Simple Bayesian

Richard Xu

August 15, 2022

1 Introduction

The purpose of this note is to provide an intuitive explanation of the basic concepts of probability, Bayes' theorem, probabilitic graphical models, for students with or without a mathematics background, so they may not be as rigorous.

I also removed the integral and matrix-vector representations for the entire note. I used summation and scalar for this note. Therefore, there is little linear algebra or calculus requirements to understand this note.

2 Revision on distributions

2.1 Joint distributions

The following is a table of joint density $\Pr(X, Y)$:

	$Y=0$	$Y=1$	$Y=2$	$\Pr(X)$
$X=0$	0	$\frac{3}{15}$	$\frac{3}{15}$	$\frac{6}{15}$
$X=1$	$\frac{2}{15}$	$\frac{6}{15}$	0	$\frac{8}{15}$
$X=2$	$\frac{1}{15}$	0	0	$\frac{1}{15}$
$\Pr(Y)$	$\frac{3}{15}$	$\frac{9}{15}$	$\frac{3}{15}$	1

The above table shows:

$$\begin{aligned} \Pr(X, Y) &\equiv \Pr(X = x, Y = y) \\ \text{for example } \Pr(X = 1, Y = 1) &= \frac{6}{15} \end{aligned} \tag{1}$$

mini-exercise:

1. what is the probability $\Pr(X = 2, Y = 1)$
2. what is the probability $\Pr(X = 3, Y = 2)$
3. what is the value of:

$$\sum_{i=0}^{2} \sum_{j=0}^{2} \Pr(X = i, Y = j) \tag{2}$$

2.2 Marginal distributions

	$Y=0$	$Y=1$	$Y=2$	$\Pr(X)$
$X=0$	0	$\frac{3}{15}$	$\frac{3}{15}$	$\frac{6}{15}$
$X=1$	$\frac{2}{15}$	$\frac{6}{15}$	0	$\frac{8}{15}$
$X=2$	$\frac{1}{15}$	0	0	$\frac{1}{15}$
$\Pr(Y)$	$\frac{3}{15}$	$\frac{9}{15}$	$\frac{3}{15}$	1

Using **sum rule**, the **marginal distribution** tells us that:

$$\Pr(X=x) = \sum_{y \in \mathcal{S}_y} \Pr(x,y) \tag{3}$$

For example:

$$\begin{aligned} \Pr(Y=1) &= \sum_{i=0}^{2} \Pr(x=i, y=1) \\ &= \frac{3}{15} + \frac{6}{15} + \frac{0}{15} = \frac{9}{15} \end{aligned} \tag{4}$$

exercise what is $\Pr(X=2)$ and $\Pr(X=1)$?

2.3 Conditional distributions

	$Y=0$	$Y=1$	$Y=2$	$\Pr(X)$
$X=0$	0	$\frac{3}{15}$	$\frac{3}{15}$	$\frac{6}{15}$
$X=1$	$\frac{2}{15}$	$\frac{6}{15}$	0	$\frac{8}{15}$
$X=2$	$\frac{1}{15}$	0	0	$\frac{1}{15}$
$\Pr(Y)$	$\frac{3}{15}$	$\frac{9}{15}$	$\frac{3}{15}$	1

Conditional density:

$$\Pr(X|Y) = \frac{\Pr(X,Y)}{\Pr(Y)} = \frac{p(Y|X)p(X)}{p(Y)} = \frac{\Pr(Y|X)\Pr(X)}{\sum_X \Pr(Y|X)\Pr(X)} \tag{5}$$

What about $p(X|Y=y)$? Pick an example:

$$\begin{aligned} \Pr(X=1|Y=1) &= \frac{\Pr(X=1, Y=1)}{p(Y=1)} \\ &= \frac{6/15}{9/15} = \frac{2}{3} \end{aligned} \tag{6}$$

In another words, you're looking at the $(Y=1)$ column, which is $\begin{bmatrix} \frac{3}{15} & \frac{6}{15} & 0 \end{bmatrix}$ and to compute the ratio of the joint density $\Pr(X=1, Y=1) = \frac{3}{15}$ on the $(Y=1)$ column.

2.3.1 quick Exercise on Conditional distributions

1. **exercise**: What is $\Pr(X=0|Y=1)$?
2. **exercise**: What is $\Pr(X=1|Y=0)$?

2.3.2 reference to association mining rule

You need to remind yourself we have looked at conditional density when we discuss "confidence" in Association Rule Mining, where in order to compute the confidence of product Y given product X:

$$\begin{aligned} \text{confidence}(\{Y\} \rightarrow \{X\}) = \Pr(\{X\}|\{Y\}) &= \frac{\Pr(\{X\},\{Y\})}{\Pr(\{Y\})} \\ &= \frac{\text{support}(\{X,Y\})}{\text{support}(\{Y\})} \end{aligned} \tag{7}$$

2.4 Independence

If X and Y are independent, the both of these are true:

$$\begin{aligned} \Pr(X|Y) &= \Pr(X) \\ \Pr(X,Y) &= \Pr(X)\Pr(Y) \end{aligned} \tag{8}$$

Both of these properties are true, when X and Y being independent:

$$\begin{aligned} \Pr(X|Y) &= \frac{\Pr(X,Y)}{\Pr(Y)} \\ &= \frac{\Pr(X)\Pr(Y)}{\Pr(Y)} \quad \text{substitute } \Pr(X,Y) = \Pr(X)\Pr(Y) \\ &= \Pr(X) \quad \text{here you proved } \Pr(X|Y) = \Pr(X) \end{aligned} \tag{9}$$

	$Y=0$	$Y=1$	$Y=2$	$\Pr(X)$
$X=0$	0	$\frac{3}{15}$	$\frac{3}{15}$	$\frac{6}{15}$
$X=1$	$\frac{2}{15}$	$\frac{6}{15}$	0	$\frac{8}{15}$
$X=2$	$\frac{1}{15}$	0	0	$\frac{1}{15}$
$\Pr(Y)$	$\frac{3}{15}$	$\frac{9}{15}$	$\frac{3}{15}$	1

X and Y are **not** independent (not product of marginals)

	$Y=0$	$Y=1$	$Y=2$	$\Pr(X)$
$X=0$	$\frac{18}{225}$	$\frac{54}{225}$	$\frac{18}{225}$	$\frac{6}{15}$
$X=1$	$\frac{24}{225}$	$\frac{72}{225}$	$\frac{24}{225}$	$\frac{8}{15}$
$X=2$	$\frac{3}{225}$	$\frac{9}{225}$	$\frac{3}{225}$	$\frac{1}{15}$
$\Pr(Y)$	$\frac{3}{15}$	$\frac{9}{15}$	$\frac{3}{15}$	1

X and Y are independent (is the product of marginals)

2.4.1 reference to χ^2 testing

You need to remind yourself we have used independence of two random variables when we look at χ^2 testing, where the **null hypothesis** is that X and Y are independent, so their joint density is just $\Pr(X,Y) = \Pr(X)\Pr(Y)$.

2.4.2 reference to association mining rule

You need to remind yourself we have looked independence when we discuss "lift" in Association Rule Mining, where in order to compute the lift of product X given product Y:

$$
\begin{aligned}
\text{lift}(\{Y\} \to \{X\}) &= \frac{\text{confidence}(\{Y\} \to \{X\})}{\text{support}(\{X\})} \\
&= \frac{\frac{\Pr(\{X\},\{Y\})}{\Pr(\{Y\})}}{\Pr(\{X\})} \\
&= \frac{\Pr(\{X\},\{Y\})}{\Pr(\{X\})\Pr(\{Y\})} \quad \text{test for independence: (= 1 when independent)} \\
&= \frac{\text{support}(\{X,Y\})}{\text{support}(\{X\})\text{support}(\{Y\})}
\end{aligned}
\tag{10}
$$

2.5 Conditional Independence

Imagine we have three random variables: X, Y and Z:

Once we know Z, then knowing Y does NOT tell us any additional information about X

Therefore:

$$
\Pr(X|Y,Z) = \Pr(X|Z) \tag{11}
$$

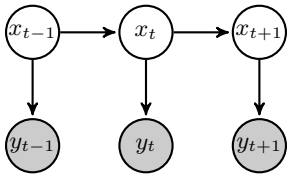
This means that X is conditionally independent of Y given Z.

If $\Pr(X|Y,Z) = \Pr(X|Z)$, then what about $\Pr(X,Y|Z)$?

$$
\begin{aligned}
\Pr(X,Y|Z) &= \frac{\Pr(X,Y,Z)}{\Pr(Z)} = \frac{\Pr(X|Y,Z)\Pr(Y,Z)}{\Pr(Z)} \\
&= \Pr(X|Y,Z)\Pr(Y|Z) \\
&= \Pr(X|Z)\Pr(Y|Z) \quad \text{apply } \Pr(X|Y,Z) = \Pr(X|Z)
\end{aligned}
\tag{12}
$$

2.6 An example of Conditional Independence

You may study **Dynamic model** in some stage:



From this model, we can see:

$$
\begin{aligned}
p(x_t|x_1,\ldots,x_{t-1},y_1,\ldots,y_{t-1}) &= p(x_t|x_{t-1}) \\
p(y_t|x_1,\ldots,x_{t-1},x_t,y_1,\ldots,y_{t-1}) &= p(y_t|x_t)
\end{aligned}
\tag{13}
$$

2.7 Expectation of Joint probabilities

Given that X, Y are two random variables:

Discrete case:

$$
\mathbb{E}[f(X,Y)] = \sum_{i=1}^{N_i}\sum_{j=1}^{N_j} f(X=i,Y=j)\Pr(X=i,Y=j) \tag{14}
$$

	$Y=1$	$Y=2$	$Y=3$
$X=1$	0	$\frac{3}{15}$	$\frac{3}{15}$
$X=2$	$\frac{2}{15}$	$\frac{6}{15}$	0
$X=3$	$\frac{1}{15}$	0	0

$\Pr(X,Y)$

	$Y=1$	$Y=2$	$Y=3$
$X=1$	6	7	8
$X=2$	3	6	2
$X=3$	1	8	6

$f(X,Y)$

$$\begin{aligned}\mathbb{E}[f(X,Y)] &= \sum_{i=1}^{N_i}\sum_{j=1}^{N_j} f(X=i,Y=j)p(X=i,Y=j) \\ &= 6\times 0 + 7\times\frac{3}{15} + 8\times\frac{3}{15} + 3\times\frac{2}{15} + 6\times\frac{6}{15} \\ &\quad + 2\times 0 + 1\times\frac{1}{15} + 8\times 0 + 6\times 0\end{aligned} \tag{15}$$

2.8 Conditional Expectation

often we use conditional expectation $\mathbb{E}[Y|X] = \sum_{Y\in\mathcal{S}_y} y\Pr(Y|X)$:

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{Y\in\mathcal{S}_y} y\Pr(Y) = \sum_{Y\in\mathcal{S}_y} y\sum_{X\in\mathcal{S}_x}\Pr(Y,X) \\ &= \sum_{X\in\mathcal{S}_x}\underbrace{\sum_{Y\in\mathcal{S}_y} y\Pr(Y|X)}_{\mathbb{E}[Y|X]}\Pr(X) \\ &= \sum_{X\in\mathcal{S}_x}\mathbb{E}[Y|X]\Pr(X)\end{aligned} \tag{16}$$

3 Revisit Bayes Theorem

Instead of using arbitrary random variable X and Y, in data mining and machine learning, very commonly:

1. θ for model parameter: $\mathcal{S}_\theta$ is the parameter space of θ:
2. $X = x_1, \ldots x_n$ for dataset: $\mathcal{S}_X$ is the sample space of X

$$\underbrace{\Pr(\theta|X)}_{\text{posterior}} = \frac{\overbrace{\Pr(X|\theta)}^{\text{likelihood}}\overbrace{\Pr(\theta)}^{\text{prior}}}{\underbrace{\Pr(X)}_{\text{normalization constant}}} = \frac{\Pr(X|\theta)\Pr(\theta)}{\sum_{\theta\in\mathcal{S}_\theta}\Pr(X|\theta)p(\theta)} \tag{17}$$

Here I will use three different cases (I was going to draw the same diagram I had drawn during classes, but it will take a long time do so so in Latex!) to illustrate Bayesian idea when:

1. $\mathcal{S}_X$ is discrete, and $\mathcal{S}_\theta$ is discrete
2. $\mathcal{S}_X$ is continuous, and $\mathcal{S}_\theta$ is discrete ()
3. $\mathcal{S}_X$ is continuous, and $\mathcal{S}_\theta$ is continuous ()

3.1 $\mathcal{S}_X$ is discrete, and $\mathcal{S}_\theta$ is discrete

3.1.1 setting

The setting: Imagine in a cyber-security problem, out of all the TCP connections (say millions), 1% of which are intrusions:

1. When there is an intrusion, the probability of system sends alarm is 87%.
2. When there is no intrusion, the probability of system sends alarm is 6%.

3.1.2 Prior probability

1% of which are intrusions

$$\begin{aligned} \Pr(\theta = \text{intrusion}) &= 0.01 \\ \Pr(\theta = \text{no intrusion}) &= 0.99 \end{aligned} \tag{18}$$

3.1.3 Likelihood probability

given intrusion occur, probability of system sends alarm is 87%:

$$\begin{aligned} \Pr(X = \text{alarm}|\theta = \text{intrusion}) &= 0.87 \\ \Pr(X = \text{no alarm}|\theta = \text{intrusion}) &= 0.13 \end{aligned} \tag{19}$$

given there is no intrusion, the probability of system sends alarm is 6%:

$$\begin{aligned} \Pr(X = \text{alarm}|\theta = \text{no intrusion}) &= 0.06 \\ \Pr(X = \text{no alarm}|\theta = \text{no intrusion}) &= 0.94 \end{aligned} \tag{20}$$

3.1.4 Computing the Posterior

Given the prior and likelihood, we can now compute **posterior probability**: $\Pr(\theta|X)$:
There 2 two possible values for parameter θ and 2 possible observation X
Therefore, there are 4 **probabilities** we need to compute, using the TP,FP,TN,FN paradigm, we have:

1. **True Positive** When system sends alarm, probability of an intrusion occurs:

$$\Pr(\theta = \text{intrusion}|X = \text{alarm}) \tag{21}$$

2. **False Positive** When system sends alarm, probability that there is no intrusion:

$$\Pr(\theta = \text{no intrusion}|X = \text{alarm}) \tag{22}$$

3. **True Negative** When system sends no alarm, probability that there is no intrusion:

$$\Pr(\theta = \text{no intrusion}|X = \text{no alarm}) \tag{23}$$

4. **False Negative** When system sends no alarm, probability that an intrusion occurs:

$$\Pr(\theta = \text{intrusion}|X = \text{no alarm}) \tag{24}$$

Apply Bayes Theorem in this setting:

$$
\begin{aligned}
&\Pr(\theta|X) \\
&= \frac{\Pr(X|\theta)\Pr(\theta)}{\Pr(X)} \\
&= \frac{\Pr(X|\theta)\Pr(\theta)}{\sum_{\theta' \in \mathcal{S}_\theta} \Pr(X|\theta')\Pr(\theta')} \quad \text{why I used } \theta' \text{ instead?} \\
&= \frac{\Pr(X|\theta)\Pr(\theta)}{\Pr(X|\theta = \text{Intrusion})\Pr(\theta = \text{Intrusion}) + \Pr(X|\theta = \text{no intrusion})\Pr(\theta = \text{no intrusion})}
\end{aligned} \tag{25}
$$

substitute numbers in!

1. **True Positive rate** When system has "alarm", what is probability of an "intrusion":

$$
\begin{aligned}
&\Pr(\theta = \text{intrusion}|X = \text{alarm}) \\
&= \frac{\Pr(X = \text{alarm}|\theta = \text{intrusion})\Pr(\theta = \text{intrusion})}{\Pr(X = \text{alarm})} \\
&= \frac{\Pr(X = \text{alarm}|\theta = \text{intrusion})\Pr(\theta = \text{intrusion})}{\Pr(X = \text{alarm}|\theta = \text{Intrusion})\Pr(\theta = \text{Intrusion}) + \Pr(X = \text{alarm}|\theta = \text{no intrusion})\Pr(\theta = \text{Intrusion})} \\
&= \frac{0.87 \times 0.01}{0.87 \times 0.01 + 0.06 \times 0.99} = 0.1278
\end{aligned} \tag{26}
$$

2. **False Positive rate** When system has "alarm", what is probability of "no intrusion":

$$
\begin{aligned}
&\Pr(\theta = \text{no intrusion}|X = \text{alarm}) \\
&= \frac{\Pr(X = \text{alarm}|\theta = \text{no intrusion})\Pr(\theta = \text{no intrusion})}{\Pr(X = \text{alarm})} \\
&= \frac{\Pr(X = \text{alarm}|\theta = \text{no intrusion})\Pr(\theta = \text{no intrusion})}{\Pr(X = \text{alarm}|\theta = \text{no intrusion})\Pr(\theta = \text{no intrusion}) + \Pr(X = \text{alarm}|\theta = \text{no intrusion})\Pr(\theta = \text{no intrusion})} \\
&= \frac{0.06 \times 0.99}{0.87 \times 0.01 + 0.06 \times 0.99} = 0.8722
\end{aligned} \tag{27}
$$

3. **False Negative** When system has "no alarm", what is probability an intrusion occurs?

$$
\begin{aligned}
&\Pr(\theta = \text{intrusion}|X = \text{no alarm}) \\
&= \frac{\Pr(X = \text{no alarm}|\theta = \text{intrusion})p(\theta = \text{intrusion})}{\Pr(X = \text{no alarm})} \\
&= \frac{\Pr(X = \text{no alarm}|\theta = \text{intrusion})p(\theta = \text{intrusion})}{\Pr(X = \text{no alarm}|\theta = \text{Intrusion})\Pr(\theta = \text{Intrusion}) + \Pr(X = \text{no alarm}|\theta = \text{no intrusion})\Pr(\theta = \text{no Intrusion})} \\
&= \frac{0.13 \times 0.01}{0.13 \times 0.01 + 0.94 \times 0.99} = 0.0014
\end{aligned} \tag{28}
$$

4. **True Negative** When system has "no alarm", what is probability there is "no intrusion"?

$$
\begin{aligned}
&\Pr(\theta = \text{no intrusion} | X = \text{no alarm}) \\
&= \frac{\Pr(X = \text{no alarm} | \theta = \text{no intrusion}) \Pr(\theta = \text{no intrusion})}{\Pr(X = \text{no intrusion})} \\
&= \frac{\Pr(X = \text{no alarm} | \theta = \text{no intrusion}) \Pr(\theta = \text{no intrusion})}{\Pr(X = \text{no alarm} | \theta = \text{no intrusion}) \Pr(\theta = \text{no intrusion}) + \Pr(X = \text{no alarm} | \theta = \text{no intrusion}) \Pr(\theta = \text{no intrusion})} \\
&= \frac{0.94 \times 0.99}{0.87 \times 0.001 + 0.06 \times 0.99} = 0.9986
\end{aligned} \tag{29}
$$

3.2 $\mathcal{S}_X$ is continuous, and $\mathcal{S}_\theta$ is discrete

let's have two classes:

$$
\begin{cases} \text{class 1: model with parameter} & \theta_1 : \quad \text{e.g. } \mathcal{N}(\theta_1, \sigma^2) \\ \text{class 2: model with parameter} & \theta_2 : \quad \text{e.g. } \mathcal{N}(\theta_2, \sigma^2) \end{cases} \tag{30}
$$

3.2.1 Prior probability

$$
\begin{aligned}
\Pr(\text{class 1}) &= \frac{1}{3} \\
\Pr(\text{class 2}) &= \frac{2}{3}
\end{aligned} \tag{31}
$$

3.2.2 Likelihood probability

$$
\begin{aligned}
p(X = x \,|\text{class 1}) &= \mathcal{N}(x; \theta_1, \sigma^2) \\
p(X = x \,|\text{class 2}) &= \mathcal{N}(x; \theta_2, \sigma^2)
\end{aligned} \tag{32}
$$

where $\mathcal{N}(x; \theta, \sigma)$ meaning the Probability Density function evaluated at x for a Gaussian with mean θ, and variance σ^2:

3.2.3 Computing the Posterior

$$
\begin{aligned}
\Pr(\text{class 1} \,|X = x) &= \frac{\frac{1}{3}\mathcal{N}(x; \theta_1, \sigma^2)}{\Pr(X = x)} \\
&= \frac{\frac{1}{3}\mathcal{N}(x; \theta_1, \sigma^2)}{\frac{1}{3}\mathcal{N}(x; \theta_1, \sigma^2) + \frac{2}{3}\mathcal{N}(x; \theta_2, \sigma^2)}
\end{aligned} \tag{33}
$$

$$
\begin{aligned}
\Pr(\text{class 2} \,|X = x) &= \frac{\frac{2}{3}\mathcal{N}(x; \theta_2, \sigma^2)}{\Pr(X = x)} \\
&= \frac{\frac{2}{3}\mathcal{N}(x; \theta_2, \sigma^2)}{\frac{1}{3}\mathcal{N}(x; \theta_1, \sigma^2) + \frac{2}{3}\mathcal{N}(x; \theta_2, \sigma^2)}
\end{aligned} \tag{34}
$$

3.3 $\mathcal{S}_X$ is continuous, and $\mathcal{S}_\theta$ is continuous

when $\mathcal{S}_\theta$ is continuous, then we have infinite number of classes (no longer a classification problem). Think about the case where we know the data is distributed from a Gaussian with unknown mean θ, but fixed variance σ^2:

$$
x \sim \mathcal{N}(\theta, \sigma^2) \tag{35}
$$

3.3.1 Prior probability

$$p(\theta) = \mathcal{N}(\theta; \mu_0, \sigma_0^2) \tag{36}$$

3.3.2 Likelihood probability

$$P(X = x \mid \theta) = \mathcal{N}(x; \theta, \sigma^2) \tag{37}$$

where $\mathcal{N}(x; \theta, \sigma)$ meaning the Probability Density function evaluated at x for a Gaussian with mean θ, and variance σ^2:

3.3.3 Computing the Posterior

$$\begin{aligned} \Pr(\theta \mid X = x) &= \frac{\mathcal{N}(x; \theta, \sigma^2)\mathcal{N}(\theta; \mu_0, \sigma_0^2)}{\Pr(X = x)} \\ &= \frac{\mathcal{N}(x; \theta, \sigma^2)\mathcal{N}(\theta; \mu_0, \sigma_0^2)}{\int_{\mathcal{S}_\theta} \mathcal{N}(x; \theta, \sigma^2)\mathcal{N}(\theta; \mu_0, \sigma_0^2) \mathrm{d}\theta} \end{aligned} \tag{38}$$

3.4 Decision based on Probabilities of multiple classes

Let us have multiple classes $\{1, 2, \ldots, m\}$, we defined our prior and likelihood:

prior: $\Pr(\theta = i)$ and of course, $\sum_{i=1}^{m} \Pr(\theta = i) = 1$ likelihood: $\Pr(X|\theta = i)$ so we can compute the posterior:

$$\begin{aligned} \Pr(\theta = i|X) &= \frac{\Pr(X|\theta_i)\Pr(\theta_i)}{\Pr(X)} \\ &= \frac{\Pr(X|\theta_i)\Pr(\theta_i)}{\sum_{j=1}^{m} \Pr(X|\theta_j)\Pr(\theta_j)} \quad \text{denominator index change to } j? \end{aligned} \tag{39}$$

how do we make an inference of a new data x^*?

$$\hat{y}^* = \arg\max\{\Pr(\theta = i|X = x^*)\}_{i=1}^{m} \tag{40}$$

it means to find which i gives largest posterior distribution.

For example, if you have the case:

$$\begin{cases} \Pr(\theta = 1|x^*) & = \frac{1}{8} \\ \Pr(\theta = 2|x^*) & = \frac{3}{4} \\ \Pr(\theta = 3|x^*) & = \frac{1}{8} \end{cases} \tag{41}$$

then you pick class 2 is the winner of the classifier for data x^*

3.5 Bayesian Predictive distribution

Combine marginal distribution and Conditional Independence together:

Very often, in machine learning, you want to compute the probability of new data y^* given training data Y, i.e., $p(y^*|Y)$. You assume there are some model explains both Y and y^*. The model parameter is θ.

$$\Pr(y^*|Y) = \sum_{\theta \in \mathcal{S}_\theta} \Pr(y^*|\theta)\Pr(\theta|Y) \tag{42}$$

Exercise: what assumption used in the above? Let me show you the derivation:

$$\begin{aligned}\Pr(y^*|Y) &= \sum_{\theta \in \mathcal{S}_\theta} \Pr(y^*, \theta|Y) \\ &= \sum_{\theta \in \mathcal{S}_\theta} \Pr(y^*|\theta, Y) \Pr(y^*|\theta, Y)\end{aligned} \tag{43}$$

now we apply the assumption $\Pr(y^*|\theta, Y) = \Pr(y^*|\theta)$, which makes sense, as given the model θ, we do not need training data Y any more to make inference on y^*:

$$\Pr(y^*|Y) = \sum_{\theta \in \mathcal{S}_\theta} \Pr(y^*|\theta) \Pr(y^*|\theta, Y) \tag{44}$$

4 Bayesian Belief Networks

Also known as belief networks, Bayesian networks, and probabilistic networks. Bayesian Belief network describes conditional dependence among subsets of variables (attributes). It combines prior knowledge about dependencies among variables with observed training data.

It is expressed in terms of a Probabilistic Graphical Model

4.1 Probabilistic Graphical Model

Express marginal distribution and Conditional Independence into a graph, using **Directed acyclic graph (DAG)**

1. **nodes**

   May be discrete- or continuous-valued, and Conditional Independent of every other nodes given its parents (or Markov Blanket):

$$\Pr\big(Y \mid \text{every variables}\big) = \Pr\big(Y \mid \text{parents}(Y)\big) \tag{45}$$

2. **edges**

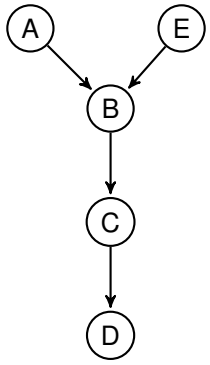
   Representation of causal knowledge, where the direction on arc representing causality

Let's see a real-life research paper (and see Figure 2 of the paper)
`https://www.ijcai.org/Proceedings/16/Papers/210.pdf`

4.2 Bayesian Belief Networks

Let's work through a simple illustrations containing the following nodes:

4.2.1 a practice example



$$\begin{aligned}&\Pr\left(A,B,C,D,E\right)\\&=\Pr\left(D|C\right)\Pr\left(C|B\right)\Pr\left(B|A,C\right)\end{aligned}\tag{46}$$

note that without this problem-specific graphical model (and all the conditional Independence it specifies), the Bayes rule tells you:

$$\begin{aligned}&\Pr\left(A,B,C,D,E\right)\\&=\Pr\left(D|A,B,C,E\right)\Pr\left(C|A,B,E\right)\Pr\left(B|A,E\right)\Pr\left(A|E\right)\Pr\left(E\right)\end{aligned}\tag{47}$$

of course, there are multiple ways (in here you have $5!$ ways) you can factorize the above.

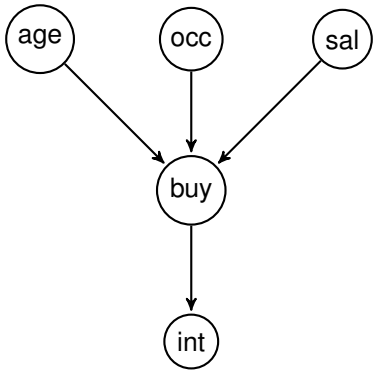
4.2.2 a somewhat "real-world" example

1. "Age", "Occupation" and "Salary" determine if customer will "buy" this product.
2. Given that customer "buys" product, whether he is interested in insurance

summarize all the abbreviation:

1. "Age" (age)
2. "Occupation" (occ)
3. "Salary" (sal)
4. "interested" (int)
5. "buy" (buy)

first, we write down its probabilistic graphical model, which illustrate all the conditional independence:



for example, we see that:

$$\Pr\left(\text{int}\mid\text{every nodes}\right)=\Pr\left(\text{int}\mid\text{buy}\right)\tag{48}$$

$$\begin{aligned}&\Pr\left(\text{age},\text{occ},\text{sal},\text{buy},\text{int}\right)\\&=\Pr\left(\text{age}\right)\Pr\left(\text{occ}\right)\Pr\left(\text{sal}\right)\Pr\left(\text{buy}\mid\text{age},\text{occ},\text{sal}\right)\Pr\left(\text{int}\mid\text{buy}\right)\end{aligned}\tag{49}$$

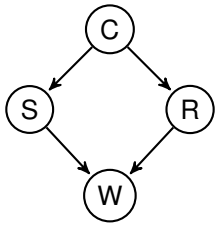
note that without this problem-specific graphical model (and all the conditional Independence it specifies), the Bayes rule tells you:

$$\begin{aligned}&\Pr\left(\text{age},\text{occ},\text{sal},\text{buy},\text{int}\right)\\&=\Pr\left(\text{age}|\text{occ},\text{sal},\text{buy},\text{int}\right)\Pr\left(\text{occ}|\text{sal},\text{buy},\text{int}\right)\Pr\left(\text{sal}|\text{buy},\text{int}\right)\Pr\left(\text{buy}|\text{int}\right)\Pr\left(\text{int}\right)\end{aligned}\tag{50}$$

4.3 a numerical example of Bayesian Belief Networks

Cloundy(C) Sprinker (S) Rain(R) Wet Grass (W)

4.3.1 corresponding probabilistic graphical model



$$\Pr\left(C,S,R,W\right)=\Pr\left(C\right)\Pr\left(S|C\right)\Pr\left(R|C\right)\Pr\left(W|S,R\right) \tag{51}$$

therefore, we must need to define the four probabilities $\Pr\left(C\right)$, $\Pr\left(S|C\right)$, $\Pr\left(R|C\right)$ and $\Pr\left(W|S,R\right)$

For illustration purposes, in this subject we will assume the use of categorical variables with only a few values. Instead of expressing them in tabular form as per lecture notes, it is much clearer to write out their equations:

4.3.2 $\Pr(\mathbf{C})$

$$\begin{aligned}\Pr(\mathrm{C}=F)&=0.5\\ \Pr(\mathrm{C}=T)&=0.5\end{aligned} \tag{52}$$

$\Pr(\mathrm{C}=F)$	$\Pr(\mathrm{C}=T)$
0.5	0.5

4.3.3 $\Pr(\mathbf{S} \mid \mathbf{C})$

$$\begin{aligned}\Pr(\mathrm{S}=F|\mathrm{C}=F)&=0.5\\ \Pr(\mathrm{S}=T|\mathrm{C}=F)&=0.5\\ \Pr(\mathrm{S}=F|\mathrm{C}=T)&=0.9\\ \Pr(\mathrm{S}=T|\mathrm{C}=T)&=0.1\end{aligned} \tag{53}$$

$\Pr(\mathrm{C})$	$\Pr(\mathrm{S}=F)$	$\Pr(\mathrm{S}=T)$
F	0.5	0.5
T	0.9	0.1

4.3.4 $\Pr(\mathbf{R} \mid \mathbf{C})$

$$\begin{aligned}\Pr(\mathrm{R}=F|\mathrm{C}=F)&=0.8\\ \Pr(\mathrm{R}=T|\mathrm{C}=F)&=0.2\\ \Pr(\mathrm{R}=F|\mathrm{C}=T)&=0.2\\ \Pr(\mathrm{R}=T|\mathrm{C}=T)&=0.8\end{aligned} \tag{54}$$

$\Pr(\mathrm{C})$	$\Pr(\mathrm{R}=F)$	$\Pr(\mathrm{R}=T)$
F	0.8	0.2
T	0.2	0.8

4.3.5 $\Pr(\mathbf{W} \mid \mathbf{S}, \mathbf{R})$

$$
\begin{aligned}
&\Pr(\mathrm{W} = F|\mathrm{S} = F, \mathrm{R} = F) = 1.0 \\
&\Pr(\mathrm{W} = F|\mathrm{S} = T, \mathrm{R} = F) = 0.1 \\
&\Pr(\mathrm{W} = F|\mathrm{S} = F, \mathrm{R} = T) = 0.1 \\
&\Pr(\mathrm{W} = F|\mathrm{S} = T, \mathrm{R} = T) = 0.01 \\
&\Pr(\mathrm{W} = T|\mathrm{S} = F, \mathrm{R} = F) = 0.0 \\
&\Pr(\mathrm{W} = T|\mathrm{S} = T, \mathrm{R} = F) = 0.9 \\
&\Pr(\mathrm{W} = T|\mathrm{S} = F, \mathrm{R} = T) = 0.9 \\
&\Pr(\mathrm{W} = T|\mathrm{S} = T, \mathrm{R} = T) = 0.99
\end{aligned}
\tag{55}
$$

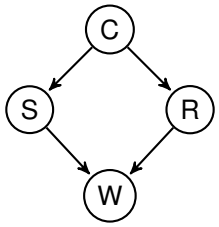
$\Pr(\mathrm{S})$	$\Pr(\mathrm{R})$	$\Pr(\mathrm{W} = F)$	$\Pr(\mathrm{W} = T)$
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

4.3.6 try plugging some numbers

$$
\begin{aligned}
&\Pr\left(C = F, S = F, R = T, W = T\right) \\
&= \Pr\left(C = F\right) \Pr\left(S = F|C = F\right) \\
&\quad \times \Pr\left(R = T|C = F\right) \Pr\left(W = T|S = F, R = T\right) \\
&= 0.5 \times 0.5 \times 0.2 \times 0.9
\end{aligned}
\tag{56}
$$

4.4 Bayesian Belief Networks Learning

using the above graphical model again, i.e.,:



However, this time we don't know what their probabilities are, i.e. we assume type of distribution, but we don't know exactly what parameters θ_C, $\theta_{S|C}$, $\theta_{R|C}$, $\theta_{W|S,R}$ are associated with these probabilities, so you need to learn what they are:

Let's first write down the joint probability:

$$
\begin{aligned}
&\Pr\big(\underbrace{C, S, R,}_{\text{input}} \underbrace{W}_{\text{output}}\big) \\
=&\Pr_{\theta_C}\left(C\right) \times \Pr_{\theta_{S|C}}\left(S|C\right) \times \Pr_{\theta_{R|C}}\left(R|C\right) \times \Pr_{\theta_{W|S,R}}\left(W|S,R\right)
\end{aligned}
\tag{57}
$$

it's obvious we need to maximize the joint probabilities of the input and output! This is similar when you perform maximum likelihood estimation:

$$
\arg\max_{\theta_C, \theta_{S|C}, \theta_{R|C}, \theta_{W|S,R}} \Big(\Pr_{\theta_C}\left(C\right) \times \Pr_{\theta_{S|C}}\left(S|C\right) \times \Pr_{\theta_{R|C}}\left(R|C\right) \times \Pr_{\theta_{W|S,R}}\left(W|S,R\right)\Big) \tag{58}
$$

Using the fact that $\log()$ is a monotonically increasing function, i.e.,

$$\Pr(X) \geq \Pr(Y) \implies \log(\Pr(X) \geq \log(\Pr(Y) \tag{59}$$

Also many common distributions such as Gaussian has $\exp(\cdot)$ term, so taking $\log(\cdot)$ actually simplifies things!

$$\implies \arg\max_{\theta_C, \theta_{S|C}, \theta_{R|C}, \theta_{W|S,R}} \Big(\log \Pr_{\theta_C}(C) + \log \Pr_{\theta_{S|C}}(S|C) + \log \Pr_{\theta_{R|C}}(R|C) + \log \Pr_{\theta_{W|S,R}}(W|S,R) \Big) \tag{60}$$

Often, there is no close-form (or analytical) solution, so you have to use numerical approximations such as gradient descent.