Introduction to Machine Learning Probability-based Learning

Prof. Chang-Chieh Cheng
Information Technology Service Center
National Chiao Tung University

Basic Statistics Methods

- Arithmetic mean, average

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

  - Example:

BMI
22.1
18.3
24.8
31.5
18.4

$$\bar{x} = \frac{22.1 + 18.3 + 24.8 + 31.5 + 18.4}{5} = 23.02$$

Basic Statistics Methods

- Median
  - The value separating the higher half from the lower half of a data sample
  - Example:
    - The median is 22.1

BMI
18.3
18.4
22.1
24.8
31.5

    - The median is $\frac{22.1+24.8}{2} = 23.45$

BMI
18.3
18.4
22.1
24.8
31.5
32.2

Basic Statistics Methods

- Standard Deviation

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Example:

BMI
22.1
18.3
24.8
31.5
18.4

$s = 5.467$

Why the denominator is n – 1 rather than n?

Because n samples only have n – 1 independent differences

Basic Statistics Methods

- Geometric mean

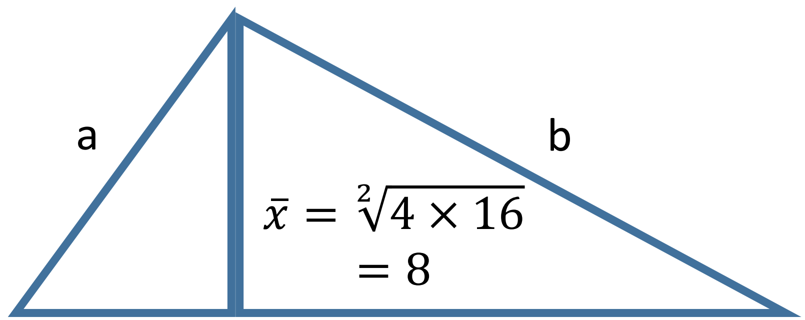
$$\bar{x} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

  - Example:

$$\bar{x} = \sqrt[5]{22.1 \times 18.3 \times 24.8 \times 31.5 \times 18.4}$$
$$= 22.53641$$

BMI
22.1
18.3
24.8
31.5
18.4

-



$x_1 = 4$ $\quad$ $x_2 = 16$

$$(x_1 + x_2)^2 = a^2 + b^2$$
$${x_1}^2 + 2x_1x_2 + {x_2}^2 = \left(\bar{x}^2 + {x_1}^2\right) + \left(\bar{x}^2 + {x_2}^2\right)$$
$$\bar{x} = \sqrt{x_1x_2}$$

Basic Statistics Methods

- Harmonic mean

$$\bar{x} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

  - Example:

BMI
22.1
18.3
24.8
31.5
18.4

$$\bar{x} = \frac{5}{\frac{1}{22.1} + \frac{1}{18.3} + \frac{1}{24.8} + \frac{1}{31.5} + \frac{1}{18.4}}$$

$$= 22.09358$$

Probability

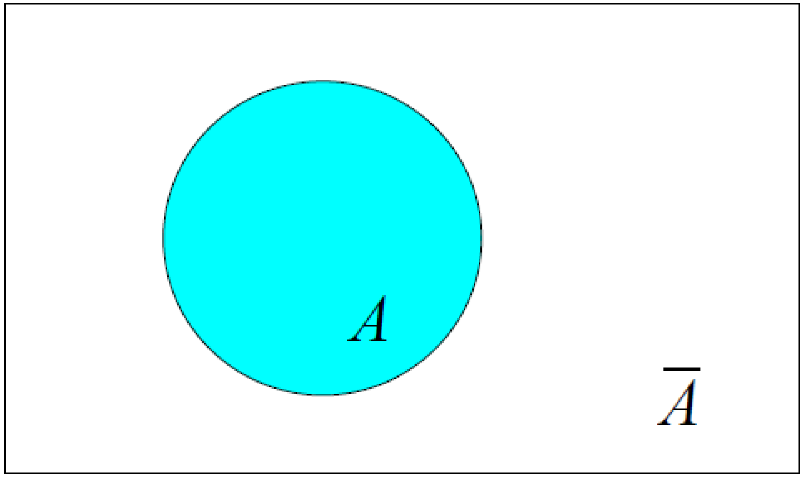
- Probability is the measure of the likelihood that an event will occur.
- The probability of an event A in a finite sample spaces
  - P(A) = the number of event A occurred / the number of total samples
  - What is the probability of headache in the following ten patients
    - P(Headache) = 7 / 10 = 0.7

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

Probability

- The complement of an event
  - What is the probability of non-headache in the ten patients
    - P(non-Headache) = 3 / 10 = 0.3 = 1.0 – P(Headache)

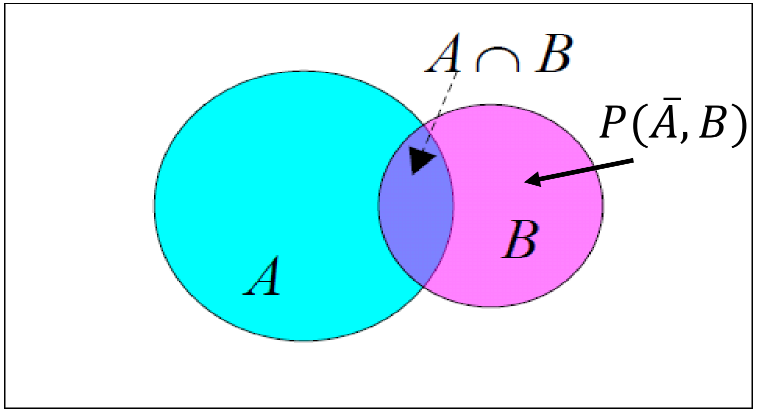


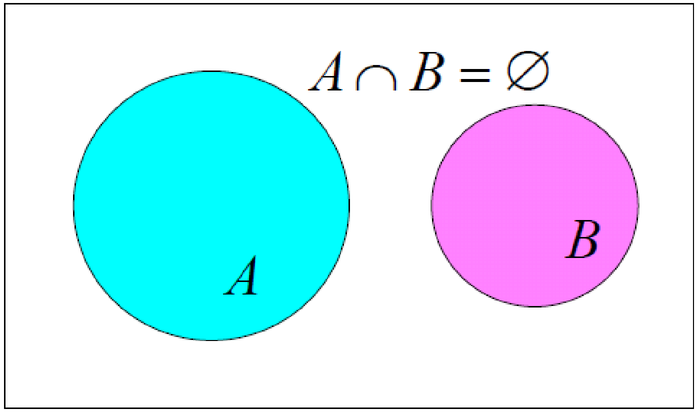
Probability

- $0.0 \leq P(A) \leq 1.0$
- Probability distribution
  - Given a random variable *X* with *n* events x_1, x_2, …, x_n,
    - $\sum_{i=1}^{n} P(X = x_i) = 1.0$
- EX:
  - Given four weather types: sunny, cloudy, shower, and rain
  - The probabilities for all weather in July 2017 are *P*(sunny), *P*(cloudy), *P*(shower), and *P*(rain) respectively.
  - *P*(sunny) + *P*(cloudy) + *P*(shower) + *P*(rain) = 1.0

Probability

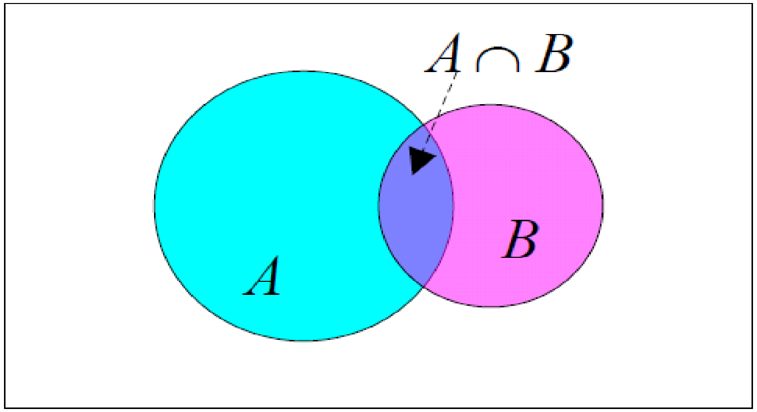
- Joint probability
  - $P(A, B)$ or $P(A \cap B)$ or $P(A \text{ and } B)$
  - if A and B are independent events: $P(A \cap B) = P(A)\,P(B)$
  - $P(A, B) = P(B, A)$
  - $P(B) = P(A, B) + P(\bar{A}, B)$

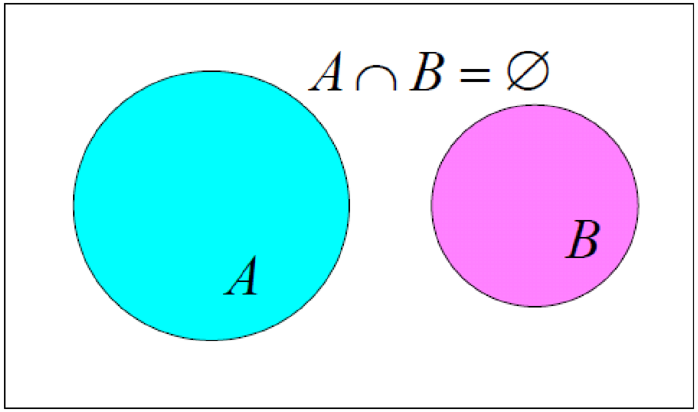




Probability

- $P(A \text{ or } B) = P(A \cup B) = \mathrm{P(A)} + \mathrm{P(B)} - P(A \cap B)$

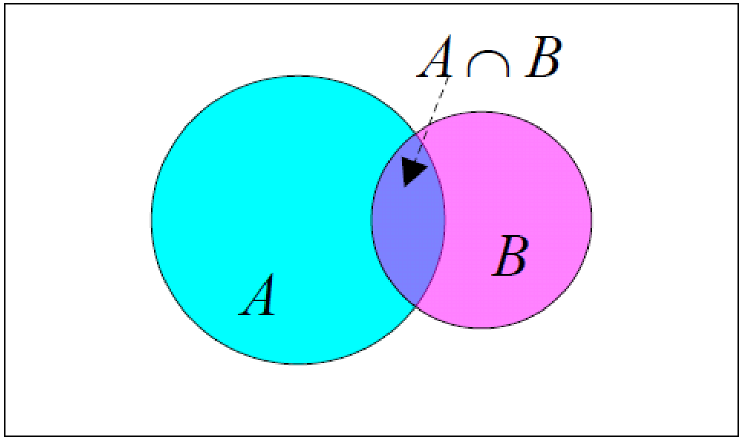




Probability

- Conditional probability
  - $P(A|B)$: the probability of event A under event B occurred

    - $P(A|B) = \frac{P(A,B)}{P(B)}$
    - $P(A|B)P(B) = P(A,B)$



Probability

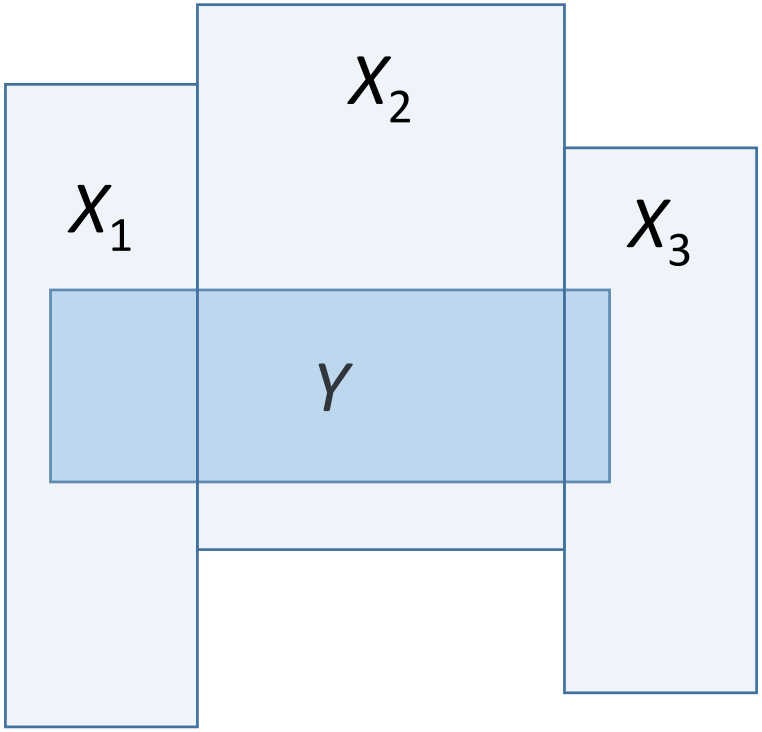
- Conditional probability
  - Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

  - $P(A)$, $P(B)$, $P(B|A)$: prior probability, they are already known.
  - $P(A|B)$: post probability, calculated form priors.
  - Proof:
    - $P(B|A) = \frac{P(B,A)}{P(A)}$
    - ➔$P(B|A)P(A) = P(B, A)$
    - ➔$\frac{P(B|A)P(A)}{P(B)} = \frac{P(B,A)}{P(B)} = \frac{P(A,B)}{P(B)} = P(A|B)$

Theorem of Total Probability

- $P(Y) = \sum_{i=1}^{n} P(Y|X_i)P(X_i)$
- where $\{X_i : i = 1,2,3 \ldots\}$ is a set of pairwise disjoint events whose union is the entire sample space



Bayes Theorem Example 1

- Assuming that a school has 60% boys and 40% girls.
- The number of girls wearing pants equals to the number of girls wearing skirts.
- All boy are wearing pants.
- What is the probability of that when you saw a person wearing pants and that person is a girl in the school?
- Let *A* is the event of girl, *B* is the event of pant wearing ➔ The answer is *P*(*A*|*B*)
  - $P(A) = 0.4 \rightarrow P(\bar{A}) = 1 - P(A) = 0.6$, which is the probability of boy
  - $P(B|A) = 0.5$, which is the probability of a girl wearing pants
  - $P(B|\bar{A}) = 1.0$, which is the probability of a boy wearing pants
  - $P(B) = P(B, A) + P(B, \bar{A})$
    $= P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) = 0.5 \times 0.4 + 1.0 \times 0.6 = 0.8$
  - $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.5 \times 0.4}{0.8} = 0.25$

Bayes Theorem Example 2

- A doctor informs a patient that he has both bad news and good news.
- The bad news is that the patient has tested positive for a serious disease and that **the test is 99% accurate**
  - the probability is 0.99 ➔ testing positive when a patient has the disease.
  - the probability is 0.01 ➔ testing positive when a patient does not have the disease.
  - the probability is also 0.99 ➔ testing negative when a patient does not have the disease.
- The good news is that the disease is extremely rare, striking **only 1 in 10,000 people**.
- What is the actual probability that the patient has the disease?
- Why is the rarity of the disease good news given that the patient has tested positive for it?

Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

Bayes Theorem Example 2

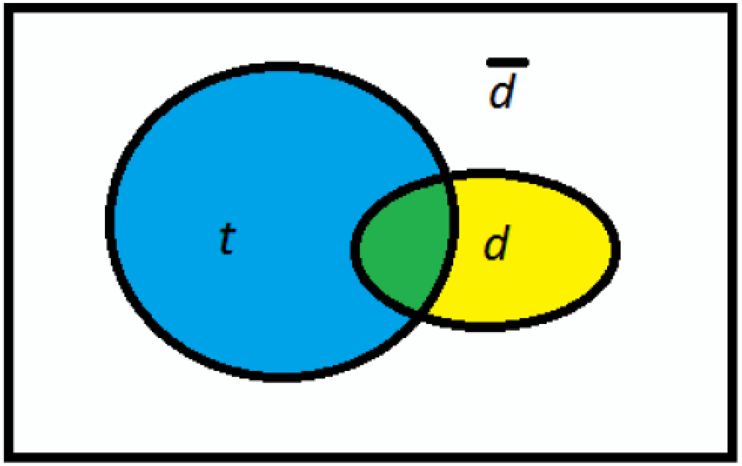
- d: a patient has the disease
- t: the test is positive
- $P(d|t) = \frac{P(t|d)P(d)}{P(t)}$
- $P(t) = P(t, d) + P(t, \bar{d})$

$$= P(t|d)P(d) + P(t|\bar{d})P(\bar{d})$$

$$= (0.99 \times 0.0001) + (0.01 \times 0.9999)$$

$$= 0.0101$$

- $P(d|t) = \frac{0.99 \times 0.0001}{0.0101} = 0.0098$



Generalized Bayes' Theorem

- Given *m* random variables, $\{X_1, X_2, \dots, X_m\}$

$$P(Y \mid X_1, X_2, \dots, X_m) = \frac{P(X_1, X_2, \dots, X_m \mid Y) P(Y)}{P(X_1, X_2, \dots, X_m)}$$

Probability-based Learning Model

- Given a query **q** with *m* features
  - $\mathbf{q} = \{X_1, X_2, \dots, X_m\}$
- And there are *n* target levels
  - $\mathbf{T} = \{Y_1, Y_2, \dots, Y_n\}$
- We want to predict which target level **q** should belong to.
  - $M(\mathbf{q}) = \underset{Y \in \mathbf{T}}{\operatorname{argmax}} P(Y \mid X_1, X_2, \dots, X_m)$

Probability-based Learning Model

- Example:

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

- Whether MENINGITIS is true if
  q = {HEADACHE = true, FEVER = false, VOMITING = true}

Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

Probability-based Learning Model

- According the generalized Bayes' theorem
  - $P(Y \mid X_1, X_2, \ldots, X_m) = \frac{P(X_1, X_2, \ldots, X_m \mid Y) P(Y)}{P(X_1, X_2, \ldots, X_m)}$
- Let Y_1 be MENINGITIS = true
  - $P(Y_1) = \frac{3}{10} = 0.3$
- Then Y_2 is MENINGITIS = false
  - $P(Y_2) = 1.0 - P(Y_1) = 0.7$
- And the probability of **q** in the training data set
  - $P(\mathbf{q}) = P(X_1, X_2, \ldots, X_m) = \frac{6}{10} = 0.6$
- So, $P(\mathbf{q}|Y) = P(X_1, X_2, \ldots, X_m \mid Y) =$?

Chain Rule

- Given *m* random variables, $\{X_1, X_2, \ldots, X_m\}$
  - $P(X_1, X_2, \ldots, X_m)$

    $= P(X_1)\, P(X_2|X_1)\ \ldots\ P(X_m|X_{m-1}, \ldots, X_2, X_1)$

    $= P(X_1) \prod_{i=2}^{m} P(X_i|X_{i-1}, \ldots, X_2, X_1)$

  - Proof:
    - $P(X_1, X_2) = P(X_1|X_2)P(X_2)$
    - $P(X_1, X_2, X_3) = P(X_1|X_2, X_3)P(X_2, X_3)$
      $= P(X_1|X_2, X_3)P(X_2|X_3)P(X_3)$
    - …

Chain Rule

- $P(X_1, X_2, \dots, X_m \mid Y) = \frac{P(Y, X_1, X_2, \dots, X_m)}{P(Y)}$

  - Or we can apply the chain rule

$$= \frac{P(Y)P(X_1|Y)\,P(X_2|X_1,Y)\ \dots\ P(X_m|X_{m-1},\dots,X_2,X_1,Y)}{P(Y)}$$

$$= P(X_1|Y)\,P(X_2|X_1,Y)\ \dots\ P(X_m|X_{m-1},\dots,X_2,X_1,Y)$$

Probability-based Learning Model

- **q** = {HEADACHE = true, FEVER = false, VOMITING = true}
- $P(\mathbf{q}|Y_1) = P(H, \bar{F}, V\,|Y_1)$

$$= P(H|Y_1) \times P(\bar{F}|H, Y_1) \times P(V\,|H, \bar{F}, Y_1)$$

$$= \frac{2}{3}$$

- $P(\mathbf{q}|Y_2) = P(H, \bar{F}, V\,|Y_1)$

$$= P(H|Y_2) \times P(\bar{F}|H, Y_2) \times P(V\,|H, \bar{F}, Y_2)$$

$$= \frac{4}{7}$$

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

Probability-based Learning Model

- Then,
  - $P(Y_1|\mathbf{q}) = \frac{P(\mathbf{q}|Y_1)P(Y_1)}{P(\mathbf{q})} = \frac{\frac{2}{3}\times\frac{3}{10}}{\frac{6}{10}} = \frac{1}{3} = 0.3333$
  - $P(Y_2|\mathbf{q}) = \frac{P(\mathbf{q}|Y_2)P(Y_2)}{P(\mathbf{q})} = \frac{\frac{4}{7}\times\frac{7}{10}}{\frac{6}{10}} = \frac{2}{3} = 0.6667$
- Therefore,
  - MENINGITIS = false if
    q = {HEADACHE = true, FEVER = false, VOMITING = true}

Probability-based Learning Model

- What if
  q = {HEADACHE = true, FEVER = true, VOMITING = true}
  - No such training data!
  - Data insufficient ➔**Model overfitting**
- **Overfitting**
  - The model is too complicated
  - The data is too simple for the model
  - Some exceptions will be considered
- **Underfitting**
  - The model is too simple
  - The data is too complicated for the model
- **Appropriate-fitting**
  - Few exceptions will be ignored

Independence

- Two events, X and Y, are independent if knowledge of Y has no effect on the probability of X.

$$P(X|Y) = P(X)$$

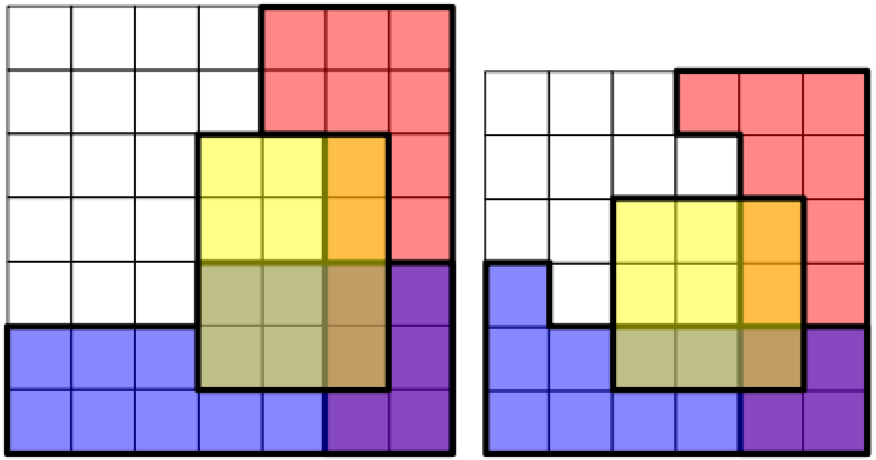
Then,

$$P(X,Y) = P(X|Y)P(Y) = P(X)P(Y)$$

Conditional Independence

- Two events *R* and *B* are conditionally independent given a third event *Y*,

$$P(R, B \mid Y) = P(R \mid Y) P(B \mid Y)$$



https://en.wikipedia.org/wiki/Conditional_independence

$$P(R, B \mid Y) = \frac{6}{12} \times \frac{4}{12} = \frac{2}{12}$$

$$P(R, B \mid Y) = \frac{3}{9} \times \frac{3}{9} = \frac{1}{9}$$

Conditional Independence

- Example 1.
  - Height (*H*) and vocabulary (*V*) are not independent
    - A taller kid could know more vocabulary than a shorter kid because the age of taller kid is larger then the shorter kid.
  - *H* and *V* are conditionally independent given a certain Age (*A*).
    - $P(H \mid A)$ and $P(V \mid A)$ are conditionally independent.
    - $P(H, V \mid A) = P(H \mid A)\, P(V \mid A)$
  - *H* and *V* are NOT conditionally independent given a gender (*G*).

Conditional Independence

- Example 2.
  - Lung cancer (*L*) and Smoking (*S*) are not independent
    - There are many people do smoking and have lung.
  - *L* and *S* are NOT conditionally independent given the condition of Regular Exercise (*E*).
    - $P(L = true \mid E = true)$ may still high if $P(S = true \mid E = true)$ is high.

  - *L* and *E* are not independent
    - Many people without lung cancer have a regular exercise.
  - *L* and *E* are conditionally independent given *S*.
    - $P(L \mid S)$ and $P(E \mid S)$ are conditionally independent.
    - $P(L, E \mid S) = P(L \mid S)\, P(E \mid S)$

Conditional Independence

- Given *m* random variables, $\{X_1, X_2, \dots, X_m\}$ and an event *Y*, if $X_1, X_2, \dots, \text{and } X_m$ are conditional independent under *Y*, then

$$P(X_1, X_2, \dots, X_m \mid Y)$$

$$= P(X_1 \mid Y) \times P(X_2 \mid Y) \times \cdots \times P(X_m \mid Y)$$

$$= \prod_{i=1}^{m} P(X_i \mid Y)$$

Conditional Independence

- Then,

$$P(Y \mid X_1, X_2, \ldots, X_m)$$

$$= \frac{P(Y) \prod_{i=1}^{m} P(X_i \mid Y)}{P(X_1, X_2, \ldots, X_m)}$$

Naive Bayes' Classifier

- Apply conditional independence to the learning model

$$M(\mathbf{q}) = \underset{Y \in \mathbf{T}}{\operatorname{argmax}} P(Y \mid X_1, X_2, \dots, X_m)$$

$$= \underset{Y \in \mathbf{T}}{\operatorname{argmax}} \frac{P(Y) \prod_{i=1}^{m} P(X_i \mid Y)}{P(X_1, X_2, \dots, X_m)}$$

Naive Bayes' Classifier

- However, the divider of $M(\mathbf{q})$, $P(X_1, X_2, \ldots, X_m)$, can be ignored in the maximum comparison.
- Therefore,

$$M(\mathbf{q}) = \underset{Y \in \mathbf{T}}{\operatorname{argmax}} P(Y) \prod_{i=1}^{m} P(X_i \mid Y) \ ,$$

  - In log-space:

$$M(\mathbf{q}) = \underset{Y \in \mathbf{T}}{\operatorname{argmax}} \left[\log P(Y) + \sum_{i=1}^{m} \log P(X_i \mid Y) \right]$$

Naive Bayes' Classifier

- What if
  q = {HEADACHE = true, FEVER = true, VOMITING = true}

- $P(\mathbf{q}|Y_1) = P(H, F, V\ |Y_1)$

$$= P(H|Y_1) \times P(F|Y_1) \times P(V\ |Y_1)$$

$$= \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3} = \frac{4}{27} = 0.1481$$

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

- $P(\mathbf{q}|Y_2) = P(H, F, V\ |Y_2)$

$$= P(H|Y_2) \times P(F|Y_2) \times P(V\ |Y_2)$$

$$= \frac{5}{7} \times \frac{3}{7} \times \frac{4}{7} = \frac{60}{343} = 0.1749$$

21

Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

Naive Bayes' Classifier

- Then,
  - $P(\mathbf{q}|Y_1)P(Y_1) = \frac{4}{27} \times \frac{3}{10} = 0.0444$
  - $P(\mathbf{q}|Y_2)P(Y_2) = \frac{60}{343} \times \frac{7}{10} = 0.1224$
- Therefore,
  - MENINGITIS = false if
    q = {HEADACHE = true, FEVER = true, VOMITING = true}

Naive Bayes' Classifier

- An example of a loan application fraud detection

ID	Credit History	Guarantor/ CoApplicant	Accomodation	Fraud
1	current	none	own	true
2	paid	none	own	false
3	paid	none	own	false
4	paid	guarantor	rent	true
5	arrears	none	own	false
6	arrears	none	own	true
7	current	none	own	false
8	arrears	none	own	false
9	current	none	rent	false
10	none	none	own	true
11	current	coapplicant	own	false
12	current	none	own	true
13	current	none	rent	true
14	paid	none	own	false
15	arrears	none	own	false
16	current	none	own	false
17	arrears	coapplicant	rent	false
18	arrears	none	free	false
19	arrears	none	own	false
20	paid	none	own	false

Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

Naive Bayes' Classifier

- Query *FRAUDULENT (FR) = ?* if
  - *CREDIT HISTORY (CH) = paid*
  - *GUARANTOR/COAPPLICANT (GC) = none*
  - *ACCOMODATION (ACC) = rent*

Naive Bayes' Classifier

- For *FR* = true
  - $P(fr) = \frac{6}{20} = 0.3$
  - $P(CH = paid \mid fr) = \frac{1}{6}$
  - $P(GC = none \mid fr) = \frac{5}{6}$
  - $P(ACC = rent \mid fr) = \frac{2}{6}$
  - $\frac{6}{20} \times \frac{1}{6} \times \frac{5}{6} \times \frac{2}{6} = 0.0139$
- For *FR* = false
  - $P(\overline{fr}) = \frac{14}{20} = 0.7$
  - $P(CH = paid \mid \overline{fr}) = \frac{4}{14}$
  - $P(GC = none \mid \overline{fr}) = \frac{12}{14}$
  - $P(ACC = rent \mid \overline{fr}) = \frac{2}{14}$
  - $\frac{14}{20} \times \frac{4}{14} \times \frac{12}{14} \times \frac{2}{14} = \mathbf{0.0245}$

Naive Bayes' Classifier

- How about that *FRAUDULENT (FR) = ?* if
  - *CREDIT HISTORY (CH) = paid*
  - *GUARANTOR/COAPPLICANT (GC) = guarantor*
  - *ACCOMODATION (ACC) = free*
-

Naive Bayes' Classifier

- For *FR* = true
  - $P(fr) = \frac{6}{20} = 0.3$
  - $P(CH = paid \mid fr) = \frac{1}{6}$
  - $P(GC = guarator \mid fr) = \frac{5}{6}$
  - $P(ACC = free \mid fr) = \frac{\mathbf{0}}{\mathbf{6}}$
- For *FR* = false
  - $P(\overline{fr}) = \frac{14}{20} = 0.7$
  - $P(CH = paid \mid \overline{fr}) = \frac{4}{14}$
  - $P(GC = guarator \mid \overline{fr}) = \frac{\mathbf{0}}{\mathbf{14}}$
  - $P(ACC = free \mid \overline{fr}) = \frac{1}{14}$

Naive Bayes' Classifier

- Smoothing
  - To take some of the probability from the events with lots of the probability share and gives it to the other probabilities in the set.
- There are several different ways to smooth probabilities.
  - Average smoothing
  - Gaussian smoothing
  - **Laplace smoothing** is commonly used to smooth categorical data.
  - Given a constant *k* and a random variable *X* with *m* events,

$$P(x|Y) = \frac{N(x|Y)+k}{N(Y)+km},$$

  - where $N(x|Y)$ is the number of samples of *x* under event *Y* and $N(Y)$ is the number of samples of *Y*.
  - Scikit-learn's **MultinomialNB** implements it

Naive Bayes' Classifier

- Let $k = 3$
- For *ACC* = free and *FR* = true
  - The number of types of *ACC* (m) is 3 (own, rent, and free)
  - $P(ACC = free|\, fr)$

$$= \frac{N(ACC = free|\, fr) + 3}{N(fr) + 3 \times 3} = \frac{0 + 3}{6 + 9}$$
$$= 0.2$$

- For *GC* = guarantor and *FR* = false
  - The number of types of *GC* (m) is 3 (none, guarantor, and coapplicant)
  - $P(GC = guarator|\overline{fr})$

$$= \frac{N(GC = guarator|\overline{fr}) + 3}{N\left(\overline{fr}\right) + 3 \times 3} = \frac{0 + 3}{14 + 9}$$
$$= 0.1304$$

Naive Bayes' Classifier

- Therefor, after applying Laplace smoothing

$$
\begin{aligned}
P(fr) &= 0.3 & P(\neg fr) &= 0.7 \\
P(CH = none|fr) &= 0.2222 & P(CH = none|\neg fr) &= 0.1154 \\
P(CH = paid|fr) &= 0.2222 & P(CH = paid|\neg fr) &= 0.2692 \\
P(CH = current|fr) &= 0.3333 & P(CH = current|\neg fr) &= 0.2692 \\
P(CH = arrears|fr) &= 0.2222 & P(CH = arrears|\neg fr) &= 0.3462 \\
P(GC = none|fr) &= 0.5333 & P(GC = none|\neg fr) &= 0.6522 \\
P(GC = guarantor|fr) &= 0.2667 & P(GC = guarantor|\neg fr) &= 0.1304 \\
P(GC = coapplicant|fr) &= 0.2 & P(GC = coapplicant|\neg fr) &= 0.2174 \\
P(ACC = own|fr) &= 0.4667 & P(ACC = own|\neg fr) &= 0.6087 \\
P(ACC = rent|fr) &= 0.3333 & P(ACC = rent|\neg fr) &= 0.2174 \\
P(ACC = Free|fr) &= 0.2 & P(ACC = Free|\neg fr) &= 0.1739
\end{aligned}
$$

Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

Naive Bayes' Classifier

- How about that *FRAUDULENT (FR) = ?* if
  - *CREDIT HISTORY (CH) = paid*
  - *GUARANTOR/COAPPLICANT (GC) = guarantor*
  - *ACCOMODATION (ACC) = free*
- For *FR* = true
  - $P(fr) \times P(CH = paid \mid fr) \times P(GC = guarator \mid fr) \times P(ACC = free \mid fr)$
  - $= 0.3 \times 0.2222 \times 0.2667 \times 0.2 = \mathbf{0.016}$
- For *FR* = false
  - $P(fr) \times P(CH = paid \mid \overline{fr}) \times P(GC = guarator \mid \overline{fr}) \times P(ACC = free \mid \overline{fr})$
  - $= 0.7 \times 0.2692 \times 0.1304 \times 0.1739 = 0.0042$

Continuous Features

- Categorical feature ➔ Discrete random variable
  - $X = \{X_1, X_2, \ldots, X_m\}$
  - $P(X_1) + P(X_2) + \cdots + P(X_m) = 1.0$

- Continuous feature ➔ Continuous random variable
  - $X \in \mathbf{R}$

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx \leq 1.0$$

$$P(X) = \int_{-\infty}^{\infty} f(x)\,dx = 1.0$$

Continuous Features

- **Probability density function** (PDF)
- If f is a PDF

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.0$$

- A PDF can be used to represent the probability distribution of a continuous random variable.
- Using a PDF to fit a probability distribution
- Five standard PDFs
  - Exponential
  - Normal
  - Student-t
  - Mixture Gaussians
  - Gamma

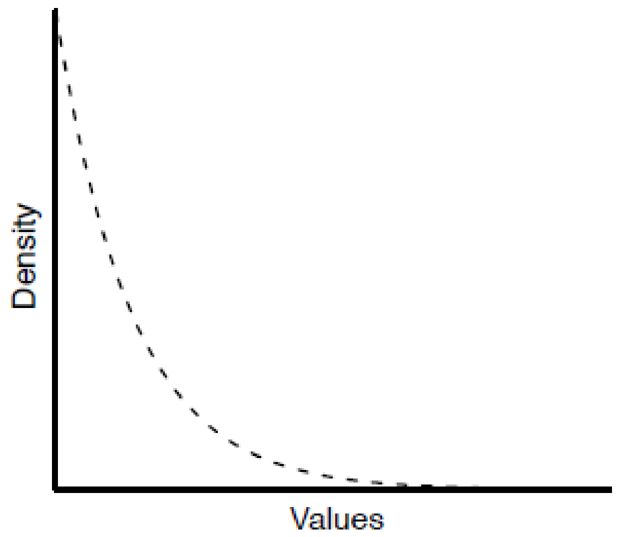
Standard PDF

- Exponential

$$E(x, \lambda) = \lambda e^{-\lambda x} \text{ if } x > 0\text{, otherwise} = 0$$

$x \in \mathbf{R}$

$\lambda \in \mathbf{R}$ and $\lambda > 0$



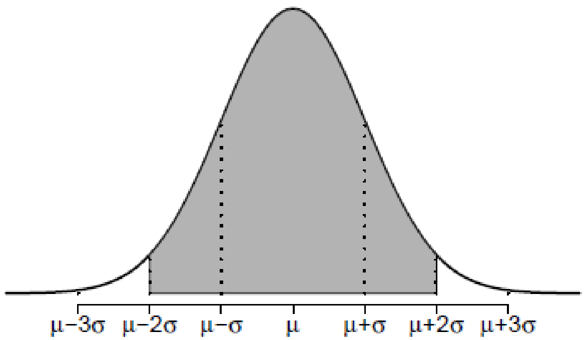
Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

Standard PDF

- Normal distribution
  - Gaussian function
  - Scikit-learn's **GassianNB** implements it

$$N(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$x \in \mathbf{R}$$
$$\mu \in \mathbf{R}$$
$$\sigma \in \mathbf{R} \text{ and } \sigma > 0$$



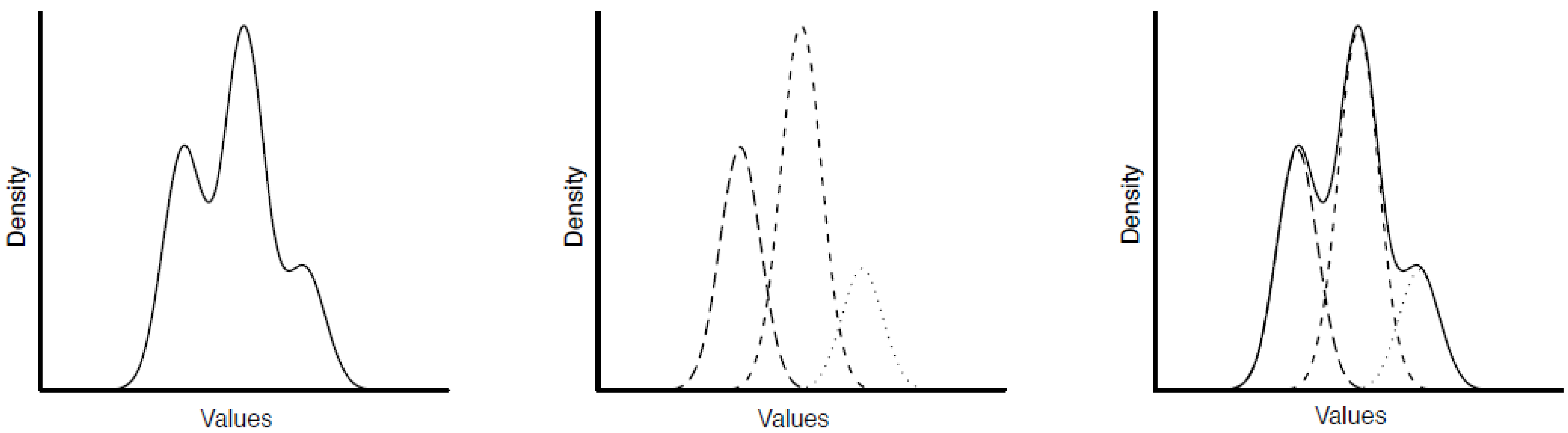
Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

Standard PDF

- Mixture Gaussians

$$N(x, \mathbf{u}, \boldsymbol{\sigma}, \mathbf{w}) = \sum_{i=1}^{n} \frac{w_i}{\sigma_i \sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2{\sigma_i}^2}}$$

$$x \in \mathbf{R}$$
$$\mathbf{u} = \{\mu_1, \mu_2, \dots, \mu_n | \mu_i \in \mathbf{R}\}$$
$$\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \dots, \sigma_n | \sigma_i \in \mathbf{R} > 0\}$$
$$\mathbf{w} = \{w_1, w_2, \dots, w_n | \, w_i \in \mathbf{R} > 0\}$$



Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

Standard PDF

- Student-t

$$\tau(x,k) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\,\Gamma(\frac{k}{2})}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$

$x \in \mathbf{R}$

$k \in \mathbf{N}$ and $k > 0$

$$\Gamma(n) = (n-1)!$$

where $n \in \mathbf{N} > \mathbf{0}$

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$$

where $z \in \mathbf{C} > \mathbf{and\ real(z)} > \mathbf{0}$

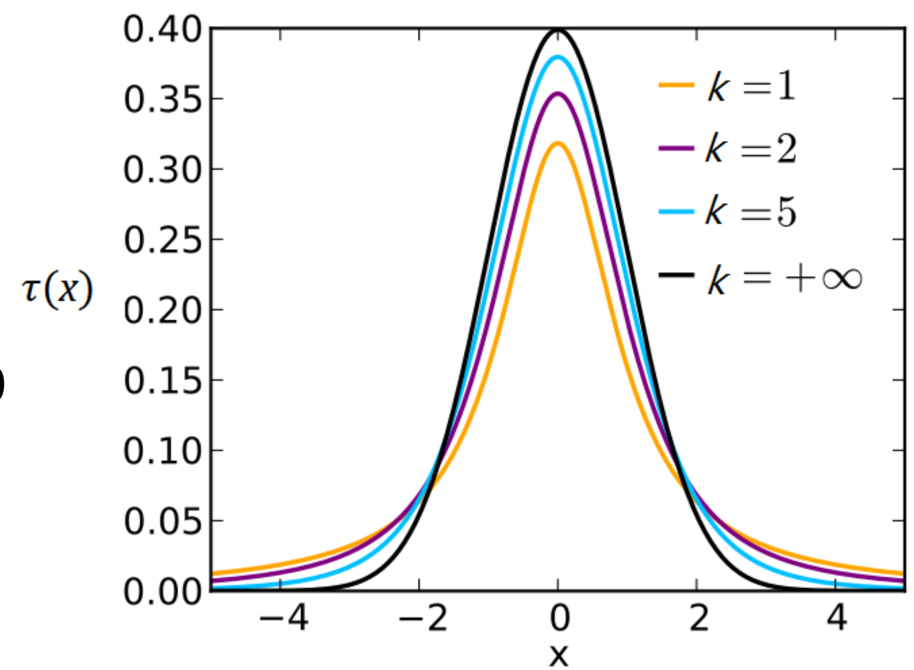


Figure from: https://en.wikipedia.org/wiki/Student%27s_t-distribution

Standard PDF

- Student-t
  - if k is even

$$\frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\,\Gamma(\frac{k}{2})} = \frac{(k-1)(k-3)\ldots 5\cdot 3}{2\sqrt{k}(k-2)(k-4)\ldots 4\cdot 2}$$

  - Otherwise

$$\frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\,\Gamma(\frac{k}{2})} = \frac{(k-1)(k-3)\ldots 4\cdot 2}{\pi\sqrt{k}(k-2)(k-4)\ldots 5\cdot 3}$$

$$\begin{aligned}
\Gamma\left(-\tfrac{3}{2}\right) &= \tfrac{4}{3}\sqrt{\pi}\\
\Gamma\left(-\tfrac{1}{2}\right) &= -2\sqrt{\pi}\\
\Gamma\left(\tfrac{1}{2}\right) &= \sqrt{\pi}\\
\Gamma\left(\tfrac{3}{2}\right) &= \tfrac{1}{2}\sqrt{\pi}\\
\Gamma\left(\tfrac{5}{2}\right) &= \tfrac{3}{4}\sqrt{\pi}\\
\Gamma\left(\tfrac{7}{2}\right) &= \tfrac{15}{8}\sqrt{\pi}
\end{aligned}$$

Standard PDF

- Gamma distribution

$$G(x,k,\theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

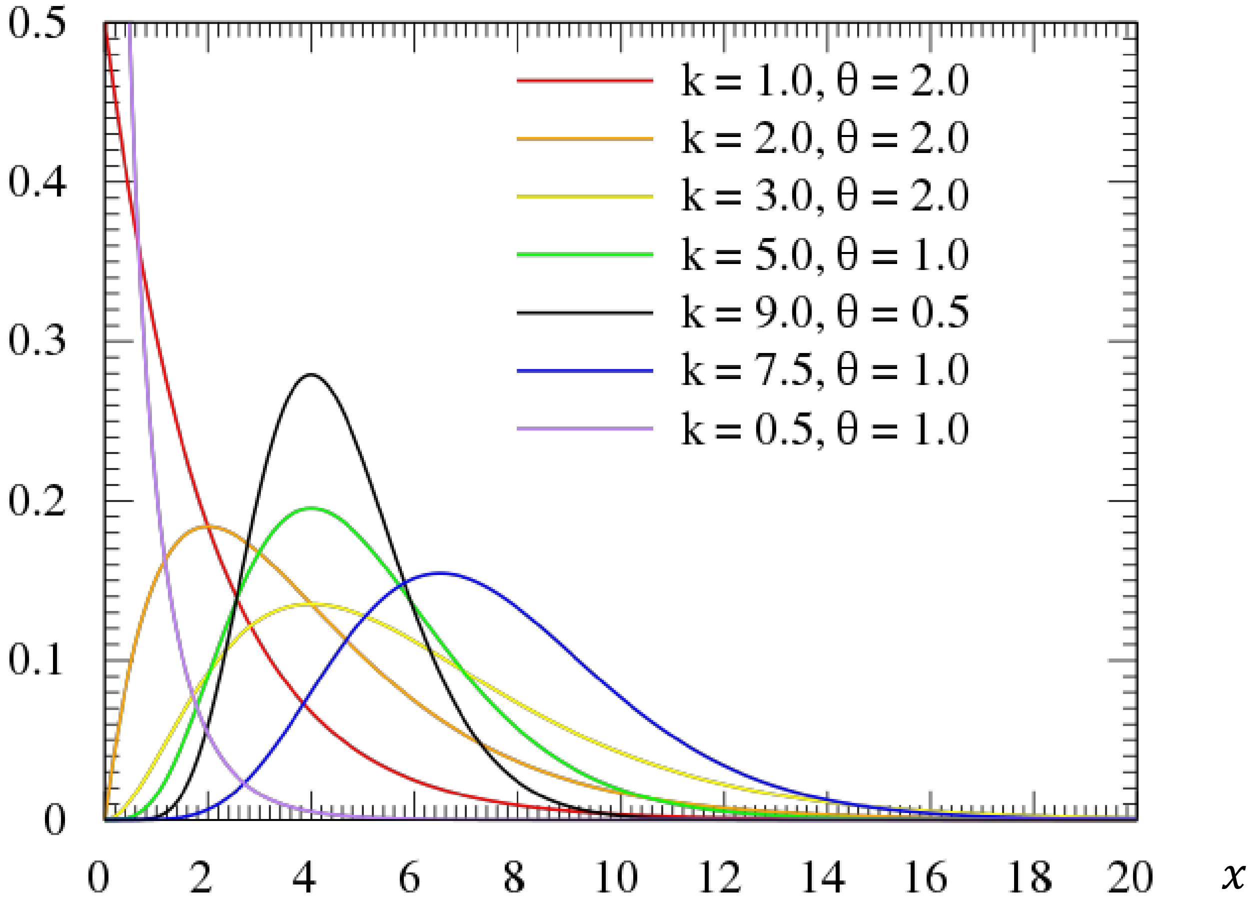
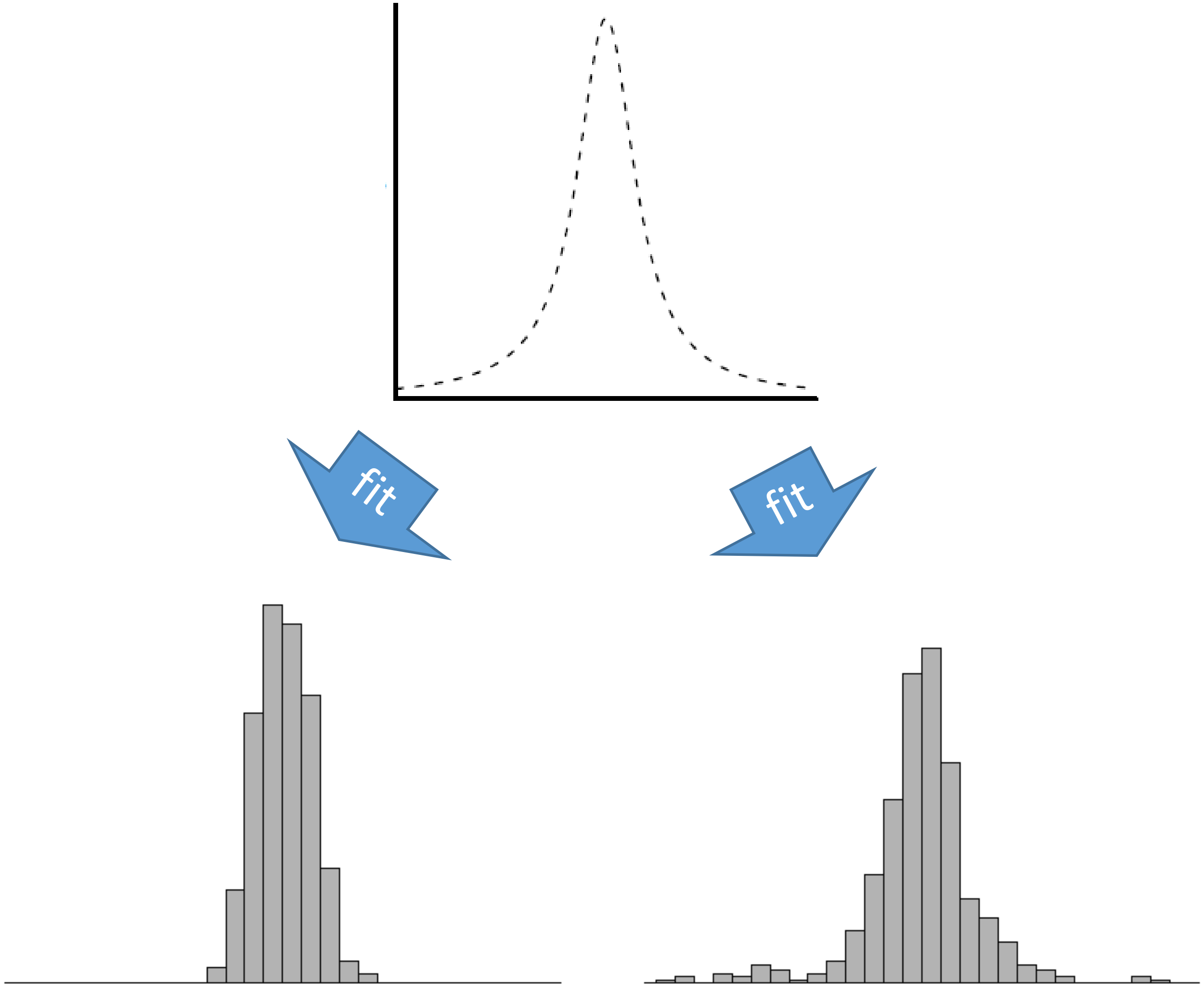


Figure from: https://en.wikipedia.org/wiki/Gamma_distribution

PDF Fitting

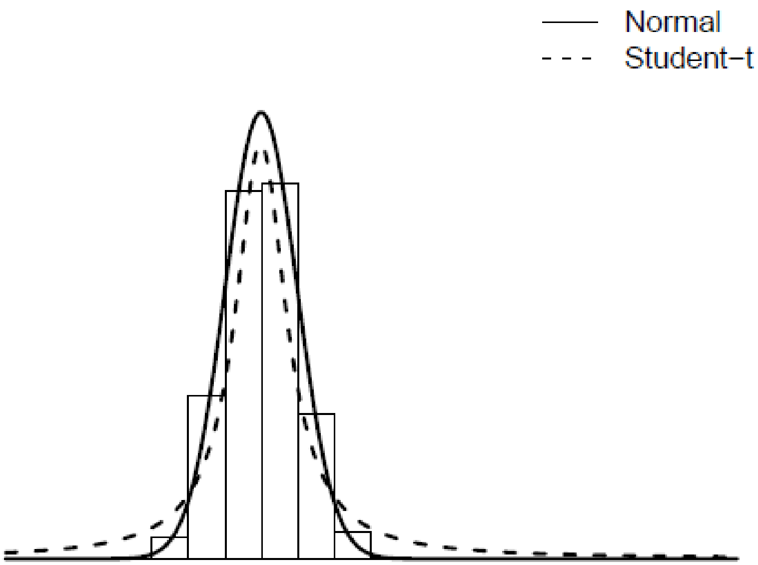
- Fitting a PDF to different histograms

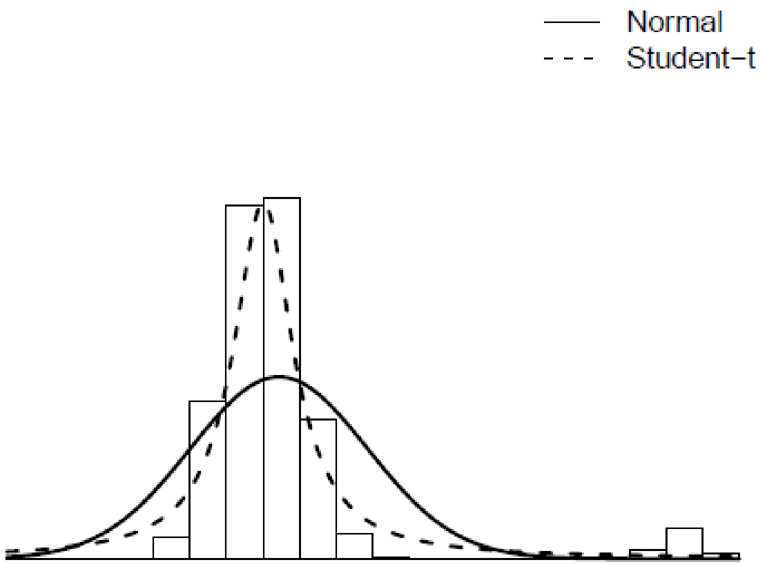


Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

PDF Fitting

- Fitting different PDFs to a histogram



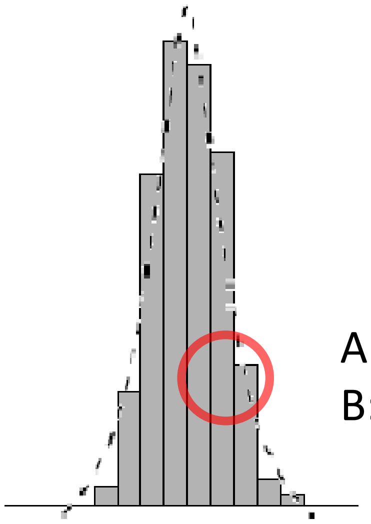


the same dataset

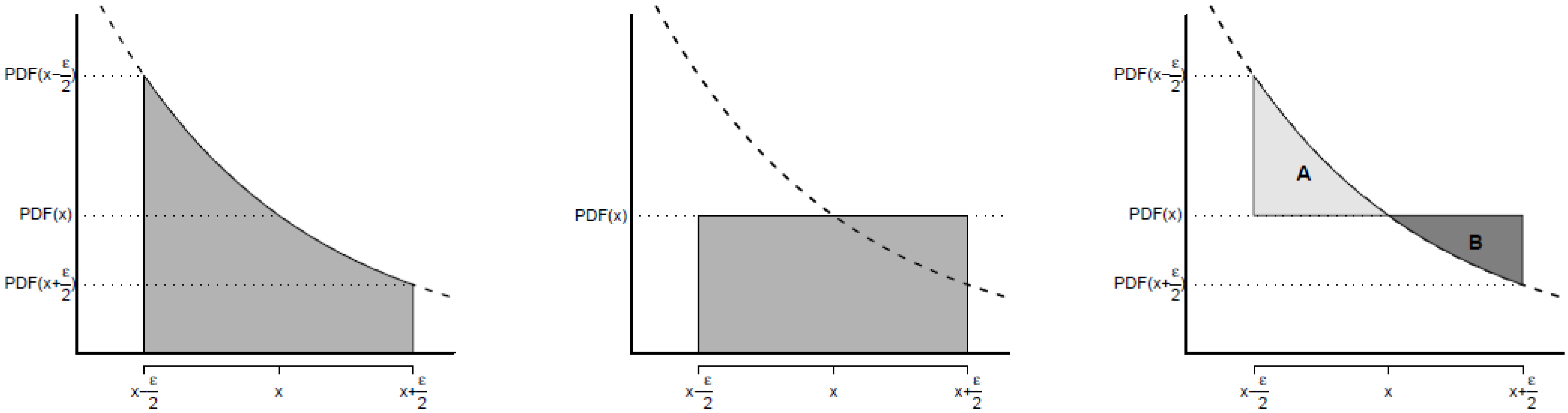
Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

PDF Fitting

- Interval error
  - Errors produced by the interval size
  - There is no hard and fast rule for deciding on interval size
  - By case



A: + error
B: - error



Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

PDF & Naive Bayes' Classifier

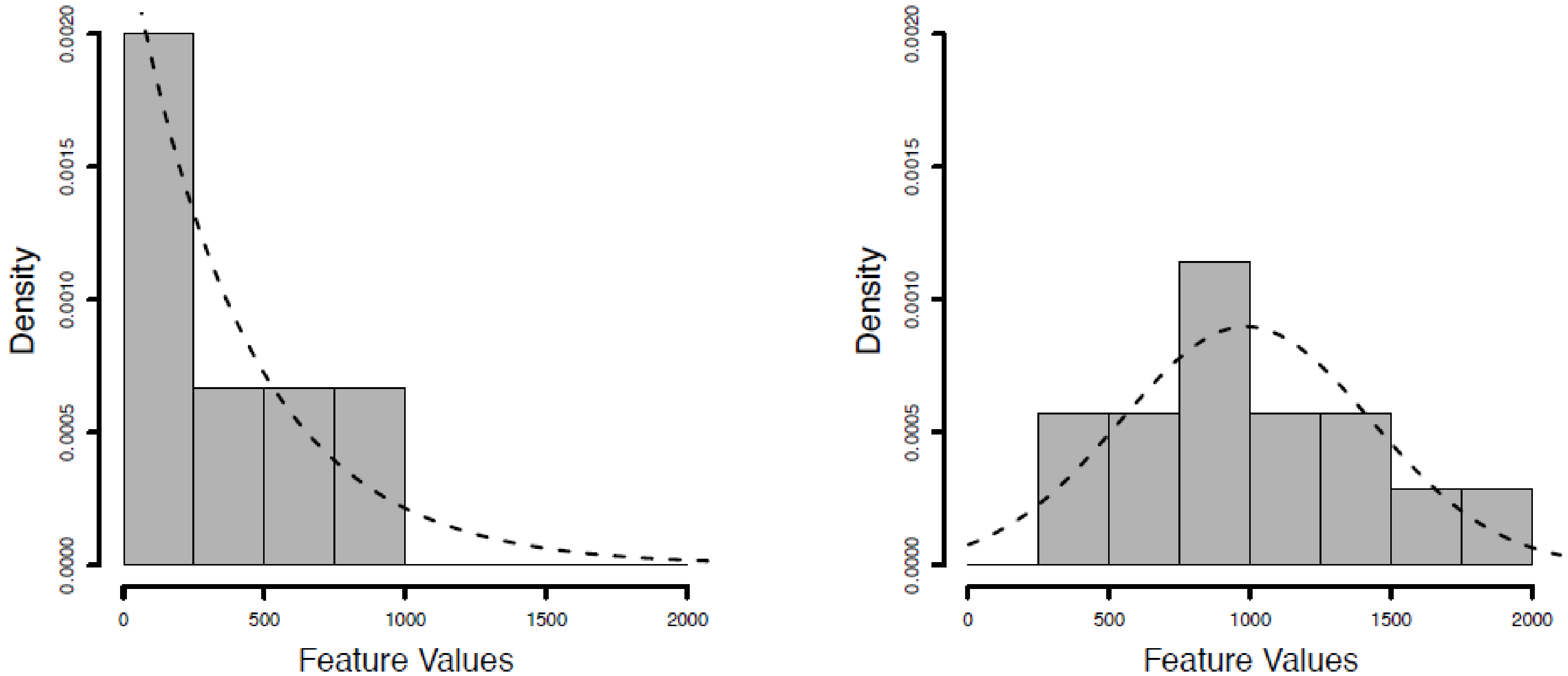
- An example of loan application fraud detection with **account balance (*AB*)**

ID	Credit History	Guarantor/ CoApplicant	Accommodation	Account Balance	Fraud
1	current	none	own	56.75	true
2	current	none	own	1,800.11	false
3	current	none	own	1,341.03	false
4	paid	guarantor	rent	749.50	true
5	arrears	none	own	1,150.00	false
6	arrears	none	own	928.30	true
7	current	none	own	250.90	false
8	arrears	none	own	806.15	false
9	current	none	rent	1,209.02	false
10	none	none	own	405.72	true
11	current	coapplicant	own	550.00	false
12	current	none	free	223.89	true
13	current	none	rent	103.23	true
14	paid	none	own	758.22	false
15	arrears	none	own	430.79	false
16	current	none	own	675.11	false
17	arrears	coapplicant	rent	1,657.20	false
18	arrears	none	free	1,405.18	false
19	arrears	none	own	760.51	false
20	current	none	own	985.41	false

Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

PDF & Naive Bayes' Classifier

- Binning for continuous data ➔Histogram
- Choose a PDF to fit each histogram



$P(AB = x | fr)$

Bin size: 250

$P\left(AB = x \middle| \overline{fr}\right)$

Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

PDF & Naive Bayes' Classifier

- A simple method to fit the exponential distribution
  - Compute the sample mean, μ, of the ACCOUNT BALANCE where FRAUDULENT = 'True'
  - Let $\lambda = \frac{1}{\mu}$
  - Then,

$$E(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}$$

PDF & Naive Bayes' Classifier

- A simple method to fit the normal distribution
  - Compute the sample mean, μ, and standard deviation, σ, of the ACCOUNT BALANCE where FRAUDULENT = 'False'
  - Then,

$$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

PDF & Naive Bayes' Classifier

- To implement a probability-based learning model, you have to do that
  - applying the Laplace smoothing for each categorical feature, and
  - fitting a PDF for each continuous feature

PDF & Naive Bayes' Classifier

- For example, how about that *FRAUDULENT (FR) = ?* if
  - *CREDIT HISTORY (CH) = paid*
  - *GUARANTOR/COAPPLICANT (GC) = guarantor*
  - *ACCOMODATION (ACC) = free*
  - *ACCOUNT BALANCE (AB)* = 759.07

$$
\begin{aligned}
P(fr) &= 0.3 & P(\neg fr) &= 0.7 \\
P(CH = paid|fr) &= 0.2222 & P(CH = paid|\neg fr) &= 0.2692 \\
P(GC = guarantor|fr) &= 0.2667 & P(GC = guarantor|\neg fr) &= 0.1304 \\
P(ACC = free|fr) &= 0.2 & P(ACC = free|\neg fr) &= 0.1739 \\
P(AB = 759.07|fr) & & P(AB = 759.07|\neg fr) & \\
\approx E\left(\begin{array}{c} 759.07, \\ \lambda = 0.0024 \end{array}\right) &= 0.00039 & \approx N\left(\begin{array}{c} 759.07, \\ \mu = 984.26, \\ \sigma = 460.94 \end{array}\right) &= 0.00077
\end{aligned}
$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k]|fr)\right) \times P(fr) = 0.0000014$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k]|\neg fr)\right) \times P(\neg fr) = 0.0000033$$

Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

Binning & Naive Bayes' Classifier

- The loan application fraud detection with a second continuous descriptive feature added: LOAN AMOUNT (LA)

ID	CREDIT HISTORY	GUARANTOR/ COAPPLICANT	ACCOMMODATION	ACCOUNT BALANCE	LOAN AMOUNT	FRAUD
1	current	none	own	56.75	900	true
2	current	none	own	1 800.11	150 000	false
3	current	none	own	1 341.03	48 000	false
4	paid	guarantor	rent	749.50	10 000	true
5	arrears	none	own	1 150.00	32 000	false
6	arrears	none	own	928.30	250 000	true
7	current	none	own	250.90	25 000	false
8	arrears	none	own	806.15	18 500	false
9	current	none	rent	1 209.02	20 000	false
10	none	none	own	405.72	9 500	true
11	current	coapplicant	own	550.00	16 750	false
12	current	none	free	223.89	9 850	true
13	current	none	rent	103.23	95 500	true
14	paid	none	own	758.22	65 000	false
15	arrears	none	own	430.79	500	false
16	current	none	own	675.11	16 000	false
17	arrears	coapplicant	rent	1 657.20	15 450	false
18	arrears	none	free	1 405.18	50 000	false
19	arrears	none	own	760.51	500	false
20	current	none	own	985.41	35 000	false

Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

Binning & Naive Bayes' Classifier

- Bin size

Bin Thresholds		
	Bin1	$\leq 9,925$
$9,925 <$	Bin2	$\leq 19,250$
$19,225 <$	Bin3	$\leq 49,000$
$49,000 <$	Bin4	

ID	Loan Amount	Binned Loan Amount	Fraud	ID	Loan Amount	Binned Loan Amount	Fraud
15	500	bin1	false	9	20,000	bin3	false
19	500	bin1	false	7	25,000	bin3	false
1	900	bin1	true	5	32,000	bin3	false
10	9,500	bin1	true	20	35,000	bin3	false
12	9,850	bin1	true	3	48,000	bin3	false
4	10,000	bin2	true	18	50,000	bin4	false
17	15,450	bin2	false	14	65,000	bin4	false
16	16,000	bin2	false	13	95,500	bin4	true
11	16,750	bin2	false	2	150,000	bin4	false
8	18,500	bin2	false	6	250,000	bin4	true

Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

Binning & Naive Bayes' Classifier

- *FRAUDULENT (FR) = ?* if
  - *CREDIT HISTORY (CH) = paid*
  - *GUARANTOR/COAPPLICANT (GC) = guarantor*
  - *ACCOMODATION (ACC) = free*
  - *ACCOUNT BALANCE (AB)* = 759.07
  - LOAN AMOUNT(LA) = 8000

$$
\begin{aligned}
P(fr) &= 0.3 & P(\neg fr) &= 0.7 \\
P(CH = paid \mid fr) &= 0.2222 & P(CH = paid \mid \neg fr) &= 0.2692 \\
P(GC = guarantor \mid fr) &= 0.2667 & P(GC = guarantor \mid \neg fr) &= 0.1304 \\
P(ACC = free \mid fr) &= 0.2 & P(ACC = free \mid \neg fr) &= 0.1739 \\
P(AB = 759.07 \mid fr) & & P(AB = 759.07 \mid \neg fr) & \\
\approx E\begin{pmatrix} 759.07, \\ \lambda = 0.0024 \end{pmatrix} &= 0.00039 & \approx N\begin{pmatrix} 759.07, \\ \mu = 984.26, \\ \sigma = 460.94 \end{pmatrix} &= 0.00077 \\
P(BLA = bin1 \mid fr) &= 0.3333 & P(BLA = bin1 \mid \neg fr) &= 0.1923
\end{aligned}
$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid fr)\right) \times P(fr) = 0.000000462$$

$$\left(\prod_{k=1}^{n} P(\mathbf{q}[k] \mid \neg fr)\right) \times P(\neg fr) = 0.000000633$$

Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

Target Prior

- So far, $P(Y)$ is estimated from the training dataset
- However, we can assign a prior probability for $P(Y)$
- For example,

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

- If **q** = {HEADACHE = true, FEVER = false, VOMITING = true}, MENINGITIS = ?
- Priors of MENINGITIS:
  - *P*(MENIGITIS = true) = 0.6
  - *P*(MENIGITIS = false) = 0.4

Data from: John D. Kelleher, et al, "Fundamentals of Machine Learning for Predictive Data Analytics - Algorithms, Worked Examples, and Case Studies," *MIT Press*, 2015.

Target Prior

- Without priors of targets
  - P(M = true) = 0.3
    - $M(\mathbf{q}) = P(\text{M = true})\ P(\text{H = true})\ P(\text{F = false})\ P(\text{V = true})$
      $= 0.3 \times 0.7 \times 0.6 \times 0.6$
      $= 0.0756$
  - P(M = false) = 0.7
    - $M(\mathbf{q}) = P(\text{M = true})\ P(\text{H = true})\ P(\text{F = false})\ P(\text{V = true})$
      $= 0.7 \times 0.7 \times 0.6 \times 0.6$
      $\mathbf{= 0.1764}$
- Without priors of targets
  - P(M = true) = 0.6
    - $M(\mathbf{q}) = P(\text{M = true})\ P(\text{H = true})\ P(\text{F = false})\ P(\text{V = true})$
      $= \mathbf{0.6} \times 0.7 \times 0.6 \times 0.6$
      $\mathbf{= 0.1512}$
  - P(M = false) = 0.4
    - $M(\mathbf{q}) = P(\text{M = true})\ P(\text{H = true})\ P(\text{F = false})\ P(\text{V = true})$
      $= \mathbf{0.4} \times 0.7 \times 0.6 \times 0.6$
      $= 0.1008$

Multinomial Distribution

- Let a set of random variates $X_1, X_2, \dots, X_m$ have a probability function
- *N* data instances, *m* features.

$$P(X_1 = x_1, X_2 = x_2, \dots, X_m = x_m) = P(\mathbf{x}) = \frac{N!}{x_1!\, x_2! \dots x_m!} p_1^{x_1} p_2^{x_2} \dots p_m^{x_m}$$

$$= \frac{N!}{\prod_{i=1}^{m} x_i!} \prod_{i=1}^{m} p_i^{x_i},$$

where x_i is the number of samples of event *i*, p_i is the probability of event *i*, and

$$\sum_{i=1}^{m} x_i = N$$

Multinomial Distribution

- For conditional probability
- *N* data instances, *m* features, under a condition *Y*.

$$P(\mathbf{x}|Y) = \frac{N!}{\prod_{i=1}^{m} x_i!} \prod_{i=1}^{m} p_{yi}^{x_i}$$

where p_{yi} is the probability of event *i* under *Y*

Multinomial Distribution

- Naïve Bays model:

$$P(Y|\mathbf{x}) = P(Y)P(\mathbf{x}|Y) \propto P(Y)\prod_{i=1}^{m} p_{yi}^{x_i},$$

where ∝ means "is proportional to"

Multinomial Naïve Bayes (MNB)

- MNB model:

$$M(\mathbf{q}) = \operatorname*{argmax}_{Y \in \mathbf{T}} P(Y)\, P(\mathbf{x}|Y)$$

$$= \operatorname*{argmax}_{Y \in \mathbf{T}} P(Y) \prod_{i=1}^{m} p_{yi}^{x_i}$$

- Applying the Laplace smoothing

$$M(\mathbf{q}) = \operatorname*{argmax}_{Y \in \mathbf{T}} P(Y) \prod_{i=1}^{m} \left(\frac{N(x_i|\, Y) + k}{N(Y) + km} \right)^{x_i}$$

- In log-space:

$$M(\mathbf{q}) = \operatorname*{argmax}_{Y \in \mathbf{T}} \left[\log P(Y) + \sum_{i=1}^{m} x_i \log \frac{N(x_i|\, Y) + k}{N(Y) + km} \right]$$

Complement Naïve Bayes (CNB)

- J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the poor assumptions of Naïve Bayes text classifiers," *ICML*, vol. 3, pp. 616-623, 2003.
- Estimates each feature's probabilites of **all targets except Y**.
- CNB model:
  - Let $\bar{Y}$ be of the set of all targets except *Y*.

$$M(\mathbf{q}) = \underset{Y \in \mathbf{T}}{\operatorname{argmax}} \left(\frac{P(Y)}{P(\mathbf{x}|\bar{Y})} \right),$$

  - In log-space:

$$M(\mathbf{q}) = \underset{Y \in \mathbf{T}}{\operatorname{argmax}} \left[\log P(Y) - \sum_{i=1}^{m} x_i \log \frac{N(x_i|\bar{Y}) + k}{N(\bar{Y}) + km} \right]$$

Bernoulli Distribution

- For a binary random variable X

$$P(X=1)=p$$
$$P(X=0)=q$$
$$p=1-q$$
$$q=1-p$$

  - PDF of binary variable of $k=\{0,1\}$:

$$B(k,p)=\begin{cases} p & \text{if } k=1 \\ q=1-p & \text{if } k=0 \end{cases}$$

    - or

$$B(k,p)=p^k(1-p)^{1-k}$$

    - or

$$B(k,p)=p^k+(1-p)\,(1-k)$$

Bernoulli Naïve Bayes

- BNB model:

$$M(\mathbf{q}) = \underset{Y \in \mathbf{T}}{\operatorname{argmax}} \, P(Y) \, P(\mathbf{x}|Y),$$

  - where

$$P(\mathbf{x}|Y) = \prod_{i=1}^{m} p_{yi}^{x_i} (1 - p_{yi})^{(1-x_i)}$$

  - In log-space:

$$M(\mathbf{q}) = \underset{Y \in \mathbf{T}}{\operatorname{argmax}} \left[\log P(Y) + \sum_{i=1}^{m} \left(x_i \log p_{yi} + (1 - x_i) \log\left(1 - p_{yi}\right) \right) \right]$$