Learning with Maximum Likelihood: Linear Regression and Logistic Regression

René Magritte, “La reproduction interdite” (1937)

CSC321: Intro to Machine Learning and Neural Networks, Winter 2016

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

Review: Maximum Likelihood

- Maximum Likelihood: the parameter θ for which the data is the most plausible
  - $argmax_\theta P(data|\theta)$
  - E.g.:

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

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

    is maximized for a value of θ for which

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

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:

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

odds

(exercise)

Logistic Regression: Likelihood

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

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

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

Logistic Regression: Learning and Testing

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

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

- For x in the test set, compute

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

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

Logistic Regression: Decision Surface

Logistic Regression: Decision Surface

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

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

$$\Leftrightarrow \quad \theta^T x > 0$$

- 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:

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

  $y^{(i)} = 0$ 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