

*Machine learning*

---

# Naïve Bayes

Lecture IV

---

פיתוח:  
ד"ר יהונתן שלר  
משה פרידמן



---

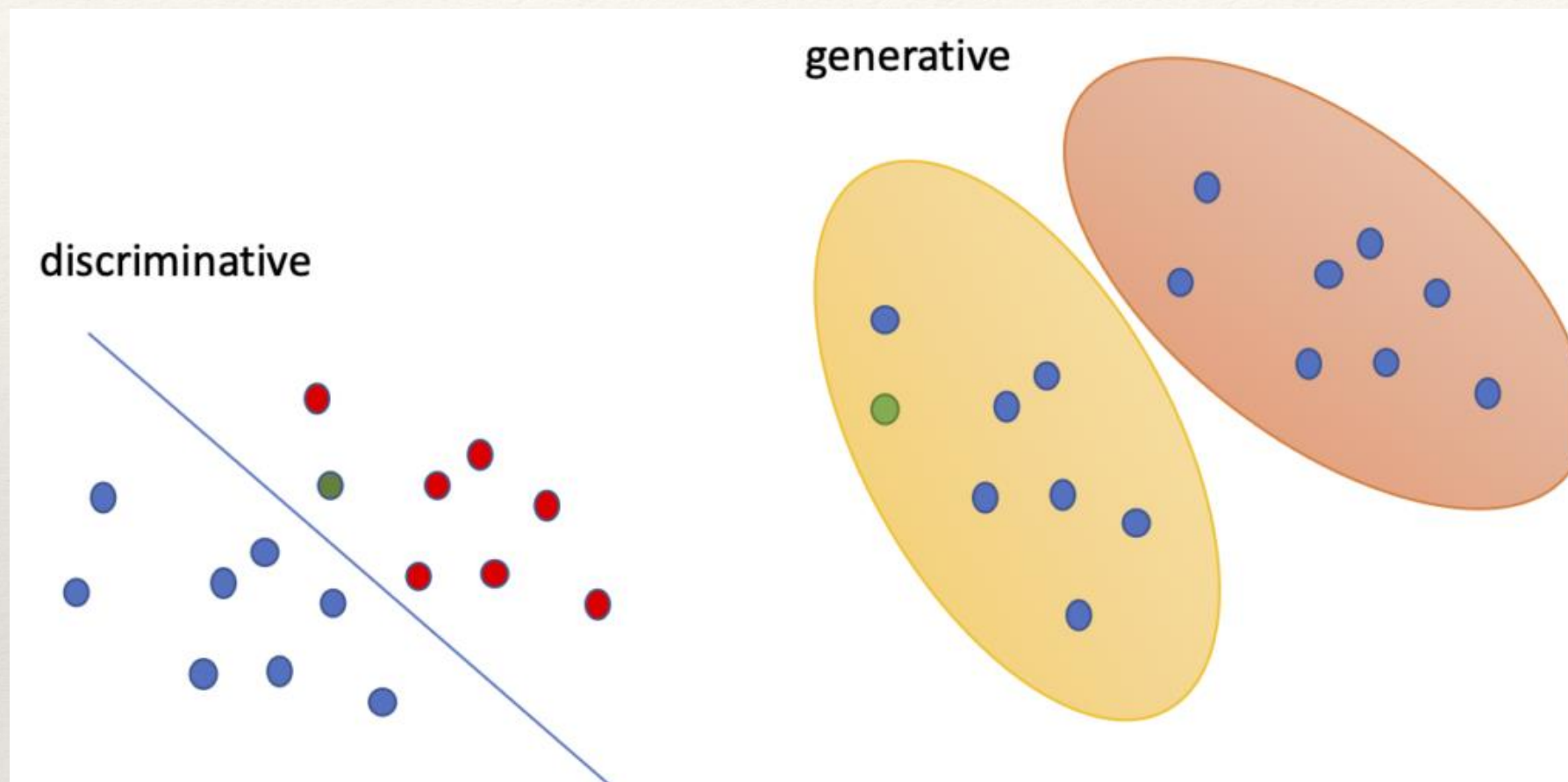
# מה נלמד על Naïve Bayes

---

- ❖ מודל גנרטיבי לסיווג
- ❖ מבוסס על התפלגות הנתונים במדגם (train-set-ב)
- ❖ המודל שייך למודלים סטטיסטים בלמידת מכונה
- ❖ משתמש בהנחת ביים



# Generative vs. discriminative models

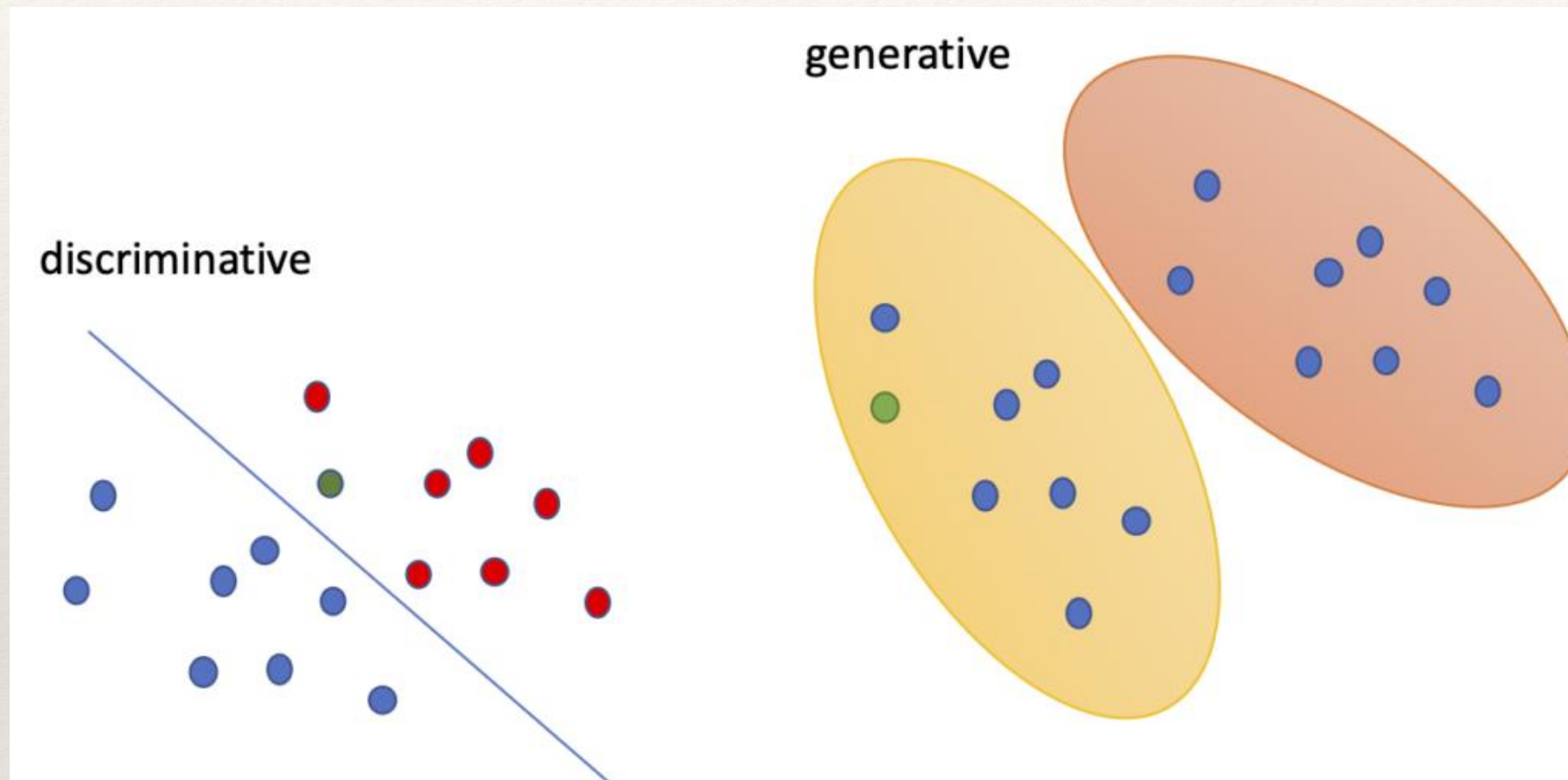


מודל גנרטיבי - מודל בו  
מנסים ללמוד את  
ההתפלגות משותפת בין  
המחלקה והמאפיינים.  
יש ביכולתם ליצר  
דוגמאות חדשות  
(generate).

מודל דיסקרמינטיבי - מודל בו מחפשים הפרדה  
בה יש להחליט איפה מתחילה ונגמרת מחלקה אחת  
ואיפה מתחילה ונגמרת מחלקה שניה.

# – Generative vs. discriminative models

## הגדרה חלופית



מודל גנרטיבי - מודל  
המתבסס על ההסתברות  
המותנת של האפיינים  
בהנתן ערך המחלקה  
 $\Pr(X | Y=y)$

מודל דיסקרמינטיבי -

מודל המתבסס על ההסתברות המותנת של  
המחלקה בהנתן ערכי המאפיינים.  
 $\Pr(Y | X=x)$



---

# שאלת סקר –

## מודלים דיסקרמנטיבים ומודלים גנרטיבים

---

מה מבין הבאים נכון:

א. עץ החלטה הוא מודל דיסקרמנטיבי

ב. עץ החלטה הוא מודל גנרטיבי



---

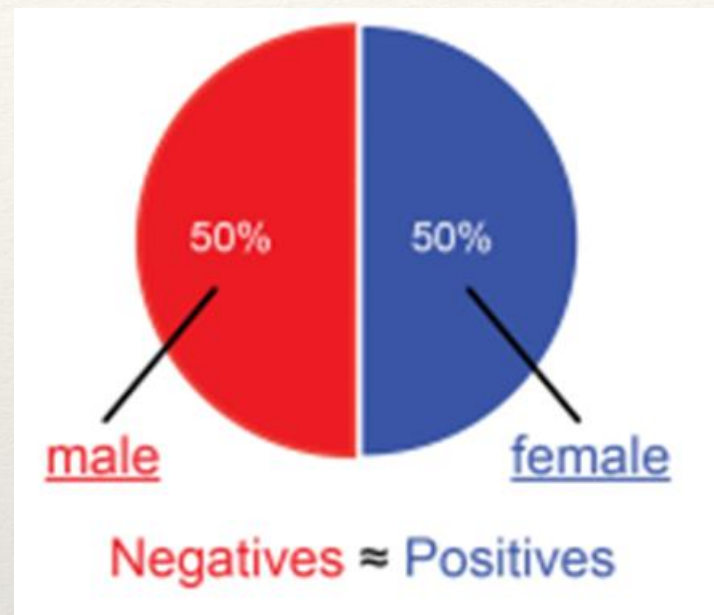
# Naïve Bayes - רקע

---

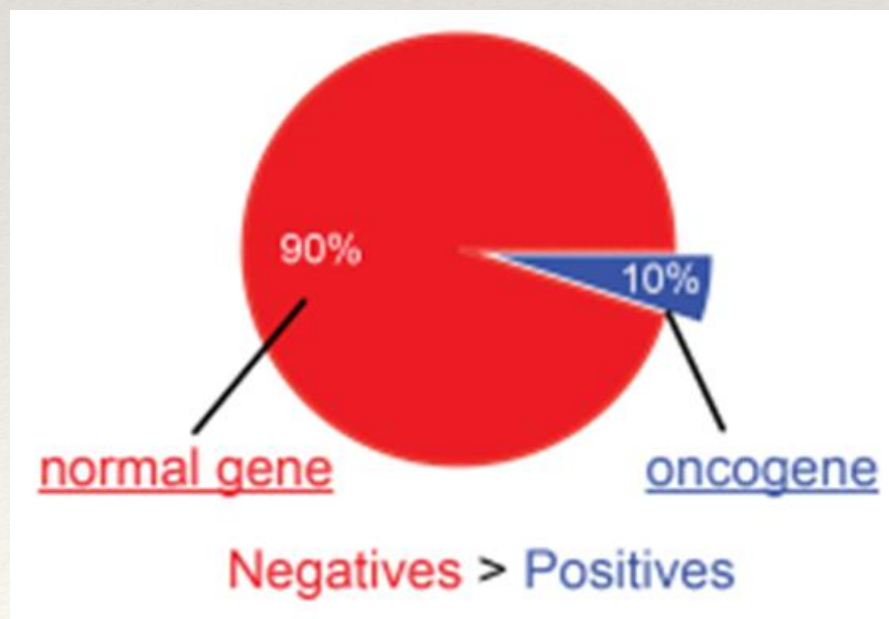
- ❖ מבוסס על ידע מוקדם של התפלגות הקטגורית / מחלקות. מתחשב בשאלות כמו למשל:
- ❖ איזו קטגוריה יותר נפוצה
- ❖ בהינתן מאפיין מסוים איזו קטגוריה יותר סבירה
- ❖ ....
- ❖ בסוף נקבל נוסחה שבהתחשב בכל המאפיינים ושבשכיחות הקטגוריה, תגיד איזו קטגוריה סבירה יותר.



# התפלגות המחלקות



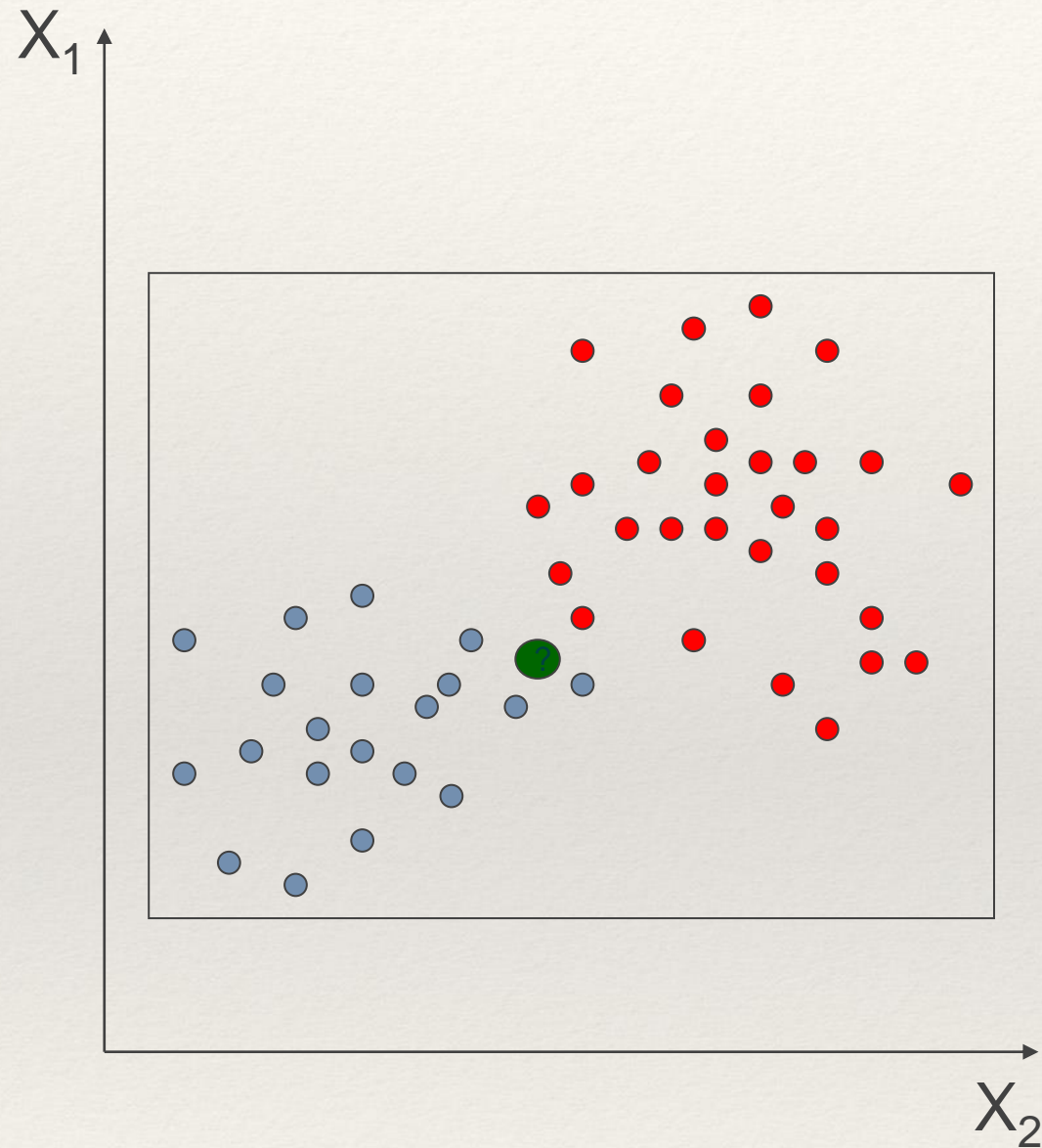
במגדר, למשל,  
הנתונים מתפלגים  
בערך בצורה מאוזנת  
(balanced).



משתנים אחרים, בהם  
יתכן ונרצה להתייחס  
(בבעיות סיווג), אינם  
מתפלגים בצורה  
מאוזנת  
(imbalanced)



# Classify according prior knowledge



Num. of observations (train-set) = 50

Prior(Blue) = 20/50

Prior(Red) = 30/50

With **only** the prior distribution

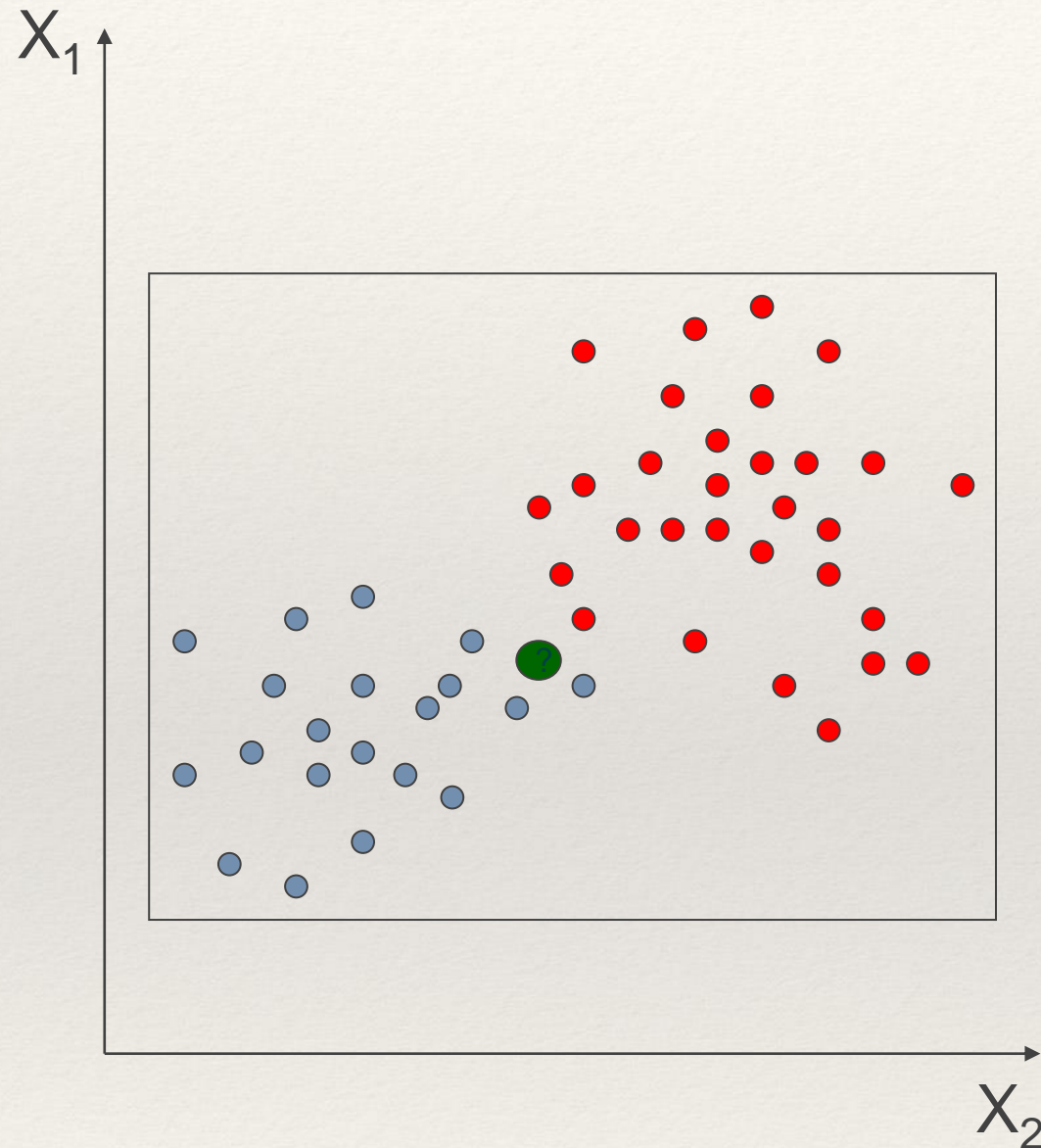
Classify a new example as “Red”



# Bayes classifier

*Prior* probabilities indicate that our new example may belong to **RED**

(More **RED** objects than **BLUE**)



In the Bayesian analysis, we combine the *prior* and the *likelihood*, to find a *posterior* probability using the Bayes' rule



---

# שאלת סקר - Naïve Bayes

---

❖ איזו מהטענות הבאות נכונה עבור Naïve Bayes?

א. במודל Naïve Bayes מחפשים הפרדה בה יש להחליט איפה מתחילה ונגמרת מחלקה אחת ואיפה מתחילה ונגמרת המחלקה שניה

ב. אלגוריתם Naïve Bayes מתחשב בהטיות סטטיסטיות, כמו, הקטגוריה היותר נפוצה.



# מוטיבציה – מכירות

- צופה מתבונן על חנות מחשבים ומגיע ל"אבחנות" הבאות על-פני קבוצת אימון גדולה.
  - 40% מבין הנכנסים לחנות קונים מחשב.
  - מבין אלו הקונים מחשב – 50% לבושים בחליפה ו-50% אינם לבושים בחליפה.
  - מבין אלו שלא קנו מחשב – כולם לבושים בחליפה.
- שאלה: הצופה רואה כעת "אדם לבוש בחליפה נכנס לחנות" – צריך לתת חיזוי האם אדם זה יקנה מחשב או לא.
- דרך הפתרון:
  - אנו צריכים לחשב את ההסתברויות הבאות:
    - הסתברות מותנית ש"אדם יקנה מחשב" בהינתן ש"לבוש בחליפה"
    - הסתברות מותנית ש"אדם לא יקנה מחשב" בהינתן ש"לבוש בחליפה"
  - ניתן חיזוי לפי ההסתברות הגבוהה מבין השתיים.



# הרעיון הכללי

❖ בהינתן ווקטור לסיווג  $(x_1, x_2, x_3, \dots, x_n)$  – ננסה להעריך את ההסתברות עבור כל סיווג  $c_i$  השייך לקבוצה  $C$  ולבחור את הסיווג עם ההסתברות הגבוהה ביותר.

$$P(c_1 | x_1, x_2, x_3, \dots, x_n) \quad \diamond$$

$$P(c_2 | x_1, x_2, x_3, \dots, x_n) \quad \diamond$$

$$P(c_3 | x_1, x_2, x_3, \dots, x_n) \quad \diamond$$

$$h_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c|X) \quad \dots \quad \diamond$$

*MAP = Maximum a posteriori (estimation) (will be explained later)*

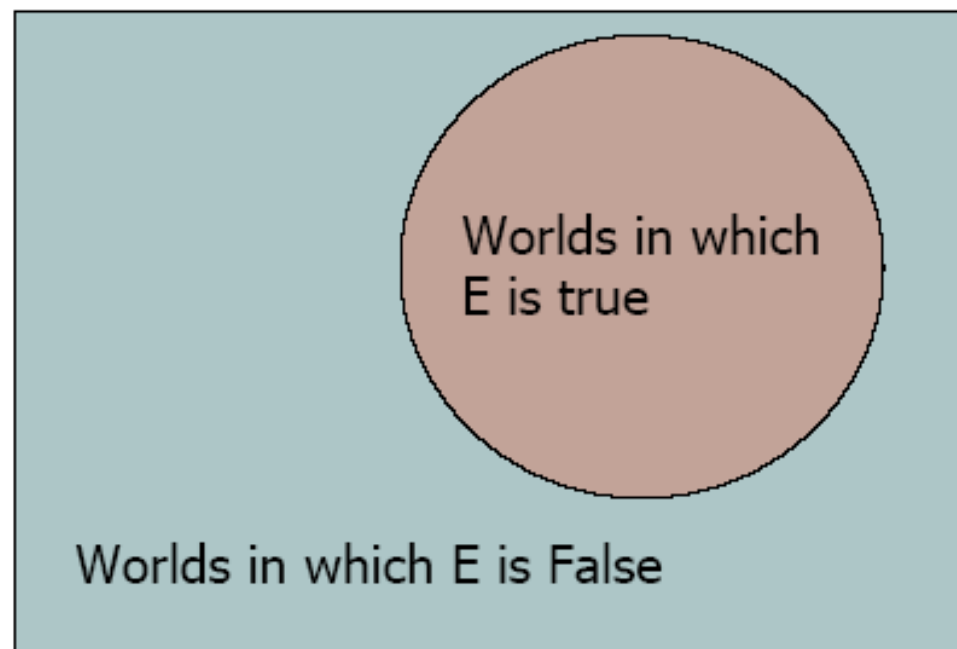


# הסתברות - דיאגרמת ואן

- We write  $P(E)$  as "the fraction of possible worlds in which  $E$  is true"

Event space of  
all possible  
worlds

Its area is 1

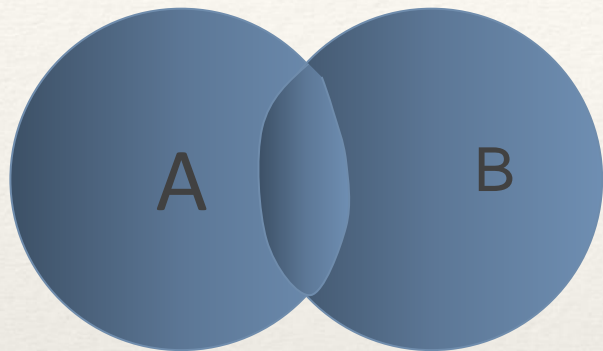


$P(E)$  = Area of  
brown circle

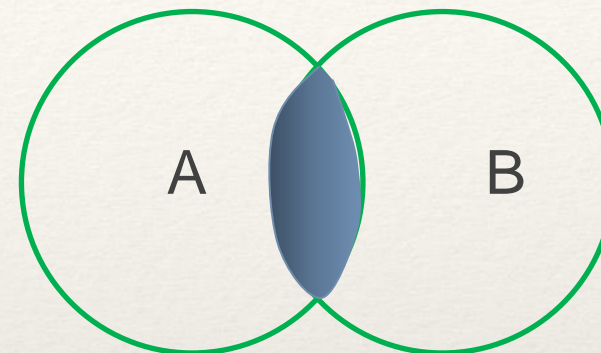


# דיאגרמות ואן

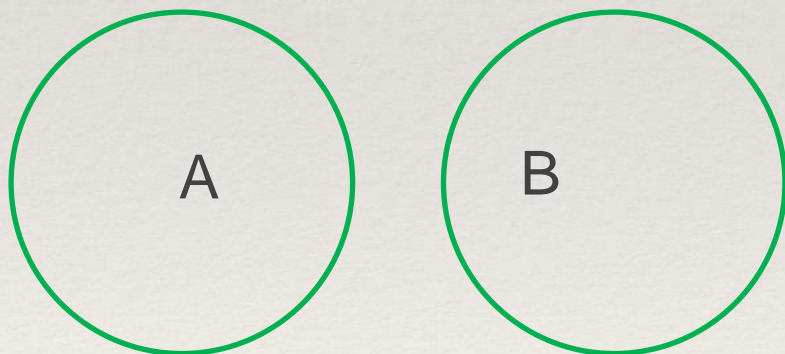
$$A \cup B$$



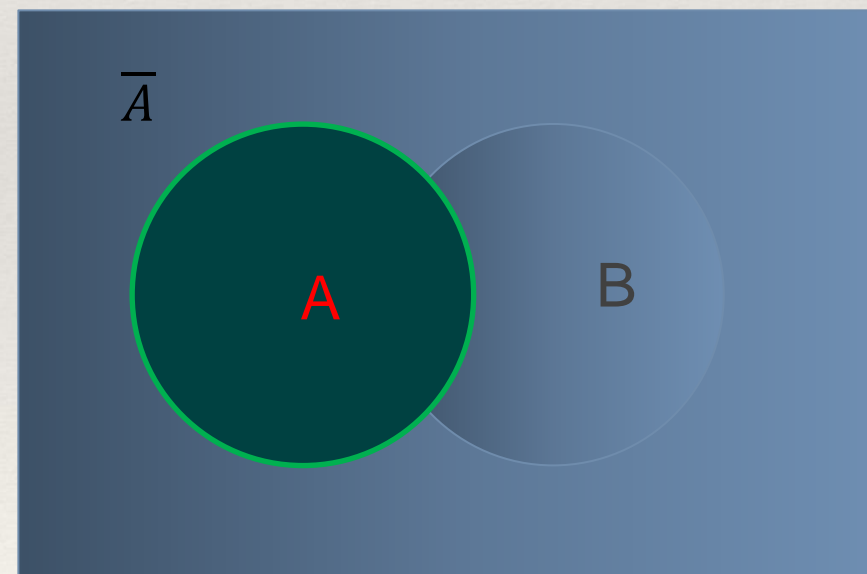
$$A \cap B$$



$$A \cap B = \varnothing$$



$$\overline{A}$$





# Bayes Classifier

- ❖ Given a new vector

$$(x_1, x_2, \dots, x_N)$$



$$C_i = ?$$

- ❖ We need to calculate

$$P(c_i | x_1, x_2, \dots, x_N) \quad \forall c_i$$

- And selecting the  $c_i$  that gives the maximum probability



---

# שאלת סקר – posterior probability , prior probability

---

❖ מה ההבדל בין prior probability לבין posterior probability של המחלקה?

א. posterior probability היא הסתברות של המאפיינים ו-prior probability , היא הסתברות של המחלקה.

ב. posterior probability היא הסתברות של המחלקה ו-prior probability , היא הסתברות של המאפיינים.

ג. prior probability היא הסתברות ראשונית של המחלקה ללא תלות במאפיינים ו-posterior probability היא הסתברות של המחלקה התלויה במאפיינים.

ד. prior probability היא הסתברות דסקרימינטיבית ו-posterior probability היא הסתברות גנרטיבית.



---

# שאלת סקר – posterior probability , prior probability

---

❖ מה ההבדל בין prior probability לבין posterior probability של המחלקה?

א. posterior probability היא הסתברות של המאפיינים ו-prior probability, היא הסתברות של המחלקה.

ב. posterior probability היא הסתברות של המחלקה ו-prior probability, היא הסתברות של המאפיינים.

ג. prior probability היא הסתברות ראשונית של המחלקה ללא תלות במאפיינים ו-posterior probability היא הסתברות של המחלקה התלויה במאפיינים.

ד. prior probability היא הסתברות דסקרימינטיבית ו-posterior probability היא הסתברות גנרטיבית.



# דוגמה 1 - הסתברות מותנית

❖ בזריקת קוביה נגדיר את  $X$  כתוצאת הזריקה.

❖ נגדיר שני מאורעות:

❖  $F$  הינו המאורע  $X=6$

❖  $E$  המאורע  $X>4$

❖ נניח שזרקו מטבע ואמרו לנו שמאורע  $E$  התרחש. מה ההסתברות שגם מאורע  $F$  התרחש?

$$P(F|E) = \frac{P(F \wedge E)}{P(E)} = \frac{\frac{1}{6}}{\frac{2}{6}} = \frac{1}{2}$$



# דוגמה 2 - הסתברות מותנית

❖  $P(B|A)$  - ההסתברות שיקרה מאורע B בתנאי שידוע כי מאורע A כבר קרה.

❖ A – בהטלת קוביה ראשונה קיבלנו "6"

❖ B – הסכום בשתי הקוביות הוא לפחות 10

❖ חשבו  $P(B|A)$

❖ הצירופים האפשריים ל-B: {46, 56, 66, 64, 65, 55} מתוך 36 אפשרויות סה"כ ולכן  $P(B) = \frac{6}{36} = \frac{1}{6}$

❖  $P(A) = \frac{1}{6}$

❖ לפי A לבד מרחב המדגם הוא: {61, 62, 63, 64, 65, 66}

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \left[ = \frac{\frac{3}{36}}{\frac{1}{6}} = \frac{1}{2} \right]$$

❖  $P(B|A) = \frac{3}{6} = \frac{1}{2}$



# דוגמה 3 – הסתברות מותנית

מרחב המדגם:

{בן, בת}

{בת, בן}

{בת, בת}

{בן, בן}

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$$

❖ נתונה משפחה לה שני ילדים

❖ נתון לנו שאחד הילדים הוא בן

❖ מהי ההסתברות ששני הילדים הם בנים ?

❖ א. שני בנים

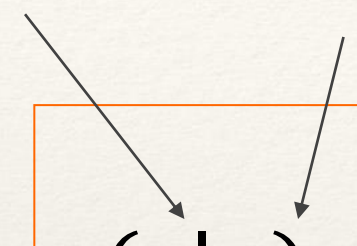
❖ ב. יש לפחות בן אחד





# Bayes' Rule

Class      observation


$$p(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$P(c)$  – **prior** probability of class  $c$  before any vector is seen

$P(x|c)$  – **likelihood** of the observed data if the class is  $c$

$P(x)$  – **evidence** probability of the data

$P(c|x)$  – **posterior** Probability of class  $c$  after the data is seen

# דוגמה – קניית מחשב

- צופה מתבונן על חנות מחשבים ומגיע ל"אבחנות" הבאות על-פני קבוצת אימון גדולה.
  - 40% מבין הנכנסים לחנות קונים מחשב.
  - מבין אלו הקונים מחשב – 50% לבושים בחליפה ו-50% אינם לבושים בחליפה.
  - מבין אלו שלא קנו מחשב – כולם לבושים בחליפה.
- שאלה: הצופה רואה כעת "אדם לבוש בחליפה נכנס לחנות" – צריך לתת חיזוי האם אדם זה יקנה מחשב או לא.
- דרך הפתרון:
- אנו צריכים לחשב את ההסתברויות הבאות:
  - הסתברות מותנית ש"אדם יקנה מחשב" בהינתן ש"לבוש בחליפה"
  - הסתברות מותנית ש"אדם לא יקנה מחשב" בהינתן ש"לבוש בחליפה"
  - ניתן חיזוי לפי ההסתברות הגבוהה מבין השתיים.



# דוגמה – קניית מחשב – פתרון על ידי כלל בייס

❖ נגדיר את המאורעות הבאים:

❖  $A$  – אדם לבוש בחליפה

❖  $B$  – קונה מחשב

לבוש בחליפה

קונה מחשב

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{0.5 \times 0.4}{0.4 \times 0.5 + 0.6 \times 1} = 0.25$$

לא קונה מחשב

לבוש בחליפה

$$P(\bar{B}|A) = \frac{P(A|\bar{B})P(\bar{B})}{P(A)} = \frac{1 \times 0.6}{0.4 \times 0.5 + 0.6 \times 1} = 0.75$$

מכיוון שקיבלנו הסתברות גבוהה יותר שלא יקנה מחשב – נעדיף  
את ההערכה הזו וזה החיזוי שלנו

# Bayes classifier in a nutshell

1. Learn the distribution over inputs for each value  $Y$ .
2. This gives  $P(X_1, X_2, \dots, X_m \mid Y=v_i)$ .
3. Estimate  $P(Y=v_i)$  as fraction of records with  $Y=v_i$ .
4. For a new prediction:

$$\begin{aligned} Y^{\text{predict}} &= \operatorname{argmax}_v P(Y = v \mid X_1 = u_1 \cdots X_m = u_m) \\ &= \operatorname{argmax}_v P(X_1 = u_1 \cdots X_m = u_m \mid Y = v) P(Y = v) \end{aligned}$$



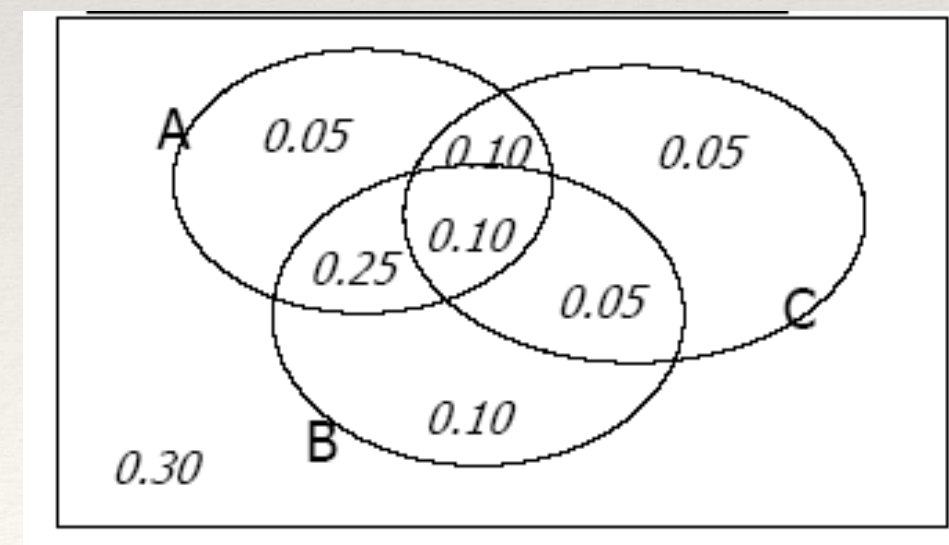
# The Joint Probability Table

Recipe for making a joint distribution of  $M$  variables:

1. Make a truth table listing all combinations of values of your variables (if there are  $M$  boolean variables then the table will have  $2^M$  rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

*Example:  $P(A, B, C)$*

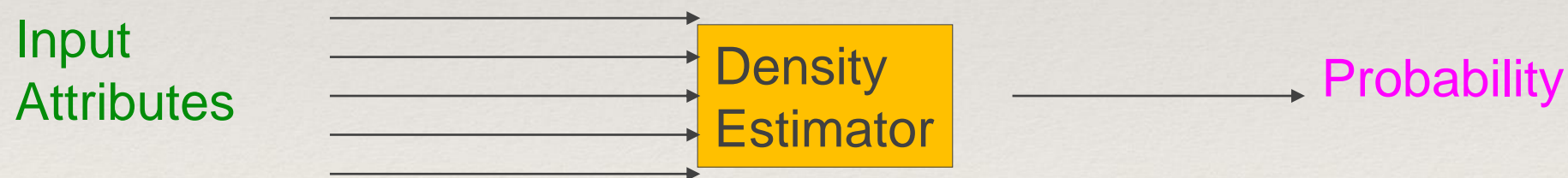
A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10





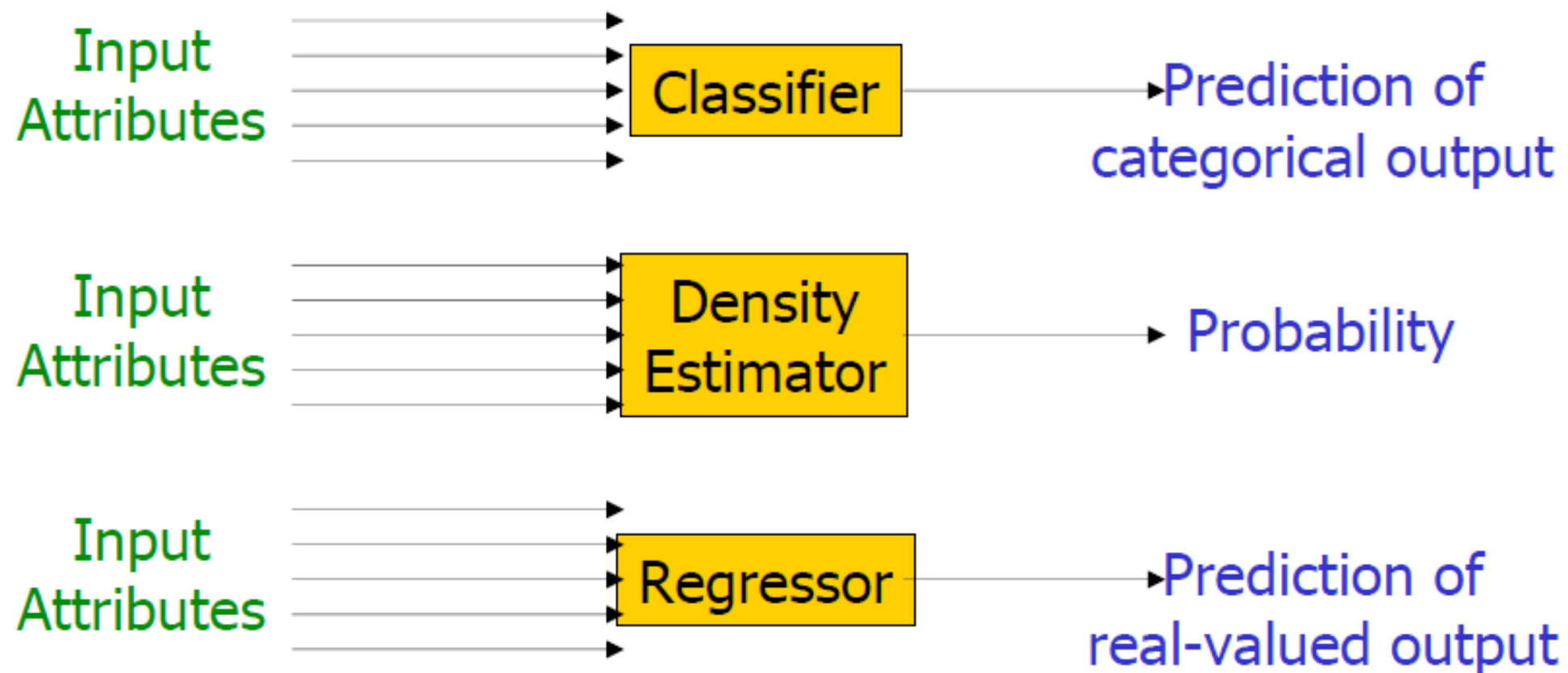
# Density Estimation

- ❖ Our Joint Probability Table (JPT) learner is our first example of something called Density Estimation
- ❖ A Density Estimator learns a mapping from a set of attributes to a Probability



# Density Estimation

- Compare it against the two other major kinds of models:





# Naïve density estimation

## Independence assumption

אבל... לימוד בצורה כזו של ה-Joint Distribution עלול להיות טריוויאלי ו"מסוכן" מדי בהיבט של overfitting

We need something which generalizes more usefully.

The **naïve model** generalizes strongly:

Assume that each attribute is distributed independently of any of the other attributes.

# Independently distributed data

- Let  $x[i]$  denote the  $i$ th field of record  $x$ .
- The independently distributed assumption says that for any  $i, v, u_1 u_2 \dots u_{i-1} u_{i+1} \dots u_M$

$$P(x[i] = v \mid x[1] = u_1, x[2] = u_2, \dots, x[i-1] = u_{i-1}, x[i+1] = u_{i+1}, \dots, x[M] = u_M) \\ = P(x[i] = v)$$

- Or in other words,  $x[i]$  is independent of  $\{x[1], x[2], \dots, x[i-1], x[i+1], \dots, x[M]\}$
- This is often written as

$$x[i] \perp \{x[1], x[2], \dots, x[i-1], x[i+1], \dots, x[M]\}$$



# A note about independent

- Assume A and B are Boolean Random Variables. Then

“A and B are independent”

if and only if

$$P(A|B) = P(A)$$

- “A and B are independent” is often notated as

$$A \perp B \longrightarrow$$

As vectors are  
orthogonal

# Independence Theorem

- Assume  $P(A|B) = P(A)$
- Then  $P(A \wedge B) =$

$$\begin{aligned} P(A \wedge B) &= \\ P(A|B) \times P(B) &= \\ P(A) \times P(B) \end{aligned}$$

$$= P(A) P(B)$$

- Assume  $P(A|B) = P(A)$
- Then  $P(B|A) =$

$$= P(B)$$



# Independence Theorem

- Assume  $P(A|B) = P(A)$
- Then  $P(\sim A|B) =$

$$= P(\sim A)$$

- Assume  $P(A|B) = P(A)$
- Then  $P(A|\sim B) =$

$$= P(A)$$

# Multivalued Independence

For multivalued Random Variables  $A$  and  $B$ ,

$$A \perp B$$

if and only if

$$\forall u, v : P(A = u \mid B = v) = P(A = u)$$

from which you can then prove things like...

$$\forall u, v : P(A = u \wedge B = v) = P(A = u)P(B = v)$$

$$\forall u, v : P(B = v \mid A = u) = P(B = v)$$



# Naïve distribution general case

- Suppose  $x[1], x[2], \dots, x[M]$  are independently distributed.

$$P(x[1] = u_1, x[2] = u_2, \dots, x[M] = u_M) = \prod_{k=1}^M P(x[k] = u_k)$$

- So if we have a Naïve Distribution we can construct any row of the implied Joint Distribution on demand.
- So we can do any inference
- But how do we learn a Naïve Density Estimator?

# Naïve Bayes Classifier

---

- Using Bayes rule:

$$P(c|x_1, x_2, \dots, x_D) = \frac{P(c)P(x_1, x_2, \dots, x_D|c)}{P(x_1, x_2, \dots, x_D)}$$

- Select the feature set such that each feature  $x_i$  is **independent** of every other feature  $x_j$ .

$$P(x_1, x_2, \dots, x_D|c) = P(x_1|c)P(x_2|c)P(x_3|c)\dots P(x_D|c) = \prod_{i=1}^D P(x_i|c)$$



# How to build a Bayes classifier

- Assume you want to predict output  $Y$  which has arity  $n_Y$  and values  $V_1, V_2, \dots, V_{n_Y}$
- Assume there are  $m$  input attributes called  $X_1, X_2, \dots, X_m$
- Break dataset into  $n_Y$  smaller datasets called  $DS_1, DS_2, \dots, DS_{n_Y}$
- Define  $DS_i =$  Records in which  $Y = V_i$
- For each  $DS_i$ , learn Density Estimator  $M_i$  to model the input distribution among the  $Y = V_i$  records.
- $M_i$  estimates  $P(X_1, X_2, \dots, X_m \mid Y = V_i)$
- Idea: When a new set of input values ( $X_1 = u_1, X_2 = u_2, \dots, X_m = u_m$ ) come along to be evaluated predict the value of  $Y$  that makes  $P(X_1, X_2, \dots, X_m \mid Y = V_i)$  most likely

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} P(Y = v \mid X_1 = u_1 \cdots X_m = u_m)$$

# דוגמה – קניית מחשב

נכנס גבר מעל גיל 30  
לחנוות. האם יקנה מחשב  
או לא?

קנה או לא	גבר / אשה	מעל 30
קנה	גבר	כן
קנה	גבר	כן
קנה	גבר	לא
קנה	אשה	לא
לא קנה	אשה	כן
לא קנה	אשה	כן
לא קנה	אשה	כן
לא קנה	גבר	כן
לא קנה	גבר	כן
לא קנה	גבר	כן



# דוגמה – קניית מחשב – פתרון על ידי Naïve Bayes

נחשב קודם כל הסתברות  
אפריוורית של קניה או  
לא:

$$40\% = 4/10 \text{ קנו}$$
$$60\% = 6/10 \text{ לא קנו}$$

קנה או לא	גבר / אשה	מעל 30
קנה	גבר	כן
קנה	גבר	כן
קנה	גבר	לא
קנה	אשה	לא
לא קנה	אשה	כן
לא קנה	אשה	כן
לא קנה	אשה	כן
לא קנה	גבר	כן
לא קנה	גבר	כן
לא קנה	גבר	כן



# דוגמה – קניית מחשב – פתרון על ידי Naïve Bayes

נחשב את הערך של כל מאפיין  
בהנתן קנה או לא קנה.

קנה:

מעל 30:  $50\% = 2/4$

גבר:  $75\% = 3/4$

לא קנה:

מעל 30:  $100\% = 6/6$

גבר:  $50\% = 3/6$

קנה או לא	גבר / אשה	מעל 30
קנה	גבר	כן
קנה	גבר	כן
קנה	גבר	לא
קנה	אשה	לא
לא קנה	אשה	כן
לא קנה	אשה	כן
לא קנה	אשה	כן
לא קנה	גבר	כן
לא קנה	גבר	כן
לא קנה	גבר	כן



# דוגמה – קנית מחשב – פתרון על ידי Naïve Bayes

נחשב את הערך של כל מאפיין בהנתן קנה או לא קנה.

קנה:  
מעל 30:  $50\% = \frac{2}{4}$   
גבר:  $75\% = \frac{3}{4}$

לא קנה:  
מעל 30:  $100\% = \frac{6}{6}$   
גבר:  $50\% = \frac{3}{6}$

נחזה עבור גבר מעל 30:  
קנה:  $0.15 = 0.75 * 0.5 * 0.4$   
לא קנה:  $0.3 = 1 * 0.5 * 0.6$

מסקנה: יותר סביר שלא יקנה

קנה או לא	גבר / אשה	מעל 30
קנה	גבר	כן
קנה	גבר	כן
קנה	גבר	לא
קנה	אשה	לא
לא קנה	אשה	כן
לא קנה	אשה	כן
לא קנה	אשה	כן
לא קנה	גבר	כן
לא קנה	גבר	כן
לא קנה	גבר	כן

# Terminology

- Maximum Likelihood estimation (MLE)

Choose value that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta)$$

- Maximum *a posteriori* (MAP) estimation

Choose value that is most probable given observed data and prior belief

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg \max_{\theta} P(\theta|D) \\ &= \arg \max_{\theta} P(D|\theta)P(\theta)\end{aligned}$$



# Bayes classifier in a nutshell

1. Learn the distribution over inputs for each value  $Y$ .
2. This gives  $P(X_1, X_2, \dots, X_m \mid Y=v_i)$ .
3. Estimate  $P(Y=v_i)$  as fraction of records with  $Y=v_i$ .
4. For a new prediction:

$$\begin{aligned} Y^{\text{predict}} &= \operatorname{argmax}_v P(Y = v \mid X_1 = u_1 \cdots X_m = u_m) \\ &= \operatorname{argmax}_v P(X_1 = u_1 \cdots X_m = u_m \mid Y = v) P(Y = v) \end{aligned}$$

# Naïve Bayes classifier

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} P(X_1 = u_1 \cdots X_m = u_m \mid Y = v) P(Y = v)$$

In the case of the naive Bayes Classifier this can be simplified:

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} P(Y = v) \prod_{j=1}^{n_Y} P(X_j = u_j \mid Y = v)$$

$$\begin{aligned} P(x_1, x_2, \dots, x_D \mid Y = v) &= \\ P(x_1 \mid Y = v) P(x_2 \mid Y = v) P(x_3 \mid Y = v) \dots P(x_m \mid Y = v) &= \\ \prod_{i=1}^m P(x_i \mid Y = v) \end{aligned}$$



# Bayes classifier Pseudo Code

- **Train Naïve Bayes** (given data for X and Y)

for each\* value  $y_k$

estimate  $\pi_k \equiv P(Y = y_k)$

for each\* value  $x_{ij}$  of each attribute  $X_i$

estimate  $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

- **Classify** ( $X^{new}$ )

$$Y^{new} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg \max_{y_k} \pi_k \prod_i \theta_{ijk}$$

\* probabilities must sum to 1, so need estimate only n-1 of these...

# Underflow Prevention

- ❖ הכפלה של הרבה איברים שכולם בין 0 ל-1 (הסתברויות) יכול להוביל אותנו ל-underflow
- ❖ מה נעשה כאשר יש לנו מאות מאפיינים ?
- ❖  $\log(xy) = \log(x) + \log(y) \rightarrow$  summing logs of probabilities rather than multiplying probabilities.
- ❖ Class with highest final un-normalized log probability score is still the most probable.

$$Y^{\text{predict}} = \underset{v}{\operatorname{argmax}} \left( \log P(Y = v) + \sum_{j=1}^{n_Y} \log P(X_j = u_j | Y = v) \right)$$



# Log computation

“Yes”

$$\left[ \log P(\text{"yes"}) + \sum_{i \in \text{תכונות}} \log P(x_i | \text{"yes"}) \right] =$$
$$\log(0.75 \cdot 0.5 \cdot 0.4) =$$
$$\log(0.75) + \log(0.5) + \log(0.4) \approx -0.823$$



“No”

$$\left[ \log P(\text{"no"}) + \sum_{i \in \text{תכונות}} \log P(x_i | \text{"no"}) \right] =$$
$$\log(0.75 \cdot 0.5 \cdot 0.6) =$$
$$\log(0.75) + \log(0.5) + \log(0.6) \approx -0.647$$

מסקנה: יותר סביר שלא יקנה



---

# Why to use Bayesian classifier?

---

- ❖ Combine Prior knowledge and observed data
- ❖ It is a generative (model based) approach – outputs a probability distribution over all classes
- ❖ Tends to work well despite strong assumption of conditional independence.
- ❖ Does not perform any search of the hypothesis space.
- ❖ Easy to implement
- ❖ Be careful when multiple dependent attributes!



# שערוך המודל - תזכורת

Confusion matrix:

	Predicted Yes	Predicted No
Actual Yes	True Positive (TP)	False Negative (FN)
Actual No	False Positive (FP)	True Negative (TN)

$$\text{accuracy} = \frac{\#correct\ predictions = \#TP + \#TN}{\#test\ instances = \#TP + \#TN + \#FP + \#FN}$$

$$\text{Error} = 1 - \text{accuracy} = \frac{\#incorrect\ predictions = \#FP + \#FN}{\#test\ instances = \#TP + \#TN + \#FP + \#FN}$$

# Precision and Recall

- ❖ Precision = How accurate is the classifier in labelling an example as Positive
- ❖ Recall – What is the coverage on the positive examples

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

	Predicted Yes	Predicted No
Actual Yes	9	2
Actual No	3	16

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{9}{9 + 3} = 0.75$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{9}{9 + 2} = 0.82$$



---

# נאיב בייס - סיכום

---

## ❖ יתרונות:

- ❖ קל להבנה/"למידה"
- ❖ קל למימוש
- ❖ אינטואיטיבי ומבוסס על סטטיסטיקה וסבירות
- ❖ קל לשימוש/"הפעלת ה"מכונה" על נתונים חדשים"
- ❖ זול (יחסית) חישובית

## ❖ חסרונות:

- ❖ להזהר מ-underflow
- ❖ זכרו את הנחת אי-התלות – במידה ולא נכונה יש לחשוב שוב..

## ❖ השלמות בשיעור הבא:

- ❖ smoothing
- ❖ מאפיינים רציפים

נראה בשבוע הבא ☺