Tutorial 2, Jan 26, 2024

Convergence and Weak Law

- *Markov inequality*: Let $X > 0$, then $P[X \geq a] \leq \frac{E[X]}{a}$
  - Let $Z(w) = \begin{cases} a & X(w) \geq a \\ 0 & \text{elsewhere} \end{cases}$
  - Observe that $Z \leq X$, so $E[Z] \leq E[X]$
  - Also, since Z is zero everywhere except where $X \geq a$ where it has a constant value a, we can easily find its expectation
  - $E[Z] = aP[X \geq a] \leq E[X] \implies P(X \geq a) \leq \frac{E[X]}{a}$
- *Chebyshev inequality*: $P[|X - m_X| \geq \epsilon] \leq \frac{\sigma^2}{\epsilon^2}$
  - This can be proven from the Markov inequality and does not require $X > 0$
  - Apply Markov to $(X - m_X)^2$
  - $P[(X - m_X)^2 \geq \epsilon^2] \leq \frac{E[(X - m_X)^2]}{\varepsilon^2}$
  - The numerator is the definition of variance and we can square root both sides in the probability
  - This gives exactly $P[|X - m_X| \geq \epsilon] \leq \frac{\sigma^2}{\epsilon^2}$

Definition

Convergence: Let $X, X_1, X_2, \ldots$ be a sequence of RVs; X_n converges to X $(X_n \to X)$ in probability if

$$\forall \epsilon > 0, \lim_{n \to \infty} P[|X_n - X| \geq \epsilon] = 0$$

- X is the "target", which can be constant but need not be

Definition

Weak Law of Large Numbers: Let $X_1, \ldots, X_n$ be a sequence of IID RVs and let $E[X_i] = m_X, \text{Var}(X_i) = \sigma^2$ both be finite, and

$$M_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

then M_n converges to m_X in probability as $n \to \infty$.

- Proof:
  - Notice $E[M_n] = \frac{1}{n} m_X$ and $\text{Var}(M_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$ due to IID implying covariance is zero
  - By Chebyshev, $P[|M_n - E[M_n]| \geq \epsilon] = P[|M_n - E[X_i]| \geq \epsilon] \leq \frac{\text{Var}(M_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$
  - Therefore for any fixed ϵ we have that as $n \to \infty$ this probability goes to zero
  - We can construct a stronger proof that shows this holds even for infinite variance

Parameter Estimation

- Suppose we have IID observations $x_1, \ldots, x_n$ coming from a model $f(\boldsymbol{x}; \theta)$; we wish to estimate θ
- The frequentist approach assumes that the parameter θ is a constant, while the Bayesian approach assumes that θ comes from a distribution
  - MLE is a frequentist approach while MAP, MLS are Bayesian approaches
- $\hat{\theta}_{\text{MLE}} = \underset{\theta}{\text{argmax}} \prod_{i=1}^{n} f(x_i; \theta) = \underset{\theta}{\text{argmax}} \sum_{i=1}^{n} \log(f(x_i; \theta))$

- Example: $X_1, \ldots, X_n$ are IID Gaussians with unit variance and unknown mean m_X: $X_i \sim \mathcal{N}(m_X, 1)$; estimate the mean using MLE
  - Multiplying the distributions adds the terms int he exponent
  - Maximizing the likelihood is equivalent to maximizing $\displaystyle\sum_{i=1}^{n} -\frac{1}{2}(x_i - m_X)^2$
  - Taking the derivative and setting to zero we obtain $m_X = \displaystyle\frac{1}{n}\sum_{i=1}^{n} x_i$
- Example: $f(x;\theta) = \begin{cases} \frac{1}{\theta}x^{\frac{1}{\theta}-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$
  - $$\begin{aligned} \hat{\theta}_{\text{MLE}} &= \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \log\left(\frac{1}{\theta}x_i^{\frac{1}{\theta}-1}\right) \\ &= \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \left(-\log\theta + \left(\frac{1}{\theta} - 1\right)\log x_i\right) \\ &= \operatorname*{argmin}_{\theta} n\log\theta + \left(1 - \frac{1}{\theta}\right)\sum_{i=1}^{n}\log(x_i) \end{aligned}$$
  - Differentiate: $\displaystyle\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n}\log x_i = 0 \implies \theta = -\frac{1}{n}\sum_{i=1}^{n}\log x_i$