CHAPTER 3:

Bayesian Decision Theory

Basic Probability

Probability and Inference

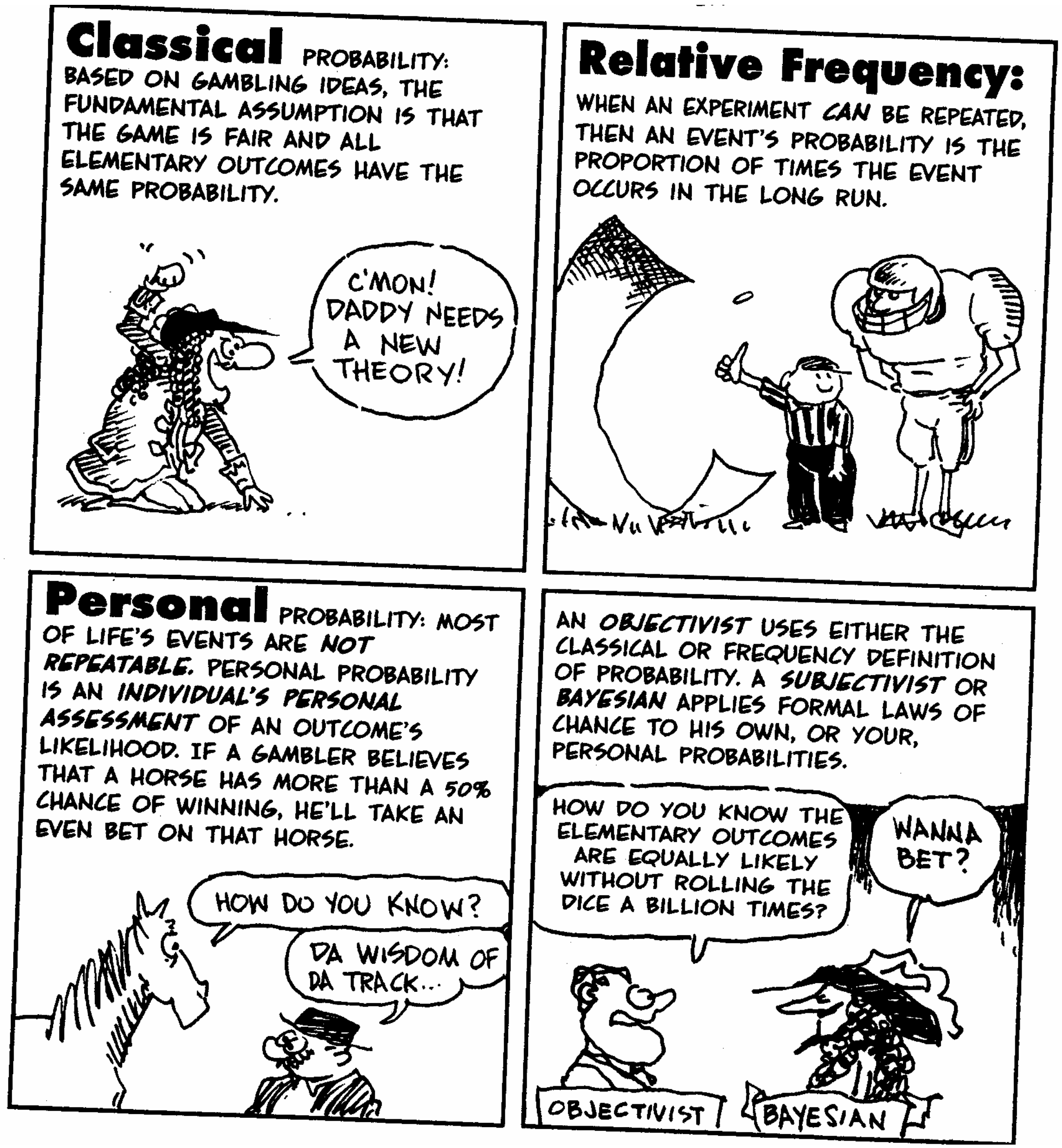
- Result of tossing a coin is $\in$ {Heads,Tails}
- Random var $X \in \{1,0\}$

  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

Probability



Probability

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 $E_i \cap E_j = \emptyset, j \neq i$, where $\emptyset$ is the *null event* that does not contain any possible outcomes) we have

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

Probability

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

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

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

If the intersection of E and F is not empty, we have

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

Conditional Probability

$P(E|F)$ is the probability of the occurrence of event E given that F occurred and is given as

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

Probability

Knowing that F occurred reduces the sample space to F, and the part of it where E also occurred is $E \cap F$. Note that equation A.3 is well-defined only if $P(F) > 0$. Because $\cap$ is commutative, we have

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

which gives us *Bayes' formula:*

$$P(F|E) = \frac{P(E|F)P(F)}{P(E)} \tag{A.4}$$

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

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

$$P(E) = \sum_{i=1}^{n} P(E \cap F_i) = \sum_{i=1}^{n} P(E|F_i)P(F_i) \tag{A.5}$$

Bayes' formula allows us to write

$$P(F_i|E) = \frac{P(E \cap F_i)}{P(E)} = \frac{P(E|F_i)P(F_i)}{\sum_j P(E|F_j)P(F_j)} \tag{A.6}$$

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

$$P(E \cap F) = P(E)P(F) \tag{A.7}$$

Joint probability.

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

Probability

A.2 Random Variables

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

A.2.1 Probability Distribution and Density Functions

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

$$F(a) = P\{X \leq a\} \tag{A.8}$$

and we have

$$P\{a < X \leq b\} = F(b) - F(a) \tag{A.9}$$

If X is a discrete random variable

$$F(a) = \sum_{\forall x \leq a} P(x) \tag{A.10}$$

where $P(\cdot)$ is the *probability mass function* defined as $P(a) = P\{X = a\}$. If X is a *continuous* random variable, $p(\cdot)$ is the *probability density function* such that

$$F(a) = \int_{-\infty}^{a} p(x)dx \tag{A.11}$$

A.2.2 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

$$F(x, y) = P\{X \le x, Y \le y\} \tag{A.12}$$

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

$$F_X(x) = P\{X \le x\} = P\{X \le x, Y \le \infty\} = F(x, \infty) \tag{A.13}$$

In the discrete case, we write

$$P(X = x) = \sum_j P(x, y_j) \tag{A.14}$$

Probability

In the discrete case, we write

(A.14) $$P(X = x) = \sum_j P(x, y_j)$$

and in the continuous case, we have

(A.15) $$p_X(x) = \int_{-\infty}^{\infty} p(x, y)dy$$

If X and Y are *independent*, we have

(A.16) $$p(x, y) = p_X(x)p_Y(y)$$

These can be generalized in a straightforward manner to more than two random variables.

A.2.3 Conditional Distributions

When X and Y are random variables

(A.17) $$P_{X|Y}(x|y) = P\{X = x|Y = y\} = \frac{P\{X = x, Y = y\}}{P\{Y = y\}} = \frac{P(x, y)}{P_Y(y)}$$

Probability

A.2.4 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:*

$$P(y|x) = \frac{P(x|y)P_Y(y)}{P_X(x)} = \frac{P(x|y)P_Y(y)}{\sum_y P(x|y)P_Y(y)} \tag{A.18}$$

Or, in words

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \tag{A.19}$$

Note that the denominator is obtained by summing (or integrating if y is continuous) the numerator over all possible y values. The "shape" of $p(y|x)$ depends on the numerator with denominator as a normalizing factor to guarantee that $p(y|x)$ sum to 1. Bayes' rule allows us to modify a prior probability into a posterior probability by taking information provided by x into account.

Probability

Bayes' rule inverts dependencies, allowing us to compute $p(y|x)$ if $p(x|y)$ is known. Suppose that y is the "cause" of x, like y going on summer vacation and x having a suntan. Then $p(x|y)$ is the probability that someone who is known to have gone on summer vacation has a suntan. This is the *causal* (or predictive) way. Bayes' rule allows us a *diagnostic* approach by allowing us to compute $p(y|x)$: namely, the probability that someone who is known to have a suntan, has gone on summer vacation. Then $p(y)$ is the general probability of anyone's going on summer vacation and $p(x)$ is the probability that anyone has a suntan, including both those who have gone on summer vacation and those who have not.

A.2.5 Expectation

Expectation, expected value, or *mean* of a random variable X, denoted by $E[X]$, is the average value of X in a large number of experiments:

$$E[X] = \begin{cases} \sum_i x_i P(x_i) & \text{if } X \text{ is discrete} \\ \int x p(x) dx & \text{if } X \text{ is continuous} \end{cases} \tag{A.20}$$

It is a weighted average where each value is weighted by the probability that X takes that value. It has the following properties ($a, b \in \Re$):

$$\begin{aligned} E[aX + b] &= aE[X] + b \\ E[X + Y] &= E[X] + E[Y] \end{aligned} \tag{A.21}$$

Probability

For any real-valued function $g(\cdot)$, the expected value is

$$\text{(A.22)} \qquad E[g(X)] = \begin{cases} \sum_i g(x_i)P(x_i) & \text{if } X \text{ is discrete} \\ \int g(x)p(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

A special $g(x) = x^n$, called the nth moment of X, is defined as

$$\text{(A.23)} \qquad E[X^n] = \begin{cases} \sum_i x_i^n P(x_i) & \text{if } X \text{ is discrete} \\ \int x^n p(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

Mean is the first moment and is denoted by μ.

A.2.6 Variance

Variance measures how much X varies around the expected value. If $\mu \equiv E[X]$, the variance is defined as

$$\text{(A.24)} \qquad \mathrm{Var}(X) = E[(X-\mu)^2] = E[X^2] - \mu^2$$

Variance is the second moment minus the square of the first moment. Variance, denoted by σ^2, satisfies the following property ($a, b \in \Re$):

$$\text{(A.25)} \qquad \mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$$

$\sqrt{\mathrm{Var}(X)}$ is called the *standard deviation* and is denoted by σ. Standard deviation has the same unit as X and is easier to interpret than variance.

Probability

... and is easier to interpret than variance.

Covariance indicates the relationship between two random variables. If the occurrence of X makes Y more likely to occur, then the covariance is positive; it is negative if X's occurrence makes Y less likely to happen and is 0 if there is no dependence.

$$\text{(A.26)} \quad \text{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X\mu_Y$$

where $\mu_X \equiv E[X]$ and $\mu_Y \equiv E[Y]$. Some other properties are

$$\begin{aligned}
\text{Cov}(X,Y) &= \text{Cov}(Y,X) \\
\text{Cov}(X,X) &= \text{Var}(X) \\
\text{Cov}(X+Z,Y) &= \text{Cov}(X,Y) + \text{Cov}(Z,Y)
\end{aligned}$$

$$\text{(A.27)} \quad \text{Cov}\left(\sum_i X_i, Y\right) = \sum_i \text{Cov}(X_i, Y)$$

$$\text{(A.28)} \quad \text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$$

$$\text{(A.29)} \quad \text{Var}\left(\sum_i X_i\right) = \sum_i \text{Var}(X_i) + \sum_i \sum_{j\neq i} \text{Cov}(X_i, X_j)$$

Probability

If X and Y are independent, $E[XY] = E[X]E[Y] = \mu_X \mu_Y$ and $\text{Cov}(X,Y) = 0$. Thus if X_i are independent

(A.30) $$\text{Var}\left(\sum_i X_i\right) = \sum_i \text{Var}(X_i)$$

Correlation is a normalized, dimensionless quantity that is always between -1 and 1:

(A.31) $$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

A.2.7 Weak Law of Large Numbers

Let $\mathcal{X} = \{X^t\}_{t=1}^N$ be a set of independent and identically distributed (iid) random variables each having mean μ and a finite variance σ^2. Then for any $\epsilon > 0$

(A.32) $$P\left\{\left|\frac{\sum_t X^t}{N} - \mu\right| > \epsilon\right\} \to 0 \text{ as } N \to \infty$$

That is, the average of N trials converges to the mean as N increases.

Probability

A.3 Special Random Variables

There are certain types of random variables that occur so frequently that names are given to them.

A.3.1 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

(A.33) $P\{X = 1\} = p$ and $P\{X = 0\} = 1 - p$

which can equivalently be written as

(A.34) $P\{X = i\} = p^i(1-p)^{1-i}, i = 0, 1$

If X is Bernoulli, its expected value and variance are

(A.35) $E[X] = p, \ \mathrm{Var}(X) = p(1-p)$

Probability

A.3.2 Binomial Distribution

If N identical independent Bernoulli trials are made, the random variable X that represents the number of successes that occurs in N trials is binomial distributed. The probability that there are i successes is

$$P\{X = i\} = \binom{N}{i} p^i (1-p)^{N-i}, i = 0 \ldots N \tag{A.36}$$

If X is binomial, its expected value and variance are

$$E[X] = Np, \ \mathrm{Var}(X) = Np(1-p) \tag{A.37}$$

A.3.3 Multinomial Distribution

Consider a generalization of Bernoulli where instead of two states, the outcome of a random event is one of K mutually exclusive and exhaustive states, each of which has a probability of occurring p_i where $\sum_{i=1}^{K} p_i = 1$. Suppose that N such trials are made where outcome i occurred N_i times with $\sum_{i=1}^{k} N_i = N$. Then the joint distribution of $N_1, N_2, \ldots, N_K$ is multinomial:

$$P(N_1, N_2, \ldots, N_K) = N! \prod_{i=1}^{K} \frac{p_i^{N_i}}{N_i!} \tag{A.38}$$

Probability

(A.38) $$P(N_1, N_2, \ldots, N_K) = N! \prod_{i=1}^{K} \frac{p_i^{N_i}}{N_i!}$$

A special case is when $N = 1$; only one trial is made. Then N_i are 0/1 indicator variables of which only one of them is 1 and all others are 0. Then equation A.38 reduces to

(A.39) $$P(N_1, N_2, \ldots, N_K) = \prod_{i=1}^{K} p_i^{N_i}$$

A.3.4 Uniform Distribution

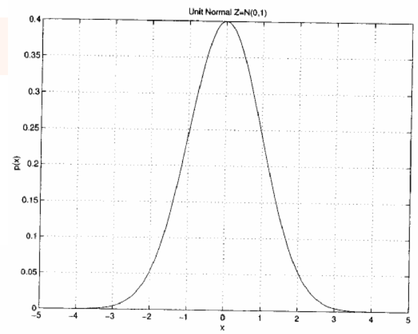
X is uniformly distributed over the interval $[a, b]$ if its density function is given by

(A.40) $$p(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

If X is uniform, its expected value and variance are

(A.41) $$E[X] = \frac{a+b}{2}, \ \mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

Probability



Normal (Gaussian) Distribution

X is normal or Gaussian distributed with mean μ and variance σ^2, denoted as $\mathcal{N}(\mu, \sigma^2)$, if its density function is

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], -\infty < x < \infty$$

Many random phenomena obey the bell-shaped normal distribution, at least approximately, and many observations from nature can be seen as a continuous, slightly different versions of a typical value—that is probably why it is called the *normal* distribution. In such a case, μ represents the typical value and σ defines how much instances vary around the prototypical value.

68.27 percent lie in $(\mu - \sigma, \mu + \sigma)$, 95.45 percent in $(\mu - 2\sigma, \mu + 2\sigma)$ and 99.73 percent in $(\mu - 3\sigma, \mu + 3\sigma)$. Thus $P\{|x - \mu| < 3\sigma\} \approx .99$. For practical purposes, $p(x) \approx 0$ if $x < \mu - 3\sigma$ or $x > \mu + 3\sigma$. $\mathcal{Z}$ is unit normal, namely, $\mathcal{N}(0, 1)$ (see figure A.1) and its density is written as

$$p_Z(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right]$$

Probability

If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$, then $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$. The sum of independent normal variables is also normal with $\mu = \sum_i \mu_i$ and $\sigma^2 = \sum_i \sigma_i^2$. If X is $\mathcal{N}(\mu, \sigma^2)$, then

(A.44) $$\frac{X - \mu}{\sigma} \sim \mathcal{Z}$$

This is called z-normalization.

CENTRAL LIMIT THEOREM

Let $X_1, X_2, \ldots, X_N$ be a set of iid random variables all having mean μ and variance σ^2. Then the *central limit theorem* states that for large N, the distribution of

(A.45) $$X_1 + X_2 + \ldots + X_N$$

is approximately $\mathcal{N}(N\mu, N\sigma^2)$. For example, if X is binomial with parameters (N, p), X can be written as the sum of N Bernoulli trials and $(X - Np)/\sqrt{Np(1-p)}$ is approximately unit normal.

Central limit theorem is also used to generate normally distributed random variables on computers. Programming languages have subroutines that return uniformly distributed (pseudo-)random numbers in the range $[0, 1]$. When U_i are such random variables, $\sum_{i=1}^{12} U_i - 6$ is approximately $\mathcal{Z}$.

Let us say $X^t \sim \mathcal{N}(\mu, \sigma^2)$. The estimated sample mean

(A.46) $$m = \frac{\sum_{t=1}^{N} X^t}{N}$$

is also normal with mean μ and variance σ^2/N.

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 $\in \{0,1\}$
- Prediction:

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

or equivalently

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

Bayes' Rule

$$\underbrace{P(C \mid \boldsymbol{x})}_{\textit{posterior}} = \frac{\overbrace{P(C)}^{\textit{prior}}\,\overbrace{p(\boldsymbol{x} \mid C)}^{\textit{likelihood}}}{\underbrace{p(\boldsymbol{x})}_{\textit{evidence}}}$$

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

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

Posterior probabilities satisfy

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

Bayes' Rule: K>2 Classes

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

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

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

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

Highest Posterior Probability

Losses and Risks

- Decisions not equally good or costly
- Actions: α_i – assign input to C_i
- Loss of α_i when the state is C_k : λ_{ik}
- Expected risk (Duda and Hart, 1973)

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

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

Choose the action which minimizes the risk

Losses and Risks: 0/1 Loss

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

Good decisions have no loss,
Bad decisions are equally costly

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

$$= \sum_{k \neq i} P(C_k \mid \boldsymbol{x})$$

$$= 1 - P(C_i \mid \boldsymbol{x})$$

For minimum risk, choose the most probable class

Losses and Risks: Reject

When misclassifications are costly!

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

<= Reject action α_{K+1}

<= Cost of misclassification

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

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

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

reject otherwise

Discriminant Functions

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

$$g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})$$

Max discriminant corresponds to min risk

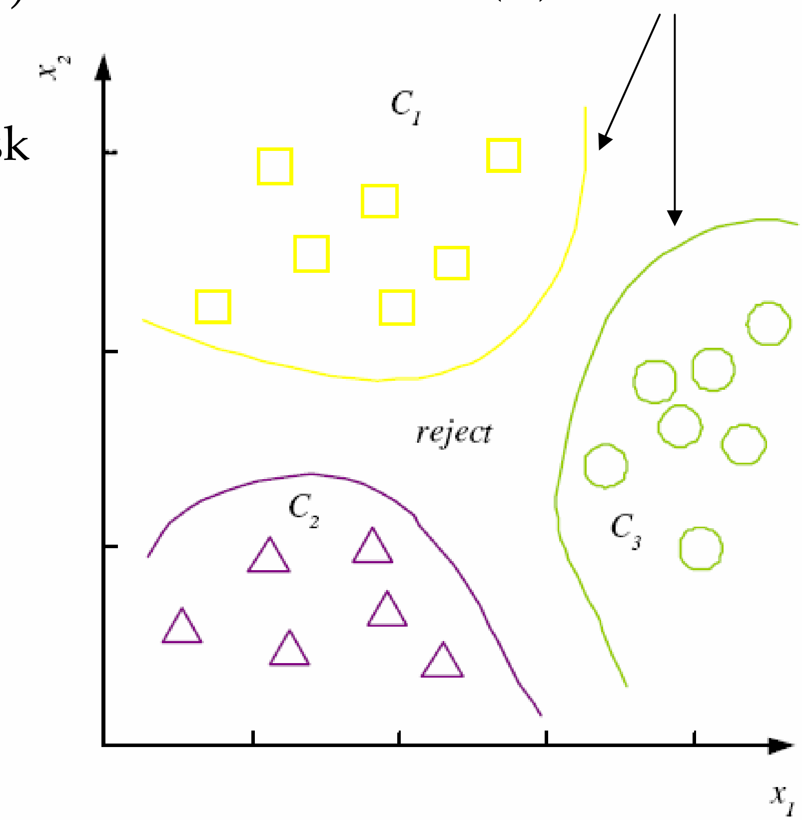
$$g_i(\mathbf{x}) = P(C_i \mid \mathbf{x})$$

For 0/1 loss function

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

When neglecting common evidence

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



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

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

K=2 Classes

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

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

- *Log odds:*

$$\log \frac{P(C_1 \mid \boldsymbol{x})}{P(C_2 \mid \boldsymbol{x})}$$

Utility Theory

Make rational decisions in case of uncertainty

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

$$EU(\alpha_i \mid \boldsymbol{x}) = \sum_k U_{ik} P(S_k \mid \boldsymbol{x})$$

$$\text{Choose } \alpha_i \text{ if } EU(\alpha_i \mid \boldsymbol{x}) = \max_j EU(\alpha_j \mid \boldsymbol{x})$$

Value of Information

- Expected utility using $\boldsymbol{x}$ only

$$EU(\boldsymbol{x}) = \max_i \sum_k U_{ik} P(S_k \mid \boldsymbol{x})$$

- Expected utility using $\boldsymbol{x}$ and new feature z

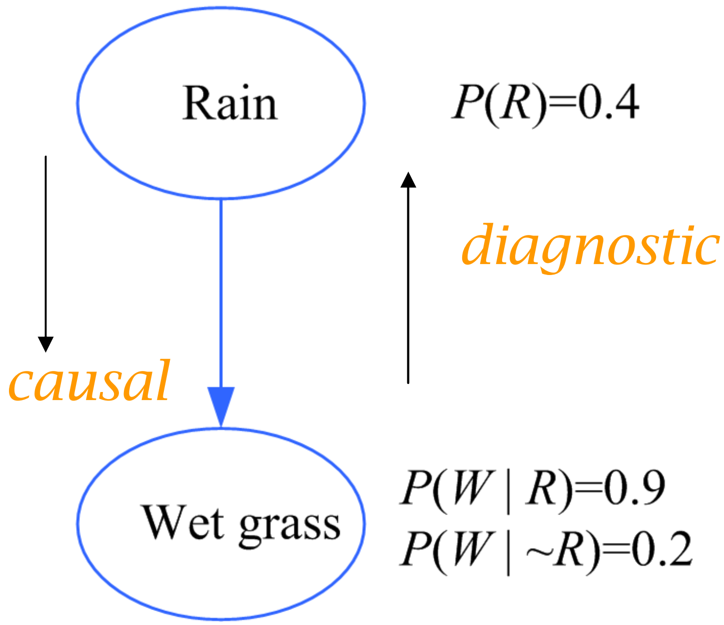
$$EU(\boldsymbol{x}, z) = \max_i \sum_k U_{ik} P(S_k \mid \boldsymbol{x}, z)$$

- z is useful if $EU(\boldsymbol{x}, z) > EU(\boldsymbol{x})$

Bayesian Networks

- Aka graphical models, probabilistic networks
- Nodes are hypotheses (random vars) and the prob corresponds to our belief in the truth of the hypothesis
- Arcs are direct direct influences between hypotheses
- The structure is represented as a directed acyclic graph (DAG)
- The parameters are the conditional probs in the arcs
- (Pearl, 1988, 2000; Jensen, 1996; Lauritzen, 1996)

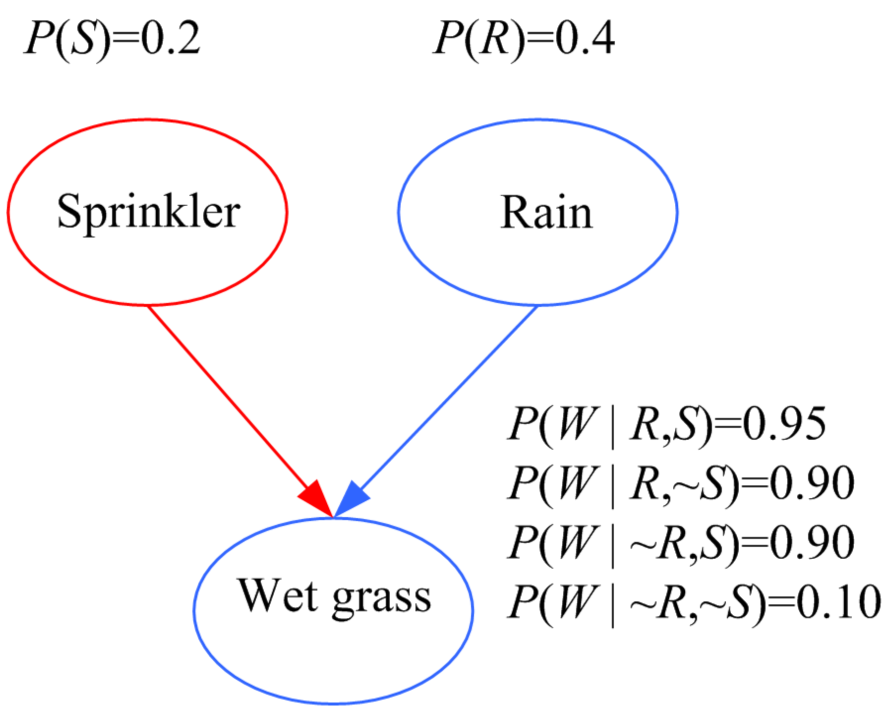
Causes and Bayes' Rule



Diagnostic inference:
Knowing that the grass is wet, what is the probability that rain is the cause?

$$
\begin{aligned}
P(R \mid W) &= \frac{P(W \mid R)P(R)}{P(W)} \\
&= \frac{P(W \mid R)P(R)}{P(W \mid R)P(R) + P(W \mid \sim R)P(\sim R)} \\
&= \frac{0.9 \times 0.4}{0.9 \times 0.4 + 0.2 \times 0.6} = 0.75
\end{aligned}
$$

Causal vs Diagnostic Inference



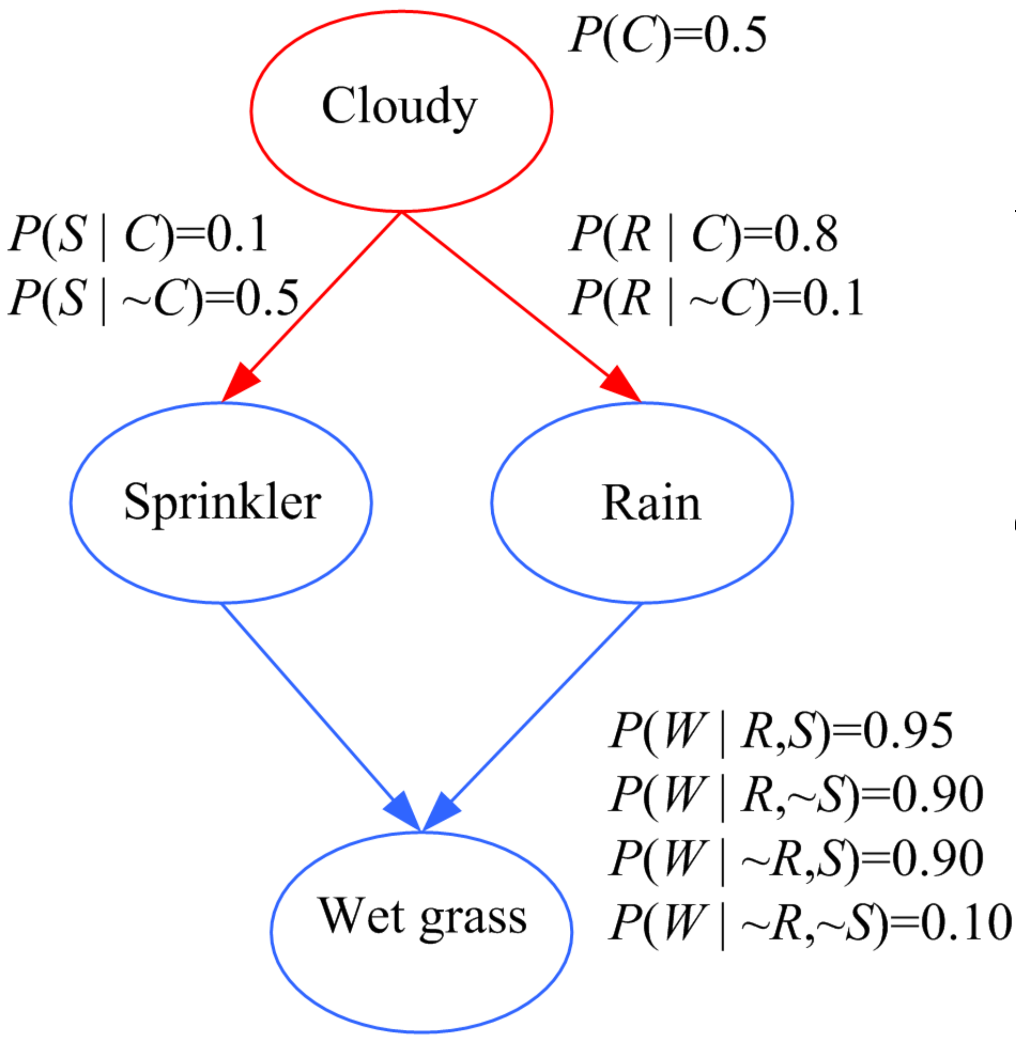
Causal inference: If the sprinkler is on, what is the probability that the grass is wet?

$$P(W|S) = P(W|R,S)\,P(R|S) + P(W|\sim R,S)\,P(\sim R|S)$$
$$= P(W|R,S)\,P(R) + P(W|\sim R,S)\,P(\sim R)$$
$$= 0.95\ 0.4 + 0.9\ 0.6 = 0.92$$

Diagnostic inference: If the grass is wet, what is the probability that the sprinkler is on? $P(S|W) = 0.35 > 0.2\ P(S)$

$P(S|R,W) = 0.21$ *Explaining away: Knowing that it has rained decreases the probability that the sprinkler is on.*

Bayesian Networks: Causes



Causal inference:

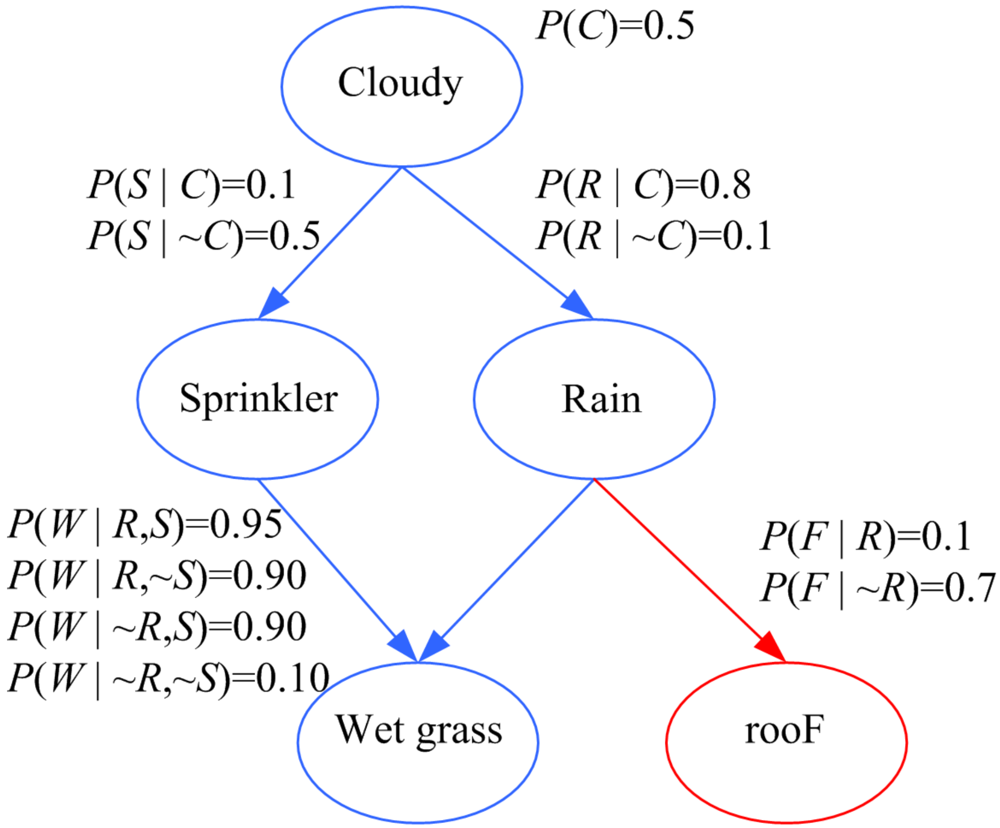
$$
\begin{aligned}
P(W|C) = {} & P(W|R,S)\, P(R,S|C) + \\
& P(W|\sim R,S)\, P(\sim R,S|C) + \\
& P(W|R,\sim S)\, P(R,\sim S|C) + \\
& P(W|\sim R,\sim S)\, P(\sim R,\sim S|C)
\end{aligned}
$$

and use the fact that

$$P(R,S|C) = P(R|C)\, P(S|C)$$

Diagnostic: $P(C|W) = ?$

Bayesian Nets: Local structure



$$P(F \mid C) = ?$$

$$P(C,S,R,W,F) = P(C)P(S \mid C)P(R \mid C)P(W \mid S,R)P(F \mid R)$$

$$P(X_1,\ldots X_d) = \prod_{i=1}^{d} P(X_i \mid \text{parents}(X_i))$$

Bayesian Networks: Inference

$P(C,S,R,W,F) = P(C)\, P(S|C)\, P(R|C)\, P(W|R,S)\, P(F|R)$

$P(C,F) = \Sigma_S\, \Sigma_R\, \Sigma_W\, P(C,S,R,W,F)$

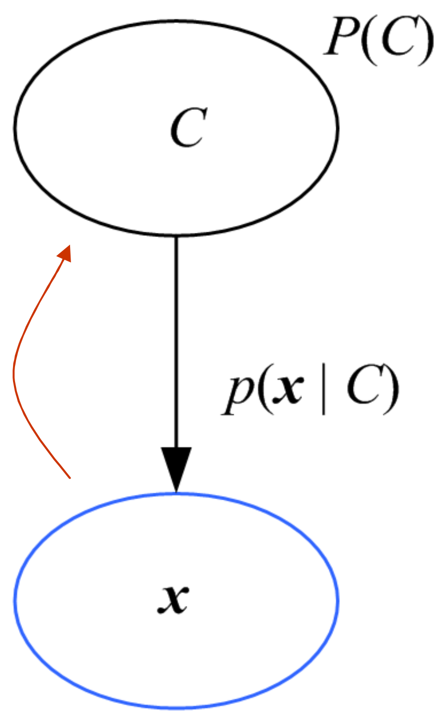
$P(F|C) = P(C,F) \,/\, P(C)$ *Not efficient!*

Belief propagation (Pearl, 1988)
Junction trees (Lauritzen and Spiegelhalter, 1988)

Bayesian Networks: Classification

diagnostic

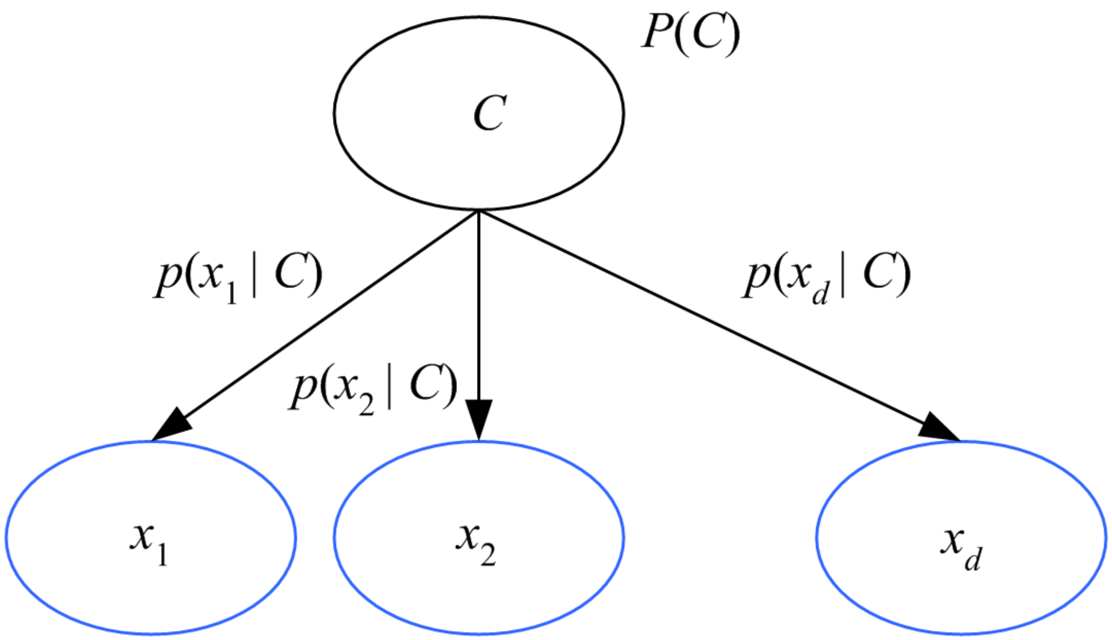
$P(C \mid x)$



Bayes' rule inverts the arc:

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

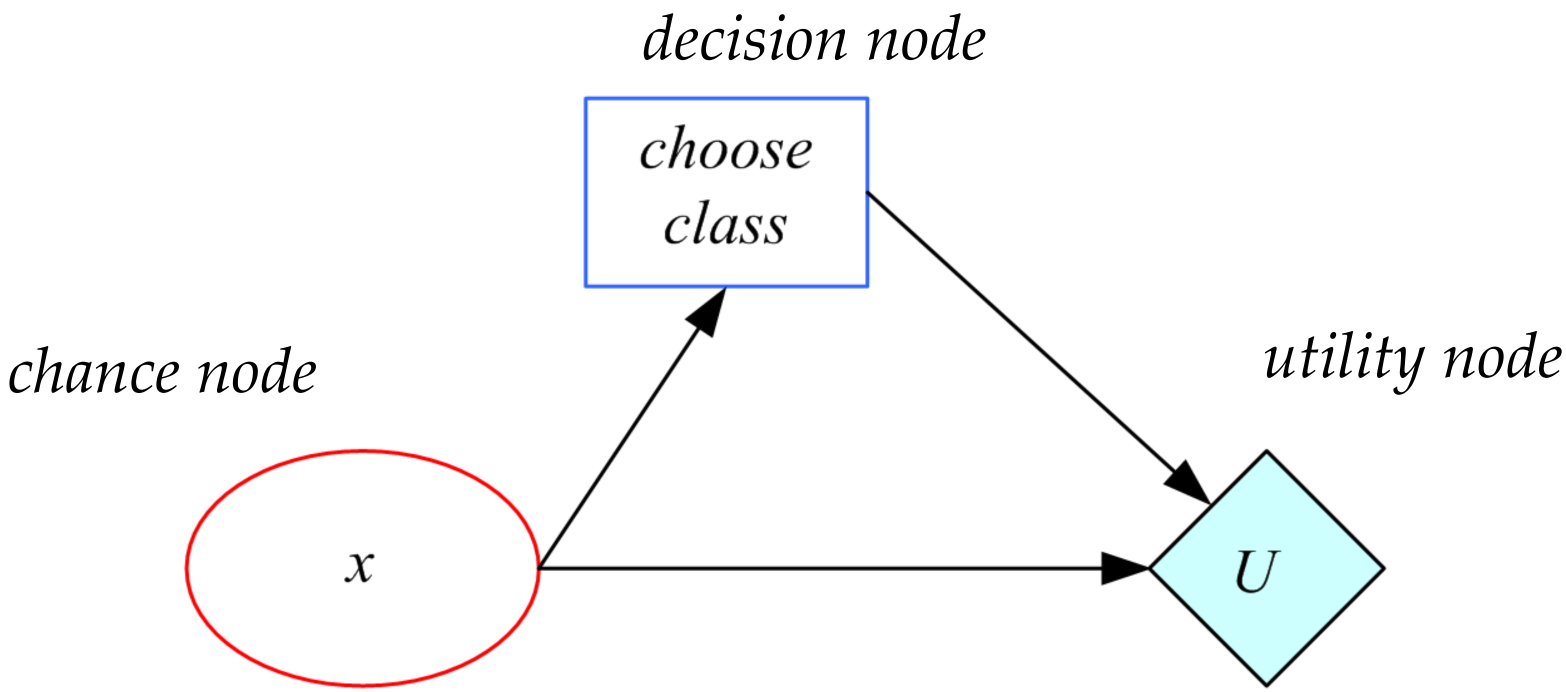
Naive Bayes' Classifier



Given C, x_j are independent:

$$p(\mathbf{x}|C) = p(x_1|C)\ p(x_2|C)\ \ldots\ p(x_d|C)$$

Influence Diagrams



Association Rules

- Association rule: $X \rightarrow Y$
- Support ($X \rightarrow Y$): Conditional probability

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

- Confidence ($X \rightarrow Y$): Statistical significance

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

$$= \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers who bought } X\}}$$

Apriori algorithm (Agrawal et al., 1996)