Maximum Likelihood Estimation

Wei Wang @ CSE, UNSW

March 19, 2020

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 best model under a criterion.

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

MLE

- Proposed by R. A. Fisher in the 1920s.
  - Write out the **likelihood function** $L(\mathbf{y} \mid \boldsymbol{\theta}) = P(\mathbf{y} \mid \boldsymbol{\theta})$.
    - It is not a distribution!
  - 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

- Biased coin with head probability of p_M. Toss n times, and the observed results are: $\{ H, T, H, H, H \}$.
  - Observed probability: $p_O = 0.8$
- Understanding first:
  - 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 the observed outcomes heads are the maximum $\Longrightarrow$ MLE
- e.g.,

$$P(\{ H, T, H, H, H \} \mid p_M = 0.1) = (0.1)^4 \cdot (1 - 0.1)^1 = 9.0 \times 10^{-5}$$
$$P(\{ H, T, H, H, H \} \mid p_M = 0.8) = (0.8)^4 \cdot (1 - 0.8)^1 = 8.1 \times 10^{-2}$$

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:
    - $\frac{\partial \ell}{\partial \boldsymbol{\theta}_i} = 0 \implies \frac{p_O n}{p_M} + \frac{-(1-p_O)n}{1-p_M} = 0$ (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)

$$P(y = 1 \mid t; \mathbf{w}) = \frac{1}{Z} \cdot w_1 \cdot t^{-w_2}$$

- 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

- $p_M(y = 1 \mid t; \mathbf{w}) = \frac{1}{Z} \cdot w_1 \cdot t^{-w_2}$
- 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.

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

$$= \sum_i \left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_O - y_M)^2}{2\sigma^2}\right)$$

  - 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.

KL Divergence

- How to measure the difference between two probability distributions?
  - One popular method is based on *divergence*, in particular, the Kullback–Leibler divergence (or simply KL-divergence).

$$D_{KL}(P(x) \| Q(x)) = \mathbf{E}_{x \sim P(x)} \left[\!\left[\log \frac{P(x)}{Q(x)} \right]\!\right] = \sum_{x \sim P(x)} P(x) \cdot \log \frac{P(x)}{Q(x)}$$

  - It is asymmetric and means the average number of *extra* bits use by an encoder based on $Q(x)$ to encode a message generated from the distribution $P(x)$.
    - The optimal number of bits to code a symbol x in a message is $-\log P(x)$.
    - If we have an estimated distribution $Q(x)$, and encode messages using the optimal number of bits according to $Q(x)$, that's exactly KL.
  - Obviously, KL is asymmetric; KL is non-negative; ...

Application 1: MLE

- Apply KL to machine learing.
  - $D_{KL}(P(x;\theta^*)\|P(x;\theta))$
  - θ^* is the true parameter value, and θ is the current model parameter

$$D_{KL}(P(x;\theta^*)\|P(x;\theta)) = \mathbf{E}_{x\sim P(x\|\theta^*)}\left[\!\left[\log P(x \mid \theta^*)\right]\!\right] - \mathbf{E}_{x\sim P(x\|\theta^*)}\left[\!\left[\log P(x \mid \theta)\right]\!\right]$$

- To minimize KL, only need to maximize the second term. Note that if we draw N samples x_i following $P(x\|\theta^*)$, according to *the Law of Large Numbers*,

$$\mathbf{E}_{x\sim P(x\|\theta^*)}\left[\!\left[\log P(x \mid \theta)\right]\!\right] \approx \left(\sum_{x_i\sim P(x\|\theta^*)} \log P(x \mid \theta)\right) / N$$

The red part is the log-likelihood.

- Exercise: what are the key assumptions for this to hold?

Application 2: Cross-entropy Loss

- k-class classification problem.
  - Ground truth distribution over the classes: $p(y)$
  - Model predicted distribution over the classes: $\hat{p}(y) = f(y \mid \theta)$
  - Both are discrete distributions:

$$
\begin{aligned}
D_{KL}(p(y) \| f(y \mid \theta)) &= -\sum_{i=1}^{k} p(y_i) \log \hat{p}(y_i) + \sum_{i=1}^{k} p(y_i) \log p(y_i) \\
&= H(p, \hat{p}) - H(p)
\end{aligned}
$$

- The two terms are called *cross-entropy* and *entropy*
- For hard classification tasks, only one of the y_i has probability 1 (denote this as y_c and $p(y_c) = 1$). Then

$$H(p, \hat{p}) = p(y_c) \log \hat{p} = \log \hat{p}$$

c.f., `https://ml-cheatsheet.readthedocs.io/en/latest/loss_functions.html`

Summary

- the Model + Loss Paradigm:
    - Real data can be deemed as generated from a very complex model M^*, which is **not** within the common model/function families.
    - In our modelling, we fix a model/function family $\mathcal{F}$, and find the best $m^* = \arg\max_{m \in \mathcal{F}} J(m)$ to approximate M^*
    - We measure the "goodness of fit" of m by a loss function J.
        - For probabilitistic distributions, KL or other well-known distance functions can be used.
- KL-divergence is related to both f-divergence and Bregmen divergence.