Machine Learning Assignment 3
Neural Networks

Alan Copa

Submission deadline: December 9, 2024

Please submit your solution in PDF format (preferably, but not necessarily, LaTeX— this .tex file can be found on iCorsi). Handwriting and scanned documents are not allowed. In case you need further help, please write on iCorsi or contact me at mikhail.andronov@idsia.ch.

1 Estimating the parameters of a statistical model (26 points)

You are given a data set of N measurements $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$, and every measurement $\mathbf{x}^{(n)}$ contains D numbers $(x_1^{(n)}, \ldots, x_D^{(n)})$, such as $x_d^{(n)} \in \mathbb{N} \cup \{0\}$ for all $n \in \{1, \ldots, N\}$ and $d \in \{1, \ldots, D\}$. You decide to model the true distribution of this dataset with an independent multivariate Poisson distribution with the parameter vector $\lambda = (\lambda_1, \ldots, \lambda_D)$, which has the form

$$p(\mathbf{x}|\lambda) = \prod_{d=1}^{D} \frac{\lambda_d^{x_d}}{x_d!} e^{-\lambda_d} \tag{1}$$

You want to estimate the optimal parameters of the model given the data.

1.1 Likelihood (3 points)

What is the likelihood function of λ given the data set of N measurements $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$? (3 points)
The likelihood function is the probability of the observed data given the parameters λ.

$$L = \prod_{n=1}^{N} p(\mathbf{x}^{(n)} \mid \lambda)$$

$$L = \prod_{n=1}^{N} \prod_{d=1}^{D} \frac{\lambda_d^{x_d^{(n)}}}{x_d^{(n)}!} e^{-\lambda_d}$$

1.2 Log-likelihood (3 points)

Derive the log-likelihood. Include all intermediate steps and simplify the final result.

To find the log-likelihood, we take the natural logarithm of the likelihood function:

$$\log(L) = \log \prod_{n=1}^{N} \prod_{d=1}^{D} \frac{\lambda_d^{x_d^{(n)}}}{x_d^{(n)}!} e^{-\lambda_d}$$

$$\log(L) = \sum_{n=1}^{N} \sum_{d=1}^{D} \left(x_d^{(n)} \log \lambda_d - \log x_d^{(n)}! - \lambda_d \right)$$

$$\log(L) = \sum_{n=1}^{N} \sum_{d=1}^{D} \left(x_d^{(n)} \log \lambda_d - \lambda_d \right) - \sum_{n=1}^{N} \sum_{d=1}^{D} \log x_d^{(n)}!$$

1.3 MLE (10 points)

Derive the maximum likelihood estimate (MLE) of λ. You can assume the critical point to be the maximum, no second derivatives are required. Include all intermediate steps and simplify the final result.
To find the critical point, we take the derivative of $\log(L)$ with respect to λ_d and set to zero:

$$\frac{\partial \log(L)}{\partial \lambda_d} = 0$$

$$\frac{\partial \log(L)}{\partial \lambda_d} = \frac{\partial}{\partial \lambda_d} \sum_{n=1}^{N} \left(x_d^{(n)} \log \lambda_d - \lambda_d \right) - \sum_{n=1}^{N} \log x_d^{(n)}!$$

$$\frac{\partial \log(L)}{\partial \lambda_d} = \sum_{n=1}^{N} \left(\frac{x_d^{(n)}}{\lambda_d} \right) - N \cdot 1 - 0$$

$$\frac{\partial \log(L)}{\partial \lambda_d} = \sum_{n=1}^{N} \frac{x_d^{(n)}}{\lambda_d} - N$$

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

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

The MLE for λ_d is:

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

1.4 MAP (10 points)

You place a constraint on the parameters of the model by introducing a prior distribution on them. You assume independent exponential priors on the parameters λ_d

$$p(\lambda) = \prod_{d=1}^{D} p(\lambda_d) = \prod_{d=1}^{D} \beta_d e^{-\beta_d \lambda_d}$$

where $\beta_i > 0$. What is the maximum a posteriori (MAP) estimate of λ? Include all intermediate steps and simplify the final result.

$$p(\lambda \mid \mathbf{x}) \ = p(\mathbf{x}|\lambda) \cdot p(\lambda)$$

$$\log p(\lambda \mid \mathbf{x}) = \log(L) + \log p(\lambda)$$
$$\frac{\partial}{\partial \lambda_d} \log p(\lambda \mid \mathbf{x}) = \frac{\partial}{\partial \lambda_d} \log(L) + \frac{\partial}{\partial \lambda_d} \log p(\lambda)$$

Compute maximum of $\log p(\lambda)$:

$$\frac{\partial}{\partial \lambda_d} \log p(\lambda) = \frac{\partial}{\partial \lambda_d} \log \prod_{d=1}^{D} \beta_d e^{-\beta_d \lambda_d}$$

$$\frac{\partial}{\partial \lambda_d} \log p(\lambda) = \frac{\partial}{\partial \lambda_d} \sum_{d=1}^{D} \log \beta_d e^{-\beta_d \lambda_d}$$

$$\frac{\partial}{\partial \lambda_d} \log p(\lambda) = \frac{\partial}{\partial \lambda_d} (\sum_{d=1}^{D} \log \beta_d - \sum_{d=1}^{D} \beta_d \lambda_d)$$

$$\frac{\partial}{\partial \lambda_d} \log p(\lambda) = 0 - \beta_d$$

$$\frac{\partial \log p(\lambda)}{\partial \lambda_d} = -\beta_d$$

with the log-likelihood computed earlier:

$$\frac{\partial}{\partial \lambda_d} \log p(\lambda \mid \mathbf{x}) = \sum_{n=1}^{N} \frac{x_d^{(n)}}{\lambda_d} - N - \beta_d = 0$$

$$\sum_{n=1}^{N} \frac{x_d^{(n)}}{\lambda_d} = N + \beta_d$$

$$\lambda_d = \frac{\sum_{n=1}^{N} x_d^{(n)}}{N + \beta_d}$$

2 Additional questions (7 points)

Give answers to the following questions.

2.1 Different prior (3 points)

What would be the MAP estimate of λ if we chose the uniform prior, i.e., the prior that treats all parameter values as equally likely? Explain your reasoning.

2.1.1 Solution

A uniform prior means that the prior distribution function is constant for all parameters λ_d, thus:

$$\log p(\lambda) = \text{constant}$$

so the MAP estimate would consist only on the MLE since the derivative of a constant is zero

$$\log p(\lambda \mid \mathbf{x}) = \log(L) + \log p(\lambda)$$

$$\frac{\partial}{\partial \lambda_d} \log p(\lambda \mid \mathbf{x}) = \frac{\partial}{\partial \lambda_d} \log(L) + \frac{\partial}{\partial \lambda_d} \log p(\lambda)$$

$$\frac{\partial}{\partial \lambda_d} \log p(\lambda \mid \mathbf{x}) = \frac{\partial}{\partial \lambda_d} \log(L) + 0$$

$$\frac{\partial}{\partial \lambda_d} \log p(\lambda \mid \mathbf{x}) = \sum_{n=1}^{N} \frac{x_d^{(n)}}{\lambda_d} - N = 0$$

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

2.2 Choice of prior (2 points)

When would the exponential prior on λ be a good choice? What kind of our belief about the model parameters are we expressing in this choice of prior?

2.2.1 Solution

An exponential prior on λ would introduce a penalization for big λs, thus a belief that the values of λ should be small. This fact can help to avoid possible situations of overfitting so it would be a good idea. A bigger β_d would increase the penalization.

2.3 Prior parameters (2 points)

If we make the β parameters of the prior smaller and smaller, how will the shape of the prior and the MAP estimate change?

2.3.1 Solution

As $\beta \to 0$ the penalization term will be less strong and the influence of the prior on the MAP estimation would become irrelevant. The prior will become more flat and less informative for the MAP estimation that will converge to the MLE estimation shape and behaviour.