Maximum Likelihood Estimation

Wei Wang @ CSE, UNSW

April 9, 2019

Inference Problem for a Model

- Model prediction:
  - A model $M(\mathbf{x}; \boldsymbol{\theta})$ usually predicts the $\mathbf{y}_M$ associated with a given $\mathbf{x}$ under a given model parameter $\boldsymbol{\theta}$.
- However, the observed/labelled $\mathbf{y}_O$ usually do not always agree with $\mathbf{y}_M$ for any $\boldsymbol{\theta}$.[1]
  - We need a principled way to choose the best $\boldsymbol{\theta}$ (within its domain). This is the inference problem.
- Candidate inference principles:
  - Least squared: find the most accurate model
  - Maximum likelihood (MLE): find the most likely model
  - Maximum a posteriori (MAP): find the model that appears most often in the posterior distribution (i.e., achieving the maximum $P(\mathbf{x}, \boldsymbol{\theta})$).
  - Based on a **loss function**: find the most accurate model

[1] We do talk about a special case where there are many $\boldsymbol{\theta}$ that will fit perfectly with the yy_O for every training data.

MLE

- Proposed by R. A. Fisher in the 1920s.
  - Write out the **likelihood function** $L(\mathbf{y} \mid \boldsymbol{\theta}) = P(\mathbf{y} \mid \boldsymbol{\theta})$.
  - Find $\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} L(\mathbf{y} \mid \boldsymbol{\theta})$.
- MLE has a few nice statistical properties: sufficiency, consistency, efficiency, and parameter invariance.
  - Consistency: when the number of samples grows to ∞, $\boldsymbol{\theta}_{MLE}$ converges to the true parameter.
  - Won't go into the formal technical details.
- Common tricks:
  - Almost always work in the log space: log-likelihood function $\ell()$.
    - (1) log here is ln. Base does not matter.
    - Also taking log still gives the same arg max solutions.
  - (Assume) all training instances are i.i.d., hence $\ell(\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \log P(\mathbf{y}_i \mid \boldsymbol{\theta})$.

MLE Example 1 /1

正面朝上的概率为PM类似于只有正 两面一面的概率为0.5

- Biased coin with head probability of p_M. Toss n times, and observed the empirical head probability as p_O.
- Understanding first: 为观察正面朝上的概率
  - p_M could be any number in $(0, 1) \Longrightarrow$ even $p_M = 0.000001$ is possible, c.f., *Murphy's law*.
  - Yet, in the absence of any other source of information/belief, a sensible choice is to choose p_M such that the probability of observing $p_O \cdot n$ heads are the maximum $\Longrightarrow$ MLE
- e.g., $p_M = 0.1$, $p_O = 0.6$, $n = 10$.

$$P(p_O = 0.6 \mid p_M = 0.1, n = 10) = \binom{10}{6} \cdot (0.1)^6 \cdot (1 - 0.1)^4$$

MLE Example 1 /2

- Biased coin with head probability of p_M. Toss n times, and observed the empirical head probability as p_O.
- Write out the log-likelihood function: $\ell(\mathbf{y} \mid \boldsymbol{\theta}) = \log P(\mathbf{y} \mid \boldsymbol{\theta})$.

$$\log P(p_O \mid p_M) = \log \left(\binom{n}{p_O n} \cdot {p_M}^{p_O n} \cdot (1 - p_M)^{(1-p_O)n} \right)$$

  Note: p_M is the only variable (i.e., view others as constants)
- Finding the maximum
  - For such a simple case, we can obtain the analytical solution by requiring:
    - $\frac{\partial \ell}{\partial \boldsymbol{\theta}_i} = 0 \Longrightarrow \frac{p_O n}{p_M} + \frac{-(1-p_O)n}{1-p_M} = 0$ (note: n does not matter)
    - $\frac{\partial^2 \ell}{\partial^2 \boldsymbol{\theta}_i} < 0$
  - Otherwise, find the $\arg\max$ solution numerically. (Might not be global maximum or non-unique/non-deterministic, esp. in the non-linear or high-dimensional cases).

MLE Example 2 /1

- Memory retention model based on power law. $y = 1$ means one still remember a given fact. It is a function over time t. (Z is the normalizing constant)

$$P(y = 1 \mid t; \mathbf{w}) = \frac{1}{Z} \cdot \mathbf{w}_1 \cdot t^{-\mathbf{w}_2}$$

- At each timestamp t_i, we recruit some volunteers to conduct the experiments, and obtain the corresponding empirical retention probability p_O.
- MLE:
  - Write out the log-likelihood function
  - Do the arg max

MLE Example 2 /2

- $p_M(y = 1 \mid t; \mathbf{w}) = \frac{1}{Z} \cdot \mathbf{w}_1 \cdot t^{-\mathbf{w}_2}$
- Data: $(t^{(i)}, p_O^{(i)})$
- MLE:
  - Write out the log-likelihood function for a given $t^{(i)}$

$$\ell^{(i)} = \log \left(\binom{n}{p_O n} \cdot {p_M}^{p_O n} \cdot (1 - p_M)^{(1-p_O)n} \right)$$

$$\ell = \sum_i \ell^{(i)}$$

    Note: the p_M and p_O (and n) in $\ell^{(i)}$ are all conditioned on i.
  - Do the arg max
    - In general, there is *no* analytical solution. Why?

MLE Example 2 /3

- The big picture:
  - Model predicted distribution $(t^{(i)}, p_M^{(i)})$
  - Observed distribution: $(t^{(i)}, p_O^{(i)})$
- MLE will give its best $\mathbf{w}$
- In general, a different $\mathbf{w}$ will be obtained if we define a **loss function**, $\sum_i J(p_M^{(i)}, p_O^{(i)})$, and find its best $\mathbf{w}$ that minimizes the loss
- In general, MAP will give a different $\mathbf{w}$ as well, as it considers not only the likelihood function, but also the prior on $\mathbf{w}$.
  - Could be useful in some cases, e.g., one already obtained a posterior distribution of $\mathbf{w}$ based on samples from volunteers in one state, and now doing the inference on volunteers from another state.

MLE Example 3: Linear Regression

- Model: $y_M = \mathbf{w}^\top \mathbf{x}$
- Observed: y_O
- Log-likelihood function:
  - As both y_M and y_O are numerical measurents, we need to come up with a different model to derive the likelihood function.
  - Without any other knowledge/info, we can assume $P(y_O \mid y_M)$ follows a *fixed* Guassian distribution $\mathcal{N}(0, \sigma^2)$ (i.e., σ is fixed for all $(\mathbf{x}^{(i)}, y^{(i)})$s.

$$\begin{aligned}
\ell = \sum_i \log P(y_O \mid y_M; \sigma^2) &= \sum_i \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_O - y_M)^2}{2\sigma^2}\right) \\
&= \sum_i \left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_O - y_M)^2}{2\sigma^2}\right)
\end{aligned}$$

  - Note that maximiming ℓ above means minimizing $(y_O - y_M)^2$! Hence, MLE inference is equivalent to Least Squared inference (or inference based on SSE as the loss function).
  - In many case, this is interpreted as $y_O = y_M + \epsilon$, where ϵ is a Guassian noise. This is the additive Gaussian noise model, but there are many cases where such modelling does not work, yet MLE (and other inference methods) still works.

Final Remarks on MLE

- It is just *one* of the model selection criteria.
  - Not always applicable
  - Could easily overfit the data (c.f., smoothing)
  - Should not be used to perform model selection (i.e., choose between two models based on their log-likelihood values on a given training data). Think why?
    - Instead, generalization (impossible to measure) is the right criteria).
    - In ML/DL, the *usually* approaches are based on Bayesian models or *structured risk minimization*
    - In pratice, typically done via a separate validation/development set.