Assignment 4: Probabilistic Modeling

Fabian Flores Gobet

3rd of December, 2023

Problem Consider a data set of binary (black and white) images. Each image is arranged into a vector of pixels by concatenating the columns of pixels in the image. The data set has N images $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ and each image has D pixels, where D is (number of rows $\times$ number of columns) in the image. For example, image $\mathbf{x}^{(n)}$ is a vector $(x_1^{(n)}, \ldots, x_D^{(n)})$ where $x_d^{(n)} \in \{0, 1\}$ for all $n \in \{1, \ldots, N\}$ and $d \in \{1, \ldots, D\}$.

Question 1. Explain why a multivariate Gaussian would not be an appropriate model for this data set of images. Give at least three reasons. (5 points) **Solution:**

- In binary images, pixel values are typically 0 or 1, representing black and white, respectively. A Gaussian distribution assumes a continuous range of values, and it's probability density function is not well suited for modelling discrete binary data.
- In a multivariate Gaussian distribution the variables are jointly normally distributed. In the context of images, neighboring pixel values are often highly dependent on each other due to patterns and structures in the images.
- Binary images have a limited range of possible pixel values (0 or 1). A Gaussian distribution has 'tails' that extend to infinity, which is not suitable for modeling a distribution with a finite range of discrete values.

Question 2. Now assume that the images were modeled as independently and identically distributed samples from a D-dimensional multivariate Bernoulli distribution with parameter vector $\mathbf{p} = (p_1, \ldots, p_D)$, which has the form

$$P(\mathbf{x} \mid \mathbf{p}) = \prod_{d=1}^{D} p_d^{x_d} \, (1 - p_d)^{(1-x_d)} \tag{1}$$

where both $\mathbf{x} \in \{0, 1\}^D$ and $\mathbf{p} \in [0, 1]^D$ are D-dimensional vectors. Here, the values 0 and 1 taken by $\mathbf{x}$ are for black and white pixels respectively.

What is the likelihood function of $\mathbf{p}$ given the data set of N images $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$? (3 points)

Solution:

- Likelihood $\mathcal{L}(p; X)$

$$\mathcal{L}(p; X) = P(X|p) = P(x^{(1)}, \cdots, x^{(N)}|p) \overset{\text{idd}}{=} \prod_{n=1}^{N} P(x^{(n)}|p)$$

$$= \prod_{n=1}^{N} \left[\prod_{d=1}^{D} p_d^{x_d^{(n)}} \cdot (1 - p_d)^{1-x_d^{(n)}} \right]$$

Question 3. Derive the log-likelihood. Include all intermediate steps and simplify the final result. (3 points)
Solution:

- Log-likelihood $\ell(p; X)$

$$
\begin{aligned}
\ell(p;X) \;=\; -\log\left(\mathcal{L}(p;X)\right) \;&=\; -\log\left(\prod_{n=1}^{N}\left[\prod_{d=1}^{D} p_d^{x_d^{(n)}}\cdot(1-p_d)^{1-x_d^{(n)}}\right]\right) \\
&=\; -\sum_{n=1}^{N}\left[\log\left(\prod_{d=1}^{D} p_d^{x_d^{(n)}}\cdot(1-p_d)^{1-x_d^{(n)}}\right)\right] \\
&=\; -\sum_{n=1}^{N}\sum_{d=1}^{D}\left[x_d^{(n)}\cdot\log\left(p_d\right)+(1-x_d^{(n)})\cdot\log\left(1-p_d\right)\right] \\
&=\; -\sum_{d=1}^{D}\sum_{n=1}^{N}\left[x_d^{(n)}\cdot\log\left(p_d\right)-x_d^{(n)}\cdot\log\Big(1-p_d\Big)+\log\left(1-p_d\right)\right] \\
&=\; -\sum_{d=1}^{D}\left[N\cdot\log\left(1-p_d\right)+\log\left(\frac{p_d}{1-p_d}\right)\cdot\sum_{n=1}^{N}x_d^{(n)}\right] \qquad (1)
\end{aligned}
$$

Let

$$
\xi_d \;=\; \sum_{n=1}^{N} x_d^{(n)}
$$

From (1) comes

$$
\begin{aligned}
&=\; -\sum_{d=1}^{D}\left[N\cdot\log\left(1-p_d\right)+\xi_d\cdot\log\left(\frac{p_d}{1-p_d}\right)\right] \\
&=\; -\sum_{d=1}^{D}\log\left((1-p_d)^N\cdot\left(\frac{p_d}{1-p_d}\right)^{\xi_d}\right) \\
&=\; -\sum_{d=1}^{D}\log\left((1-p_d)^{N-\xi_d}\cdot p_d^{\xi_d}\right) \\
&=\; -\log\left(\prod_{d=1}^{D}(1-p_d)^{N-\xi_d}\cdot p_d^{\xi_d}\right), \quad \xi_d \;=\; \sum_{n=1}^{N}x_d^{(n)}
\end{aligned}
$$

Question 4. What is the equation for the maximum likelihood (ML) estimate of $\mathbf{p}$? You can assume the critical point to be the maximum, no second derivatives are required. Include all intermediate steps and simplify the final result. (10 points)

Solution:

- $\frac{\partial \ell}{\partial p_i}$

$$\frac{\partial \ell}{\partial p_i} = \frac{\partial}{\partial p_i}\Big(-\log\Big(\prod_{d=1}^{D}(1-p_d)^{N-\xi_d}\cdot p_d^{\xi_d}\Big)\Big)$$
$$= -\frac{\prod_{d=1,d\neq i}^{D}(1-p_d)^{N-\xi_d}\cdot p_d^{\xi_d}}{\prod_{d=1}^{D}(1-p_d)^{N-\xi_d}\cdot p_d^{\xi_d}}\cdot\Big(-(N-\xi_i)(1-p_i)^{N-\xi_i-1}p_i^{\xi_i}+\xi_i(1-p_i)^{N-\xi_i}p_i^{\xi_i-1}\Big)$$
$$= -\Big(\frac{N-\xi_i}{1-p_i}-\frac{\xi_i}{p_i}\Big) = \frac{Np_i-\xi_i}{(1-p_i)p_i}$$

- $\frac{\partial \ell}{\partial p_i} = 0$

$$\frac{\partial \ell}{\partial p_i} = 0 \Leftrightarrow \frac{Np_i-\xi_i}{(1-p_i)p_i} = 0 \Leftrightarrow p_i = \frac{\xi_i}{N}$$

- $\frac{\mathrm{d}\ell}{\mathrm{d}p} = 0$

$$\frac{\partial \ell}{\partial p} = 0 \Leftrightarrow p_{_{MLE}} = \frac{\vec{1}\cdot X}{N}$$

Where X is the matrix with $x^{(i)}$ on line i and $\vec{1}$ a row vector of 1s of size D.

Question 5. Assuming independent Beta priors on the parameters p_d

$$P(p_d) = \frac{1}{B(\alpha, \beta)} p_d^{\alpha-1} (1 - p_d)^{\beta-1}$$

and $P(\mathbf{p}) = \prod_d P(p_d)$, what is the maximum a posteriori (MAP) estimate of $\mathbf{p}$? $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the normalization constant. Include all intermediate steps and simplify the final result. Hint: Maximize the log posterior with respect to $\mathbf{p}$. Like before, you can assume the critical point to be the maximum. (12 points)

Solution:

- $P(p|X)$

$$P(p|X) = \frac{P(X|p)P(p)}{P(X)} \propto_p P(X|p)P(p)$$

- $\frac{\partial \ell(P(X|p)P(p))}{\partial p_i}$

$$\begin{aligned}
\frac{\partial \ell(P(X|p)P(p))}{\partial p_i} &= \frac{\partial}{\partial p_i}\Big(- \log \Big(P(X|p)P(p)\Big)\Big) \\
&= \frac{\partial}{\partial p_i}\Big(- \log (P(X|p)) - \log (P(p))\Big) \\
&= \frac{Np_i - \xi_i}{(1-p_i)p_i} - \frac{\partial}{\partial p_i}\Big(\log (P(p))\Big) \qquad (2)
\end{aligned}$$

- $\frac{\partial}{\partial p_i}\Big(- \log (P(p))\Big)$

$$\begin{aligned}
\frac{\partial}{\partial p_i}\Big(- \log (P(p))\Big) &= \frac{\partial}{\partial p_i}\Big(- \log \Big(\prod_{d=1}^{D} P(p_d)\Big)\Big) \\
= \frac{\partial}{\partial p_i}\Big(- \sum_{d=1}^{D} \log (P(p_d)\Big) &= -\frac{1}{P(p_i)} \frac{\partial}{\partial p_i}\Big(\frac{1}{B(\alpha, \beta)} p_i^{\alpha-1}(1-p_i)^{\beta-1}\Big) \\
&= -\frac{1}{P(p_i)B(\alpha, \beta)}\Big((\alpha-1)p_i^{\alpha-2}(1-p_i)^{\beta-1} + p_i^{\alpha-1}(1-p_i)^{\beta-1}\Big) \\
&= \frac{p_i^{\alpha-1}(1-p_i)^{\beta-1}}{P(p_i)B(\alpha, \beta)}\Big(\frac{\alpha-1}{p_i} - \frac{\beta-1}{1-p_i}\Big) \\
&= \frac{P(p_i)}{P(p_i)}\Big(\frac{p_i(\beta-1) - (1-p_i)(\alpha-1)}{p_i(1-p_i)}\Big)
\end{aligned}$$

- $\frac{\partial \ell(P(X|p)P(p))}{\partial p_i} = 0$

  Plugging the prior result to (2) we have

$$\begin{aligned} &\frac{\partial \ell(P(X|p)P(p))}{\partial p_i} = 0 \\ \Leftrightarrow &\frac{Np_i - \xi_i}{(1-p_i)p_i} + \frac{p_i(\beta - 1) - (1-p_i)(\alpha - 1)}{p_i(1-p_i)} = 0 \\ \Leftrightarrow &Np_i - \xi_i + (\beta - 1)p_i - \alpha + 1p_i = 0 \\ \Leftrightarrow &p_i(N + \beta - 1 + \alpha - 1) = \xi_i + \alpha - 1 \\ \Leftrightarrow &p_i = \frac{\xi_i + \alpha - 1}{N + \alpha + \beta - 2} \end{aligned}$$

- $\frac{\mathrm{d}\ell(P(X|p)P(p))}{\mathrm{d}p} = 0$

$$\frac{\mathrm{d}\ell(P(X|p)P(p))}{\mathrm{d}p} = 0 \Leftrightarrow p_{_{MAP}} = \frac{\vec{1} \cdot X + \alpha - 1}{N + \alpha + \beta - 2}$$