Assignment 4: Probabilistic Modeling

The Swiss AI Lab IDSIA, USI, SUPSI

Student Name : Krunal Rathod | email: krunal.rathod@usi.ch

December 2023

1 Why a multivariate Gaussian would not be an appropriate model

- Binary images have pixels that are either 0 or 1, making them inherently discrete, while a Gaussian appropriates continuous data.
- Gaussian models works with independent variables, but in images, pixel values are often correlated, which violating this assumption.
- Gaussian distributions cover the entire real number line, which not suitable for the limited range of binary value, 0s or 1s.
- A key assumption of the Gaussian model, it should follow normal distribution, which is not in case of binary image.
- Binary images are often sparse, with many 0s, while Gaussian models expect data to be concentrated around the mean, which does not hold in sparse binary data.

2 Likelihood function of 'p'

From the question, the probability of a single image $x^{(n)}$ under the multivariate Bernoulli distribution is given by the expression you provided (Equation 1):

$$P(x^{(n)}|p) = \prod_{d=1}^{D} p_d^{x_d^{(n)}} \cdot (1 - p_d)^{1-x_d^{(n)}}$$

Now, for the likelihood function for the entire dataset is the product of these probabilities for all N images:

$$L(p|X) = \prod_{n=1}^{N} P(x^{(n)}|p)$$

From equation (1), substituting the expression for $P(x^{(n)}|p)$ into the likelihood function, we get:

$$L(p|X) = \prod_{n=1}^{N} \prod_{d=1}^{D} p_d^{x_d^{(n)}} \cdot (1 - p_d)^{1 - x_d^{(n)}}$$

The above function represents the likelihood function for the parameter vector p given the dataset X of N images.

3 Derive the log-likelihood

Starting from the likelihood function:

$$L(p|X) = \prod_{n=1}^{N} \prod_{d=1}^{D} p_d^{x_d^{(n)}} \cdot (1 - p_d)^{1 - x_d^{(n)}}$$

Taking the logarithm both sides:

$$\ell(p|X) = \log \left(\prod_{n=1}^{N} \prod_{d=1}^{D} p_d^{x_d^{(n)}} \cdot (1 - p_d)^{1 - x_d^{(n)}} \right)$$

Using the product rule $\log(ab) = \log a + \log b$:

$$\ell(p|X) = \sum_{n=1}^{N} \sum_{d=1}^{D} \left(\log p_d^{x_d^{(n)}} + \log(1 - p_d)^{1 - x_d^{(n)}} \right)$$

Simplifying further:

$$\ell(p|X) = \sum_{n=1}^{N} \sum_{d=1}^{D} \left(x_d^{(n)} \log p_d + (1 - x_d^{(n)}) \log(1 - p_d) \right)$$

Now, combining the summations:

$$\ell(p|X) = \sum_{n=1}^{N} \left(\sum_{d=1}^{D} x_d^{(n)} \log p_d + \sum_{d=1}^{D} (1 - x_d^{(n)}) \log(1 - p_d) \right)$$

Above equation is the general form of the log-likelihood for the multivariate Bernoulli distribution.

4 Equation for the maximum likelihood (ML) estimate of 'p'

The log-likelihood function is given by:

$$\ell(p|X) = \sum_{n=1}^{N} \left(\sum_{d=1}^{D} x_d^{(n)} \log p_d + \sum_{d=1}^{D} (1 - x_d^{(n)}) \log(1 - p_d) \right)$$

Taking the derivative with respect to p_d and assigning to 0:

$$\frac{\partial \ell}{\partial p_d} = \sum_{n=1}^{N} \left(\frac{x_d^{(n)}}{p_d} - \frac{1 - x_d^{(n)}}{1 - p_d} \right) = 0$$

Solving for p_d:

$$\sum_{n=1}^{N} \frac{x_d^{(n)}}{p_d} - \frac{1 - x_d^{(n)}}{1 - p_d} = 0$$

Now multiplying by $p_d(1 - p_d)$ to remove the fraction parts:

$$\sum_{n=1}^{N} x_d^{(n)}(1 - p_d) - (1 - x_d^{(n)})p_d = 0$$

Expanding further and rearranging:

$$\sum_{n=1}^{N} x_d^{(n)} - \sum_{n=1}^{N} p_d = 0$$

Solving again for p_d:

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

Dividing both sides by N:

$$p_d = \frac{1}{N} \sum_{n=1}^{N} x_d^{(n)}$$

The Maximum Likelihood (ML) estimate for each p_d is the average pixel value across all images in the dataset.

5 Maximum a posteriori (MAP) estimate of 'p'

The log prior for p_d is given by the logarithm of the Beta distribution:

$$\log P(p_d) = \log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_d^{\alpha-1}(1-p_d)^{\beta-1}\right)$$

The log posterior is the sum of log likelihood and the log prior, So:

$$\log P(p|X) = \ell(p|X) + \sum_{d=1}^{D} \log P(p_d)$$

From the derived log likelihood:

$$\ell(p|X) = \sum_{n=1}^{N}\sum_{d=1}^{D}\left(x_d^{(n)} \log p_d + (1-x_d^{(n)})\log(1-p_d)\right)$$

Adding the log prior term, we get:

$$\log P(p|X) = \sum_{n=1}^{N}\sum_{d=1}^{D}\left(x_d^{(n)} \log p_d + (1-x_d^{(n)})\log(1-p_d)\right) + \sum_{d=1}^{D} \log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_d^{\alpha-1}(1-p_d)^{\beta-1}\right)$$

By taking the derivatives with respect to each p_d and assigning it to 0, we get the critical points:

$$\frac{\partial \log P(p|X)}{\partial p_d} = \frac{\partial \ell(p|X)}{\partial p_d} + \frac{\partial \log P(p_d)}{\partial p_d} = 0$$

Solve for p_d in this system of equations to find the MAP estimate. The solution might be complex, and numerical methods may be needed for practical computation.

where $\frac{\partial \log P(p|X)}{\partial p_d}$=0 because it does not depend on p_d.

From answer 4:

$$\frac{\partial \log(p|X)}{\partial p_d} = \frac{\sum_{n=1}^{N}(X)}{p_d} - \frac{\sum_{n=1}^{N}(1-X)}{1-p_d}$$

For second term $\frac{\partial \log(p)}{\partial p_d}$, we start with $P(p)$, assume each pixel have an independent prior:

$$P(p) = \prod_{d=1}^{D} P(p_d)$$

Assume a Beta prior on each p_d:

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

Taking the log, we get:

$$\log(p) = \sum_{d=1}^{D} -\log(B(\alpha,\beta)) + (\alpha-1)\log p_d + (\beta-1)\log(1-p_d)$$

Now taking the derivative with respect to p_d:

$$\frac{\log(p)}{\partial p_d} = \frac{(\alpha-1)}{p_d} - \frac{(\beta-1)}{1-p_d}$$

As we are only concerned with p_d, we are left with a single element of the summation to p_d.

By combining we have have an expression for $\frac{\partial \log(p|X)}{\partial p_d}$:

$$\frac{\partial \log(p|X)}{\partial p_d} = \frac{\sum_{n=1}^{N}(X)}{p_d} - \frac{\sum_{n=1}^{N}(1-X)}{1-p_d} + \frac{(\alpha-1)}{p_d} - \frac{(\beta-1)}{1-p_d}$$

$$\frac{\partial \log(p|X)}{\partial p_d} = \frac{(\alpha-1)+\sum_{n=1}^{N}(X)}{p_d} - \frac{(\beta-1)+\sum_{n=1}^{N}(1-X)}{1-p_d}$$

To find the maximum a posteriori (MAP) estimate $\frac{\partial \log(p|X)}{\partial p_d} = 0$ and solve:

$$0 = \frac{(\alpha-1)+\sum_{n=1}^{N}(X)}{\hat{p_d}} - \frac{(\beta-1)+\sum_{n=1}^{N}(1-X)}{1-\hat{p_d}}$$

$$0 = (1-\hat{p_d})(\alpha-1) + (1-\hat{p_d})\left(\sum_{n=1}^{N}(X)\right) - \hat{p_d}(\beta-1) - \hat{p_d}\left(\sum_{n=1}^{N}(1-X)\right)$$

$$0 = (\alpha-\alpha\hat{p_d}+\hat{p_d}-1)+\left(\sum_{n=1}^{N}(X)-\hat{p_d}\sum_{n=1}^{N}(X)\right)-(\hat{p_d}\beta-\hat{p_d})-\left(\hat{p_d}\cdot N-\hat{p_d}\sum_{n=1}^{N}(X)\right)$$

Cancelling the $\hat{p}_d \sum_{n=1}^{N}(X)$ terms, we get:

$$0 = \alpha - \alpha\hat{p_d} + \hat{p_d} - 1 + \sum_{n=1}^{N}(X) - \hat{p_d}\beta + \hat{p_d} - \hat{p_d}\cdot N$$

$$0 = \hat{p_d}(2-\alpha-\beta-N) + \alpha - 1 + \sum_{n=1}^{N}(X)$$

$$\hat{p_d} = \frac{\alpha-1+\sum_{n=1}^{N}(X)}{(N+\alpha+\beta-2)}$$