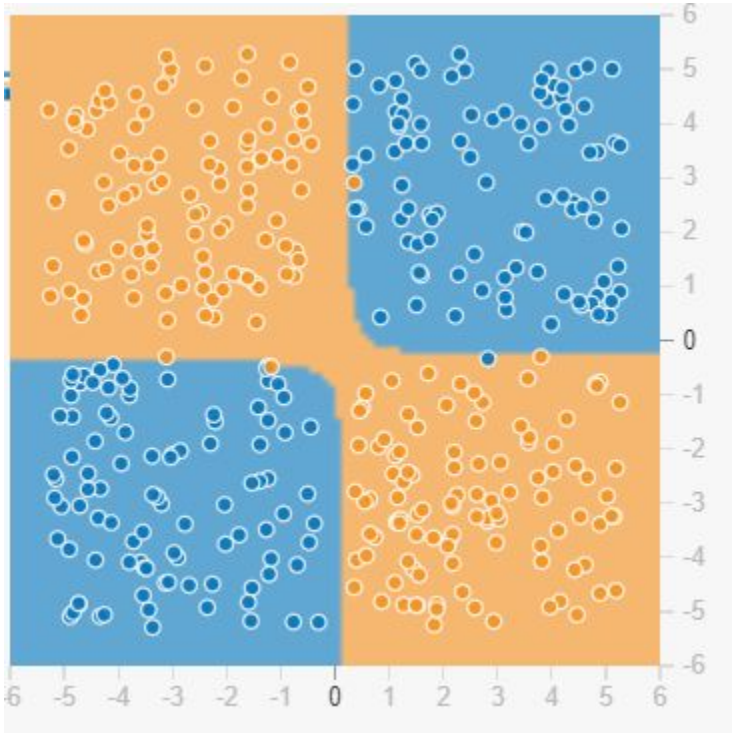
דוגמה שנייה: מציאת קו גבול (linear separation)



Training loss 0.538

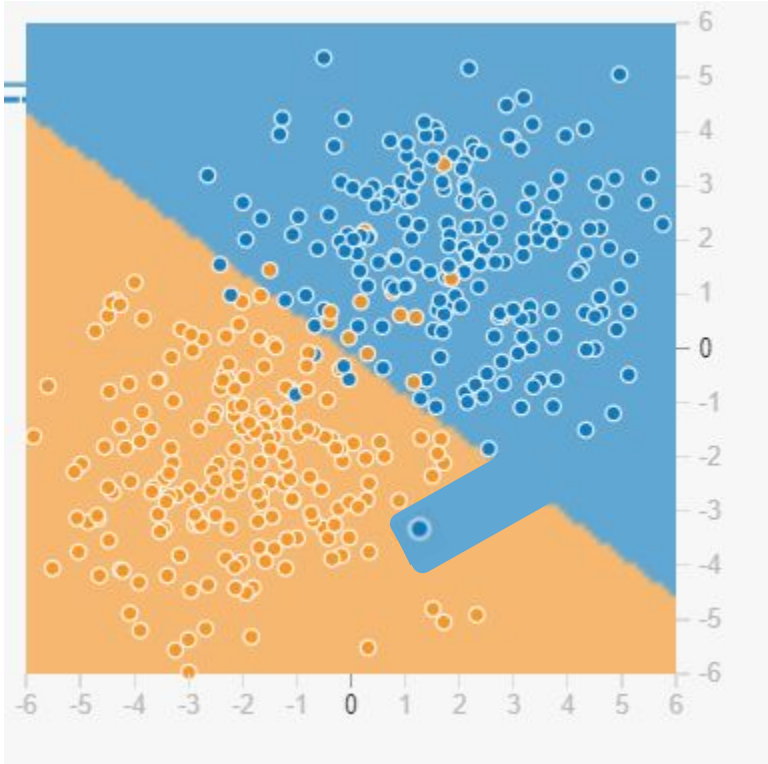
מודל "חלש" מדי
Underfitting

דוגמה שנייה: מציאת קו גבול (linear separation)



Training loss 0.012

דוגמה שנייה: מציאת קו גבול (linear separation)



מודל "חזק" מדי
Overfitting

תכונה של מרחב ההשערה
Hypothesis class

פיצ׳רים (features)

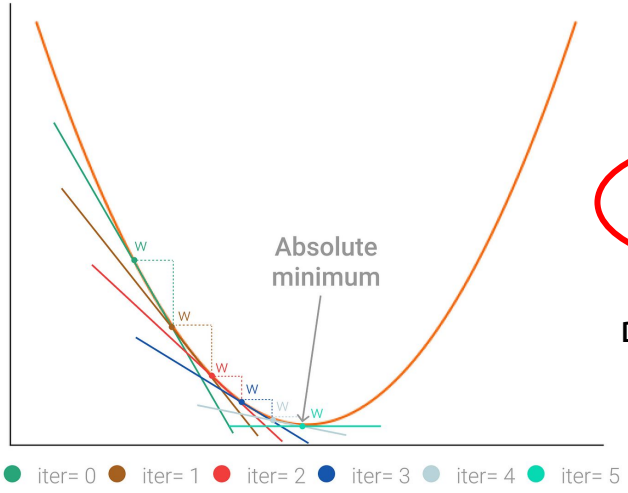
- הדוגמאות $x^{(i)}$ מגיעות מאיזשהו עולם, וצריך לייצג אותן בדרך ש״מדברת״ עם אלגוריתם חישובי
- במקרים הבסיסיים המצב ברור: כל נקודה נתונה ע״י שתי קואורדינטות, $x^{(i)}_1$ ו-$x^{(i)}_2$.
- ברוב המקרים, ובטח בשפה, צריך לחלץ ייצוגים כאלה, ולדברים שחילצנו נקרא features.
  - לדוגמא, בטקסט ביקורת סרט, פיצ׳ר יכול להיות מספר הפעמים בהם הופיעה בטקסט המילה *טוב*.
    - (כמה פיצ׳רים מהסוג הזה צריך?)
  - לדוגמא, בזיהוי ישויות, פיצ׳ר יכול להיות ״האם המילה נגמרת ב-*ברג*?״
    - (כמה פיצ׳רים צפויים כאן?)
    - (כמה ערכים יש לכל פיצ׳ר?)
- כל פיצ׳ר יצטרך התייחסות מהפרמטרים (θ) במודל (אחרת הוא מיותר)
- לכן נגדיר אותם תמיד כוקטור (feature vector)

פיצ'רים לדוגמא במערכת סיווג

- ספירת מילים (Bag of words)
- לקסיקונים (Lexicons)
  - רשימות מילים מקוטלגות כבעלות "אופי" או "תכונה" - ניתן לספור הופעות
- "צורת המסמך" (surface features)
  - מספר מילים
  - מספר types
  - סימני פיסוק
  - מבנים תחביריים
- פיצ'רים חוץ-טקסטואליים
  - (לביקורות) שם המשתמש
  - זמן הפרסום

ירידה בגרדיאנט (Gradient Descent, GD)

- רוב משפחות הפונקציות שניתקל בהן אינו פתירות באופן סגור
- עבור פונקציית מטרה קעורה וקצב "מתאים", **מובטחת התכנסות למינימום**.
- ירידה סטוכאסטית - Stochastic GD, SGD - מעבר על הדאטא דוגמא-דוגמא
  - או באצוות - batches
  - מצריך **אתחול** כלשהו של הפרמטרים
  - מצריך הגדרת **קריטריון התכנסות**



$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla J(\theta^{(t)})$$

פרמטרים חדשים

"קצב למידה" (לא חייב להיות קבוע)

פרמטרים קודמים

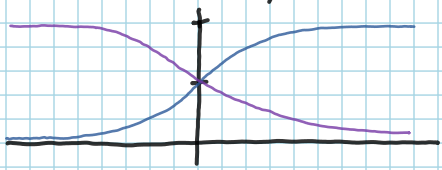
גרדיאנט ההפסד

רגרסיה לוגיסטית - לוח

רגרסיה לוגיסטית

חישוב ישיר של ההסתברות $P(y|x)$

פונקציית הסיגמויד σ:



$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$$\sigma(-z) = 1-\sigma(z)$$

$$\Rightarrow \sigma(z)+\sigma(-z) = 1$$

$$\Rightarrow \sigma'(z) = \sigma(z)\left(1-\sigma(z)\right)$$

פונקציית ההיפותזה:

$$\hat{y} = \sigma(\vec{F}\cdot\vec{\Theta}+b)$$