

Learning with Maximum Likelihood: Linear Regression and Logistic Regression



René Magritte, "La reproduction interdite" (1937)

CSC411: Machine Learning and Data Mining, Winter 2017

Michael Guerzhoy

Review: Likelihood

- 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 single datapoint

- I.e., the probability of the point given θ .

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

- We can then compute the likelihood for the entire training set $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$ (assuming each point is independent)

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

Review: Maximum Likelihood

- Maximum Likelihood: the parameter θ for which the data is the most plausible

- $\operatorname{argmax}_{\theta} P(\text{data}|\theta)$

- E.g.:

$$\begin{aligned} P(\text{data}|\theta) &= P(y; x|\theta) \\ &= \prod_1^m \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \end{aligned}$$

- $\log P(\text{data}|\theta) = \sum -\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} + 2m/\log(2\pi\sigma^2)$

is maximized for a value of θ for which

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

Logistic Regression

- Assume the data is generated according to

$$y^{(i)} = 1 \text{ with probability } \frac{1}{1 + \exp(-\theta^T x^{(i)})}$$

$$y^{(i)} = 0 \text{ with probability } \frac{\exp(-\theta^T x^{(i)})}{1 + \exp(-\theta^T x^{(i)})}$$

- This can be written concisely as:

$$\frac{P(x^{(i)}, y^{(i)} = 1 | \theta)}{P(x^{(i)}, y^{(i)} = 0 | \theta)} = \exp(\theta^T x^{(i)})$$

odds

(exercise)

Logistic Regression: Likelihood

- $$P(x^{(i)}, y^{(i)} | \theta) = \left(\frac{1}{1 + \exp(-\theta^T x^{(i)})} \right)^{y^{(i)}} \left(\frac{\exp(-\theta^T x^{(i)})}{1 + \exp(-\theta^T x^{(i)})} \right)^{1 - y^{(i)}}$$

(just a trick that works because $y^{(i)}$ is either 1 or 0)

- $$P(data | \theta) = \prod_{i=1}^m \left(\frac{1}{1 + \exp(-\theta^T x^{(i)})} \right)^{y^{(i)}} \left(\frac{\exp(-\theta^T x^{(i)})}{1 + \exp(-\theta^T x^{(i)})} \right)^{1 - y^{(i)}}$$

- $$\log P(data | \theta) = \sum_{i=1}^m y^{(i)} \log \left(\frac{1}{1 + \exp(-\theta^T x^{(i)})} \right) + (1 - y^{(i)}) \log \left(\frac{\exp(-\theta^T x^{(i)})}{1 + \exp(-\theta^T x^{(i)})} \right)$$

Logistic Regression: Learning and Testing

- Learning: find the best θ that maximizes the log-likelihood:

$$\sum_{i=1}^m y^{(i)} \log\left(\frac{1}{1 + \exp(-\theta^T x^{(i)})}\right) + (1 - y^{(i)}) \log\left(\frac{\exp(-\theta^T x^{(i)})}{1 + \exp(-\theta^T x^{(i)})}\right)$$

- For x in the test set, compute

$$P(x, y = 1 | \theta) = \frac{1}{1 + \exp(-\theta^T x)}$$

- Predict $y=1$ if $P(y = 1; x | \theta) > .5$

Logistic Regression: Decision Surface

Logistic Regression: Decision Surface

- Predict $y=1$ if $\frac{1}{1+\exp(-\theta^T x)} > .5$



$$-\theta^T x < 0$$



$$\theta^T x > 0$$

- So the decision surface is $\theta^T x = 0$, a hyperplane

Logistic Regression

- Outputs the probability of the datapoint's belonging to a certain class:

$$y^{(i)} = 1 \text{ with probability } \frac{1}{1 + \exp(-\theta^T x^{(i)})}$$

$$y^{(i)} = 0 \text{ with probability } \frac{\exp(-\theta^T x^{(i)})}{1 + \exp(-\theta^T x^{(i)})}$$

(compare with linear regression)

- Linear decision surface
- Probably the first thing you would try in a real-world setting for a classification task