CHAPTER 3:

Bayesian Decision Theory

Contents

Introduction

- Programming computers *to make inference from data* is a cross between *statistics* and *computer science*,

  where *statisticians* provide the *mathematical framework of making inference from data* and *computer scientists work on the efficient implementation of the inference methods*.

- Data comes *from a process* that is *not completely known*.

- This *lack of knowledge* is indicated by *modeling the process* as a *random process*.

- Maybe *the process is actually deterministic*, but because we do not have *access to complete knowledge* about it, we *model it as random* and *use probability theory to analyze it*.

Introduction

Elements of Probability

- A *random experiment* is one whose *outcome is not predictable* with certainty in advance (Ross 1987; Casella and Berger 1990).

- The *set of all possible outcomes* is known as the *sample space S*.

- A *sample space* is *discrete* if it consists of a *finite* (or countably infinite)
  -set of outcomes; otherwise it is *continuous*.

- Any *subset E of S* is an event.

- Events are *sets*, and we can talk about their *complement*, *intersection*, *union*, and so forth.

Introduction

Elements of Probability

- One interpretation of probability is as a ***frequency***.

- When an experiment is ***continually repeated*** under the exact same conditions, for any ***event E***, the ***proportion of time that the outcome is in E approaches some constant value***.

- This ***constant limiting frequency is the probability of the event***, and we denote it as ***P(E)***.

Introduction

Elements of Probability

- Probability sometimes is interpreted as a *degree of belief*.

- For example, when we speak of Turkey's probability of winning the World Soccer Cup in 2010, we do not mean a frequency of occurrence, since the championship will happen only once and it has not yet occurred (at the time of the writing of this book).

- What we mean in such a case is a *subjective degree* of belief in the occurrence of the event.

- Because it is **subjective**, *different individuals may assign different probabilities to the same event*.

Introduction

Axioms of Probability

Axioms *ensure that the probabilities assigned in a random experiment* can be interpreted as *relative frequencies* and that the *assignments are consistent* with our *intuitive understanding* of relationships among relative frequencies:

1. *$0 \leq P(E) \leq 1$*.

   If E_1 is an event that cannot possibly occur, then $P(E_1) = 0$.

   If E_2 is sure to occur, $P(E_2) = 1$.

2. *S* is the sample space containing all possible outcomes, $P(S) = 1$.

3. If E_i, $i = 1, \ldots, n$ are *mutually exclusive* (i.e., if they cannot occur at the same time, as in $E_i \cap E_j = \emptyset$, $j = i$, where ∅ is the *null event* that does not contain any possible outcomes),

 we have

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) \qquad \text{Eq. 1}$$

Introduction

A box contains three marbles — one red, one green, and one blue. Consider an experiment that consists of taking one marble from the box, then replacing it in the box and drawing a second marble from the box. Describe the sample space. Repeat for the case in which the second marble is drawn without first replacing the first marble.

S = {rr, rb, rg, br, bb, bg, gr, gb, gg} when done with replacement and
S = {rb, rg, br, bg, gr, gb} when done without replacement**,**
where rb means, for instance, that the first marble is red and the second green.

Introduction

An experiment consists of tossing a coin three times. What is the sample space of this experiment? Which event corresponds to the experiment resulting in more heads than tails?

S = {hhh, hht, hth, htt, thh, tht, tth, ttt}.

The event {hhh, hht,hth, thh} corresponds to more heads than tails.

Introduction

Axioms of Probability

we have,

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) \qquad \text{Eq. 2}$$

For example, letting E^c denote the *complement of E*, consisting of all possible outcomes in *S* that are not in *E*, we have $E \cap E^C = \emptyset$ and

$P(E \cup Ec) = P(E) + P(E^c) = 1$

$P(E^c) = 1 - P(E)$

If the *intersection of E and F* is *not empty*,

we have, $P(E \cup F) = P(E) + P(F) - P(E \cap F)$

Introduction

Let $S = \{1, 2, 3, 4, 5, 6, 7\}$, $E = \{1, 3, 5, 7\}$, $F = \{7, 4, 6\}$, $G = \{1, 4\}$. Find

(a) EF; **(c)** EG^c; **(e)** $E^c(F \cup G)$;

(b) $E \cup FG$; **(d)** $EF^c \cup G$; **(f)** $EG \cup FG$.

(a) $\{7\}$, (b) $\{1, 3, 4, 5, 7\}$, (c) $\{3, 5, 7\}$, (d) $\{1, 3, 4, 5\}$, (e) $\{4, 6\}$, (f) $\{1, 4\}$

Introduction

BASIC PRINCIPLE OF COUNTING

Suppose that two experiments are to be performed. Then if experiment 1 can result in any one of m possible outcomes and if, for each outcome of experiment 1, there are n possible outcomes of experiment 2, then together there are mn possible outcomes of the two experiments.

Ex: Two balls are “randomly drawn” from a bowl containing 6 white and 5 black balls. What is the probability that one of the drawn balls is white and the other black?

If we regard the order in which the balls are selected as being significant, then as the first drawn ball may be any of the 11 and the second any of the remaining 10, it follows that the sample space consists of $11 \cdot 10 = 110$ points. Furthermore, there are $6 \cdot 5 = 30$ ways in which the first ball selected is white and the second black, and similarly there are $5 \cdot 6 = 30$ ways in which the first ball is black and the second white. Hence, assuming that “randomly drawn” means that each of the 110 points in the sample space is equally likely to occur, then we see that the desired probability is $30+30/110 = 6/11$

Introduction

Generalized Basic Principle of Counting

If r experiments that are to be performed are such that the first one may result in any of n1 possible outcomes, and if for each of these n1 possible outcomes there are n2 possible outcomes of the second experiment, and if for each of the possible outcomes of the first two experiments there are n3 possible outcomes of the third experiment, and if, . . . , then there are a total of n1 · n2 · · · nr possible outcomes of the r experiments.

Mr. Ravi has 10 books that he is going to put on his bookshelf. Of these, 4 are mathematics books, 3 are chemistry books, 2 are history books, and 1 is a language book. Ravi wants to arrange his books so that all the books dealing with the same subject are together on the shelf. How many different arrangements are possible?

There are *4! 3! 2! 1!* arrangements such that the mathematics books are first in line, then the chemistry books, then the history books, and then the language book. Similarly, for each possible ordering of the subjects, there are *4! 3! 2! 1!* possible arrangements. Hence, as there are 4! possible orderings of the subjects, the desired answer is ***4! 4! 3! 2! 1! = 6,912***.

Introduction

Conditional Probability

- *P(E|F)* is the probability of the occurrence of event *E* given that *F* occurred and is given as $$P(E|F) = \frac{P(E \cap F)}{P(F)} \qquad \text{Eq. 3}$$
- Knowing that F occurred reduces the sample space to F, and the part of it where E also occurred is E∩F.
- Eq. 3 is well-defined only if P(F) > 0.
- Because ∩ is commutative, we have

$$P(E \cap F) = P(E|F)P(F) = P(F|E)P(E)$$

which gives us *Bayes' formula*:
$$P(F|E) = \frac{P(E|F)P(F)}{P(E)}$$

Introduction

You might be interested in finding out a patient's probability of having liver disease if they are an alcoholic. "Being an alcoholic" is the **test** (kind of like a litmus test) for liver disease.

- **A** could mean the event "Patient has liver disease." Past data tells you that 10% of patients entering your clinic have liver disease. P(A) = 0.10.
- **B** could mean the litmus test that "Patient is an alcoholic." Five percent of the clinic's patients are alcoholics. P(B) = 0.05.
- You might also know that among those patients diagnosed with liver disease, 7% are alcoholics. This is your **B|A:** the probability that a patient is alcoholic, given that they have liver disease, is 7%.

Bayes' theorem tells you: $P(A|B)=P(B|A)P(A)/P(B)$

$$\mathbf{P(A|B) = (0.07 * 0.1)/0.05 = 0.14}$$

In other words, if the patient is an alcoholic, their chances of having liver disease is 0.14 (14%). This is a large increase from the 10% suggested by past data. But it's still unlikely that any particular patient has liver disease.

Introduction

A bin contains 5 defective (that immediately fail when put in use), 10 partially defective (that fail after a couple of hours of use), and 25 acceptable transistors. A transistor is chosen at random from the bin and put into use. If it does not immediately fail, what is the probability it is acceptable?

Since the transistor did not immediately fail, we know that it is not one of the 5 defectives and so the desired probability is:

$$P\{\text{acceptable}|\text{not defective}\}$$

$$= \frac{P\{\text{acceptable, not defective}\}}{P\{\text{not defective}\}}$$

$$= \frac{P\{\text{acceptable}\}}{P\{\text{not defective}\}}$$

Introduction

where the last equality follows since the transistor will be both acceptable and not defective if it is acceptable. Hence, assuming that each of the 40 transistors is equally likely to be chosen, we obtain that

$$P\{\text{acceptable}|\text{not defective}\} = \frac{25/40}{35/40} = 5/7$$

- It should be noted that we could also have derived this probability by working directly with the reduced sample space.
- That is, since we know that the chosen transistor is not defective, the problem reduces to computing the probability that a transistor, chosen at random from a bin containing 25 acceptable and 10 partially defective transistors, is acceptable.
- This is clearly equal to **25/35** .

Introduction

Bayes Formula

$$P(F|E) = \frac{P(E|F)P(F)}{P(E)}$$

When F_i are mutually exclusive and exhaustive, namely, $\bigcup_{i=1}^{n} F_i = S$

$$E = \bigcup_{i=1}^{n} E \cap F_i$$

$$P(E) = \sum_{i=1}^{n} P(E \cap F_i) = \sum_{i=1}^{n} P(E|F_i)P(F_i)$$

Bayes' formula allows us to write

$$P(F_i|E) = \frac{P(E \cap F_i)}{P(E)} = \frac{P(E|F_i)P(F_i)}{\sum_j P(E|F_j)P(F_j)}$$

If E and F are *independent*, we have $P(E|F) = P(E)$ and thus

$$P(E \cap F) = P(E)P(F)$$

That is, knowledge of whether F has occurred does not change the probability that E occurs.

Introduction

1% of people have a certain genetic defect.
90% of tests for the gene detect the defect (true positives).
9.6% of the tests are false positives.
If a person gets a positive test result, **what are the odds they actually have the genetic defect?**

The first step into solving Bayes' theorem problems is to assign letters to events:

- A = chance of having the faulty gene. That was given in the question as 1%.
- That also means the probability of not having the gene (~A) is 99%.
- X = A positive test result.

So:

1. P(A|X) = Probability of having the gene given a positive test result.
2. P(X|A) = Chance of a positive test result given that the person actually has the gene. That was given in the question as 90%.
3. P(X|~A) = Chance of a positive test if the person doesn't have the gene. That was given in the question as 9.6%

Now we have all of the information we need to put into the equation:

$P(A|X) = P(X|A)P(A)/(P(X/A)P(A)+P(X/\sim A)P(\sim A))$

$P(A|X) = (.9 * .01) / (.9 * .01 + .096 * .99) = 0.0865$ (8.65%).

The probability of having the faulty gene on the test is 8.65%.

Introduction

Q. Given the following statistics, what is the probability that a woman has cancer if she has a positive mammogram result?
One percent of women over 50 have breast cancer.
Ninety percent of women who have breast cancer test positive on mammograms.
Eight percent of women will have false positives.

Step 1: Assign events to A or X. You want to know what a woman's probability of having cancer is, given a positive mammogram. For this problem, actually having cancer is A and a positive test result is X.
Step 2: List out the parts of the equation (this makes it easier to work the actual equation):

$P(A)=0.01$
$P(\sim A)=0.99$
$P(X|A)=0.9$
$P(X|\sim A)=0.08$

Step 3: Insert the parts into the equation and solve.
$P(A|X) = P(X|A)P(A)/(P(X/A)P(A)+P(X/\sim A)P(\sim A))$
$P(A|X) = (0.9 * 0.01) / ((0.9 * 0.01) + (0.08 * 0.99) = 0.10.$
The probability of a woman having cancer, given a positive test result, is 10%.

Introduction

Random Variables

A random variable is *a function that assigns a number to each outcome in the sample space of a random experiment*.

Probability Distribution and Density Functions

The probability distribution function F(·) of *a random variable X for any real number a* is

$$F(a) = P\{X \le a\}$$

and we have

$$P\{a < X \le b\} = F(b) - F(a)$$

If X is a discrete random variable

$$F(a) = \sum_{\forall x \le a} P(x)$$

where *P(·) is the probability mass function defined as P(a) = P{X = a}.*

If X is a continuous random variable, p(·) is the probability density function such that

$$F(a) = \int_{-\infty}^{a} p(x)dx$$

Introduction

Joint Distribution and Density Functions

In certain experiments, we may be interested in the relationship between two or more random variables, and we use the *joint* probability distribution and density functions of *X* and *Y* satisfying

$$F(x, y) = P\{X \le x, Y \le y\}$$

Individual marginal distributions and densities can be computed by marginalizing, namely, summing over the free variable:

$$F_X(x) = P\{X \le x\} = P\{X \le x, Y \le \infty\} = F(x, \infty)$$

In the discrete case, we write $P(X = x) = \sum_j P(x, y_j)$

and in the continuous case, we have $p_X(x) = \int_{-\infty}^{\infty} p(x, y)dy$

If X and Y are *independent*, we have

$$p(x, y) = p_X(x)p_Y(y)$$

Introduction

Examples

- *A civil engineer may not be directly concerned with the daily risings and declines of the water level of a reservoir (which we can take as the experimental result) but may only care about the level at the end of a rainy season. These quantities of interest that are determined by the result of the experiment are known as random variables.*

- Suppose that *an individual purchases two electronic components* each of which may be either *defective or acceptable*. In addition, suppose that the *four possible results — (d, d), (d, a), (a, d), (a, a)* — have respective probabilities *.09, .21, .21, .49*.

  [where (d, d) means that both components are defective, (d, a) that the first component is defective and the second acceptable, and so on].

  Let X denote the *number of acceptable components obtained in the purchase*, then X is a random variable taking on one of the values 0, 1, 2 with respective probabilities.

$$P\{X = 0\} = .09$$
$$P\{X = 1\} = .42$$
$$P\{X = 2\} = .49$$

Introduction

If we were mainly concerned with whether there was at least one acceptable component, we could define the random variable I by

$$I = \begin{cases} 1 & \text{if } X = 1 \text{ or } 2 \\ 0 & \text{if } X = 0 \end{cases}$$

If A denotes the event that at least one acceptable component is obtained, then the random variable I is called the indicator random variable for the event A, since I will equal 1 or 0 depending upon whether A occurs.

The probabilities attached to the possible values of I are

$$P\{I = 1\} = .91$$

$$P\{I = 0\} = .09$$

Introduction

Conditional Distributions

When *X* and *Y* are random variables

$$P_{X|Y}(x|y) = P\{X = x|Y = y\} = \frac{P\{X = x, Y = y\}}{P\{Y = y\}} = \frac{P(x, y)}{P_Y(y)}$$

Bayes' Rule

When two random variables are jointly distributed with the value of one known, the probability that the other takes a given value can be computed using *Bayes' rule*:

$$P(y|x) = \frac{P(x|y)P_Y(y)}{P_X(x)} = \frac{P(x|y)P_Y(y)}{\sum_y P(x|y)P_Y(y)}$$

Or, in words

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Introduction

Properties of Probability Distribution

- Expectation
- Median
- Mode
- Variance
- Standard Deviation
- Covariance
- Skewness
- Kurtosis
- Moments

Special Random Variables

- Bernoulli Distribution
  - A trial is performed whose outcome is either a "success" or a "failure."
  - The random variable X is a 0/1 indicator variable and takes the value 1 for a success outcome and is 0 otherwise. p is the probability that the result of trial is a success.
  - Then $P\{X = 1\} = p$ and $P\{X = 0\} = 1 - p$

    which can equivalently be written as

    $P\{X = i\} = p^i(1 - p)^{1-i}, i = 0, 1$

Probability and Inference

- Result of tossing a coin is $\in$ {Heads,Tails}
- Random var $X \in \{1,0\}$

  Bernoulli: $P\{X=1\} = p_o^{X}(1 - p_o)^{(1-X)}$
- Sample: $\boldsymbol{X} = \{x^t\}_{t=1}^{N}$

  Estimation: $p_o = \#\{\text{Heads}\}/\#\{\text{Tosses}\} = \sum_t x^t / N$
- Prediction of next toss:

  Heads if $p_o > ½$, Tails otherwise

Classification

- Credit scoring: Inputs are income and savings.
  Output is low-risk vs high-risk
- Input: $\boldsymbol{x} = [x_1,x_2]^T$,Output: C Î {0,1}
- Prediction:

$$\text{choose} \begin{cases} C = 1 \text{ if } P(C = 1 | x_1, x_2) > 0.5 \\ C = 0 \text{ otherwise} \end{cases}$$

or

$$\text{choose} \begin{cases} C = 1 \text{ if } P(C = 1 | x_1, x_2) > P(C = 0 | x_1, x_2) \\ C = 0 \text{ otherwise} \end{cases}$$

Bayes' Rule

prior *likelihood*

posterior

$$P(C \mid \mathbf{x}) = \frac{P(C)\,p(\mathbf{x} \mid C)}{p(\mathbf{x})}$$

evidence

$$P(C=0) + P(C=1) = 1$$

$$p(\mathbf{x}) = p(\mathbf{x} \mid C=1)P(C=1) + p(\mathbf{x} \mid C=0)P(C=0)$$

$$p(C=0 \mid \mathbf{x}) + P(C=1 \mid \mathbf{x}) = 1$$

Bayes' Rule: *K*>2 Classes

$$P(C_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_i)P(C_i)}{p(\mathbf{x})}$$

$$= \frac{p(\mathbf{x} \mid C_i)P(C_i)}{\sum_{k=1}^{K} p(\mathbf{x} \mid C_k)P(C_k)}$$

$P(C_i) \geq 0$ and $\sum_{i=1}^{K} P(C_i) = 1$

choose C_i if $P(C_i \mid \mathbf{x}) = \max_k P(C_k \mid \mathbf{x})$

Losses and Risks

- Actions: α_i
- Loss of α_i when the state is C_k : λ_{ik}
- Expected risk (Duda and Hart, 1973)

$$R(\alpha_i \mid \mathbf{x}) = \sum_{k=1}^{K} \lambda_{ik} P(C_k \mid \mathbf{x})$$

$$\text{choose } \alpha_i \text{ if } R(\alpha_i \mid \mathbf{x}) = \min_k R(\alpha_k \mid \mathbf{x})$$

Losses and Risks: 0/1 Loss

$$\lambda_{ik} = \begin{cases} 0 \text{ if } i = k \\ 1 \text{ if } i \neq k \end{cases}$$

$$\begin{aligned} R(\alpha_i \mid \mathbf{x}) &= \sum_{k=1}^{K} \lambda_{ik} P(C_k \mid \mathbf{x}) \\ &= \sum_{k \neq i} P(C_k \mid \mathbf{x}) \\ &= 1 - P(C_i \mid \mathbf{x}) \end{aligned}$$

For minimum risk, choose the most probable class

Losses and Risks: Reject

$$\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ \lambda & \text{if } i = K+1, \quad 0 < \lambda < 1 \\ 1 & \text{otherwise} \end{cases}$$

$$R(\alpha_{K+1} \mid \mathbf{x}) = \sum_{k=1}^{K} \lambda P(C_k \mid \mathbf{x}) = \lambda$$

$$R(\alpha_i \mid \mathbf{x}) = \sum_{k \neq i} P(C_k \mid \mathbf{x}) = 1 - P(C_i \mid \mathbf{x})$$

choose C_i if $P(C_i \mid \mathbf{x}) > P(C_k \mid \mathbf{x})\ \forall k \neq i$ and $P(C_i \mid \mathbf{x}) > 1 - \lambda$

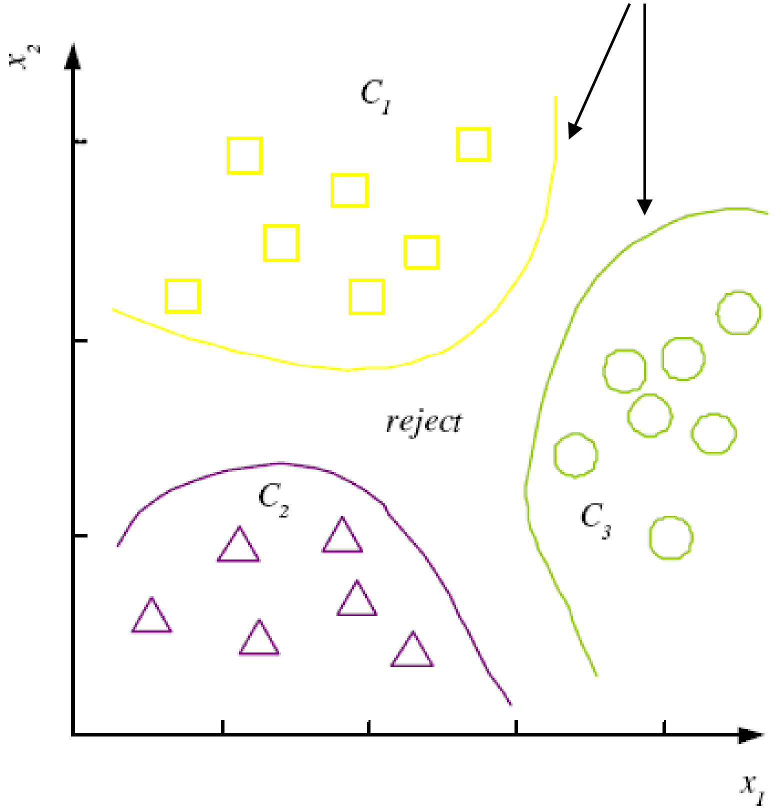
reject otherwise

Discriminant Functions

$$\text{choose } C_i \text{ if } g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})$$

$$g_i(\mathbf{x}), i = 1, \ldots, K$$

$$g_i(\mathbf{x}) = \begin{cases} -R(\alpha_i \mid \mathbf{x}) \\ P(C_i \mid \mathbf{x}) \\ p(\mathbf{x} \mid C_i) P(C_i) \end{cases}$$



K decision regions $\mathcal{R}_1, \ldots, \mathcal{R}_K$

$$\mathcal{R}_i = \{\mathbf{x} \mid g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})\}$$

K=2 Classes

- Dichotomizer (K=2) vs Polychotomizer (K>2)
- $g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$

$$\text{choose}\begin{cases} C_1 \text{ if } g(\mathbf{x}) > 0 \\ C_2 \text{ otherwise} \end{cases}$$

- *Log odds:* $\log\frac{P(C_1 \mid \mathbf{x})}{P(C_2 \mid \mathbf{x})}$

Utility Theory

- Prob of state k given exidence $\mathbf{x}$: $P(S_k|\mathbf{x})$
- Utility of α_i when state is k: U_{ik}
- Expected utility:

$$EU(\alpha_i|\mathbf{x}) = \sum_k U_{ik} P(S_k|\mathbf{x})$$

$$\text{Choose } \alpha_i \text{ if } EU(\alpha_i|\mathbf{x}) = \max_j EU(\alpha_j|\mathbf{x})$$

Association Rules

- Association rule: $X \rightarrow Y$
- *People who buy/click/visit/enjoy X are also likely to buy/click/visit/enjoy Y.*
- A rule implies association, not necessarily causation.

Association measures

- Support $(X \rightarrow Y)$:

$$P(X,Y) = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers}\}}$$

- Confidence $(X \rightarrow Y)$:

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)} = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers who bought } X\}}$$

- Lift $(X \rightarrow Y)$:

$$= \frac{P(X,Y)}{P(X)P(Y)} = \frac{P(Y \mid X)}{P(Y)}$$

Apriori algorithm (Agrawal et al., 1996)

- For (X,Y,Z), a 3-item set, to be frequent (have enough support), (X,Y), (X,Z), and (Y,Z) should be frequent.
- If (X,Y) is not frequent, none of its supersets can be frequent.
- Once we find the frequent *k*-item sets, we convert them to rules: X, Y $\rightarrow$ Z, ...

  and X $\rightarrow$ Y, Z, ...