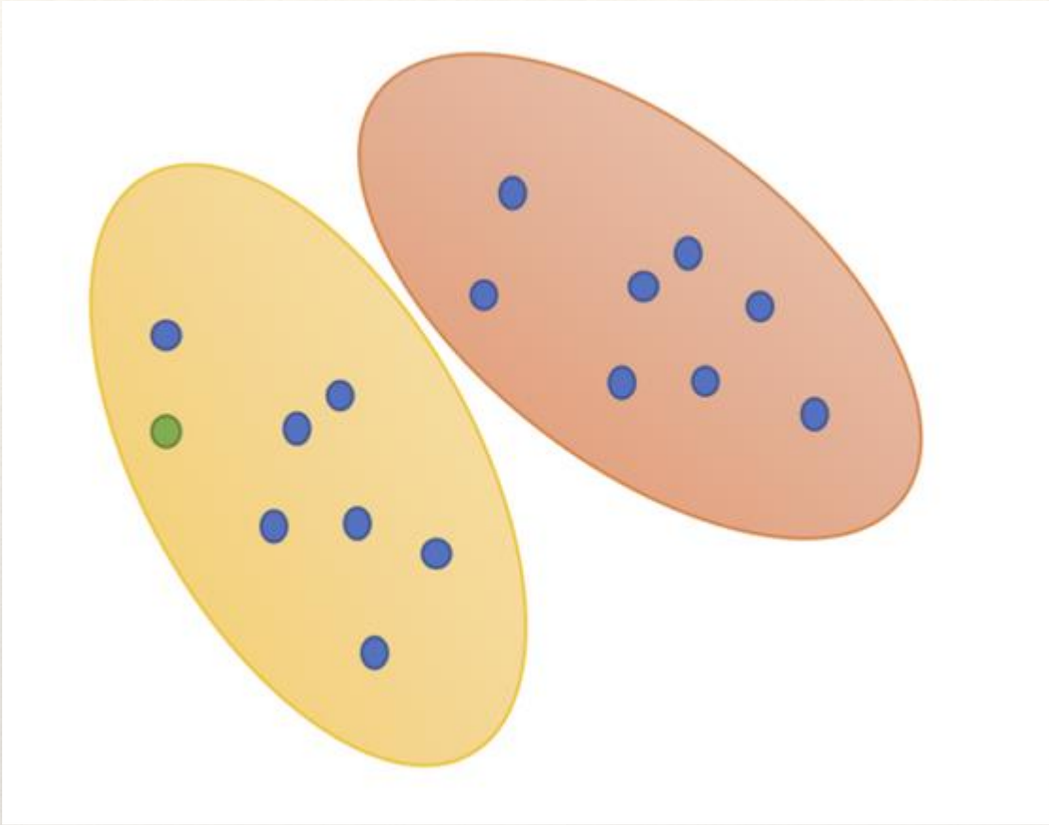
Machine learning

Naïve Bayes – Part 2
* Smoothing
* Continuous features

Lecture V

פיתוח:
ד"ר יהונתן שלר
משה פרידמן

חלק ב'- Naïve Bayes



- בעיית שכיחות האפס
- Naïve Bayes עבור מ"מ רציפים

Bayes classifier in a nutshell

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

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

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

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:

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

$$P(x_1, x_2, \ldots\ldots x_D | Y = v) = P(x_1|Y = v)P(x_2|Y = v)P(x_3|Y = v)\ldots P(x_m|Y = v) = \prod_{i=1}^{m} P(x_i|Y = v)$$

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...

Smoothing

	Age	Hobby	Weather	Buy Computer?
1	Young	Sport	Cold	“Yes”
2	Young	Sport	Cold	“Yes”
3	Young	Sport	Cold	“Yes”
4	Old	Sport	Hot	“Yes”
5	Old	Sport	Hot	“Yes”
6	Old	Paint	Hot	“No”
7	Old	Paint	Cold	“Yes”
8	Old	Paint	Cold	“Yes”
9	Young	Paint	Hot	“No”
10	Young	Sport	Hot	“No”

- נסתכל בדוגמא הבאה:
- ננסה לחזות את המקרה בו (old, sport, cold)

$$P(Weather = "cold" | "no") = P(Weather = "cold" \wedge "no")/P("no") = \frac{0}{3} = 0$$

- P(“no” | old, sport, cold) = 0

Smoothing solution

- Probability estimates are adjusted or *smoothed*.
- Assumes that each feature is given a prior probability, *p*, that is assumed to have been previously observed in a "virtual" sample of size *m*.
- Usually, in the binary case *p* is simply assumed to be 0.5

$$P(X = x|Y = y) = \frac{n_c + mp}{n + m}$$

- n = number of training examples for which Y = y
- n_c =number of examples where *X*=*x* and Y=y
- p = a prior estimation for P(X=x|Y=y)
- m = the equivalent sample size

לאור מה שראינו בדוגמה

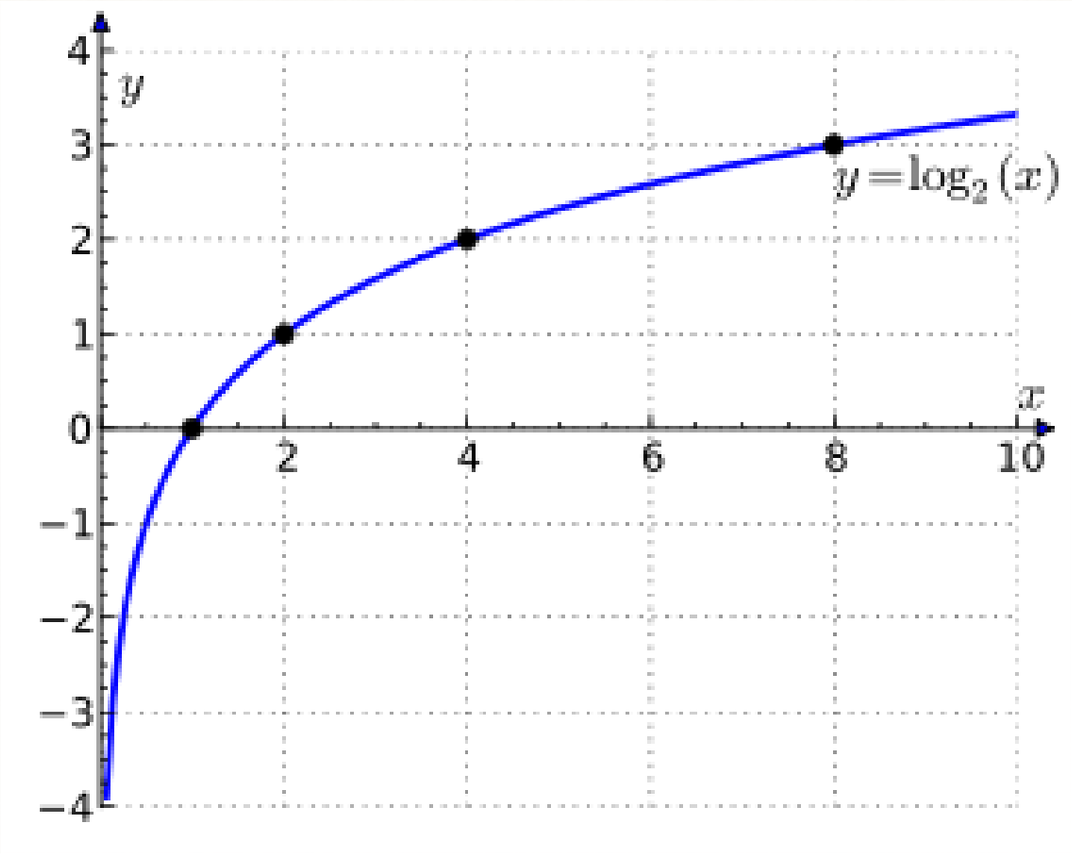
$$P(Weather = "cold" | "no") =$$
$$P(Weather = "cold" \wedge "no") / P("no") =$$

$$P(X = x | Y = y) = \frac{n_c + mp}{n + m}$$

$$n = 3$$
$$n_c = 0$$
$$p = 0.5$$
$$m = 4$$

$$P(Weather = "cold" | "no") =$$
$$P(Weather = "cold" \wedge "no") / P("no") =$$
$$\frac{0 + 4 \times 0.5}{3 + 4} = \frac{2}{7}$$

פונקצית לוג



תכונות:

- פונקציית לוג של שברים תהיה שלילית, אך היא שומרת על הסדר, והיא גם מונוטונית עולה.
- log(x*y)=log(x)+log(y).

לכן, נרצה לעבוד עם חיבור לוגים, במקום מכפלת שברים (של הסתברויות).

מדוע?

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) → summing logs of probabilities rather than multiplying probabilities.
- Class with highest final un-normalized log probability score is still the most probable.

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

תרגיל – השלימו את החישוב – בעזרת log

"Yes"

$$\left[\log P\left("yes"\right) + \sum_{i \in \text{תכונות}} \log P\left(x_i | "yes"\right)\right] =$$

"No"

$$\left[\log P\left("no"\right) + \sum_{i \in \text{תכונות}} \log P\left(x_i | "no"\right)\right] =$$

חוק בייס והנחת חוסר התלות - תזכורת

Class prior

Likelihood probability

חוק בייס:

$$P(c \mid x_1, x_2, \ldots\ldots x_D) = \frac{P(c)P(x_1, x_2, \ldots\ldots x_D \mid c)}{P(x_1, x_2, \ldots\ldots x_D)}$$

a posteriori probability

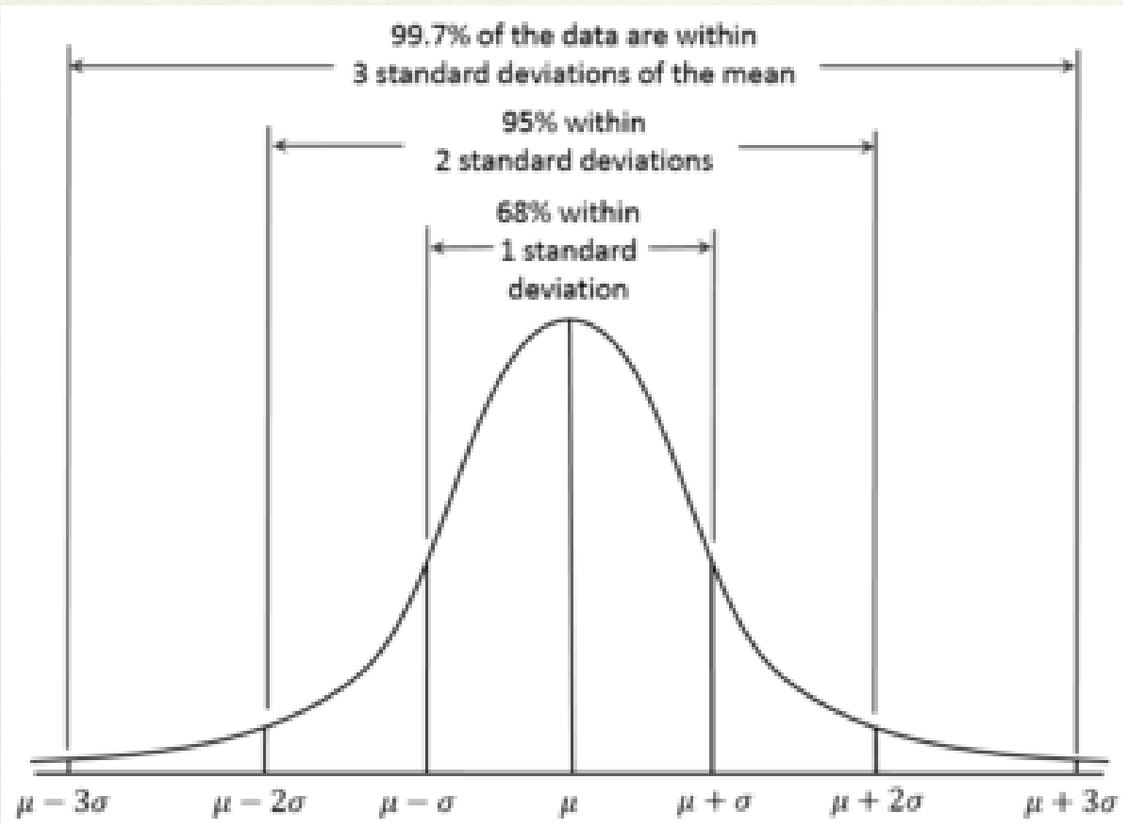
Feature (predictor) priors

בגלל הנחת חוסר התלות בין המאפיינים:

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

אבל, איך יודעים לחשב
עבור מ"מ רציף?

התפלגות נורמלית – פונקצית צפיפות



הערה חשובה: אנחנו נחשב סטיית תקן במדגם ואת ה-t-val

התפלגות נורמלית: נקראת גם גאוסיאן (Gaussian) או עקומת פעמון.

- פונקציית צפיפות סמטרית.

התפלגות z: תת קבוצה של התפלגות נורמלית בו התוחלת/הממוצע=0 וסטיית התקן=1.

- כל התפלגות נורמלית ניתן להפוך להתפלגות z
- **מהתפלגות נורמלית להתפלגות z**: z-val= (x-μ)/σ

תוחלת/ממוצע ערך מאורע

פונקצית צפיפות:

$$\frac{1}{\sigma\sqrt{2\pi}}\ \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

סטיית תקן

What if features are continuous?

Gaussian Naïve Bayes (GNB):

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \quad e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

Different mean and variance for each class K and each attribute i

Sometimes assume variance

- is independent of Y (i.e., σ_i),
- or independent of X_i (i.e., σ_k)
- or both (i.e., σ)

Example: P(A, B, C)

A	B	C	Class
0.1	0.3	0.04	1
0.12	0.4	0.99	1
0.13	0.9	0.01	1
0.44	0.86	0.93	1
0.67	0.45	0.34	0
0.77	0.55	0.75	0
.88	.79	0.09	0
.89	.82	.81	0

Continuous Params Estimation

$$
\begin{aligned}
h_{NB}(\mathbf{x}) &= \arg\max_y P(y) \prod_i P(X_i = x_i | y) \\
&\approx \arg\max_k \widehat{P}(Y = k) \prod_i \mathcal{N}(\widehat{\mu}_{ik}, \widehat{\sigma}_{ik})
\end{aligned}
$$

$$\boxed{\widehat{\mu}_{MLE} = \frac{1}{N} \sum_{j=1}^{N} x_j}$$

$$\boxed{\widehat{\sigma}^2_{unbiased} = \frac{1}{N-1} \sum_{j=1}^{N} (x_j - \widehat{\mu})^2}$$

Gaussian NB Pseudo Code

- Train Naïve Bayes (examples)

  for each value y_k

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

  for each attribute X_i estimate $P(X_i|Y = y_k)$

  - class conditional mean μ_{ik}, standard deviation σ_{ik}

- Classify (X^{new})

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

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \mathcal{N}(X_i^{new}; \mu_{ik}, \sigma_{ik})$$

* probabilities must sum to 1, so need estimate only n-1 parameters...

נאיב בייס - סיכום

- <u>יתרונות:</u>
  - קל להבנה/"למידה"
  - קל למימוש
  - אינטואיטיבי ומבוסס על סטטיסטיקה וסבירות
  - קל לשימוש/"הפעלת ה"מכונה" על נתונים חדשים"
  - זול (יחסית) חישובית
- <u>חסרונות:</u>
  - להזהר מ –underflow
  - זכרו את הנחת אי-התלות – במידה ולא נכונה יש לחשוב שוב..

נתראה בתרגול ☺