Conditional Log-Likelihood, Mean Squared Error, and Properties of Maximum Likelihood

Shariq Azeem, Nikilas John, Leo Nguyen

Conditional Log-Likelihood

The most common basic situation in supervised learning is to predict y given x. The estimation can be done based on the conditional probability $P(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta})$

- In the case when algorithm only learn to take an input x and produce an output y, the conditional maximum likelihood can be used:

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

Conditional Log-Likelihood

- In the case of infinitely large training set, we can have several training examples with the same input x but different value of y.

  The goal of learning algorithm is now to fit the distribution p(y|x) to all different y values that are all compatible with x so the algorithm can produce the same result as the conditional maximum likelihood above.

  In this case, the conditional log-likelihood will be used:

$$\sum_{i=1}^{m} \log p(y^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

$$= -m \log \sigma - \frac{m}{2} \log(2\pi) - \sum_{i=1}^{m} \frac{\left\| \hat{y}^{(i)} - y^{(i)} \right\|^2}{2\sigma^2},$$

Mean Squared Error

Mean Squared Error gives an average of the square of the errors, where error is the difference between the estimated value and the actual value.

The formula for Mean Squared Error: $\mathrm{MSE}_{\mathrm{train}} = \frac{1}{m}\sum_{i=1}^{m} ||\hat{y}^{(i)} - y^{(i)}||^2,$

If we maximize the log likelihood, given in the previous slide, with respect to ***w***, we can notice that it gives the same estimate for the parameter ***w***, as minimizing the MSE.

Therefore, we can use MSE to estimate maximum likelihood.

Properties of Maximum Likelihood

Why use it?

The Maximum Likelihood estimator can be shown to be the best estimator asymptotically as M approaches infinity

Statistical Efficiency is the metric in which we can compare the estimators

This metric is classified as the generalization error for a fixed number of samples M

This is studied through a **parametric case** (meaning we are estimating the value of a parameter rather than the value of a function)

Properties of Maximum Likelihood

How are we able to tell that we are close to the “true” parameter?

> Computing the squared difference between the estimated and true parameter values

With these properties, Maximum Likelihood is regarded as the go-to estimator for machine learning usages