Inference with Maximum Likelihood

René Magritte, "La reproduction interdite" (1937)

SML310: Research Projects in Data Science, Fall 2019

Michael Guerzhoy

Likelihood: Bernoulli Variables

- Suppose a coin is tossed n times, independently
  - $Y_i \sim Bernoulli(\theta)$
- $P(Y_i = 1) = \theta, P(Y_i = 0) = 1 - \theta$
- $Y_1, \ldots, Y_n$ are independently identically Bernoulli-distributed (i.i.d.)
- We observe the data $Y_1 = y_1, Y_2 = y_2, \ldots, Y_m = y_m$ (m i.i.d. Bernoulli variables), and would like to know what θ is
- How to do that? (Intuitively)
- How do you define what a good estimate is?

Likelihood: Bernoulli Variables

- The likelihood is* the probability of observing the dataset when the parameters are θ
  - $P(Y_i = 1|\theta) = \theta$
  - $P(Y_i = 0|\theta) = 1 - \theta$
  - $P(Y_i = y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$
  - $P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_m = y_m|\theta) = \prod_{i=1}^{m} P(Y_i = y_i|\theta)$
- **Confusing notation alert**: at this point, θ is not an event or an R.V., it's just a number.

* In the discrete case

Maximum likelihood: Bernoulli

- Suppose we observe the data $Y_1 = y_1, Y_2 = y_2, \dots, Y_m = y_m$ (m i.i.d. Bernoulli variables), and would like to know what θ is
- One possibility: find the θ that maximizes the likelihood function
    - What value of θ makes the data set that we are actually observing (i.e., the training set) the most plausible?
- $P(Y_1 = y_1, Y_2 = y_2, \dots, Y_m = y_m|\theta)$ is maximized at $\theta = \frac{1}{m}\sum_{i=1}^{m} y_i$

(Switch to R)

- Generate fake data from *n* Bernoullis
- Compute the likelihood
- Find the maximum-likelihood solution

Likelihood: Gaussian Noise

- Assume each data point is generated using some process.
  - E.g., $y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)},\ \epsilon^{(i)} \sim N(0, \sigma^2)$
- We can now compute the likelihood of single datapoint
  - I.e., the probability of the point for a set θ.
  - E.g., $P\left(\mathrm{y}^{(i)} \middle| \theta, x^{(i)} \right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\sigma^2} \right)$
  - We can then compute the likelihood for the entire set $\{\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \dots, \left(x^{(m)}, y^{(m)}\right)\}$ (assuming each point is independent)

(Switch to R)

- Generate fake data from a linear model with Gaussian residuals
- Find the maximum-likelihood slope assuming 0 intercept
  - Time-permitting

Maximum Likelihood: Least Squares

- $P(Y = y|\theta, x) = \Pi_1^m \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\sigma^2}\right)$
- $\log P(Y = y|\theta, x) = \sum -\frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\sigma^2} - \frac{m}{2}\log(2\pi\sigma^2)$

  is maximized for a value of θ for which

  $\sum_{i=1}^m \left(y^{(i)} - \theta^T x^{(i)}\right)^2$ is minimized