Lecture 21, Nov 29, 2023

Extracting Estimates from PDFs

Maximum Likelihood (ML)

- This method is often used when $\boldsymbol{x}$ is an unknown constant parameter without a known probabilistic description, i.e. we have no prior information about $\boldsymbol{x}$
  - e.g. in Bayesian estimation, we had a prior (prediction) for $\boldsymbol{x}$, but here we are assuming no knowledge of that
- For a given observation $\boldsymbol{y}$ and observation model $f(\boldsymbol{y}|\boldsymbol{x})$, the method seeks a value of $\boldsymbol{x}$ that maximizes the likelihood of observing $\boldsymbol{y}$, i.e. $\hat{\boldsymbol{x}}^{ML} = \underset{\boldsymbol{x}\in\mathcal{X}}{\operatorname{argmax}}\, f(\boldsymbol{y}|\boldsymbol{x})$
  - $f(\boldsymbol{y}|\boldsymbol{x})$ as a function of $\boldsymbol{x}$ is the *likelihood function*
  - $\boldsymbol{x}$ is a parameter of the observation model; e.g. the model can be a Gaussian, and $\boldsymbol{x}$ may denote its mean or variance, etc
- Example: Consider two measurements of a scalar quantity $x \in \mathbb{R}$: $y_1 = x + w_1, y_2 = x + w_2$ where $w_1, w_2 \sim \mathcal{N}(0,1)$
  - Note $\mathcal{N}(\mu, \sigma)$ denotes a Gaussian with mean μ and variance σ
  - $f(w_i) = \dfrac{1}{\sqrt{2\pi}} \exp\left(-\dfrac{w_i^2}{2}\right)$
  - We can consider w_1, w_2 as additive noise parameters; this essentially makes $y_i \sim \mathcal{N}(x, 1)$
    - Note formally we would use a change of variables: $y_i = x + w_i \implies w_i = y_i - x$
    - Now we can just substitute w_i into the Gaussian equation since we have a linear relationship
  - y_1, y_2 are conditionally independent on x, so $f(y_1, y_2|x) = f(y_1|x)f(y_2|x) = \dfrac{1}{2\pi} e^{-\frac{1}{2}\left((y_1-x)^2+(y_2-x)^2\right)}$
  - This is now an unconstrained optimization problem; we can differentiate with respect to x and set this to 0
  - We get $(y_1 - \hat{x}) + (y_2 - \hat{x}) = 0 \implies \hat{x} = \dfrac{y_1 + y_2}{2}$, which is just the average
- Suppose we generalize the last example to a collection of measurements $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{w}$ where $\boldsymbol{z}, \boldsymbol{w} \in \mathbb{R}^m, \boldsymbol{x} \in \mathbb{R}^n$ and $m > n$; as above $w_i \sim (0, 1)$ are independent
  - Let $\boldsymbol{H} = \begin{bmatrix} \boldsymbol{h}_1^T \\ \vdots \\ \boldsymbol{h}_m^T \end{bmatrix}$ where $\boldsymbol{h}_i^T = \begin{bmatrix} h_{i_1} & \cdots & h_{in} \end{bmatrix}$
  - Then $z_i = \boldsymbol{h}_i^T \boldsymbol{x} + \boldsymbol{w}_i$
  - As before $f(\boldsymbol{z}|\boldsymbol{x}) \propto \exp\left(-\dfrac{1}{2}\left((z_1 - \boldsymbol{h}_1^T\boldsymbol{x})^2 + \cdots + (z_m - \boldsymbol{h}_m^T\boldsymbol{x})^2\right)\right)$
  - Differentiating with respect to each x_j we have $(z_1 - \boldsymbol{h}_1^T\hat{\boldsymbol{x}})h_{1j} + \cdots + (z_m - \boldsymbol{h}_m^T\hat{\boldsymbol{x}})h_{mj} = \begin{bmatrix} h_{1j} & \cdots & h_{mj} \end{bmatrix}(\boldsymbol{z} - \boldsymbol{H}\hat{\boldsymbol{x}}) = 0$
  - With all the rows, we get $\boldsymbol{H}^T(\boldsymbol{z} - \boldsymbol{H}\hat{\boldsymbol{x}}) = \boldsymbol{0} \implies \hat{\boldsymbol{x}} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{z}$, which is the least squares solution
    - Note we can write $\boldsymbol{w}(\boldsymbol{x}) = \boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}$, so $\boldsymbol{w}$ is some error term; then $\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}}\, \boldsymbol{w}^T\boldsymbol{w}$
    - If not all the errors have the same variance, then we have weighted least squares

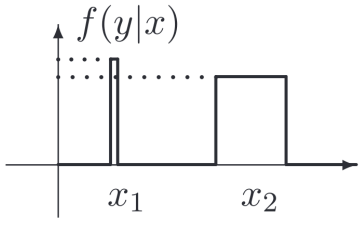

Figure 1: Undesirable case of maximum likelihood.

- Limitations of ML:
  - In general ML is more sensitive to outliers and modelling error
  - The maximum of the distribution may not always be what we want – we may lose robustness
    * In the example above, ML will give x_1 if there are measurements on it, which is very sensitive to changes in the data or model – small variations in the model might cause x_1 to have a likelihood of zero instead
    * Choosing x_2 is more robust; since the distribution is wider, we're less sensitive to changes in the data or model
    * Outliers that happen to line up with a peak can give us an incorrect estimate
  - We might also have prior knowledge about x (i.e. its PDF), which ML cannot incorporate

Maximum *a Posteriori* (MAP)

- If we have a PDF for $\boldsymbol{x}$, we can use MAP
- From Bayes's theorem: $f(\boldsymbol{x}|\boldsymbol{y}) = \dfrac{\boldsymbol{f}(\boldsymbol{y}|\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x})}{\boldsymbol{f}(\boldsymbol{y})}$
- With MAP, we have $\hat{\boldsymbol{x}}^{MAP} = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x})$
  - We want to maximize the choice of the parameter that makes both the observations and the parameter itself most likely
- If $f(\boldsymbol{x})$ is constant, then $\hat{\boldsymbol{x}}^{MAP} = \hat{\boldsymbol{x}}^{ML}$
- As with ML, we are still maximizing a function over $\boldsymbol{x}$, so the same sensitivity to outliers and modelling error still applies
- Example: consider the scalar observation $y = x + w$, where $w \sim \mathcal{N}(0,1), x \sim \mathcal{N}(\bar{x}, \sigma_x^2)$ and x, w independent
  - $f(x) \propto \exp\left(-\dfrac{1}{2}\dfrac{(x-\bar{x})^2}{\sigma_x^2}\right)$
  - $f(y|x) \propto \exp\left(-\dfrac{1}{2}(y-x)^2\right)$
  - $f(y|x)f(x) \propto \exp\left(-\dfrac{1}{2}\left(\dfrac{(x-\bar{x})^2}{\sigma_x^2} + (y-x)^2\right)\right)$
  - Differentiating with respect to x and setting to zero gives the following solution:
  - $\hat{x}^{MAP} = \dfrac{1}{1+\sigma_x^2}\bar{x} + \dfrac{\sigma_x^2}{1+\sigma_x^2}y$
    * Notice that this is a weighted sum between the mean of the prior distribution and the new measurement
  - Consider the extreme cases:
    * $\sigma_x^2 = 0 \implies \hat{x}^{MAP} = \bar{x}$ (if we're certain about x before any measurements, we just get the max of the prior)
    * $\sigma_x^2 \to \infty \implies \hat{x}^{MAP} = y$ (if we're uncertain about x, we just get the new measurement; note this is the same as maximum likelihood)
- This is most often used in state estimation

Minimum Mean Squared Error (MMSE)

- The MMSE is the a posteriori estimate that minimizes the mean squared error
- $\hat{\boldsymbol{x}}^{MMSE} = \underset{\hat{\boldsymbol{x}}}{\operatorname{argmin}}\, E_{x|y}\left[(\boldsymbol{x}-\hat{\boldsymbol{x}})^T(\boldsymbol{x}-\hat{\boldsymbol{x}})|y\right]$
  - Expand and differentiate with respect to $\hat{\boldsymbol{x}}$: $2\hat{\boldsymbol{x}} - 2E[x|y] = 0 \implies \hat{\boldsymbol{x}} = E[x|y]$
  - The MMSE estimate is the expected value of $\boldsymbol{x}$ conditioned on $\boldsymbol{y}$
- While MAP is the maximum of the posterior, MMSE is the mean of the posterior
- Note we did not constrain $\hat{\boldsymbol{x}}$ in our minimization, but for some applications we might want to introduce constraints
  - e.g. for a discrete random variable with sample space $\mathcal{X}$, we need to constrain the minimization to $\hat{\boldsymbol{x}} \in \mathcal{X}$

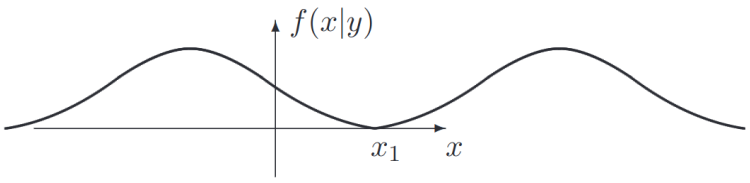

Figure 2: Undesirable case of MMSE.

- MMSE always takes the whole probability mass into consideration, whereas MAP and ML just pick the maximum probability – in some cases, this is desirable, while in other cases it is not
  - Consider the bimodal distribution of $f(x|y)$ above; MMSE would give x_1, but the probability of having x actually being near x_1 is zero
    * MAP would have picked one of the two peaks
  - On the other hand, the MMSE is typically more robust to modelling errors and outliers, since it is not as sensitive to sharp peaks