Machine Learning

Linear Models

Fabio Vandin

November 8th, 2022

Therefore, given training set $S = ((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m))$ the ERM problem for logistic regression is:

$$\arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{m} \sum_{i=1}^m \log\left(1 + e^{-y_i\langle \mathbf{w}, \mathbf{x}_i \rangle}\right)$$

Notes: logistic loss function is a *convex function* \Rightarrow ERM problem can be solved efficiently

Definition may look a bit arbitrary: actually, ERM formulation is the same as the one arising from *Maximum Likelihood Estimation*

Maximum Likelihood Estimation (MLE) [UML, 24.1]

MLE is a statistical approach for finding the parameters that maximize the joint probability of a given dataset assuming a specific parametric probability function.

Note: MLE essentially assumes a generative model for the data

General approach:

- 1 given training set $S = ((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m))$, assume each (\mathbf{x}_i, y_i) is i.i.d. from some probability distribution of parameters θ
- 2 consider $\mathbb{P}[S|\theta]$ (likelihood of data given parameters)

Maximum Likelihood Estimation (MLE) [UML, 24.1]

MLE is a statistical approach for finding the parameters that maximize the joint probability of a given dataset assuming a specific parametric probability function.

Note: MLE essentially assumes a *generative model* for the data

General approach:

- 1 given training set $S = ((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m))$, assume each (\mathbf{x}_i, y_i) is i.i.d. from some probability distribution of parameters θ
- 2 consider $\mathbb{P}[S|\theta]$ (likelihood of data given parameters)
- 3 log likelihood: $L(S; \theta) = \log(\mathbb{P}[S|\theta])$

Maximum Likelihood Estimation (MLE) [UML, 24.1]

MLE is a statistical approach for finding the parameters that maximize the joint probability of a given dataset assuming a specific parametric probability function.

Note: MLE essentially assumes a generative model for the data

General approach:

- 1 given training set $S = ((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m))$, assume each (\mathbf{x}_i, y_i) is i.i.d. from some probability distribution of parameters θ
- 2 consider $\mathbb{P}[S|\theta]$ (likelihood of data given parameters)
- 3 log likelihood: $L(S; \theta) = \log(\mathbb{P}[S|\theta])$
- **4** maximum likelihood estimator: $\hat{\theta} = \arg\max_{\theta} L(S; \theta)$

Logistic Regression and MLE

Assuming $\mathbf{x}_1, \dots, \mathbf{x}_m$ are fixed, the probability that \mathbf{x}_i has label $y_i = 1$ is

$$h_{\mathbf{w}}(\mathbf{x}_i) = \frac{1}{1 + e^{-(\mathbf{w}_i \mathbf{x}_i)}}$$

while the probability that \mathbf{x}_i has label $y_i = -1$ is

$$(1 - h_{\mathbf{w}}(\mathbf{x}_i)) = \frac{1}{1 + e^{(\mathbf{w})\mathbf{x}_i)}}$$

Logistic Regression and MLE

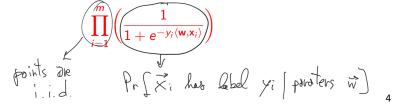
Assuming x_1, \dots, x_m are fixed, the probability that x_i has label $y_i = 1$ is

$$h_{\mathbf{w}}(\mathbf{x}_i) = \frac{1}{1 + e^{-\langle \mathbf{w}, \mathbf{x}_i \rangle}}$$

while the probability that x_i has label $y_i = -1$ is

$$(1 - h_{\mathbf{w}}(\mathbf{x}_i)) = \frac{1}{1 + e^{\langle \mathbf{w}, \mathbf{x}_i \rangle}}$$

Then the likelihood for training set S is:



Therefore the log likelihood is:

Interestic the log likelihood is.

$$\int_{a}^{b} \left(\frac{1}{1+e^{-y_{i} < w_{i} \times i}} \right)$$

$$= \sum_{i=1}^{m} \left(\int_{a}^{b} \left(1 \right) - \int_{a}^{b} \left(1 + e^{-y_{i} < w_{i} \times i} \right) \right)$$

$$= -\sum_{i=1}^{m} \left(\int_{a}^{b} \left(1 \right) - \int_{a}^{b} \left(1 + e^{-y_{i} < w_{i} \times i} \right) \right)$$

$$= -\sum_{i=1}^{m} \int_{a}^{b} \left(1 + e^{-y_{i} < w_{i} \times i} \right)$$

Therefore the log likelihood is:

$$-\sum_{i=1}^{m}\log\left(1+e^{-y_{i}\langle\mathbf{w},\mathbf{x}_{i}\rangle}\right)$$

And note that the maximum likelihood estimator for w is:

$$\arg\max_{\mathbf{w}\in\mathbb{R}^d} - \sum_{i=1}^m \log\left(1 + e^{-y_i\langle\mathbf{w},\mathbf{x}_i\rangle}\right) = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \sum_{i=1}^m \log\left(1 + e^{-y_i\langle\mathbf{w},\mathbf{x}_i\rangle}\right)$$

Therefore the log likelihood is:

$$-\sum_{i=1}^{m}\log\left(1+e^{-y_{i}\langle\mathbf{w},\mathbf{x}_{i}\rangle}\right)$$

And note that the maximum likelihood estimator for w is:

$$\arg\max_{\mathbf{w}\in\mathbb{R}^d} - \sum_{i=1}^m \log\left(1 + e^{-y_i\langle\mathbf{w},\mathbf{x}_i\rangle}\right) = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \sum_{i=1}^m \log\left(1 + e^{-y_i\langle\mathbf{w},\mathbf{x}_i\rangle}\right)$$

⇒ MLE solution is equivalent to ERM solution!

Bibliography

[UML] Chapter 9:

• no 9.1.1