Machine Learning Assignment 3
Neural Networks

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 $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$, and every measurement $\mathbf{x}^{(n)}$ contains D numbers $(x_1^{(n)}, \dots, x_D^{(n)})$, such as $x_d^{(n)} \in \mathbb{N} \cup \{0\}$ for all $n \in \{1, \dots, N\}$ and $d \in \{1, \dots, D\}$. You decide to model the true distribution of this dataset with an independent multivariate Poisson distribution with the parameter vector $\lambda = (\lambda_1, \dots, \lambda_D)$, 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 $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$? (3 points)

1.2 Log-likelihood (3 points)

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

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.

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.

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.

2.2 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?

2.3 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?

PROBLEM 1

1 $L(\lambda) := P(\{x^{(1)}, \dots, x^{(N)}\}) = \prod_{n=1}^{N} P(x^{(n)} | \lambda) =$

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

2 $\tilde{L}(\lambda) = \ln(L(\lambda)) = \ln\left(\prod_{n=1}^{N} \prod_{d=1}^{D} \frac{\lambda_d^{x_d^{(n)}}}{x_d^{(n)}!} e^{-\lambda_d}\right) =$

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

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

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

3 $MLE(\lambda) = \max(\tilde{L}(\lambda))$

$\frac{d}{d\lambda_d} \tilde{L}(\lambda) = \sum_{n=1}^{N} \left(\frac{x_d^{(n)}}{\lambda_d} - 0 - 1\right) = \sum_{n=1}^{N} \frac{x_d^{(n)}}{\lambda_d} - N = 0$

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

ALREADY CALCULATED

4 $L = P(\lambda) \cdot P(x|\lambda) \Rightarrow \tilde{L} = \ln(P(\lambda)) + \ln(P(x|\lambda))$

$\ln(P(\lambda)) = \ln\left(\prod_{d=1}^{D} \beta_d e^{-\beta_d \lambda_d}\right) = \sum_{d=1}^{D} (\ln(\beta_d) - \beta_d \lambda_d)$

$\frac{d}{d\lambda_d} \ln(P(\lambda_d)) = -\beta_d$

$\frac{d}{d\lambda_d} \tilde{L} = \sum_{n=1}^{N} \frac{x_d^{(n)}}{\lambda_d} - N - \beta_d = 0$ $\quad MAP(\lambda_d) = \frac{\sum_{n=1}^{N} x_d^{(n)}}{N + \beta_d}$

PROBLEM 2

1 WHEN CHOOSING A UNIFORM PRIOR, NO "EXTRA INFORMATIONS" ARE ADDED TO THE LIKELIHOOD $\Rightarrow$ MAP = MLE

2 EXPONENTIAL PRIOR ARE MORE SUITED WHEN EXPECTING POSITIVE AND FAST DECREASING VALUES OF λ_d, LEADING TO PREFER SMALL VALUES FOR λ_d. IT CAN BE USEFUL FOR AVOIDING OVERFITTING

3 FOR SMALLER VALUES OF β_d, PRIOR HAS LESS IMPACT ON THE MODEL AND MAP WILL BE MORE SIMILAR TO MLE

$$\lim_{\beta_d \to 0} MAP(\lambda_d) = \frac{\sum_{n=1}^{N} x_d^{(n)}}{N + \beta_d} = \frac{\sum_{n=1}^{N} x_d^{(n)}}{N} = MLE(\lambda_d)$$