Lecture 22, Dec 1, 2023

Gaussian Probability Distributions

- A 1-dimensional Gaussian PDF is given by $f(x|\mu,\sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\dfrac{1}{2}\dfrac{(x-\mu)^2}{\sigma^2}\right)$
- Why are Gaussians common?
  - Central limit theorem: averages of independently drawn random variables become normally distributed when the number of random variables is sufficiently large
    - If we don't actually know the distributions, we can usually approximate it as a Gaussian
    - Let $y_1, \dots, y_n$ be a sequence of n independent random variables drawn from a distribution with finite mean and variance and let $\bar{y} = y_1 + \cdots + y_n$, then for some $a, b \in \mathbb{R}$, $\lim_{n\to\infty} \Pr\left(a < \dfrac{\bar{y} - n\mu}{\sqrt{n}\sigma} < b\right) = \int_a^b \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\,\mathrm{d}y$, i.e. $\bar{y}$ is Gaussian distributed with mean $n\mu$ and variance $n\sigma^2$
    - This still holds if the y_i come from different distributions, provided the mean and variance are finite for each PDF
  - They are easy to handle mathematically – you only need the mean and variance
  - They remain Gaussian under many operations (summation, marginalization, conditioning, etc)
- In multiple variables, $f(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \dfrac{1}{\sqrt{(2\pi)^M \det\boldsymbol{\Sigma}}}\exp\left(-\dfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$
  - $\boldsymbol{\mu} \in \mathbb{R}^M$ is the mean and $\boldsymbol{\Sigma} \in \mathbb{R}^{M\times M}$ is the symmetric, positive definite covariance matrix
  - This can be visualized as an ellipse around the mean

n	$\Pr(-n\sigma \le x \le n\sigma)$
0.67	≈ 0.5
1	≈ 0.68
2	≈ 0.95
3	≈ 0.997
4	≈ 0.99994

Figure 1: Probability of samples lying in different intervals around the mean.

- The above table shows the probability of a sample lying within a certain number of standard deviations of the mean
  - Note that this is for 1 dimension only; in higher dimensional space we need to look at probability ellipses
- Given some multivariate Gaussian $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, we can partition it as $\boldsymbol{x} = \begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}$, then we can write $f(\boldsymbol{x}) = f(\boldsymbol{x}_1, \boldsymbol{x}_2)$
  - We can marginalize to find $f(\boldsymbol{x}_1) = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) = \int_{-\infty}^{\infty} f(\boldsymbol{x}_1, \boldsymbol{x}_2|\boldsymbol{\mu},\boldsymbol{\Sigma})\,\mathrm{d}\boldsymbol{x}_2$ and likewise for $\boldsymbol{x}_2$
    - Marginalization picks out the relevant subblocks of the partitioned Gaussian, and gives Gaussian marginals
  - For conditioning, $f(\boldsymbol{x}_1|\boldsymbol{x}_2) = \mathcal{N}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$
    - The factors are Gaussian
    - This can be derived using the Schur complement
      - Using an LDL decomposition we can find the inverse of $\boldsymbol{\Sigma}$, and then substitute into $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$
- The sum of two independent Gaussians is Gaussian: $\boldsymbol{y}_1 \sim \mathcal{N}(\boldsymbol{a}_1, \boldsymbol{B}_1), \boldsymbol{y}_2 \sim \mathcal{N}(\boldsymbol{a}_2, \boldsymbol{B}_2) \implies c_1\boldsymbol{y}_1 + c_2\boldsymbol{y}_2 \sim \mathcal{N}(c_1\boldsymbol{a}_1 + c_2\boldsymbol{a}_2, c_1^2\boldsymbol{B}_1 + c_2^2\boldsymbol{B}_2)$
  - This extends to matrix coefficients, since the transformation is linear

- In general passing a Gaussian through a linear transformation preserves the Gaussian property
  - $\boldsymbol{C}_1\boldsymbol{y}_1 + \boldsymbol{C}_2\boldsymbol{y}_2 \sim \mathcal{N}(\boldsymbol{C}_1\boldsymbol{a}_1 + \boldsymbol{C}_2\boldsymbol{a}_2, \boldsymbol{C}_1^T\boldsymbol{B}_1\boldsymbol{C}_1, \boldsymbol{C}_2^T\boldsymbol{B}_2\boldsymbol{C}_2)$
- However, Gaussians don't remain Gaussian after passing through a nonlinear mapping
  - We can approximate using a linear function around the mean, so that if $y = g(x)$, then $x \sim \mathcal{N}(\mu_x, \sigma_x^2) \implies y \sim \mathcal{N}(\mu_y, a^2\sigma_x^2)$
    * $\delta y = y - \mu_y \approx \left.\frac{\mathrm{d}g(x)}{\mathrm{d}x}\right|_{x=\mu_x}(x - \mu x) = a\delta x$
    * $\sigma_y^2 = E[\delta y^2] = a^2E[\delta x^2] = a^2\sigma_x^2$
  - If the nonlinear function can be approximated as linear in $[-3\sigma, 3\sigma]$, then the resulting Gaussian is a good approximation
  - In multiple variables: $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x), \boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x}) \implies \boldsymbol{y} \sim (\boldsymbol{g}(\boldsymbol{\mu}_x), \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^T)$, where $\boldsymbol{A} = \left.\frac{\partial \boldsymbol{g}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{\mu}_x}$ is the Jacobian of $\boldsymbol{g}(\boldsymbol{x})$, $A_{ij} = \frac{\partial g_i}{\partial x_j}$
    * To see this, note $\Delta\boldsymbol{y} = \left.\frac{\partial \boldsymbol{g}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{\mu}_x}\Delta\boldsymbol{x}$, so $\boldsymbol{\Sigma}_y = E[\Delta\boldsymbol{y}\Delta\boldsymbol{y}^T] = \boldsymbol{A}E[\Delta\boldsymbol{x}\Delta\boldsymbol{x}^T]\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^T$
- We can fuse two Gaussians by multiplying them together and then renormalizing; this gives another Gaussian
  - This comes up when we want to combine multiple sources of information with different uncertainties
  - For two Gaussians with means μ_1, μ_2 and variances σ_1^2, σ_2^2, then $\frac{1}{\sigma^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}, \frac{\mu}{\sigma^2} = \frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}$
    * Notice the means are weighted by the inverse of their variances, since a lower variance means more certainty
    * The inverse variance is sometimes referred to as the *precision* of the Gaussian
  - The direct product of two Gaussians has the exponent $\frac{(x-\mu)^2}{\sigma^2} = \frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}$
    * $\frac{x^2 - 2\mu x + \mu^2}{\sigma^2} = \frac{(\sigma_1^2+\sigma_2^2)x^2 - 2(\sigma_2^2\mu_1 + \sigma_1^2\mu_2)x + (\sigma_2^2\mu_1^2 + \sigma_2^2\mu_2^2)}{\sigma_1^2\sigma_2^2}$
    * Comparing the x^2 terms gives $\frac{1}{\sigma^2} = \frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}$
    * Comparing the x terms gives $\frac{\mu}{\sigma^2} = \frac{\sigma_2^2\mu_1 + \sigma_1^2\mu_2}{\sigma_1^2\sigma_2^2} = \frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}$
    * Note the constant terms are not equal, but this is fixed by normalization
  - Note the product of Gaussians is not normalized, so we need to find the normalization constant so that the distribution integrates to 1
    * In practice however we almost never have to compute this, since we usually only keep track of the mean and variance
  - In the multivariate case, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \beta\prod_{n=1}^{N}\mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$
    * $\boldsymbol{\Sigma}^{-1} = \sum_{n=1}^{n}\boldsymbol{\Sigma}_n^{-1}$
    * $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \sum_{n=1}^{N}\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\mu}_n$
- Since Gaussians remain Gaussian under many different useful operations, it often suffices to keep track of only the sum and (co)variance of the distributions
- Under the assumption that random variables are Gaussian, analytic results for state estimation and other applications are available (e.g. a Kalman filter); without this assumption PDFs often have to be propagated through sampling (e.g. Monte Carlo methods)