Basic Assumptions for Efficient Model Representation

Michael U. Gutmann

Probabilistic Machine Learning (INFR11298)
School of Informatics, The University of Edinburgh

Autumn Semester 2026

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)

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}) \tag{1}$$

  for all possible values of $\mathbf{x}$ and $\mathbf{y}$, where $p(\mathbf{x})$ and $p(\mathbf{y})$ are the marginals of $\mathbf{x}$ and $\mathbf{y}$, respectively.
- We say that the joint factorises into a product of $p(\mathbf{x})$ and $p(\mathbf{y})$.
- Notation: $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$

Equivalent characterisation of independence

- Equivalent characterisation: $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$ iff

$$p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x}) \tag{2}$$

  for all values of $\mathbf{x}$ and $\mathbf{y}$ where $p(\mathbf{y}) > 0$.
- The equivalency follows from the product rule: $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$.

Proof for $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \iff p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x})$

$\Rightarrow$ Assume $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$ holds. Since $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$, we have

$$p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) = p(\mathbf{x})p(\mathbf{y}) \tag{3}$$

and hence $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x})$ for all $\mathbf{y}$ where $p(\mathbf{y}) > 0$.

$\Leftarrow$ Assume $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x})$ holds for all $\mathbf{y}$ where $p(\mathbf{y}) > 0$. Then:

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) = p(\mathbf{x})p(\mathbf{y}) \tag{4}$$

which is the first characterisation of independence, and completes the proof.

 ©Gutmann, University of Edinburgh CC BY 4.0

Some intuition for statistical independence

- $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$ means that knowing $\mathbf{y}$ does not help you to predict $\mathbf{x}$, and vice versa.
- One way to predict the value of $\mathbf{x}$ from $\mathbf{y}$ is by computing the conditional expectation $\mathbb{E}[\mathbf{x}|\mathbf{y}]$
- In case of independence, we have (assuming pmfs, replace sums with integrals in case of pdfs)

$$\begin{aligned}
\mathbb{E}[\mathbf{x}|\mathbf{y}] &= \sum_{\mathbf{x}} p(\mathbf{x}|\mathbf{y})\mathbf{x} && (5)\\
&= \sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x} && (6)\\
&= \mathbb{E}[\mathbf{x}] && (7)
\end{aligned}$$

- Knowing the value of $\mathbf{y}$ does not change the value of the expectation; it doesn't help you to predict $\mathbf{x}$.
- Generalises to arbitrary functions of $\mathbf{x}$, i.e. $\mathbb{E}[g(\mathbf{x})|\mathbf{y}] = \mathbb{E}[g(\mathbf{x})]$ if $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$.

Statistical independence of multiple random variables

- Variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent iff for every partition of the index set $\{1, \ldots, n\}$ into disjoint subsets A and B, the random vectors $\mathbf{x}_A$ and $\mathbf{x}_B$ are independent.
- More actionable characterisation: Variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent iff

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} p(\mathbf{x}_i) \tag{8}$$

  where $p(\mathbf{x}_i)$ is the marginal for $\mathbf{x}_i$.
- We say that the joint factorises into a product of the marginals.
- Notation: $\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp \mathbf{x}_n$

Conditional statistical independence

- The characterisation of statistical independence extends to conditional pdfs (pmfs) $p(\mathbf{x}, \mathbf{y}|\mathbf{z})$.
- Criteria from before carry over: functions do now also depend on $\mathbf{z}$.
- $\mathbf{x}$ and $\mathbf{y}$ are conditionally independent given $\mathbf{z}$ iff, for all possible values of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$,

$$p(\mathbf{x}, \mathbf{y}|\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z}) \quad (\text{for } p(\mathbf{z}) > 0) \qquad \text{or} \tag{9}$$

$$p(\mathbf{x}|\mathbf{y}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z}) \quad (\text{for } p(\mathbf{y}, \mathbf{z}) > 0) \tag{10}$$

- Proof of equivalence analogue to unconditional case.
- Notation: $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$
- From the product rule it follows that the joint $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ factorises as $p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z})$ when $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$.

The impact of independence assumptions

- The key is that independence assumptions lead to a partial factorisation of the pdf/pmf with factors that involve fewer variables.
- Independence assumptions force $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ to take on a particular form.
- Reduces the number of directly interacting variables and thereby the amount of numbers (parameters) that specify a pdf/pmf.

Example: table representation without independence

- Let $\mathbf{x}, \mathbf{y}, \mathbf{z}$ all be one-dimensional and binary.
- Without independence, we need to specify $2^3 - 1 = 7$ non-negative parameters p_i to specify the pmf.
- Table representation

x	y	z	$p(x, y, z)$
0	0	0	p_1
0	0	1	p_2
0	1	0	p_3
0	1	1	p_4
1	0	0	p_5
1	0	1	p_6
1	1	0	p_7
1	1	1	p_8

with the constraint that $\sum_i p_i = 1$, which removes one degree of freedom so that we only need to specify 7 p_i.

Example: table representation with full independence

- With full independence, $p(x, y, z) = p(x)p(y)p(z)$.
- 3 non-negative parameters, $p(x = 1) = p_1$, $p(y = 1) = p_2$, and $p(z = 1) = p_3$, fully specify the pmf

x	y	z	$p(x, y, z) = p(x)p(y)p(z)$
0	0	0	$(1 - p_1)(1 - p_2)(1 - p_3)$
0	0	1	$(1 - p_1)(1 - p_2)p_3$
0	1	0	$(1 - p_1)p_2(1 - p_3)$
0	1	1	$(1 - p_1)p_2p_3$
1	0	0	$p_1(1 - p_2)(1 - p_3)$
1	0	1	$p_1(1 - p_2)p_3$
1	1	0	$p_1p_2(1 - p_3)$
1	1	1	$p_1p_2p_3$

- x, y, z are not interacting: the probability of joint events, e.g. $\{x = 1 \text{ and } y = 1 \text{ and } z = 1\}$, is fully determined by the marginal probabilities.

Example: table repr with conditional independence

- Assume $x \perp\!\!\!\perp y \mid z$ so that $p(x, y, z) = p(z)p(x|z)p(y|z)$
- For $p(z)$ we need 1 parameter
- For $p(x|z)$, we need 2 parameters: one for $p(x|z = 0)$ and one for $p(x|z = 1)$.
- Same for $p(y|z)$.
- Total: $1+2+2 = 5$ non-negative parameters
- With

$$p_1 = p(z = 1) \qquad p_2 = p(x = 1|z = 0), \quad p_3 = p(x = 1|z = 1),$$
$$p_4 = p(y = 1|z = 0), \quad p_5 = p(y = 1|z = 1)$$

  we can represent $p(x, y, z)$ as a table.

 ©Gutmann, University of Edinburgh CC BY 4.0

Conditional independence is often a good middle-ground

- Consider $p(\mathbf{x}) = p(x_1, \ldots, x_d)$, with each x_i taking on K different values (e.g. $d = 100$, $K = 10$).
- No independence: $K^d - 1$ parameters, e.g. $10^{100} - 1$
- Full independence (factorisation): $d(K-1)$, e.g. 900
- For conditional independence $x_{i+1} \perp\!\!\!\perp x_1, \ldots, x_{i-1} \mid x_i$ (future independent of the past given the present), we have (see later)

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_2)\ldots p(x_d|x_{d-1}) \qquad (11)$$

  The number of parameters is $K - 1 + (d-1)K(K-1)$, e.g. 8919
- While the case of no independence is not tractable and full independence often too strong an assumption, conditional independence assumptions are often a powerful middle-ground.

Program

1. Independence assumptions
2. Assumptions on form of interaction
   - Parametric models restrict how a given number of variables may interact (autoregressive models)
   - Combination with independence assumptions

Assumption 2: limiting the form of the interaction

- (Conditional) independence assumptions limit the number of variables that may directly interact with each other, e.g. x_{i+1} only directly interacted with x_i.
- *How* the variables interact, however, was not restricted.
- Assumption 2: We restrict how a given number of variables may interact with each other.
- Often corresponds to making parametric family assumptions.

Autoregressive model for binary variables

- In the chain rule $p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i|\mathrm{pre}_i)$, specify each $p(x_i|\mathrm{pre}_i)$ as a member of a parametric family. No independence assumption made.
- Defines so-called autoregressive models.
- Let the variables be binary and assume

$$p(x_i = 1|\mathrm{pre}_i) = \frac{1}{1 + \exp\left(-b_i - \sum_{j=1}^{i-1} w_{ij}x_j\right)} \tag{12}$$

- The parameters w_{ij} can be stored as a $d \times d$ matrix $\mathbf{W}$

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ w_{21} & 0 & 0 & \dots & 0 & 0 \\ w_{31} & w_{32} & 0 & \dots & 0 & 0 \\ \vdots & & & & & \\ w_{d1} & w_{d2} & w_{d3} & \dots & w_{d(d-1)} & 0 \end{pmatrix} \tag{13}$$

- Matrix contains $(d^2 - d)/2$ parameters w_{ij}.
- With d biases b_i: $(d^2 - d)/2 + d = (d^2 + d)/2$ parameters

Autoregressive models for binary variables

- Table representation without independence requires $2^d - 1$ parameters
- For $d = 100$: 5050 vs $2^{100} - 1 \approx 1.27 \cdot 10^{30}$ parameters.
- Instead of linear combination of the predecessors, $\sum_{j=1}^{i-1} w_{ij} x_j$ we may use (parameterised) nonlinear functions such as neural networks.
- Leads to deep generative modelling (see later).
- We can use the same idea for continuous variables.

Recap: the univariate Gaussian distribution

- A real-valued random variable x is said to be Gaussian (normally) distributed if it has the pdf

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{14}$$

- Properties:
  - $\mathbb{E}[x] = \mu$, $\mathbb{V}[x] = \sigma^2$
  - Any linear combination of univariate Gaussians is Gaussian.
- Bell-shaped, symmetric around μ.
- Notation: $x \sim \mathcal{N}(x; \mu, \sigma^2)$.
- Can be generated by $x = \mu + \sigma n$ where $n \sim \mathcal{N}(n; 0, 1)$.
- Libraries can generate samples from $\mathcal{N}(n; 0, 1)$

Recap: the multivariate Gaussian distribution

- A random vector $\mathbf{x} \in \mathbb{R}^d$ is multivariate Gaussian (normal) if for all projections $\mathbf{a}$, $\mathbf{a}^\top \mathbf{x}$ is univariate Gaussian.
- If $\mathbf{x}$ has a density, it equals

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \quad (15)$$

- Properties:
  - $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$, $\mathbb{V}[\mathbf{x}] = \boldsymbol{\Sigma}$
  - Isocontours, i.e. the $\mathbf{x}$ where $p(\mathbf{x}) =$ const, are ellipses.
- Notation: $\mathbf{x} \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$

Autoregressive model for continuous variables

- We could "convert" continuous variables into discrete ones by discretisation. Loses information and number of bins grows as K^d when each variable has K discretisation levels.
- Use chain rule with parametric assumptions instead.
- For $p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i|\mathrm{pre}_i)$, assume each $p(x_i|\mathrm{pre}_i)$ is a univariate Gaussian where the mean and, possibly, variance depend on pre_i.
- Simplest case: $p(x_i|\mathrm{pre}_i)$ is Gaussian with constant variance σ_i^2 and means μ_i that depend linearly on pre_i

$$\mu_1 = b_1, \qquad \mu_i = b_i + \sum_{j=1}^{i-1} w_{ij} x_j \quad (i > 1) \tag{16}$$

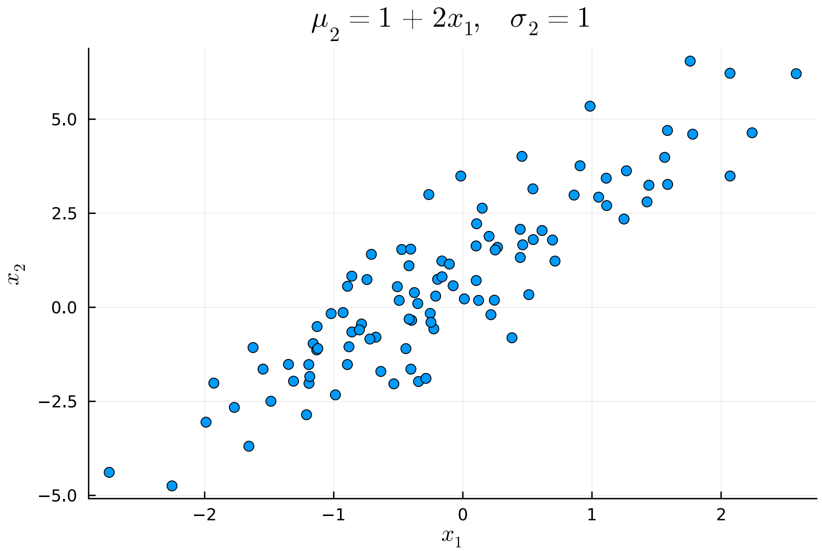
- Has $(d^2 + d)/2 + d$ parameters (same reasoning as before).
- Defines a multivariate Gaussian.

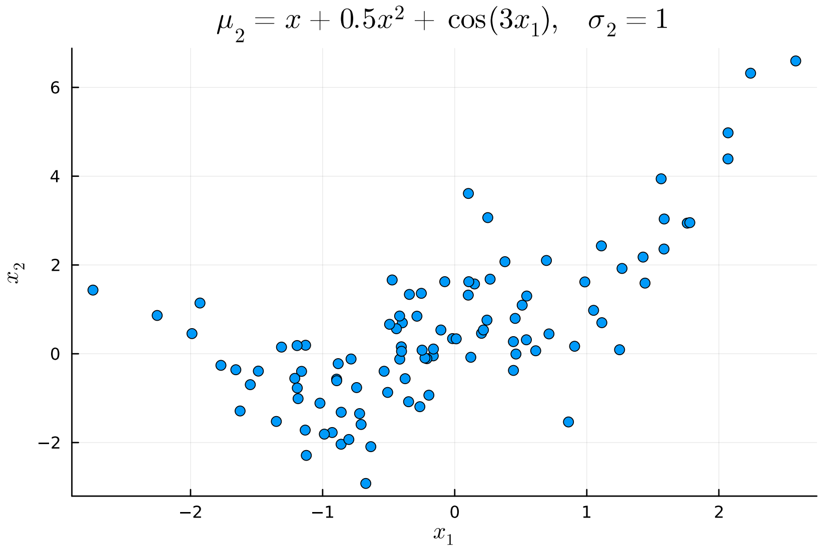
Autoregressive model for continuous variables

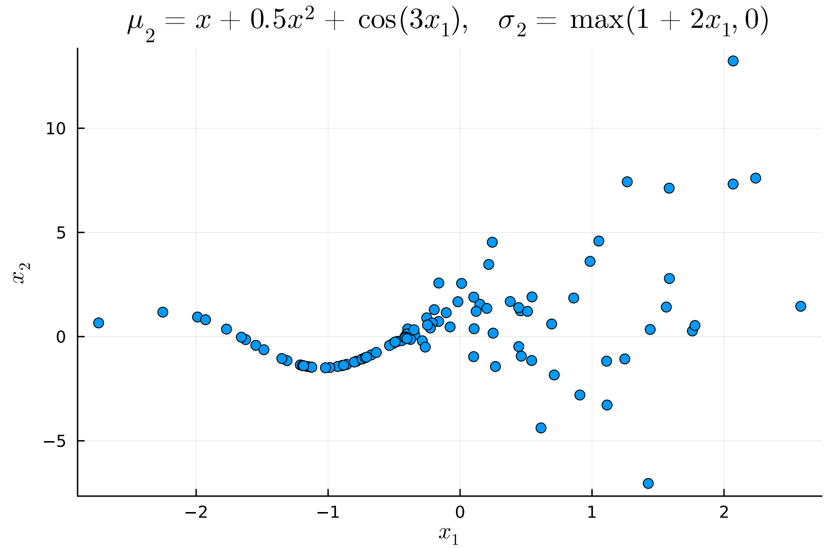
- More complex cases obtained by
  - letting variance depend on pre_i,
  - replacing the linear combination $\sum_{j=1}^{i-1} w_{ij} x_j$ with a (parameterised) nonlinear function such as a neural network.
- In the second case, each conditional mean depends nonlinearly on the predecessors.
- Each factor $p(x_i|\mathrm{pre}_i)$ defines a nonlinear regression model.
- While each factor $p(x_i|\mathrm{pre}_i)$ is conditionally Gaussian, the overall pdf $p(\mathbf{x})$ is not multivariate Gaussian.

Autoregressive model for continuous variables

x_1 is standard normal, and x_2 is Gaussian with different conditional means and variances.







Combining independence and parametric assumptions

- Reconsider the case where $x_{i+1} \perp\!\!\!\perp x_1, \ldots, x_{i-1} \mid x_i$ (future independent of the past given the present), so that

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_2)\ldots p(x_d|x_{d-1}) \tag{17}$$

- Before, we discussed the case of discrete random variables with a table representation.
- For continuous random variables, we can represent $p(x_{i+1}|x_i)$ with a parametric distribution, e.g. a Gaussian with a nonlinear mean function.
- We then make two assumptions: independence assumptions and parametric assumptions.
- These two assumptions are main workhorses to specify models in probabilistic machine learning (an additional one are latent variables, see later).

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 models restrict how a given number of variables may interact (autoregressive models)
   - Combination with independence assumptions