Learning with Maximum Likelihood

René Magritte, “La reproduction interdite” (1937)

CSC411/2515: Machine Learning and Data Mining, Winter 2018

Michael Guerzhoy and Lisa Zhang

Likelihood: Bernoulli Variables

- Suppose a fair coin is tossed n times, independently
  - $Y \sim Bernoulli(\theta)$
- The likelihood (discrete case) is the probability of observing the dataset when the parameters are θ)
  - $P(Y_i = 1|\theta) = \theta$
  - $P(Y_i = 1|\theta) = \theta$
  - $P(Y_i = y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$

  - $P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_m = y_m|\theta) = \prod_{i=1}^{m} P(Y_i = y_i|\theta)$

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 $\theta = \frac{1}{m}\sum_{i=1}^{m} y_i$

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 training 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)
  - E.g., $P(\, y | \theta, x) = \Pi_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y^{(i)} - \theta^T x^{(i)}\right)^2}{2\sigma^2}\right)$

Maximum Likelihood

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

  is maximized for a value of θ for which

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