Estimating Probabilities from data

Suppose you have a dataset $\mathcal{D} = \{\bar{x}_i\}_{i=1}^n \in \mathbb{R}^d$. Each data point $\bar{x}_i$ is drawn independently from $\mathcal{N}(\bar{\mu}, I)$ where I is the $d \times d$ identity matrix.

1. Find the MLE for $\bar{\mu}$.
2. Assume a standard Gaussian prior on u, namely, $P(\bar{\mu}) = \mathcal{N}(0, I)$, find the MAP for $\bar{\mu}$.
3. Assume the same Gaussian prior on $\bar{\mu}$ in (2), find the posterior distribution $P(\bar{\mu}|\mathcal{D})$. (Hint: the posterior is a Gaussian distribution.)

Solution:

1. We first find $P(\mathcal{D}|\bar{\mu})$. Since the data is i.i.d., we have

$$\begin{aligned}
P(\mathcal{D}|\bar{\mu}) &= \prod_{i=1}^{n} P(\bar{x}_i|\bar{\mu}) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d |I|}} \exp\left(-\frac{(\bar{x}_i - \bar{\mu})^T I^{-1} (\bar{x}_i - \bar{\mu})}{2}\right) \\
&= \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{(\bar{x}_i - \bar{\mu})^T (\bar{x}_i - \bar{\mu})}{2}\right)
\end{aligned}$$

where we use the fact that $\det I = 1$, and $I^{-1} = I$. We take the log-likelihood (making the derivative calculation much easier), giving us

$$\begin{aligned}
\log P(\mathcal{D}|\bar{\mu}) &= \log \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{(\bar{x}_i - \bar{\mu})^T (\bar{x}_i - \bar{\mu})}{2}\right) \\
&= \sum_{i=1}^{n} \log\left(\frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{(\bar{x}_i - \bar{\mu})^T (\bar{x}_i - \bar{\mu})}{2}\right)\right) \\
&= \sum_{i=1}^{n} \log\left(\frac{1}{(2\pi)^{\frac{d}{2}}}\right) - \frac{(\bar{x}_i - \bar{\mu})^T (\bar{x}_i - \bar{\mu})}{2} \\
&= \sum_{i=1}^{n} -\frac{d}{2}\log(2\pi) - \frac{(\bar{x}_i - \bar{\mu})^T (\bar{x}_i - \bar{\mu})}{2} \\
&= -\frac{nd}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(\bar{x}_i - \bar{\mu})^T (\bar{x}_i - \bar{\mu})}{2}
\end{aligned}$$

We then find the partial derivative with respect to $\bar{\mu}$:

$$\begin{aligned}
\nabla_{\bar{\mu}} \log P(\mathcal{D}|\bar{\mu}) &= \nabla_{\bar{\mu}} \left(-\frac{nd}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(\bar{x}_i - \bar{\mu})^T (\bar{x}_i - \bar{\mu})}{2} \right) \\
&= \sum_{i=1}^{n} \bar{x}_i - \bar{\mu}
\end{aligned}$$

To find the derivative of $(\bar{x}_i - \bar{\mu})^T(\bar{x}_i - \bar{\mu})$ we note that $(\bar{x}_i - \bar{\mu})^T(\bar{x}_i - \bar{\mu}) = ||\bar{x}_i - \bar{\mu}||_2^2$, and $\nabla_x ||x||_2^2 = 2x$, then we can apply the chain rule. Finally, setting the derivative to zero gets us

$$\begin{aligned}
&\nabla_{\bar{\mu}} \log P(\mathcal{D}|\bar{\mu}) = 0 \\
\implies &\sum_{i=1}^{n} \bar{x}_i - \bar{\mu} = 0 \\
\implies &n\bar{\mu} = \sum_{i=1}^{n} \bar{x}_i \\
\implies &\boxed{\bar{\mu} = \frac{\sum_{i=1}^{n} \bar{x}_i}{n}}
\end{aligned}$$

To verify that this is indeed a maximum, we perform the second derivative test

$$\nabla_{\bar{\mu}}^2 \log P(\mathcal{D}|\bar{\mu}) = -nI$$

Since the second derivative is the negative definite, we have a local maximum.

2. To find the MAP, we simply want to find $\arg\max_{\bar{\mu}} P(\bar{\mu}|\mathcal{D})$. We know

$$\begin{aligned}
\arg\max_{\bar{\mu}} P(\bar{\mu}|\mathcal{D}) &= \arg\max_{\bar{\mu}} \log P(\bar{\mu}|\mathcal{D}) \\
&= \arg\max_{\bar{\mu}} \log \left(\frac{P(\mathcal{D}|\bar{\mu})P(\bar{\mu})}{P(\mathcal{D})} \right) \\
&= \arg\max_{\bar{\mu}} \log(P(\mathcal{D}|\bar{\mu})P(\bar{\mu})) \\
&= \arg\max_{\bar{\mu}} \log P(\mathcal{D}|\bar{\mu}) + \log P(\bar{\mu})
\end{aligned}$$

where we use the monotonicity of the log function in the first step above. We have already computed $\log P(\mathcal{D}|\bar{\mu})$, leaving us only to find $\log P(\bar{\mu})$. We have

$$\begin{aligned}
\log P(\bar{\mu}) &= \log \frac{1}{\sqrt{(2\pi)^d |I|}} \exp\left(-\frac{\bar{\mu}^T I^{-1} \bar{\mu}}{2} \right) \\
&= -\frac{d}{2} \log 2\pi - \frac{\bar{\mu}^T \bar{\mu}}{2}
\end{aligned}$$

Trivially, we have the partial derivative of $\log P(\bar{\mu})$ with respect to $\bar{\mu}$ is $\bar{\mu}$. Thus,

$$
\begin{aligned}
&\nabla_{\bar{\mu}}(\log P(\mathcal{D}|\bar{\mu}) + \log P(\bar{\mu})) = 0 \\
\implies &\sum_{i=1}^{n}(\bar{x}_i - \bar{\mu}) - \bar{\mu} = 0 \\
\implies &(n+1)\bar{\mu} = \sum_{i=1}^{n} \bar{x}_i \\
\implies &\boxed{\bar{\mu} = \frac{\sum_{i=1}^{n} \bar{x}_i}{n+1}}
\end{aligned}
$$

Again, we test this point by taking the second derivative

$$\nabla^2_{\bar{\mu}}(\log P(\mathcal{D}|\bar{\mu}) + \log P(\bar{\mu})) = -I(n+1)$$

The second derivative is negative definite so it is a maximum.

3. Intuitively, we can expect the mean to be $\frac{\sum_{i=1}^{n} \bar{x}_i}{n+1}$ and the variance to be $\frac{1}{n+1}I$, but let us see if we can prove it. We have

$$
\begin{aligned}
P(\bar{\mu}|\mathcal{D}) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d\,|I|}} \exp\left(-\frac{(\bar{x}_i - \bar{\mu})^T\, I^{-1}\,(\bar{x}_i - \bar{\mu})}{2}\right) \frac{1}{\sqrt{(2\pi)^d\,|I|}} \exp\left(-\frac{\bar{\mu}^T I^{-1} \bar{\mu}}{2}\right) \\
&= \left(\frac{1}{\sqrt{(2\pi)^d}}\right)^{n+1} \exp\left(-\frac{(n+1)\bar{\mu}^T\bar{\mu} - 2\bar{\mu}^T\sum_{i=1}^{n}\bar{x}_i + \sum_{i=1}^{n}\bar{x}_i^T\bar{x}_i}{2}\right)
\end{aligned}
$$

We know the above has to be a Gaussian distribution. We note that a Gaussian distribution has the form of $c \exp -\frac{(\bar{\mu}-m)^T \Sigma^{-1} (\bar{\mu}-m)}{2}$, which when expanding the numerator, comes out to be

$$c \exp\left(-\frac{1}{2}\bar{\mu}^T\Sigma^{-1}\bar{\mu} - \bar{\mu}^T\Sigma^{-1}m + \text{const}\right)$$

where Σ^{-1} is the inverse covariance matrix.

We can see that the above is in the same form, and from there we derive that

$$\Sigma^{-1} = (n+1)I \implies \Sigma = \frac{1}{n+1}I$$

and

$$\Sigma^{-1}m = \sum_{i=1}^{n} \bar{x}_i \implies m = \frac{\sum_{i=1}^{n} \bar{x}_i}{n+1}$$

Thus,

$$\boxed{P(\bar{\mu}|\mathcal{D}) \sim \mathcal{N}\left(\frac{\sum_{i=1}^{n} \bar{x}_i}{n+1}, \frac{1}{n+1}I\right)}$$