Chapter 2

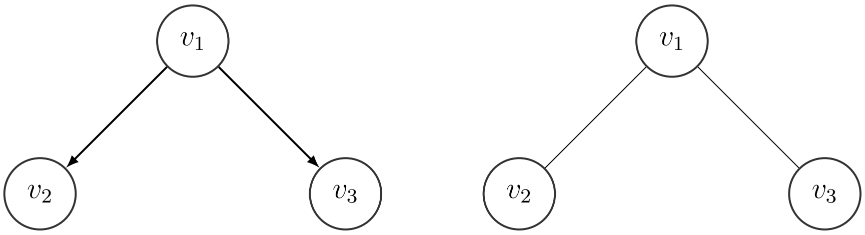
Bayesian networks

2.1 Graphs

A graph $\mathcal{G}$ consists of a set of vertices (or nodes) V and an edge set E. Roughly speaking, an edge is a link or an arrow connecting two nodes. A first classification of graphs is between **directed** and **undirected** graphs: in the former the edges are directed pairs (Figure 2.1a), in the latter they are undirected pairs (Figure 2.1b). If two nodes are connected they are said to be **adjacent** and the set of all the adjacent nodes to a given one is its **neighbourhood** (of order 1). By using nothing but the definitions introduced so far, we see that up to a permutation of the node labels, a graph can be uniquely represented by its **adjacency** matrix: a square matrix of order $|V|$ whose entry (i, j) is one if v_i is connected with v_j, zero otherwise. For instance, the adjacency matrix associated with the directed graph in Figure 2.1a is

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{2.1}$$

where the firs row/column corresponds to v_1, the second to v_2 and the third to v_3. Clearly the adjacency matrix of an undirected graph is symmetric while the one of a directed graph is not necessarily. In all the applications we consider, self loops are not allowed, meaning that the main diagonal of the adjacency matrix is zero. The sum of the elements on each row (respectively columns) is the out (in) **degree** of the corresponding node. In undirected graphs, in and out degree are the same thing, namely the number of neighbours of a node.



(a) Directed graph. (b) Undirected graph.

Figure 2.1: Two simple graphs.

If all the nodes in a graph are connected to each other the graph is **fully-connected**. A fully-connected sub-graph in a graph is called a **clique**. A sequence of adjacent vertices in a graph, starting from v_i and ending with v_j is called a **path**, either directed or undirected. In directed graphs, we have some additional and important notions to take into account. If an arrow connects v_i *to* v_j, we call v_i **parent** of v_j and v_j **child** of v_i. The set of all the parents of v_j is denoted by $\pi(v_j)$. In case a directed path (longer than one) connects v_i to v_j we say that v_i is an **ancestor** of v_j (the **descendant**). An ancestor of itself gives rise to a **cycle**.

Graphs in machine learning

Graphs are ubiquitous in machine learning. In a single course, It would be extremely difficult to inspect all the areas of machine learning in which graphs appear. Two main uses of graphs are considered here.

First, the **nodes** of the graphs are seen as random variables and the links used to model some dependency relations between them. In particular graphs are very suited to describe some features of the joint probability distributions over random variables. Here, we are in the realm of **graphical models**. Depending on whether directed or undirected graphs are employed, graphical models divide into **Bayesian networks** or **Random Markov Fields**. In this course we focus on the formers. Bayesian networks adopt directed acyclic graphs (**DAG**) to model probability distributions and we look at them in more details in the next sub-sections. Notice that, in this framework, the observed data typically is *not* a graph, but rather a feature vector.

Second, the observed data can be directly seen as one or more graphs (think for instance at social networks). In this case, we work either at the

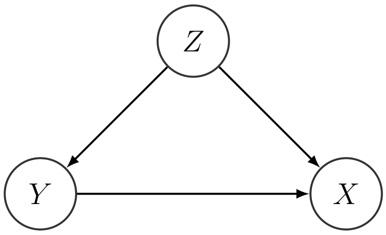


Figure 2.2: A DAG for p_θ.

instance-level and consider the **edges** of the graph as random variables (i.e. the adjacency matrix is the observed data) or at the graph-level, basically working with sequences of graphs. These scenarios will be discussed in more details in later chapters.

Apart from Random Markov Fields (Bishop and Nasrabadi, 2006, Ch.8) other notable fields of ML involving graphs include (not extensively!) Conditional Markov Fields (Sutton et al., 2012), spectral clustering (Von Luxburg, 2007) and statistical relational learning with knowledge graphs (Nickel et al., 2015)

2.2 DAGs and probability distributions

Consider three random variables X, Y and Z whose joint probability (density or mass) function $p_\theta(x, y, z)$ *always* factorizes as

$$p_\theta(x, y, z) = p_\theta(x, y|z)p_\theta(z) = p_\theta(x|y, z)p_\theta(y|z)p_\theta(z),$$

where the subscript θ indicates a general parameter set that usually varies across the joint distribution and its conditionals/marginals. This factorization can be represented by the graph in Figure 2.2. Notice that the Bayesian net or DAG in that figure "fits" the factorization independently on the functional form of p_θ. In other terms, it doesn't matter whether the random variables are discrete or continuous, Gaussian or exponentially distributed, etc. Instead the order of the marginalization matters. For instance, the decomposition

$$p_\theta(x, y, z) = p_\theta(z|y, x)p_\theta(y|x)p_\theta(x)$$

holds true but it is no longer represented by the DAG in Figure 2.2. This allows us to introduce the following

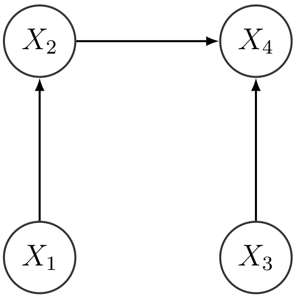


Figure 2.3: Another DAG.

Definition 1. *Given N random variables $X_1, \ldots, X_N$ with joint probability*[1] *p_θ, we say that the DAG G* represents *p_θ if*

$$p_\theta(x_1, \ldots x_N) = \prod_{i=1}^{N} p_\theta(x_i | \pi(x_i)), \tag{2.2}$$

where $\pi(x_i)$ are the parents of x_i according to $\mathcal{G}$. The set of distributions represented by $\mathcal{G}$ is denoted by $M(\mathcal{G})$.

Example. A joint distribution over the random variables X_1, X_2, X_3, X_4 is represented by the DAG in Figure 2.3 iff its probability (density or mass) function p_θ satisfies

$$p_\theta(x_1, x_2, x_3, x_4) = p_\theta(x_1)p_\theta(x_2|x_1)p_\theta(x_4|x_2, x_3)p_\theta(x_3).$$

Since a likelihood function is nothing but a joint probability function, we can use DAGs to describe entire statistical models.

Example. The linear Gaussian model introduced in Section 1.2 can be represented graphically as in Figure 2.4. Here, the black dots are additional features representing the model parameters. And β, σ^2 are condensed in a single parameter. In order to avoid the writing down of $Y_1, \ldots, Y_N$ multiple times, we can adopts a condensed notation that adopt red *plates*, as in Figure 2.4. The N above the plate indicates that the random variables inside the plate are sampled N times independently.

[1] With a slight abuse of language, we use the expression "with joint probability" to mean "whose joint distribution admits a probability (density or mass) function".

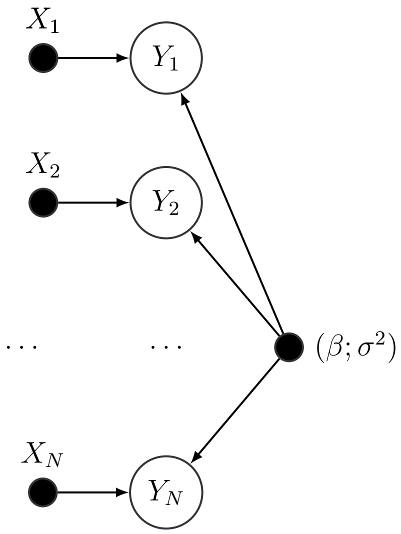


Figure 2.4: Linear regression model.

In order to illustrate a useful feature that will be massively used later in this course, consider a Bayesian variant of the Gaussian linear model, in which β is now see as a random variable assumed to samples from a Gaussian distribution $\mathcal{N}(0, \eta I_P)$, with I_P being the identity matrix of order P . So, additionally to the **observed** random variables $Y_1, \ldots, Y_N$, for all i, the model parameters $X_1, \ldots, X_N$, η and σ^2 we now also have an **hidden** random vector β, which is not observed but whose posterior probability we would like to infer. The joint density of $(Y_1, \ldots, Y_N, \beta)$ given the model parameters looks like[2]

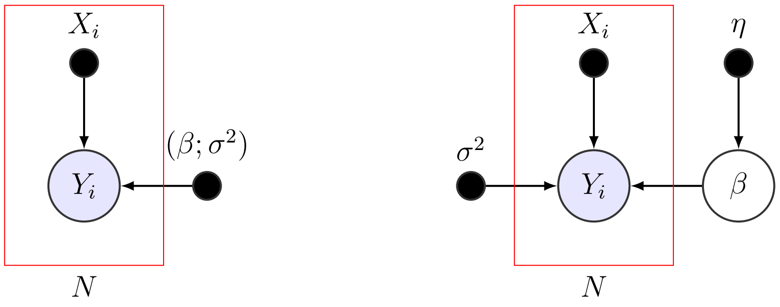
$$p(y_1, \ldots, y_N, \beta | X, \eta, \sigma^2) = p(\beta|\eta) \prod_{i=1}^{N} p(y_i | \beta, X_i, \sigma^2), \tag{2.3}$$

where

$$\begin{aligned} p(\beta|\eta) &:= \phi(\beta; 0, \eta I_P) \\ p(y_i|\beta, X_i, \sigma^2) &:= \phi(y_i; X_i\beta, \sigma^2 I_N), \end{aligned}$$

and $\phi(\cdot; \mu, \Sigma)$ denotes the multivariate Gaussian pdf of a random vector with mean μ and covariance matrix Σ. The likelihood in Eq. (2.3) is represented

[2] In this notes, the notations $p_\theta(\cdot)$ and $p(\cdot|\theta)$ are used interchangeably, depending on the context. So, for instance $p(y_1, \ldots, y_N, \beta|X, \eta, \sigma^2)$ is the same as $p_{X,\eta,\sigma^2}(y_1, \ldots, y_N)$.



(a) Linear regression.

(b) Bayesian linear regression.

Figure 2.5: Graphical models. Observed random variables are represented by light-blue circles, hidden random variables by white circles and parameters by black small circles. Everything contained in the red plate is repeated N times.

by the DAG in Figure 2.5b, where the white circle represents the hidden random vector β and Y_i is in light-blue to emphasize the fact that it is an observed random variable.

Via Eq. (2.2), Bayesian networks are related with a simple yet powerful sampling technique, known as **ancestral sampling**. Before looking at it we need to state the following

Proposition 1. *In a DAG there exists an ordering of the nodes (a.k.a.* topological *ordering) such that there are no links that go from any node to any lower numbered node.*

Proof. The ordering $v_1, \dots v_N$ we are looking for is such that, if $v_i \in \pi(v_j)$ then $i < j$. We proceed by recurrence. With two nodes things are easy, since either they are not connected (no links at all) or we call v_1 the parent and v_2 the child. Consider now the case of $N > 2$ nodes and assume the proposition true for any subgraph with $N - 1$ nodes. The key point is that in DAG we can always find a source node (i.e. node without parents). Indeed starting from any node and proceeding from parents to parents, whenever possible, we must end somewhere otherwise we have a cycle. So we can find a source of our graph of order N, remove the source, find the desired ordering for the $(N-1)$-order sub-graph without the source, set the source equal to v_1 and increase by one the order of the other nodes. □

We can now go back to Eq. (2.2). In order to sample $X_1^{(s)}, \dots, X_N^{(s)}$ from p_θ we can start from the source node $X_1^{(s)} \sim p_\theta(x_1)$ and proceed node after node in increasing order:

$$X_i^{(s)} \sim p_\theta(x_i | \pi(x_{i-1})), \qquad \forall i \leq N.$$

Thanks to the topological ordering, at the i-th step all the parents of node X_i have already been sampled.

Exercise (from Wasserman, 2004). Consider three random variables (X,Y,Z) with the following joint distribution

$$\begin{aligned} X &\sim \text{Ber}\left(\frac{1}{2}\right) \\ Y|X = x &\sim \text{Ber}\left(\frac{e^{4x-2}}{1+e^{4x-2}}\right) \\ Z|X = x, Y = y &\sim \text{Ber}\left(\frac{e^{2(x+y)-2}}{1+e^{2(x+y)-2}}\right) \end{aligned} \quad .$$

Use ancestral sampling in order to estimate $\mathbb{P}(Z = 1)$ and $\mathbb{P}(Z = 1|Y = 1)$

Exercise. Consider the following random pair (X, Z) with $Z \sim \text{Ber}(\pi)$ and X such that $\mathbb{P}(X = 0|Z = 1) = 1$ whereas

$$X|Z = 0 \sim \mathcal{N}(\mu, \sigma^2).$$

1. Sketch the graphical model corresponding the joint distribution of (X, Z).
2. Compute $\mathbb{E}(X)$ analytically.
3. Say $\pi = 0.9$, $\mu = 1.0$ and $\sigma^2 = .02$. Use ancestral sampling in order to sample $N = 1000$ outcomes from the marginal distribution of X and compute (an estimate of) $\mathbb{E}(X)$ numerically.

2.3 Conditional independence

Definition 2. *Given three random variables X, Y and Z, we say that X and Y are **conditionally independent** given Z, written $X \amalg Y | Z$ if*

$$p_\theta(x, y|z) = p_\theta(x|z)p_\theta(y|z)$$

for all y, x and z.

A first important remark is that conditional independence can be equivalently formulated (**exercise**) as

$$p_\theta(x|y,z) = p_\theta(x|z).$$

Now, one of the most interesting features about DAGs is that if $X_1, \dots, X_N$ follow a joint distribution have joint probability p_θ represented by a graph $\mathcal{G}$, then a number of conditional independence relationships between groups of those random variables can be read (almost) straight forward on the graph. First, with respect to Definition 1, we can state the following

Theorem 1. *(Markov Condition) The joint probability $p_\theta \in M(\mathcal{G})$ iff*

$$X_i \amalg \overline{X}_i | \pi(X_i), \qquad \forall i \in 1, \dots, N, \tag{2.4}$$

where $\overline{X}_i$ denotes all the remaining random variables except the parents and descendants of X_i.

Proof. Assume first that Eq. (2.4) holds. Assuming also that a topological ordering of the nodes of the DAG is adopted. Then

$$p_\theta(x_1, \dots, x_N) = \prod_{i=1}^{N} p_\theta(x_i|x_1, \dots, x_{i-1})$$

by repeated marginalisation. The topological ordering guarantees that among $X_1, \dots, X_{i-1}$ we only find the ancestors of X_i and/or other nodes *not* being descendants of X_i. Thanks to Eq. (2.4) the r.h.s. of the above equation reduces to $\prod_{i=1}^{N} p_\theta(x_i|\pi(x_i))$.

Still assuming a topological ordering, consider X_i and X_j with $i < j$. So X_i is not a descendant of X_j and we further assume that neither it is a parent of X_j. We want to show that

$$p_\theta(x_j|\pi(x_j), x_i) = p_\theta(x_j|\pi(x_j)), \tag{2.5}$$

i.e. that $X_i \amalg X_j|\pi(X_j)$. Indeed, by both marginalization and Eq. (2.2) it holds that

$$\prod_{j=1}^{N} p_\theta(x_j|x_1, \dots, x_{j-1}) = p_\theta(x_1, \dots, x_N) = \prod_{j=1}^{N} p_\theta(x_j|\pi(x_j)). \tag{2.6}$$

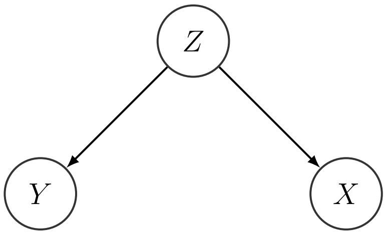


Figure 2.6: Tail-to-tail configuration.

By working on induction on j, it's easy to see that $p_\theta(x_j|x_1,\ldots,x_{j-1}) = p_\theta(x_j|\pi(x_j))$ for all $j \leq N$. In particular, by definition of conditional independence, it means that $X_j \amalg X_i|\pi(X_j)$ for all $i < j$ and such that $X_i \notin \pi(X_j)$ and Eq. (2.5) holds true. □

The above theorem allows us to immediately uncover some conditional independence relations by just looking at the graph. For instance, when looking at the DAG in Figure 2.3 we know that X_1 is independent from X_3 and X_4 is independent from X_1 given (i.e. conditionally to) (X_2, X_3). However, other conditional independence relations may exist although been less trivial to uncover. For instance: is X_4 conditionally independent from X_1 given X_2 alone? The answer is yes and the reason is **d-separation**. In order to illustrate this important notion we focus on three "minimal" DAGs.

Consider first the tail-to-tail DAG in Figure 2.6. Without loss of generality we assume that (X, Y, Z) are three continuous random variables with real support and consider the joint density of (Y, X)

$$p_\theta(y,x) = \int_{\mathbb{R}} p_\theta(y,x,z)dz = \int_{\mathbb{R}} p_\theta(y,x|z)f(z)dz = \int_{\mathbb{R}} p_\theta(y|z)p_\theta(x|z)p_\theta(z)dz$$

where the last equality follows from the Markov condition. Since in general

$$\int_{\mathbb{R}} p_\theta(y,x|z)f(z)dz = \int_{\mathbb{R}} p_\theta(y|z)p_\theta(x|z)p_\theta(z)dz \neq p_\theta(y)p_\theta(x),$$

Y and X are not independent and we say that they are **d-connected**. However, by the same Markov property follows that $X \amalg Y|Z$ so that Y and X are **d-separated** given Z.

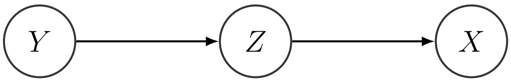


Figure 2.7: Head-to-tail configuration.

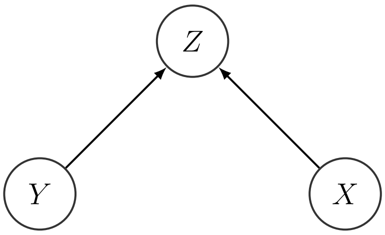


Figure 2.8: Head-to-head configuration.

Exercise. Show that in the head-to-tail configuration in Figure 2.7 the same conclusions about the d-connectedness and d-separateness of (X, Y) can be reached.

The third configuration is the head-to-head one in Figure 2.8. Here Z is also said to be a **collider**. Now, X and Y are independent or d-separated as a consequence of the the Markov property but

$$p_\theta(x, y|z) = \frac{p_\theta(x, y, z)}{p_\theta(z)} = \frac{p_\theta(z|x, y)p_\theta(x)p_\theta(y)}{p_\theta(z)}$$

which is in general different from $p_\theta(x|z)p_\theta(y|z)$. Thus X and Y are d-connected given Z.

We can now give a more general definition of what d-separation is in DAGs. Consider a DAG and three disjoint sets of nodes A, B and C whose union does not necessarily include all nodes of the DAG. We aim at assessing whether $A \amalg B|C$or not. We check all the paths from any node in A any node in B. A path is said to be *blocked* if it includes at least one node such that either

1. the directed arrows meet at that node head-to-tail or tail-to-tail and the node is in set C;

2. the node is a collider neither it nor its descendants are in C.

If all paths from A to B are blocked, then $A \amalg B|C$ and A and B are d-separated given C.

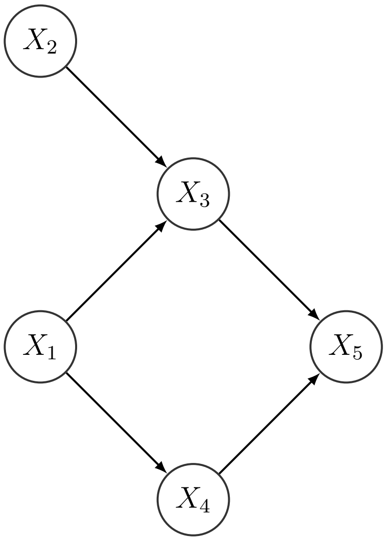


Figure 2.9: A slightly more difficult DAG.

For instance consider the graph in Figure 2.9. Some (conditional) independence relations follow directly from the Markov property, for instance $X_5 \amalg \{X_2, X_1\} | \{X_3, X_4\}$. However can we say (for instance) that $X_5 \amalg X_2 | \{X_1, X_3\}$? We consider the two paths from X_5 to X_2. The first passes solely through X_3 in a head-to-tail manner and X_3 is in the set we are conditioning to. So the path is blocked. The other path passes through X_1 in a tail-to-tail manner and still X_1 is in the conditioning set. The two paths are blocked so the answer is yes, X_5 is d-separated from X_2 given $\{X_1, X_3\}$.

2.4 Inference in Bayesian networks

Learning is related with Bayesian Networks in two different ways: i) either we have N (usually independent) observations $X_1, \ldots, X_N$ each in dimension D *and* a DAG $\mathcal{G}$ defining the dependencies between $X_{i1}, \ldots, X_{iD}$ for all i. Then we choose a probability density or mass function p_θ, Markovian with respect to $\mathcal{G}$, and our aim is to infer the model parameters θ from the data. Otherwise, ii) we aim to directly infer $\mathcal{G}$ from the data. In this course we only focus on case i). For more details on learning in Bayesian networks the reader is referred to Heckerman (2008) and for a complete overview on how to

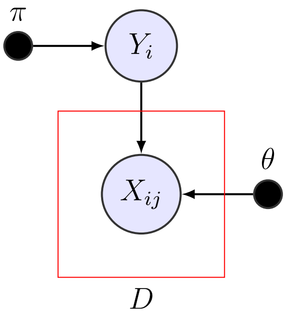


Figure 2.10: Graphical representation of the Bayesian naive model (i-th observation).

implement Bayesian networks in R, we recommend Nagarajan et al. (2013).

So, the typical scenario we deal with in the next chapters is the one described in i), above. Our standard choice to infer θ from the data $X_1, \dots, X_N$ is **maximum likelihood**

$$\max_{\theta} \left(\prod_{i=1}^{N} \prod_{j=1}^{D} p(X_{ij}|\pi_j, \theta_j) \right),$$

where θ_j are the parameters for the j-th conditional density (i.e. $\theta = \cup_j \theta_j$) and $\pi_j := \pi(X_{\cdot j})$ denotes the parents set of the j-th node of the graph. Solving the above maximization problem can be more or less difficult based on the nature of $\mathcal{G}$. In order to illustrate the idea, we now inspect in more details an example of a simple Bayesian network.

Naive Bayesian classifier

We are given a training data set of feature observations $X_1, \dots, X_N$ in $\mathbb{R}^D$ and labels $Y_1, \dots, Y_N$, where Y_i is the class of the i-th observation. Whether Y_i is binary (0-1) we are in very same framework described in Seciton 1.3, otherwise we face a multiclass classification problem and Y_i is a K-dimensional binary vector in a 1-to-K encoding scheme. What it means is that $Y_{ik} = 1$ iff the i-th observation belongs to class k (and $Y_{ij} = 0 \quad \forall \quad j \neq k$), zero otherwise[3]. First, $Y_1, \dots, Y_N$ are assumed to be N independent outcomes

[3] As it is common in the latent variable models literature, we adopt here the following convention: when no confusion arises Y_i will both denote a categorical random variable taking values $1, 2, \dots, K$ or the 1-to-K vector described above.

from a categorical distribution of parameter $\pi := (\pi_1, \ldots, \pi_K)$

$$\mathbb{P}_\pi(Y_{ik} = 1) = \pi_k,$$

with $\sum_{k=1}^K \pi_k = 1$. Equivalently, $p_\pi(Y_i) = \prod_{k=1}^K \pi_k^{Y_{ik}}$

Then, the main assumption in the Naive Bayesian classifier is that the distribution of the input variables $X_{i1}, \ldots, X_{iD}$ factorizes conditionally to the class

$$p(X_i|Y_i, \theta) = \prod_{j=1}^D p(X_{ij}; \theta_{Y_i}), \qquad \forall i \tag{2.7}$$

where $\theta_1, \ldots, \theta_K$ are the parameter sets corresponding to each class. This is also illustrated in Figure 2.10. The features are assumed to be independent given the class. However, given the tail-to-tail pattern through Y_i, when integrating Y_i out due to marginalisation the $X_{i1}, \ldots, X_{iD}$ are no longer independent! In order to fix the ideas, assume that X_i is a D-Gaussian vector conditional to Y_i, namely

$$X_i|Y_i = k \sim \mathcal{N}\left(\mu_k, \Sigma_k\right),$$

where $\mu_k \in \mathbb{R}^D$ is the mean vector and $\Sigma_k \in \mathbb{R}_+^{D\times D}$ the variance covariance matrix. Naive Bayesian modeling forces Σ_k to be diagonal (why?) so that

$$X_{ij}|Y_i = k \sim \mathcal{N}\left(\mu_{kj}, \sigma_{kj}^2\right),$$

independently for all $j \leq D$. Assuming independence among the observations in the training data set, the likelihood of the whole set is

$$\begin{aligned}\mathcal{L}(\pi, \theta) :&= \prod_{i=1}^N \pi_{Y_i} \prod_{j=1}^D \left(\frac{1}{C\sigma_{Y_ij}} \exp\left(-\frac{(X_{ij} - \mu_{Y_ij})^2}{2\sigma_{Y_ij}^2}\right)\right) \\ &= \prod_{i=1}^N \prod_{k=1}^K \left[\pi_k \prod_{j=1}^D \frac{1}{C\sigma_{kj}} \exp\left(-\frac{(X_{ij} - \mu_{kj})^2}{2\sigma_{kj}^2}\right)\right]^{Y_{ik}}\end{aligned} \tag{2.8}$$

where θ now denotes the whole set of the Gaussian parameters and $C = \sqrt{2\pi}$ is the normalizing constant of the Gaussian distribution where the π in the definition C should not be confused with the parameter π of the multinomial distribution. The above equation makes it clear that the likelihood of the training data factorizes over k, meaning that each class can be treated separately. Indeed, by taking the logarithm of the above likelihood, adding a Lagrange multiplier accounting for the constraint ($\sum_k \pi_k = 1$), computing the gradient with respect to θ and π and setting it equal to zero it is easy to see (**exercise**) that the ML estimates are

$$\begin{aligned}
\hat{\pi}_k &= \frac{1}{N}\sum_{i=1}^{N} Y_{ik}, \\
\hat{\mu}_{kj} &= \frac{\sum_{i=1}^{N} Y_{ik} X_{ij}}{\sum_{i=1}^{N} Y_{ik}}, \\
\hat{\sigma}_{kj}^2 &= \frac{\sum_{i=1}^{N} Y_{ik}\left(X_{ij} - \hat{\mu}_{kj}\right)^2}{\left(\sum_{i=1}^{N} Y_{ik}\right)^2}.
\end{aligned}$$

Once the model is fit to the data, let us assume that we observe a new feature vector X^* and want to assign it to a class (i.e. Y^* is unobserved). The key observation is that, thanks to the Bayes Rule we can compute the *posterior* probability

$$\mathbb{P}(Y^* = k | X^*, \hat{\pi}, \hat{\theta}) = \frac{p(X^*, Y^* = k|\hat{\pi}, \hat{\theta})}{p(X^*|\hat{\pi}, \hat{\theta})} \propto p(X^*, Y^* = k|\hat{\pi}, \hat{\theta})$$

where $p(X^*|\hat{\pi}, \hat{\theta}) = \sum_{k=1}^{K} p(X^*, Y^* = k|\hat{\pi}, \hat{\theta})$. As such the decision rule is

$$Y^* := \arg \max_{k \in \{1,\ldots,K\}} p(X^*, Y^* = k|\hat{\pi}, \hat{\theta}).$$

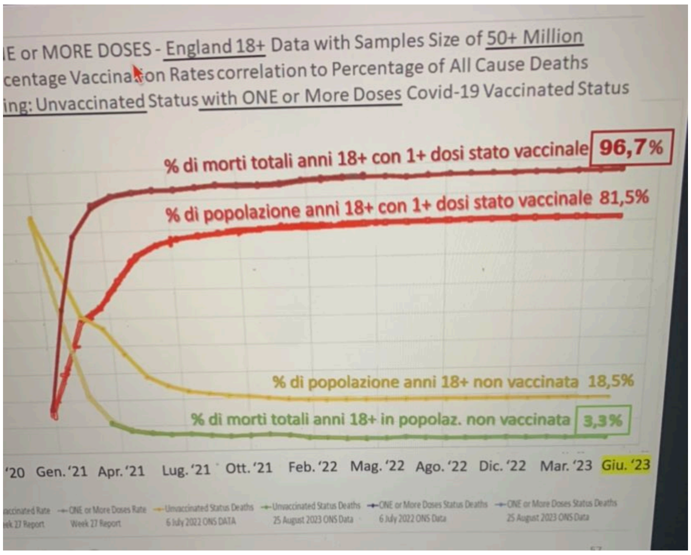


Figure 2.11: Percentage of people being vaccinated and deaths in England in the last two years.

2.A Pills of causal inference

Figure 2.11 shows a graph that I saw on Twitter sometimes ago. It reports the percentage of English people being vaccinated against Covid19, 81,5% of the entire population older than 18 in 2023. Interestingly, when looking at the percentage of people being vaccinated among the deaths in the same year (still older than 18, no matter the cause of death), this percentage is sensibly higher: 96,7%. Some users were scared by this graph since, more or less consciously, they deduced from it that vaccination against Covid19 is a *cause* of death. Is it true? Although giving an answer to this question is (far) outside the scope of this course, what we can safely state is that the graph in Figure 2.11 is *not* sufficient to conclude that vaccination against Covid19 causes death, because of the famous statement "correlation is not causation". In short, since fragile and old people are more likely both being vaccinated and dying than young and healthy people, it is reasonable to see a positive correlation between vaccinations and death (to be clear: no correlation would mean that the percentages of vaccinations would be the same for living and dead people). Let's try to better figure our this point with an example.

Consider the toy graphical model in Figure 2.12a. Here A stands for

Age, considered as a continuous random variable in $[18, 100]$, T strands for *Treatment*, a binary random variable taking value 1 in case of vaccination, 0 otherwise and D stands for *Death*, taking value 1 in case of death for any reason, 0 otherwise. Everything is assumed to be observed. A is a **confounding** variable, formally defined as a parent of both the *presumed* cause and the *presumed* effect. The arrows streaming from A represent the idea that the older people are more likely to be vaccinated or to die. For instance let us introduce the following notations $p_T := \mathbb{P}(T = 1|A = a)$ and $P_D := \mathbb{P}(D = 1|A = a)$. Then, the following equation

$$\log\left(\frac{p_T}{1 - p_T}\right) = \alpha a, \qquad \exists \alpha > 0 \tag{2.9}$$

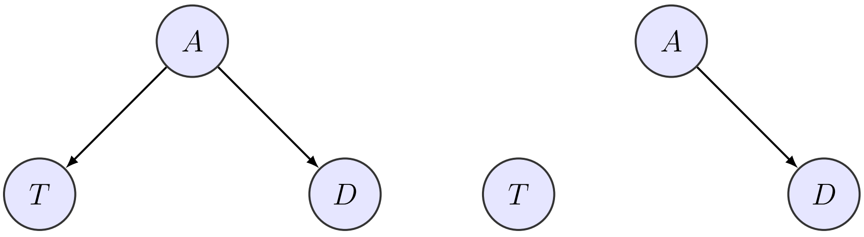
captures the positive correlation between the age of and individual and the probability that she is vaccinated, since it induces $p_T = \frac{1}{1+\exp(-\alpha a)}$. Similarly we assume that

$$\log\left(\frac{p_D}{1 - p_D}\right) = \beta a, \qquad \exists \beta > 0 \tag{2.10}$$

In this graphical model (we are just playing with) T and D are conditionally independent given A, which basically means that for any given age, death and treatment are independent. This induces us to assume (intuitively) that T does not directly cause D in this universe. This intuition can be formalized by the notion of **intervention**. In particular we intervene on the graph by fixing $T := 1$ and treating T as a parameter in the following sequence of equations:

$$\begin{aligned}\mathbb{E}[D|T := 1] &= \mathbb{P}(D = 1|T := 1)\\ &: = \int_{[18,100]} \mathbb{P}(D = 1|T = 1, a)p(a)da\\ &= \int_{[18,100]} \mathbb{P}(D = 1|a)p(a)da\\ &= \mathbb{P}(D = 1)\end{aligned} \tag{2.11}$$

where $p(a)$ denotes the pdf of A and the penultimate equality comes from conditional independence between D and T given A. The fact that $\mathbb{P}(D = 1|T =: 1) = \mathbb{P}(D = 1|T := 0) = \mathbb{P}(D = 1)$ confirms that T does not cause D in our DAG. It is important to note the difference between intervention and



(a) Original DAG. (b) Intervened DAG.

Figure 2.12: Graphical model with A being *Age*, T *Treatment* and D *Death*.

conditioning:

$$\begin{aligned}\mathbb{E}[D|T=1] &= \mathbb{P}(D=1|T=1)\\ :&= \int_{[18,100]} \mathbb{P}(D=1|T=1,a)\frac{\mathbb{P}(T=1|a)}{\mathbb{P}(T=1)}p(a)da \\ &= \int_{[18,100]} \mathbb{P}(D=1|a)p(a|T=1)da.\end{aligned} \tag{2.12}$$

As it can be seen intervening on T can be done by i) creating a new DAG obtained by removing all the arrows pointing to T (see Figure 2.12b) and ii) creating new distributions $\mathbb{P}(D=d|T:=t)$, obtained by removing $\mathbb{P}(T=t|\pi_T)$ from the joint $\mathbb{P}(D=d,\dots,T=t)$, with $d,t\in\{0,1\}$.

Now, Eq. (2.12) makes it clear that, in general, $\mathbb{P}(D=1|T=1)\neq\mathbb{P}(D=1|T=0)$ in our model (since $p(a|T=1)$ might differ from $p(a|T=0)$) and thus intervening is not equivalent to conditioning. Moreover, assume that we observe the data $\{(D_i,T_i)\}_{\{i\leq N\}}$, where N is the 50+ million people the graph in Figure 2.11 refers to. We could fit to the data the following linear model

$$D_i = c_0 + c_1 T_i + \epsilon_i \tag{2.13}$$

with ϵ_i being centred, independent noises with the same variance and (c_0,c_1) parameters to estimate. It is easy to show (**exercise**) that the OLS estimates of c_0 and c_1 are

$$\begin{aligned}\hat{c}_0 &= \bar{D} - \hat{c}_1\bar{T}\\ \hat{c}_1 &= \hat{\rho}_{DT}\frac{\hat{\sigma}_D}{\hat{\sigma}_T}\end{aligned} \tag{2.14}$$

where $\hat{\rho}_{DT},\hat{\sigma}_T,\hat{\sigma}_D$ are the empirical correlation coefficient and standard deviations, respectively. We thus have consistent estimates of $\mathbb{E}(D|T=1)\simeq\hat{c}_0+\hat{c}_1$

and $\mathbb{E}(D|T=0) \simeq \hat{c}_0$. As soon as $\hat{c}_1$ (and hence $\hat{\rho}_{DT}$) is significantly different from zero we have a correlation. In Figure 2.11, *if* the difference $15,2\% = 96,7\% - 81,5\%$ is significantly different from zero we are in the case of positive correlation. The important point to keep in mind is that this correlation does **not** invalidate the DAG if Figure 2.12a . For instance, when replacing Eq. (2.9) in Eq. (2.10), after some manipulations (**exercise**) we find

$$P_D = 1 - \frac{1}{1 + (\frac{P_T}{1-P_T})^{\frac{\beta}{\alpha}}}. \tag{2.15}$$

As it can be seen as P_T tends to 1 (respectively to zero) so does P_D. This gives an intuition about why we could see a positive correlation between D and T even in a model where D does not cause T!

In any case, apart from the model in Figure 2.12a what said so far should make clear that we can always estimate a correlation between D and T, but this correlation does in general not mean causation. So the question is: how can we asses whether vaccination against Covid19 causes death? In two ways: randomization or making all confounding variables explicit. The former consists into vaccinating people randomly. With respect to our toy model this would mean randomly with respect to the age and so breaking the link from A to T. In this way the original DAG and the intervened one are the same and $\mathbb{E}[D|T=1] = \mathbb{E}[D|T:=1]$. Correlation is now causation. The other solution would be to modify Eq. (2.13) as follows:

$$D_i = c_0 + c_1 T_i + c_2 A_i + \epsilon_i. \tag{2.16}$$

Now that A is explicit (observed) we make use of Eq. (2.11) to observe that

$$\mathbb{E}[D|T:=1] = \mathbb{E}_A\left[\mathbb{E}[D|T=1,A]\right].$$

This quantity (and similarly $\mathbb{E}[D|T:=0]$) can be estimated via

$$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{c}_0 + \hat{c}_1 + \hat{c}_2 A_i\right),$$

where $\hat{c}_0, \hat{c}_1$ and $\hat{c}_2$ are the OLS estimates of the parameters c_0, c_1 and c_2, respectively, in Eq. (2.16).