Maximum Likelihood Estimation (MLE)

CHEN Si

Contents

1 Motivation

- Need some principle from which we can derive specific functions that are good estimators for different models

2 Maximum Likelihood Estimator

Problem Settings

- Consider a set of m examples $\mathbb{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ drawn **independently** from the **true but unknown** data-generating distribution $p_{\text{data}}(\boldsymbol{x})$
- Let $p_{\text{model}}(\boldsymbol{x}; \boldsymbol{\theta})$ be a parametric family of probability distributions over the same space indexed by $\boldsymbol{\theta}$

Definition

$$\begin{aligned}\boldsymbol{\theta}_{\mathrm{ML}} &= \arg\max_{\boldsymbol{\theta}} p_{\mathrm{model}}(\mathbb{X};\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{\mathrm{model}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta})\end{aligned}$$

- Disadvantage: Prone to numerical underflow

Logarithm of the Likelihood

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log p_{\mathrm{model}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta})$$

Rescale by m

$$\begin{aligned}\boldsymbol{\theta}_{\mathrm{ML}} &= \arg\max_{\boldsymbol{\theta}} \frac{1}{m}\sum_{i=1}^{m} \log p_{\mathrm{model}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}\sim\hat{p}_{\mathrm{data}}} \left[\log p_{\mathrm{model}}(\boldsymbol{x};\boldsymbol{\theta})\right]\end{aligned}$$

- This is an expectation with respect to the empirical distribution $\hat{p}_{\mathrm{data}}$ defined by the training data

2.1 Interpretation

- Minimize the cross-entropy

$$H(\hat{p}_{\mathrm{data}}, p_{\mathrm{model}}) = -\mathbb{E}_{\mathbf{x}\sim\hat{p}_{\mathrm{data}}} \left[\log p_{\mathrm{model}}(\boldsymbol{x})\right]$$

  between the empirical distribution difined by the training set and the probability distribution defined by model

- Equivalently, minimize the dissimilarity between $\hat{p}_{\mathrm{data}}$ and $\hat{p}_{\mathrm{model}}$

$$D_{\mathrm{KL}}(\hat{p}_{\mathrm{data}} \| p_{\mathrm{model}}) = \mathbb{E}_{\mathbf{x}\sim\hat{p}_{\mathrm{data}}} \left[\log \hat{p}_{\mathrm{data}}(\boldsymbol{x}) - \log p_{\mathrm{model}}(\boldsymbol{x};\boldsymbol{\theta})\right]$$

  , whose minimum is 0

- Make the model distribution match the empirical distribution

2.2 Properties

It can be shown to be the best estimator **asymptotically**, as the number of examples $m \to \infty$, in terms of its rate of convergence.

1. Consistency, if under appropriate conditions:
    - The true distribution p_{data} must lie within the model family $p_{\text{model}}(\cdot; \boldsymbol{\theta})$
      Otherwise, no estimator can recover p_{data}
    - The true distribution p_{data} must correspond to exactly one value of $\boldsymbol{\theta}$
      Otherwise, maximum liklihood can recover the correct p_{data} but will not be able to determine which value of $\boldsymbol{\theta}$ was used by the data-generating process
2. Statistical Efficiency
    - The Cramér-Rao lower bound (Rao, 1945; Cramér, 1946) shows that no consistent estimator has a lower MSE than the maximum likelihood estimator (Appendix A)

For its consistency and efficiency, maximum likelihood is often considered the preferred estimator to use of machine learning.
When the number of examples is small enough to yield overfitting behavior, regularization strategies such as weight decay may be used to obtain a biased version of maximum likelihood that has less variance when training data is limited

3 Conditional Log-Likelihood

3.1 Goal

- Estimate a conditional probability $P(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta})$
- In order to predict $\mathbf{y}$ given $\mathbf{x}$ (This forms the basis for most supervised learning)

3.2 Conditional Maximum Likelihood Estimator

Notations

- $\boldsymbol{X}$: All our inputs
- $\boldsymbol{Y}$: All our observed targets

Definition

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{Y} \mid \boldsymbol{X}; \boldsymbol{\theta})$$

If the examples are assumed to be **i.i.d.**, then this can be decomposed into

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log P(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

3.3 Example: Linear Regression as Maximum Likelihood

Problem Settings

- The model produces a conditional distribution $p(y \mid \boldsymbol{x})$
- We define $p(y \mid \boldsymbol{x}) = \mathcal{N}(y; \hat{y}(\boldsymbol{x}; \boldsymbol{w}), \sigma^2)$
- The function $\hat{y}(\boldsymbol{x}; \boldsymbol{w})$ gives the prediction of the mean of the Gaussian
- We assume that the variance is fixed to some constant σ^2 chosen by the user

Mean Squared Error Since the examples are assumed to be **i.i.d.**, the conditional log-likelihood if given by

$$\sum_{i=1}^{m} \log p(y^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{w}) = -m \log \sigma - \frac{m}{2} \log(2\pi) - \sum_{i=1}^{m} \frac{\|\hat{y}^{(i)} - y^{(i)}\|^2}{2\sigma^2}$$

Thus, maximizing the log-likelihood with respect to $\boldsymbol{w}$ yields the same estimate of the parameters $\boldsymbol{w}$ as minimizing the mean squared error

$$\mathrm{MSE}_{\mathrm{train}} = \frac{1}{m} \sum_{i=1}^{m} \|\hat{y}^{(i)} - y^{(i)}\|^2$$

Remark MSE is the cross-entropy between the empirical distribution defined by the training set and a Gaussian model defined by model

Appendix A Statistical Efficiency of MLE

There are other inductive principles besides the maximum likelihood estimator,many of which share the property of being consistent estimators. Consistentestimators can differ, however, in theirstatistical efficiency, meaning that oneconsistent estimator may obtain lower generalization error for a fixed number ofsamplesm, or equivalently, may require fewer examples to obtain a fixed level ofgeneralization error.

Statistical efficiency is typically studied in theparametric case(as in linear-regression), where our goal is to estimate the value of a parameter (assuming itis possible to identify the true parameter), not the value of a function. A way tomeasure how close we are to the true parameter is by the expected mean squarederror, computing the squared difference between the estimated and true parameter values, where the expectation is overmtraining samples from the data-generatingdistribution. That parametric mean squared error decreases asmincreases, andformlarge, the Cramér-Rao lower bound (Rao, 1945; Cramér, 1946) shows thatno consistent estimator has a lower MSE than the maximum likelihood estimator.