Probabilist models and Maximum Likelihood

Pierre Gaillard and Alessandro Rudi

April 2018

1 Introduction

In probabilistic modeling, we are given a set of observations $D_n = (y_1, \dots, y_n)$ in $\mathcal{Y}$ that we assume to be generated from some unknown i.i.d. distribution. The objective is to find a probabilistic model that explains well the data. For instance by estimating the density of the underlying distribution. If possible, we would like the model to predict well new data and to be able to incorporate prior knowledge and assumptions.

Let μ denote some reference measure on the output set $\mathcal{Y}$. Typically, μ is the counting measure if $\mathcal{Y} \subset \mathbb{N}$ or the Lebesgue measure if $\mathcal{Y} \subset \mathbb{R}^d$.

Definition 1 (Parametric model)**.** *Let $p \geq 1$ and $\Theta \subset \mathbb{R}^p$ be a set of parameters. A parametric model $\mathcal{P}$ is a set of probability distributions taking value in $\mathcal{Y}$ with a density with respect to μ and indexed by Θ: $\mathcal{P} = \{p_\theta d\mu | \theta \in \Theta\}$.*

Example 1.1. Here are a few examples of statistical parametric models based on well known family distributions:

- Binomial model: $\mathcal{Y} = \mathbb{N}$, $\Theta = [0,1]$ and $p_\theta(k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$;
- Gaussian model: $\mathcal{Y} = \mathbb{R}$, $\Theta = \{(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+\}$ and $p_{(\mu,\sigma)}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Multidimensional Gaussian model: $\mathcal{Y} = \mathbb{R}^d$, $\Theta = \{(\mu, \Sigma) \in \mathbb{R}^d \times \mathcal{M}_d(\mathbb{R})\}$ and

$$p_{(\mu,\Sigma)}(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}.$$

- Exponential model on $\mathcal{Y} = \mathbb{R}_+$, Bernoulli model on $\mathcal{Y} = \{0,1\}$,...

Now, we assume that we are given some model $\mathcal{P}$ indexed by $\theta \in \Theta$ and we assume that the data D_n is generated independently from $p_{\theta_*} \in \mathcal{P}$ for some unknown parameter θ_*. We would like to recover the best parameter θ_* from the data. Note that in practice the data might come from a distribution which is not in $\mathcal{P}$: we call this misspecification but we will not enter into this details in this class.

2 Maximum likelihood estimation

The idea behind maximum likelihood estimation is to choose the most probable parameter $\theta \in \Theta$ for the observed data. Assume that $\mathcal{Y}$ is discrete and that $Y \sim p_{\theta_*} d\mu$ for some $\theta_* \in \Theta$. Then, for any observation y_i the probability that $Y = y_i$ equals $p_{\theta_*}(y_i)$. Similarly, the probability of observing $(y_1, \dots, y_n) \in \mathcal{Y}^n$ if all the samples were sampled independently from p_θ is $\prod_{i=1}^m p_\theta(y_i)$. Hence, the high level idea of maximum likelihood estimation will be to maximize this probability over $\theta \in \Theta$. This is formalized by the definition of the likelihood which also holds for non-discret set $\mathcal{Y}$.

Definition 2 (Likelihood)**.** *Let $\mathcal{P} = \{p_\theta, \theta \in \Theta\}$ a parametric model and $y \in \mathcal{Y}$. Given the outcome $y \in \mathcal{Y}$, the likelihood is the function $\theta \mapsto p_\theta(y)$. The likelihood $L(.|D_n)$ of a data set $D_n = (y_1, \dots, y_n)$ is the function*

$$L(\cdot|D_n) : \theta \mapsto \prod_{i=1}^{m} p_\theta(y_i)\,.$$

The maximum likelihood estimator (MLE) is then the parameter which maximizes the likelihood, i.e.,

$$\widehat{\theta}_n \in \arg\max_{\theta\in\Theta} \left\{ \prod_{i=1}^{n} p_\theta(y_i) \right\}.$$

This principle was proposed by Ronald Fisher in 1922 and was validated since with good theoretical properties. It is worth pointing out that since log is an increasing function, the maximum likelihood estimator can also be obtained by maximizing the log-likelihood:

$$\widehat{\theta}_n \in \arg\max_{\theta\in\Theta} \left\{ \sum_{i=1}^{n} \log(p_\theta(y_i)) \right\}. \tag{MLE}$$

This turns out to be much more convenient in practice because it is easier to maximize a sum than a product. Convince yourself by computing the gradients!

Examples

- Bernoulli model: $\mathcal{Y} = \{0, 1\}$, $\Theta = [0, 1]$, $p_\theta(y) = \theta^y(1-\theta)^{(1-y)}$. We assume that D_n was generated from a Bernoulli distribution of parameter θ_*, the the maximum likelihood estimator is:

  $$\widehat{\theta}_n = \arg\min_{0\le\theta\le 1} \frac{1}{n} \sum_{i=1}^{n} \big(y_i \log\theta + (1-y_i)\log(1-\theta) \big)\,.$$

  Denoting $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$ the empirical average and solving $d \log L(\widehat{\theta}_n|D_n)/d\theta = 0$ yields

  $$\frac{\bar{y}_n}{\widehat{\theta}_n} - \frac{1-\bar{y}_n}{1-\widehat{\theta}_n} = 0 \quad \Rightarrow (1-\bar{y}_n)\widehat{\theta}_n = (1-\widehat{\theta}_n)\bar{y}_n \quad \Rightarrow \quad \widehat{\theta}_n = \bar{y}_n\,.$$

  Therefore the maximum likelihood estimator is in this case the empirical mean.

- As an exercise, compute the maximum likelihood estimator for the models seen in Example 1.1.

Link with empirical risk minimization In density estimation, the goal is to find the density of the distribution which generated the data. Assuming that the density belongs to the model $\mathcal{P}$, the possible densities are p_θ, for $\theta \in \Theta$. A standard loss function in this setting is the negative log-likelihood: $\ell : (\theta, y) \in \Theta \times \mathcal{Y} \mapsto -\log\big(p_\theta(y)\big)$. The risk (or generalization error) is then:

$$\mathcal{R}(\theta) = -\mathbb{E}_Y\big[\log(p_\theta(Y))\big]\,.$$

In particular, if $Y \sim p_{\theta_*} d\mu$ for some $\theta_* \in \Theta$, θ_* minimizes the risk and the objective is to recover θ_*. The empirical risk is then by definition

$$\widehat{\mathcal{R}}_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\big(p_\theta(y_i)\big)\,.$$

Therefore, the empirical risk minimizer matches the estimator obtained from maximum likelihood in Equation (MLE).

Link with Kullback-Leibler divergence The Kulback-Leibler divergence is a measure of dissimilarity two between probability distributions. It was introduced by Kullback and Leibler in 1951.

Definition 3 (Kullback-Leibler divergence)**.** *Let $pd\mu$ and $qd\mu$ be two probability distributions. The Kullback-Leibler divergence from p to q is defined as*

$$KL(p||q) := \mathbb{E}_{Y\sim pd\mu}\left[\log\frac{p(Y)}{q(Y)}\right] = \int_{\mathcal{Y}} p(y)\log\frac{p(y)}{q(y)}d\mu(y)\,.$$

The KL divergence has various interpretations. As we will see now, it can be interpreted as the excess risk of the measure $p_\theta d\mu$ when the data follows distribution $p_{\theta_*}d\mu$ when the loss function is the negative log-likelihood. Assume that the data D_n were generated from p_{θ_*}. Then, the excess risk can be written

$$\begin{aligned}\mathcal{R}(\theta) - \mathcal{R}(\theta_*) &= -\mathbb{E}_{Y\sim\theta_*}\big[\log(p_\theta(Y))\big] + \mathbb{E}_{Y\sim\theta_*}\big[\log(p_{\theta_*}(Y))\big] \\ &= \mathbb{E}_{\theta_*}\Big[\log\Big(\frac{p_{\theta_*}(Y)}{p_\theta(Y)}\Big)\Big] =: KL\big(p_{\theta_*}||p_\theta\big)\end{aligned}$$

where $\mathbb{E}_{\theta_*}[f(Y)]$ denotes $\mathbb{E}_{Y\sim p_{\theta_*}d\mu}[f(Y)]$ the expectation of $f(Y)$ when Y follows $p_{\theta_*}d\mu$.
Another interpretation comes from information theory. It can be seen as the difference of bits needed to encode D_n under a code optimized for $p_\theta d\mu$ compared to a code optimized for $p_{\theta_*}d\mu$.

Properties and remarks about the KL-divergence:

- $KL(P||Q) \geq 0$ by Jensen's inequality
- $KL(p||p) = 0$. Therefore, we see that p_{θ_*} minimizes the the risk and thus maximizes the likelihood.
- If the distributions are discrete and μ is the counting measure, we have in particular $KL(p||q) := \sum_{i\in\mathcal{Y}} p(i)\log\big(\frac{p(i)}{q(i)}\big)$.
- The Kullback–Leibler divergence is defined only if for all $A \subset \mathcal{Y}$, $q(A) = 0$ implies $p(A) = 0$, i.e., if q is absolutely continuous with respect to p.
- Though KL is often seen as a distance, it does not fill the requirements: it is not symmetric and it does not satisfy the triangular inequality.
- With an abuse of notation, we can rewrite the empirical risk minimization for log loss with the KL:

  $$\widehat{\theta}_n \in \arg\min_{\theta\in\Theta} KL(\widehat{p}_n||p_\theta)$$

  where $\widehat{p}_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ is the empirical measure (which does not have any density with respect to the Lebesgue measure though).

Conditional modeling

Until now, we considered the problem of density estimation when the data set has only outputs $y_i \in \mathcal{Y}$. However, the principle of maximum likelihood can be extended to couples of input outputs $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ in $\mathcal{X}\times\mathcal{Y}$. We can then distinguish two different modeling:

- generative modeling: we aim at estimating the density of couples of input outputs (X, Y) among a family of densities $(x, y) \in \mathcal{X}\times\mathcal{Y} \mapsto p_\theta(x, y)$ on $\mathcal{X}\times\mathcal{Y}$. Then the risk and the empirical risks are:

  $$\mathcal{R}(\theta) = -\mathbb{E}\big[\log(p_\theta(X, Y)\big] \qquad \widehat{\mathcal{R}}_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log\big(p_\theta(x_i, y_i)\big)\,.$$

  This can be useful to generate some new samples (see what is obtained with GANs).
- conditional modeling: we aim at estimating the density of an output Y given an input X. The family of densities are now densities $y \in \mathcal{Y} \mapsto p_\theta(.|x)$ on $\mathcal{Y}$ only but that depend on the inputs. The risk and the empirical risk with negative log-likelihood are then

  $$\mathcal{R}(\theta) = -\mathbb{E}\big[\log(p_\theta(Y|X))\big] \qquad \widehat{\mathcal{R}}_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log\big(p_\theta(y_i|x_i)\big)\,.$$

This is useful if one want to predict the distribution or the value of a new output Y given X.

Example 2.1. Linear regression. We consider a data set $D_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ of samples in $\mathcal{X} \times \mathcal{Y}$. We assume that the outputs y_i were independently generated from a Gaussian distribution of mean $w^\top x_i$ and variance σ^2. In other words, we model an output Y given an input X as

$$Y = {w_*}^\top X + \varepsilon, \qquad \text{where} \quad \varepsilon \sim \mathcal{N}(0, \sigma_*^2)\,.$$

for some unknown $\theta_* = (w_*, \sigma_*^2) \in \mathbb{R}^d \times \mathbb{R}_+$. Our family of possible conditional densities is indexed by parameters $\theta = (w, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+$

$$p_\theta(y|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - w^\top x)^2}{2\sigma^2}}\,.$$

The empirical risk (or conditional log-likelihood) is then

$$\widehat{\mathcal{R}}_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log\big(p_\theta(y_i|x_i)\big) = \frac{1}{2n\sigma^2}\sum_{i=1}^n \big(y_i - w^\top x_i\big)^2 + \frac{1}{2}\log(2\pi\sigma^2)\,.$$

Therefore, the maximum likelihood estimator $\widehat{w}_n$ of w in a Gaussian model is the estimator obtained by least square linear regression. As an exercise, you may show that the maximum likelihood estimator for σ is

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \big(y_i - \widehat{w}_n^\top x_i\big)^2\,.$$

Example 2.2. Logistic regression. A plus of logistic loss in comparison with Hinge loss is that it has a probabilist interpretation. Consider $(X, Y) \sim \mathcal{P}$ a pair of input-output random variables in $\mathbb{R}^d \times \{0, 1\}$. In binary classification, the objective is given the input X to predict the probability that $Y = 1$. In other words, we want to estimate

$$\mathbb{P}(Y = 1|X)$$

from observations $D_n := \{(x_i, y_i)\}_{1 \le i \le n} \in \big(\mathbb{R}^d \times \{0, 1\}\big)^n$ that were independently generated from $\mathcal{P}$. The issue is that linear predictions of the form $\theta^\top x_i$ belongs to $\mathbb{R}$ while our model needs to output probabilities with values in the range $[0, 1]$. A function that maps $\mathbb{R}$ to $[0, 1]$ is the sigmoid function

$$\sigma : z \in \mathbb{R} \mapsto \frac{1}{1 + e^{-z}}\,.$$

This function satisfies $\sigma(-z) = 1 - \sigma(z)$ and $\frac{\mathrm{d}\sigma(z)}{\mathrm{d}z} = \sigma(z)\sigma(-z)$. Logistic regression assumes the probabilistic model

$$\mathbb{P}(Y = 1|X) = \sigma\big(\theta^\top X\big)\,.$$

It is worth pointing out that this probabilistic model is satisfied for many natural models: for instance if $X|Y = 1$ and $X|Y = 0$ follow independent Gaussian distributions. We can define the family of possible densities (with respect to the countable measure) on $\{0, 1\}$ considered here

$$p_\theta(y|x) = \sigma\big(\theta^\top x\big)^y \big(1 - \sigma(\theta^\top x)\big)^{1-y}\,, \tag{1}$$

indexed by $\theta \in \mathbb{R}^d$. We recall that similarly to linear regression, to ease the notation, the intercept (to make afffine prediction) may be included into the input X by adding the constant 1.

Proposition 1. *Assume that $D_n = \{(x_i, y_i)\}_{1 \le i \le n}$ are realizations of i.i.d. random variables. Then the logistic regression estimator $\widehat{\theta}_{(logit)} \in \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n \ell(\theta^\top x_i, y_i)$ with $\ell(x, y) := y \log(\sigma(\theta^\top x)) + (1 - y)\log(\sigma(-\theta^\top x))$ matches the definition of the maximum likelihood estimator of $\theta \in \mathbb{R}^d$ with parametric model (1).*

Proof. The conditionnal log-likelihood can be written

$$\begin{aligned}\log L(\cdot|D_n) &= \sum_{i=1}^{n} \log\left(p_\theta(y_i|x_i)\right) \\ &= \sum_{i=1}^{n} \log\left(\sigma(\theta^\top x_i)^{y_i}\sigma(-\theta^\top x_i)^{1-y_i}\right) \\ &= \sum_{i=1}^{n} y_i \log\left(\sigma(\theta^\top x_i)\right) + (1-y_i)\log\left(\sigma(-\theta^\top x_i)\right) \\ &= \sum_{i=1}^{n} \ell({\theta_*}^\top x_i, y_i)\end{aligned}$$

where ℓ is the logistic loss. □

3 Maximum a-posteriori

If the dimension p of the parameter space Θ is too large compared to the number of samples n, the MLE may lead to poor performance. Similarly to least square linear regression without regularization, it overfits when $p > n$. A second limitation is that no prior knowledge on the parameters θ is included. Let us see this on an example.

Example 3.1. Consider the multinomial model where each observation is a discrete observation in k classes $\{1, \dots, k\}$. Each class $j \in \{1, \dots, k\}$ is sampled with a probability θ_j^* and we aim at retrieving these probabilities.
For convenience, we define the output set $\mathcal{Y} = \left\{y \in \{0,1\}^k : \sum_{i=1}^k y_i = 1\right\}$. An observation $y_i \in \{0,1\}^k$ is such that $y_i(j) = 1$ if it is in class $j \in \{1, \dots, k\}$ and 0 otherwise. The multinomial model consists of densities of the form:

$$p_\theta : y \in \mathcal{Y} \mapsto \prod_{j=1}^{k} \theta_j^{y(j)}, \quad \text{for} \quad \theta \in [0,1]^k : \sum_{j=1}^{k} \theta_j = 1\,.$$

In other words, the probability of an observation to be in class j equals θ_j. The dimension of the parameter space is $p = k - 1$. The MLE is

$$\widehat{\theta}_j = \arg\max_{\theta_j} \frac{1}{n} \sum_{i=1}^{n} y_i(j) \log \theta_j \quad = \frac{n_j}{n}$$

where $n_j = \sum_{i=1}^n y_i(j)$ is the proportion of occurrence of class j in the data set. If $k > n$ (think about the probability of words into some text, each word being a possible class, the number of possible words k can be much larger than the number of words in the text), many classes j are never observed and estimated with 0. The log-loss of these options is infinite and so is the risk (or generalization error) $\mathcal{R}(\theta)$. We say the the model is overestimating.

This problem can be solved by adding a regularization which can also be seen from a Bayesian point of view as a prior distribution over the possible distributions θ. This is what does Maximum a Posteriori (MAP). The idea behind MAP is to see the parameter θ as a random variable taking values in Θ, and to choose the most probable value $\widehat{\theta}^{MAP}$ for the observed data. Given the data set D_n, the MAP can be formalized as the solution of

$$\widehat{\theta}_n^{MAP} \in \arg\max_{\theta \in \Theta} p(\theta|D_n)$$

where $p(\theta|D_n)$ is the density of the posterior distribution of the model given the data. In discrete model space Θ, the MAP is exactly the most probable model. To calculate the posterior distribution we use the

Bayes rule:

$$p(\theta|D_n) = \frac{p(D_n|\theta)p(\theta)}{p(D_n)}\,,$$

where D_n is a random data set and

- $p(D_n|\theta)$ is the probability density of observing D_n if the distribution follows $p_\theta d\mu$. This is exactly the likelihood $L(\theta|D_n)$;
- $p(\theta)$ is the prior distribution of the model. How likely we think it is before seeing the data. In general, the simpler, the more likely!
- $p(D_n)$ is the marginal distribution of the data.

Hence, the MAP is the solution of

$$\widehat{\theta}_n^{MAP} \in \arg\max_{\theta\in\Theta} \left\{L(\theta|D_n)p(\theta)\right\} = \arg\min_{\theta\in\Theta} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(y_i) + \log\frac{1}{p(\theta)} \right\}. \tag{MAP}$$

In some situation, we may not have to prefer one model over another and one can think of $p(\theta)$ as a constant over the parameter space Θ. Then the MAP reduces to the MLE. However, this assumption that $p(\theta)$ is constant is problematic because uniform distribution cannot always be defined if Θ is not compact. Therefore it may be better to see MAP as a regularized version of MLE with a regularization of the form $\log\frac{1}{p(\theta)}$ rather than MLE as a particular case of MAP with uniform prior.