Lecture 5, Jan 22, 2024

Maximum A Posteriori (MAP) Estimation

- MAP estimation tries to maximize the probability of the posterior, using a Bayesian approach
- $\hat{\Theta}_n = \underset{\theta}{\operatorname{argmax}}\, p_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x}) = \underset{\theta}{\operatorname{argmax}} \dfrac{p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\Theta) f_\Theta(\theta)}{p_{\boldsymbol{X}}(\boldsymbol{x})}$
  - As with MLE, sometimes it is more convenient to use the log of the posterior instead
  - To simplify the computation we often pick a prior for Θ that matches the form of the likelihood function; this is known as a *conjugate prior*; important ones include:
    * Beta: binomial, geometric
    * Dirichlet: multinomial
    * Gamma: Poisson, exponential
    * Gaussian: Gaussian
  - Note the distribution $p_{\boldsymbol{X}}(\boldsymbol{x})$ usually doesn't matter since it's constant wrt θ
- Example: binomial distribution $p_{X|\Theta}(x|\theta) = \dbinom{n}{k}\theta^k(1-\theta)^{n-k} = \dfrac{n!}{k!(n-k)!}\theta^k(1-\theta)^{n-k}$
  - There are many possible shapes of priors
  - These are all represented by the *beta distribution* $f_\Theta(\theta) = c\theta^{\alpha-1}(1-\theta)^{\beta-1}$ where $\alpha, \beta > 0, 0 \leq \theta \leq 1$ and c is a normalization constant
    * When $\alpha = \beta = 1$ this is uniform
    * $c = \dfrac{1}{B(\alpha,\beta)}$ where $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
      - Note $\Gamma(m+1) = m!$ for integer m
    * If α, β are integers then $\dfrac{1}{B(\alpha,\beta)} = \dfrac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}$
    * This has mean at $E[\Theta] = \dfrac{1}{B(\alpha,\beta)}$

$$\begin{aligned} &= \int_0^1 \theta f_\Theta(\theta)\,\mathrm{d}\theta \\ &= \int_0^1 \theta^\alpha(1-\theta)^{\beta-1}\,\mathrm{d}\theta \\ &= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \\ &= \frac{\alpha}{\alpha+\beta} \end{aligned}$$

    * Maximum at $\theta = \dfrac{\alpha-1}{\alpha+\beta-2}$
  - The beta distribution is the conjugate prior of the binomial distribution
  - $p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta) f_\Theta(\theta) = \dfrac{\binom{n}{k}}{B(\alpha,\beta)}\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$
  - $p_{\boldsymbol{X}}(\boldsymbol{x}) = \dfrac{\binom{n}{k}}{B(\alpha,\beta)}\displaystyle\int_0^1 \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}\,\mathrm{d}\theta$
    * Note that the integral is just $B(k+\alpha, n-k+\beta)$
    * Therefore $p_{\boldsymbol{X}}(\boldsymbol{x}) = \dfrac{n!}{k!(n-k)!}\dfrac{\Gamma(\alpha+\beta)\Gamma(k+\alpha)\Gamma(n-k+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+n+\beta)}$
  - Solve $\dfrac{\mathrm{d}}{\mathrm{d}\theta}\log f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x}) = 0$
    * $\dfrac{\mathrm{d}}{\mathrm{d}\theta}\log\left(c\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}\right) = \dfrac{k+\alpha-1}{\theta} - \dfrac{n-k+\beta-1}{1-\theta} = 0$
    * $\hat{\theta} = \dfrac{k+\alpha-1}{n+\alpha+\beta-2}$

- The choice of α and β depends on our knowledge of the prior, e.g. where it peaks, how much variance it has, etc
  * Notice that $\lim_{n\to\infty} \hat{\theta}_{\text{MAP}} = \dfrac{k}{n} = \hat{\theta}_{\text{ML}}$
  * As we take more and more trials, the prior distribution of θ becomes irrelevant since the estimate converges by the weak law

Least Mean Square and Conditional Expectation

- We want to find an estimator that minimizes the mean squared difference between the true value and the estimated value
  - This is another Bayesian approach since we need the prior
- $\hat{\theta}_{\text{LMS}} = \operatorname*{argmin}_{\hat{\theta}} E[(\hat{\theta} - \Theta)^2] = E[\Theta|\boldsymbol{X} = \boldsymbol{x}]$
- Suppose we have no data, so we estimate Θ by a constant $\hat{\theta}$:
  - $E[(\hat{\theta} - \Theta)^2] = E[\Theta^2 - 2\Theta\hat{\theta} + \hat{\theta}^2] = \hat{\theta}^2 - 2\hat{\theta}E[\Theta] + E[\Theta]$
  - Differentiate: $2\hat{\theta} - 2E[\Theta] = 0$
  - So in this case the best estimate is $\hat{\theta} = E[\Theta]$
- If we do have data:
  - $E[(\hat{\theta} - \Theta)^2] = E[E[(\hat{\theta} - \Theta)^2|\boldsymbol{x}]] = \displaystyle\int_{-\infty}^{\infty} E[(\Theta - \hat{\theta})|\boldsymbol{X} = \boldsymbol{x}]f_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$
  - This can then be minimized by taking $\hat{\theta} = E[\Theta|\boldsymbol{X} = \boldsymbol{x}]$ following the same derivation as the case above

Comparison of MLE, MAP, and LMS Estimation

- Let Θ have a prior uniform on $[0, 1]$ and let X be distributed as uniformly on $[0, \Theta]$
  - The joint distribution covers a triangular area
  - $f(x|\theta)$ is uniform from 0 to θ with value $\dfrac{1}{\theta}$
  - $f(x, \theta) = f(x|\theta)f(\theta) = \dfrac{1}{\theta}\dfrac{1}{1} = \dfrac{1}{\theta}, 0 < x < \theta < 1$
- For ML:
  - Maximize $f(x|\theta)$
  - We need $\theta \geq x$ because otherwise the value of x couldn't possibly occur
  - And note $f(x|\theta) = \dfrac{1}{\theta}$ on $x \in [0, \theta]$ so to maximize this we take θ as small as possible
  - Therefore $\hat{\theta}_{\text{ML}} = x$
- For MAP:
  - $f(\theta|x) = \dfrac{f(x|\theta)f(\theta)}{f(x)} = \dfrac{f(\theta, x)}{\int_x^1 f(\theta, x)\,\mathrm{d}\theta} = \dfrac{1}{\theta \ln \frac{1}{x}}, 0 < x < \theta < 1$
  - To maximize this we again take $\hat{\theta}_{\text{MAP}} = x$
  - For this problem, the MAP and ML estimates are the same
- For LMS:
  - $\hat{\theta}_{\text{LMS}} = E[\Theta|x] = \displaystyle\int_x^1 \theta f(\theta|x)\,\mathrm{d}\theta = \int_x^1 \frac{\theta}{\theta \ln \frac{1}{x}}\,\mathrm{d}\theta = \frac{1 - x}{\ln \frac{1}{x}}$
  - In this case LMS is less conservative