Class n°2

Macéo Ottavy

10/17/2025

1 Bayesian learning

When flipping a coin 5000 times, suppose someone reports observing 2000 heads and 3000 tails. The maximum likelihood estimator (MLE) for the probability of getting a tail would then be $\frac{3000}{5000} = 0.6$. Now consider a much smaller experiment: if we only flip the coin twice and both outcomes are tails, the MLE would give a probability of 1.0, implying 100% chance of getting tails. This is clearly suspicious and illustrates a core issue with MLE: it relies entirely on the observed data, without incorporating any prior knowledge or uncertainty. When the sample size is small, MLE is extremely prone to overfitting, producing unrealistic or overly confident estimates (see figure 1).

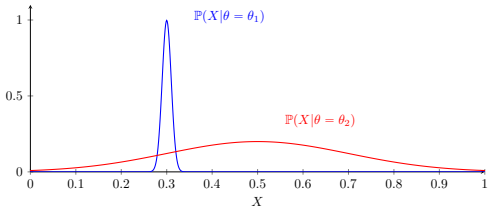

Figure 1: probability example

This motivates Bayesian methods, which introduce a prior distribution over parameters to regularize the solution and quantify uncertainty. Instead of finding a single point estimate like MLE, Bayesian inference computes the posterior distribution: $\mathbb{P}(\theta|X)$ and not the MLE $\mathbb{P}(X|\theta)$.

Bayes formula:

$$\mathbb{P}(X|\theta) = \frac{\mathbb{P}(\theta)}{\mathbb{P}(X)}\mathbb{P}(X|\theta)$$

Therefore:

$$\begin{aligned}\max_\theta \log \mathbb{P}(\theta|X) &= \max_\theta \log \mathbb{P}(X|\theta) + \log \mathbb{P}(\theta) - \mathbb{P}(X) \\ &= \max_\theta \log \mathbb{P}(X|\theta) + \log \mathbb{P}(\theta) \\ &= \max_\theta \log \mathcal{L}(X|\theta) + \log \text{Prior}\end{aligned}$$

Prior $\mathbb{P}(\theta)$ is a probability distribution over the model parameters θ that represents our beliefs about the parameters before observing any data. It is independent of the observed dataset. A uniform prior assumes that all parameter values are equally likely a priori. Therefore, it implies that $\max_\theta \log \mathbb{P}(\theta|X)$ is equivalent to computing the classical MLE.

Bayesian learning takes a fundamentally different view of parameters, instead of treating θ as a fixed but unknown quantity, it models θ as a random variable and uses probability distributions to represent uncertainty or lack of knowledge. *Baysians* uses probability to model (lack of) knowledge. They something optimizes meaningful and always keep probability estimates.

2 Variationnal Inference

In expectation maximization (EM), we decided to put a probability on Z which is the the probability of Z being "good". In variationnal inference, we are doing the same thing on both Z and θ. Then we fuse all parameters to get a new θ. However, it is impossible to find $\mathbb{P}(\theta|X)$ in the full space of probability distribution on θ. Then, we restrict to a parametrized space and find a probability distribution q^* that minimize $KL\left(q(\theta)||p(\theta|X)\right)$ (see figure 2).

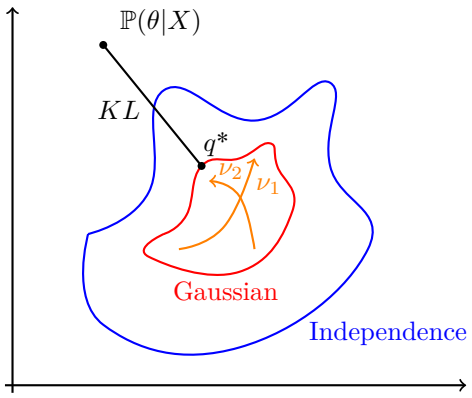

Figure 2: Space of probability distribution on θ

The goal is:

$$\underset{\nu_1,\nu_2,\ldots}{arg\,\min}\ KL\left(q_\nu(\theta)||p(\theta|X)\right)$$

$\nu_1, \nu_2, \ldots$ are called *varationnel parameters*. Note that sometimes, researchers uses ELBO instead of KL.

3 Autoencoder

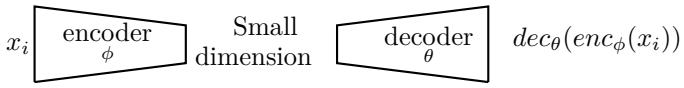

Figure 3: Autoencoder representation

The goal is to reduce the dimensionnality of x_i thank to the encoder and then to retrieve x_i from its reduced form thanks to the decoder. The loss is $\mathcal{L}(\phi,\theta) = \|dec_\theta(enc_\phi(x_i)) - x_i\|_2^2$.

$$\begin{aligned}\underset{\theta}{arg\,\min}\ \|f_\theta(x) - y\|_2^2 &= \underset{\theta}{arg\,\max}\ \frac{1}{\sqrt{2\pi 1^2}} \exp\left(-\frac{1}{2}\frac{\|f_\theta(x) - y\|_2^2}{1^2}\right) \\ &= \underset{\theta}{arg\,\max}\ \mathcal{N}\left(f_\theta(x), 1\right)(y)\end{aligned}$$

In this form, we see the MLE form. Then all the problem related to MLE appear in this situation too.

4 Variationnal Autoencoder (VAE)

In this part, we assume that $\forall i, X_i \sim \mathcal{N}(dec_\theta(Z_i), 1^2)$. Then, the prior is: $Z_i \simeq p(Z) \sim \mathcal{N}(0,1)$.
The variable of interest are the unknown $\{Z_i\}$. The goal is $\mathbb{P}(\{Z_i\}_i|\{X_i\}_i)$. There are two variationnal assumptions (see figure 2):

$$\begin{aligned}\mathbb{P}(\{Z_i\}_i|X) &= \prod_i \mathbb{P}(Z_i|X) && \text{Independence assumption} \\ &= \prod_i \mathcal{N}(\mu_i, \sigma_i^2)(Z_i) && \text{Gaussian assumption}\end{aligned}$$

KL definition:

$$KL(q||p) := \mathbb{E}_{q_\nu}\left[log\frac{q}{p}\right] = \int_\theta q(\theta) log\frac{q(\theta)}{p(\theta)} d\theta$$

$$\begin{aligned}
\arg\min_{\theta,\nu} KL\left(q_\nu(z)||p(z|X)\right) &= \arg\min_{\theta,\nu} \mathbb{E}_{q_\nu}\left[\log q(z) - \log p(z|X)\right] \\
&= \arg\min_{\theta,\nu} \mathbb{E}_{q_\nu}\left[\log q(z)\right] - \mathbb{E}_{q_\nu}\left[\log\left(\frac{p(X|z)p(z)}{p(X)}\right)\right] \\
&= \arg\min_{\theta,\nu} \mathbb{E}_{q_\nu}\left[\log q(z) - \log p(z)\right] - \mathbb{E}_{q_\nu}\left[\log p(X|z)\right] \\
&= \arg\min_{\theta,\nu} KL(q(z)||p(z)) - \mathbb{E}_{q_\nu}\left[\log p(X|z)\right] \\
&= \arg\min_{\theta,\nu} \sum_i KL\left(q_{\mu_i,\sigma_i}(z_i)||\mathcal{N}(0,1)(z_i)\right) - \sum_i \mathop{\mathbb{E}}_{z_i\sim\mathcal{N}(\mu_i,\sigma_i)}\left[K\left\|dec_\theta(z_i) - x_i\right\|_2^2\right] \\
&= \arg\min_{\theta,\nu} \sum_i KL\left(\mathcal{N}(\mu_i,\sigma_i)(z_i)||\mathcal{N}(0,1)(z_i)\right) - \sum_i \frac{1}{M}\sum_k \left\|dec_\theta(z_i^{(k)}) - x_i\right\|_2^2
\end{aligned}$$

With K and M being some constants. We are using stochastic methods with $M = 1$ and $z_i^{(k)} \sim \mathcal{N}$.

The problem of this methods is that it involves an enormous amount of parameters. Therefore, we are using *amortization*: instead of having ν parameters, we learn a function $x_i \mapsto \mu_i, \sigma_i$ thanks to the encoder$_\phi$.