1 Introduction

In probabilistic modeling, we are given a set of observations $D_n = (y_1, \ldots, y_n)$ 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 $\mu$ denote some reference measure on the output set $\mathcal{Y}$. Typically, $\mu$ 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 $\mu$ and indexed by $\Theta$:

$$\mathcal{P} = \{ p_\theta \, d\mu | \theta \in \Theta \}.$$

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

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

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

- Many others: exponential model on $\mathcal{Y} = \mathbb{R}_+$, ...
Definition 2 (Likelihood). Let \( \mathcal{P} = \{ p_\theta, \theta \in \Theta \} \) a parametric model and \( y \in \mathcal{Y} \). The likelihood is the function \( \theta \mapsto p_\theta(x) \). The likelihood \( L(\cdot|D_n) \) of a data set \( D_n = (y_1, \ldots, y_n) \) 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.,

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

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:

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

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!

Exercise 1. Assume the Bernoulli model: \( \mathcal{Y} = \{0, 1\} \), \( \Theta = [0, 1] \), \( p_\theta(y) = \theta^y (1-\theta)^{1-y} \). We observe \( n \) samples \( y_1, \ldots, y_n \). What is the maximum likelihood estimator \( \hat{\theta}_n \)?

Exercise 2. Assume the Gaussian model: \( \mathcal{Y} = \mathbb{R} \), \( \Theta = \{ (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+ \} \) and \( p_{(\mu, \sigma)}(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2 / 2\sigma^2} \). We observe \( n \) samples \( y_1, \ldots, y_n \). What is the maximum likelihood estimator \( \hat{\theta}_n \)?

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 \( p_\theta, \) for \( \theta \in \Theta \). A standard loss function in this setting is the negative log-likelihood: \( \ell : (\theta, y) \in \Theta \times \mathcal{Y} \mapsto -\log(p_\theta(y)) \). The risk (or generalization error) is then:

\[
\mathcal{R}(\theta) = -E_{Y \sim \theta}[\log(p_\theta(Y))].
\]

In particular, if \( Y \sim p_\theta \cdot d\mu \) for some \( \theta^* \in \Theta \), \( \theta^* \) minimize the risk and the objective is to recover \( \theta^* \). The empirical risk is then by definition

\[
\hat{\mathcal{R}}_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log(p_\theta(y_i)).
\]

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

Link with Kullback-Leibler divergence The Kullback-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 \( p \cdot d\mu \) and \( q \cdot d\mu \) be two probability distributions. The Kullback-Leibler divergence from \( p \) to \( q \) is defined as

\[
KL(p||q) := E_{Y \sim p \cdot d\mu} \left[ \log \frac{p(Y)}{q(Y)} \right] = \int_Y p(y) \log \frac{p(y)}{q(y)} d\mu(y).
\]
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 set follows distribution $p_\theta d\mu$ when the loss function is the negative log-likelihood. Assume that the data set $D_n$ were generated from $p_\theta$. Then, the excess risk can be written
\[ R(\theta) - R(\theta^*) = -E_{\theta^*} E_{\theta^*} \left[ \log(p_{\theta}(Y)) \right] + E_{Y \sim p_{\theta}} \left[ \log(p_{\theta^*}(Y)) \right] \]
\[ = E_{\theta^*} \left[ \log \left( \frac{p_{\theta}(Y)}{p_{\theta^*}(Y)} \right) \right] =: KL(p_{\theta^*} || p_{\theta}) \]
where $E_{\theta^*}[f(Y)]$ denotes $E_{Y \sim p_{\theta^*}}[f(Y)]$ 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 $D$.

Properties and remarks about the KL-divergence:
- $KL(P||Q) \geq 0$ by Jensen’s inequality.
- $KL(p||p) = 0$. Therefore, we see that $p_{\theta}$ minimize the the risk and thus maximize the likelihood.
- If the distributions are discrete and $\mu$ is the counting measure, we have in particular $KL(p||q) := \sum_{i \in \mathcal{X}} p(i) \log \left( \frac{p(i)}{q(i)} \right)$.
- 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:
\[ \hat{\theta}_n \in \arg \min_{\theta \in \Theta} KL(\hat{p}_n || p_{\theta}) \]
where $\hat{p}_n = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$ 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:
\[ R(\theta) = -E \left[ \log(p_{\theta}(X, Y)) \right] \quad \hat{R}_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \log \left( p_{\theta}(x_i, y_i) \right) . \]

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 $y \in \mathcal{Y} \mapsto p_{\theta}(\cdot|x)$ on $\mathcal{Y}$ only but that depend on the inputs. The risks are then
\[ R(\theta) = -E \left[ \log(p_{\theta}(Y|X)) \right] \quad \hat{R}_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \log \left( p_{\theta}(y_i|x_i) \right) . \]

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

**Exercise 3.** We consider a data set $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of samples in $\mathcal{X} \times \mathcal{Y}$. We assume that the outputs $y_i$ were independently generated from a Gaussian distribution of mean $w^T x_i$ and variance $\sigma^2$.

In other words, we model an output $Y$ given an input $X$ as
\[ Y = w^* x + \varepsilon, \quad \text{where} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) . \]

for some unknown $\theta^* = (w^*, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+$. What is the maximum likelihood estimator of $w^*, \sigma^2$ from the data?
3 Maximum a-posteriori

The maximum likelihood has however limitations if the dimension \( p \) of the parameter space \( \Theta \) is too large compared to the number of samples \( n \). Similarly to least square linear regression without regularization, it overfits when \( p > n \). A second limitation is that no prior knowledge on the parameters \( \theta \) is included. Let us see this on the following example.

**Exercise 4.** Consider the multinomial model where each observation is a discrete observation in \( k \) classes \( \{1, \ldots, k\} \). Each class \( j \in \{1, \ldots, k\} \) is sampled with a probability \( \theta_j^* \) and we aim at retrieving these probabilities.

For convenience, we define the output set \( \mathcal{Y} = \{ y \in \{0, 1\}^k : \sum_{j=1}^k y(j) = 1 \} \). An observation \( y_i \in \{0, 1\}^k \) is such that \( y_i(j) = 1 \) if it is in class \( j \in \{1, \ldots, k\} \) and 0 otherwise. The multinomial model consists of densities of the form:

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

In other words, the probability of an observation to be in class \( j \) equals \( \theta_j \). The dimension of the parameter space is \( p = k - 1 \).

1. What is the maximum likelihood estimator?

2. Assume that the data \( y_1, \ldots, y_n \) was generated under \( p_{\theta^*} \). Assume \( n < k \) and that the samples were generated according to some law \( p_\theta \), where \( \theta^* \) has positive coordinates. What is the excess risk of the maximum likelihood estimator?

The case \( k > n \) is realistic: 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. We say 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 \( \theta \). This is what does Maximum a Posteriori (MAP). The idea behind MAP is to see the parameter \( \theta \) as a random variable taking values in \( \Theta \), and to choose the most probable value \( \hat{\theta}^{MAP} \) for the observed data. Given the data set \( D_n \), the MAP can be formalized as the solution of

\[
\hat{\theta}^{MAP}_n \in \arg \max_{\theta \in \Theta} p(\theta|D_n)
\]

where \( p(\theta|D_n) \) is the density of the posterior distribution of the model given the data. In discrete model space \( \Theta \), 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

- \( p(D_n|\theta) \) 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

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

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 \( \Theta \). The MAP reduces to the MLE. However, this assumption that \( p(\theta) \) is constant
is problematic because uniform distribution cannot always be defined if $\Theta$ 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.

**Exercise 5.** Consider the model of 3. We have seen that the maximum likelihood estimate of $w_*$ correspond to the least-square estimate. *

1. What prior should we add of $w_*$ to obtain a ridge regression with regularization $\lambda$?

2. What prior should we add of $w_*$ to obtain a Lasso regression with regularization $\lambda$?