# Approximating distributions

Bobak Shahriari

8 March 2016

Consider a random variable $X$ distributed as a mixture of two one-dimensional Gaussians:

$$P := \rho_1 \mathcal{N}(\mu_1, \sigma_1^2) + \rho_2 \mathcal{N}(\mu_2, \sigma_2^2), \text{ where } \rho_1 + \rho_2 = 1.$$

Let us approximate $P$ with a single normal distribution $Q_\theta := \mathcal{N}(\mu, \sigma^2)$, where $\theta := (\mu, \sigma)$. There are various ways to fit $\mu$ and $\sigma$ that we explore in this document.

## 1 Variational approximation

We will fit this Gaussian using the two different direction of the Kullback-Leibler (KL) divergence:

$$\begin{aligned} J(\theta) := D_{\mathrm{KL}}(Q_\theta \,||\, P) &= \int_\Omega Q_\theta(\mathrm{d}x) \log \frac{Q_\theta(x)}{P(x)} \\ &= -H[Q_\theta] - \int_\Omega \phi_\theta(x) \log P(x) \mathrm{d}x \\ J(\theta) &= -\frac{1}{2} \log(2\pi e \sigma^2) - \mathbb{E}_u \log P(\mu + u\sigma) \end{aligned}$$

where we use that $\phi_\theta(x) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$, where $\phi_\theta$ and $\phi$ denote the normal and standard normal probability density functions (pdf), respectively; and let $u$ denote a standard normal random variable such that $x = \mu + u\sigma$ and $\mathrm{d}x = \sigma \mathrm{d}u$. The derivatives of the chosen loss can then be expressed as

$$\begin{aligned} \partial_\mu J(\theta) &= -\int_\Omega \partial_\mu \phi_\theta(x) \log P(x) \mathrm{d}x \\ &= -\int_\Omega \frac{1}{\sigma} \partial_\mu \phi\left(\frac{x-\mu}{\sigma}\right) \log P(x) \mathrm{d}x \\ &= -\int_\Omega \phi'(u) \partial_\mu u \log P(\mu + u\sigma) \mathrm{d}u \\ &= -\int_\Omega -u\phi(u) \left(-\frac{1}{\sigma}\right) \log P(\mu + u\sigma) \mathrm{d}u \\ \partial_\mu J(\theta) &= -\frac{1}{\sigma} \mathbb{E}_u \left[u \log P(\mu + u\sigma)\right] \end{aligned}$$

and with respect to the standard deviation

$$
\begin{aligned}
\partial_\sigma J(\theta) &= -\frac{1}{\sigma} - \int_\Omega \partial_\sigma \phi_\theta(x) \log P(x) \mathrm{d}x \\
&= -\frac{1}{\sigma} - \int_\Omega \partial_\sigma \left[ \frac{1}{\sigma} \phi \left( \frac{x-\mu}{\sigma} \right) \right] \log P(x) \mathrm{d}x \\
&= -\frac{1}{\sigma} - \int_\Omega \left[ -\frac{1}{\sigma^2} \phi(u) + \frac{1}{\sigma} \partial_\sigma \phi(u) \right] \log P(x) \mathrm{d}x \\
&= -\frac{1}{\sigma} - \int_\Omega \frac{1}{\sigma} \left[ -\frac{1}{\sigma} \phi(u) - u\phi(u) \partial_\sigma \frac{x-\mu}{\sigma} \right] \log P(x) \mathrm{d}x \\
&= -\frac{1}{\sigma} - \int_\Omega \left[ -\frac{1}{\sigma} + u \frac{x-\mu}{\sigma^2} \right] \log P(\mu + u\sigma) \phi(u) \mathrm{d}u \\
&= -\frac{1}{\sigma} - \frac{1}{\sigma} \int_\Omega (u^2 - 1) \log P(\mu + u\sigma) \phi(u) \mathrm{d}u \\
\partial_\sigma J(\theta) &= -\frac{1}{\sigma} \left[ 1 + \mathbb{E}_u (u^2 - 1) \log P(\mu + u\sigma) \right]
\end{aligned}
$$

## 2 Moment matching

Recall that the KL divergence is not symmetric, hence why it is not called the KL distance. With this in mind, let us now consider the same loss in the other direction.

$$
\begin{aligned}
M(\theta) := D_{\mathrm{KL}}(P \,||\, Q_\theta) - H[P] &= \int_\Omega P(\mathrm{d}x) \log Q_\theta(x) \\
&= -\frac{1}{2} \log(2\pi\sigma^2) - \int_\Omega P(\mathrm{d}x) \frac{(x-\mu)^2}{2\sigma^2} \\
&= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left( \mu^2 - 2\mu \mathbb{E}_P[X] + \mathbb{E}_P[X^2] \right)
\end{aligned}
$$

with derivatives

$$
\begin{aligned}
\partial_\mu M(\theta) &= -\frac{1}{\sigma^2} \left( \mu - \mathbb{E}_P[X] \right) \\
\partial_{\sigma^2} M(\theta) &= -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} \left( \mu^2 - 2\mu \mathbb{E}_P[X] + \mathbb{E}_P[X^2] \right)
\end{aligned}
$$

setting these to zero yields the best approximation according to $M$, namely:

$$
\begin{aligned}
\mu &= \mathbb{E}_P[X] \\
\sigma^2 &= \mathbb{V}_P[X],
\end{aligned}
$$

which corresponds to matching the first two moments of $P$.

## 3 Laplace approximation