Basic Assumptions for Efficient Model Representation

Chris Williams
(based on slides by Michael U. Gutmann)

Probabilistic Modelling and Reasoning (INFR11134)
School of Informatics, The University of Edinburgh

Spring Semester 2024

Recap

$$p(\mathbf{x}|\mathbf{y}_o) = \frac{\sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{y}_o, \mathbf{z})}{\sum_{\mathbf{x},\mathbf{z}} p(\mathbf{x}, \mathbf{y}_o, \mathbf{z})}$$

Assume that $\mathbf{x}, \mathbf{y}, \mathbf{z}$ each are $d = 500$ dimensional, and that each element of the vectors can take $K = 10$ values.

- Issue 1: To specify $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$, we need to specify $K^{3d} - 1 = 10^{1500} - 1$ non-negative numbers, which is impossible.

  Topic 1: Representation What reasonably weak assumptions can we make to efficiently represent $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$?

Two fundamental assumptions

Consider two assumptions:

1. only a limited number of variables may directly interact with each other (independence assumptions)
2. for any number of interacting variables, the form of interaction is limited or restricted (often: parametric family assumptions)

The two assumptions can be used together or separately.

Program

1. Independence assumptions

2. Assumptions on form of interaction

Program

1. Independence assumptions
   - Definition and properties of statistical independence
   - Factorisation of the pdf and reduction in the number of directly interacting variables
2. Assumptions on form of interaction

Statistical independence

- Let $\mathbf{x}$ and $\mathbf{y}$ be two disjoint subsets of random variables. Then $\mathbf{x}$ and $\mathbf{y}$ are independent of each other if and only if (iff)

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$$

  for all possible values of $\mathbf{x}$ and $\mathbf{y}$; otherwise they are said to be dependent.

- We say that the joint factorises into a product of $p(\mathbf{x})$ and $p(\mathbf{y})$.
- Equivalent definition by the product rule (or by definition of conditional probability)

$$p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x})$$

  for all values of $\mathbf{x}$ and $\mathbf{y}$ where $p(\mathbf{y}) > 0$.

- Notation: $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$
- Variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent iff

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} p(\mathbf{x}_i)$$

Conditional statistical independence

- The characterisation of statistical independence extends to conditional pdfs (pmfs) $p(\mathbf{x}, \mathbf{y}|\mathbf{z})$.
- The condition $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$ becomes $p(\mathbf{x}, \mathbf{y}|\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z})$
- The equivalent condition $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x})$ becomes $p(\mathbf{x}|\mathbf{y}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})$
- We say that $\mathbf{x}$ and $\mathbf{y}$ are conditionally independent given $\mathbf{z}$ iff, for all possible values of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ with $p(\mathbf{z}) > 0$:

$$p(\mathbf{x}, \mathbf{y}|\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z}) \quad \text{or}$$

$$p(\mathbf{x}|\mathbf{y}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z}) \quad (\text{for } p(\mathbf{y}, \mathbf{z}) > 0)$$

- Notation: $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$

The impact of independence assumptions

- The key is that the independence assumption leads to a partial factorisation of the pdf/pmf with factors that involve fewer variables.
- Reduces the number of directly interacting variables.
- For example, if $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are independent of each other, then

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x})p(\mathbf{y})p(\mathbf{z})$$

- Independence assumption forces $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ to take on a particular form.

The impact of independence assumptions

Assume $p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x})p(\mathbf{y})p(\mathbf{z})$

- If $\dim(\mathbf{x}) = \dim(\mathbf{y}) = \dim(\mathbf{z}) = d$, and each element of the vectors can take K values, factorisation reduces the numbers that need to be specified ("parameters") from $K^{3d} - 1$ to $3(K^d - 1)$.
- If all variables were independent: $3d(K - 1)$ numbers needed.

  For example: $10^{1500} - 1$ vs. $3(10^{500} - 1)$ vs. $1500(10 - 1) = 13500$
- But full independence (factorisation) assumption is often too strong and does not hold.

The impact of independence assumptions

- Conditional independence assumptions are a powerful middle-ground.
- For $p(\mathbf{x}) = p(x_1, \ldots, x_d)$, we have by the product rule:

$$p(\mathbf{x}) = p(x_d | x_1, \ldots x_{d-1}) p(x_1, \ldots, x_{d-1})$$

- If, for example, $x_d \perp\!\!\!\perp x_1, \ldots, x_{d-4} \,|\, x_{d-3}, x_{d-2}, x_{d-1}$, we have

$$p(x_d | x_1, \ldots, x_{d-1}) = p(x_d | x_{d-3}, x_{d-2}, x_{d-1})$$

- If the x_i can take K different values:

  $p(x_d | x_1, \ldots, x_{d-1})$ specified by $K^{d-1} \cdot (K-1)$ numbers

  $p(x_d | x_{d-3}, x_{d-2}, x_{d-1})$ specified by $K^3 \cdot (K-1)$ numbers

  For $d = 500, K = 10$: $10^{499} \cdot 9 \approx 10^{500}$ vs $9000 \approx 10^4$.

Program

1. Independence assumptions

2. Assumptions on form of interaction
   - Parametric model to restrict how a given number of variables may interact

Assumption 2: limiting the form of the interaction

- The (conditional) independence assumption limits the number of variables that may directly interact with each other, e.g.

  x_d only directly interacted with $x_{d-3}, x_{d-2}, x_{d-1}$.
- How x_d interacts with the three variables, however, was not restricted.
- Assumption 2: We restrict how a given number of variables may interact with each other.
- For example, for $x_i \in \{0, 1\}$, we may assume that $p(x_d | x_1, \dots, x_{d-1})$ is specified as

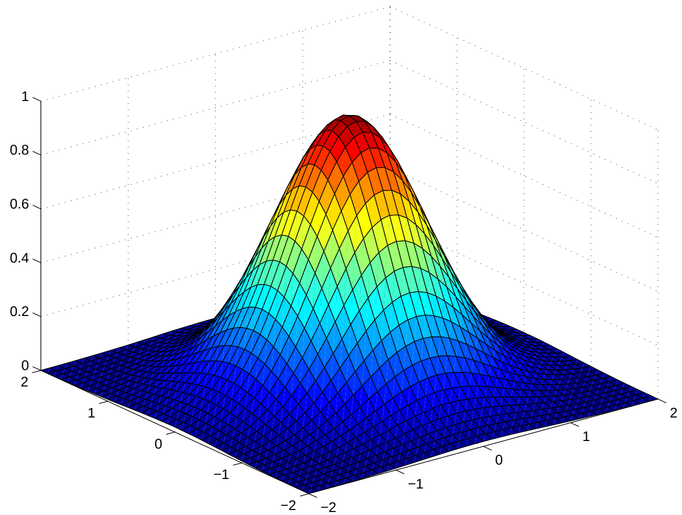
$$p(x_d = 1 | x_1, \dots, x_{d-1}) = \frac{1}{1 + \exp\left(-w_0 - \sum_{i=1}^{d-1} w_i x_i\right)}$$

  with d free numbers ("parameters") $w_0, \dots, w_{d-1}$.
- d vs 2^{d-1} parameters (for $d = 500$: 500 vs. $2^{499} \approx 10^{150}$)

Gaussian parametric assumption for real-valued variables

- Multivariate Gaussian $N(\boldsymbol{\mu}, \Sigma)$
- Has mean $\boldsymbol{\mu}$ and covariance Σ
- $\Sigma_{ij} = \Sigma_{ji} = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$
- Probability density $p(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^d$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}$$



Exact inference for Gaussian RVs

Exact inference is possible for the multivariate Gaussian $N(\boldsymbol{\mu}, \Sigma)$.
Basic rules:

- Partition variables into two groups, $\mathbf{X}_1$ and $\mathbf{X}_2$

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$$
$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

- Marginal distribution: $\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \Sigma_{11})$
- Conditional distribution

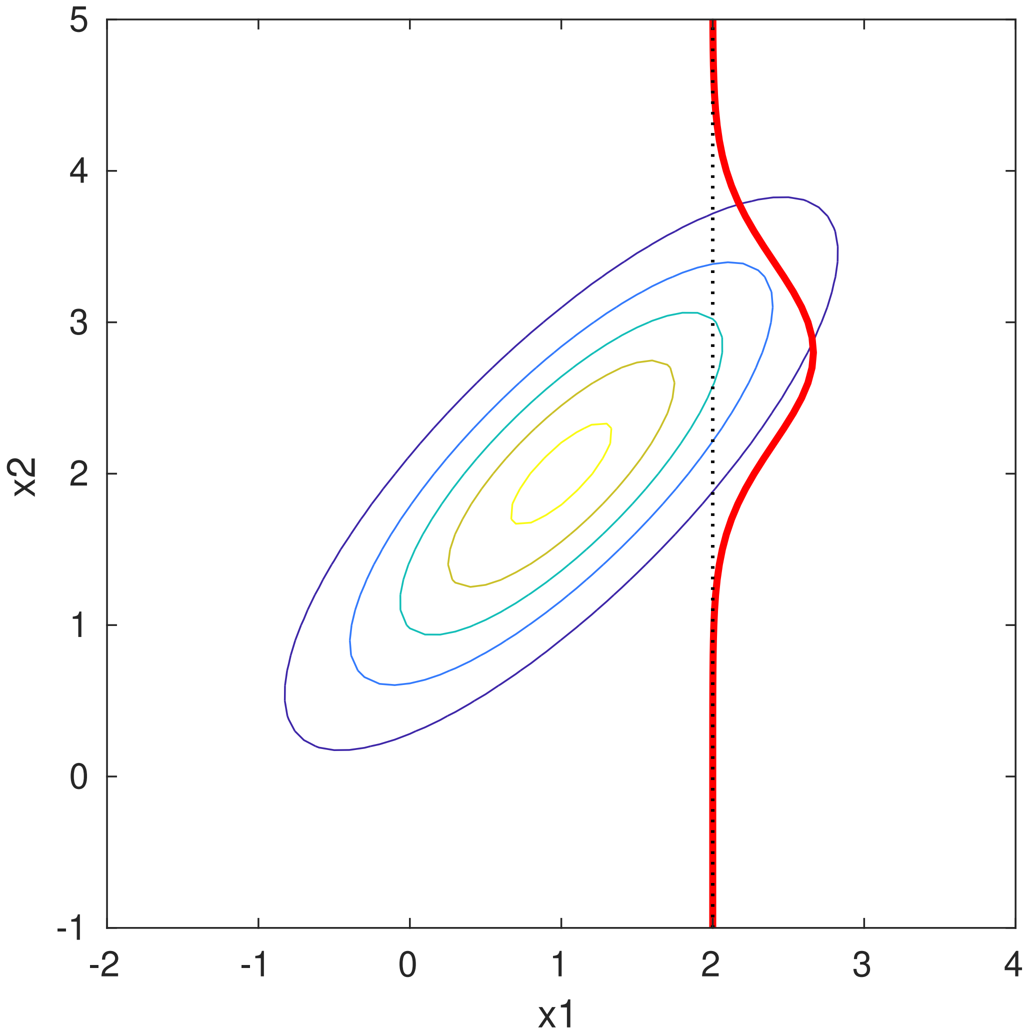
$$\boldsymbol{\mu}^c_{1|2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$
$$\Sigma^c_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

- For proof see sec. 2.3.1 of Bishop (2006) (not examinable)

- We have joint Gaussian for $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$, and want $p(\mathbf{x}|\mathbf{y}_o)$
- $\mathbf{z}$ can be marginalized out trivially (just ignore the $\mathbf{z}$ parts of the mean and covariance)
- Use the conditional distribution rule to obtain $\mathbf{x}|\mathbf{y}_o \sim N(\boldsymbol{\mu}^c_{\mathbf{x}|\mathbf{y}_o}, \Sigma^c_{\mathbf{x}|\mathbf{y}_o})$ with

$$\boldsymbol{\mu}^c_{\mathbf{x}|\mathbf{y}} = \boldsymbol{\mu}_\mathbf{x} + \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1}(\mathbf{y}_o - \boldsymbol{\mu}_\mathbf{y})$$
$$\Sigma^c_{\mathbf{x}|\mathbf{y}} = \Sigma_{\mathbf{xx}} - \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{yx}}$$

- Assume that $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ each are each d dimensional,
- Complexity is dominated by inversion of $\Sigma_{\mathbf{yy}}$ in $O(d^3)$ time
- If all variables are discretized into K bins, complexity for computing $p(\mathbf{x}|\mathbf{y}_o)$ is $O(K^d)$, even for approximate inference



- Conditional distribution of x_2 given $x_1 = 2$ shown in red

Program recap

We asked: What reasonably weak assumptions can we make to efficiently represent a probabilistic model?

1. Independence assumptions
   - Definition and properties of statistical independence
   - Factorisation of the pdf and reduction in the number of directly interacting variables

2. Assumptions on form of interaction
   - Parametric model to restrict how a given number of variables may interact

Credits

These slides are modified from ones produced by Michael Gutmann, made available under Creative Commons licence CC BY 4.0.