Machine Learning

Exercise Sheet 6

Winter Term 2023/2024
Prof. Dr. Niels Landwehr
Dr. Ujjwal

Available: 15.12.2023
Hand in until: 21.12.2023 11:59am
Exercise sessions: 08.1.2024/10.1.2024

Task 1 – MAP estimation of discrete events [15 points]

Fever (f)	**Breathing Issues (b)**	**COVID (c)**
1	1	1
1	0	0
0	1	1
0	1	0
1	1	1
0	0	0
1	0	1

The above table summarizes a set of 7 observations ($\mathcal{V}$) of 3 sets of variables a, b and c respectively. Compute the MAP estimation of probabilities corresponding to the following events based on the above observations. For MAP estimation, assume all conditional priors to be of uniform distribution (i.e uninformative prior). All probabilities are modeled as in the case of coin tosses in the lecture slides.

1. $p(f = 1|\mathcal{V})$
2. $p(b = 1|\mathcal{V})$
3. $p(c = 1|f = 1, b = 1, \mathcal{V})$
4. $p(c = 1|f = 1, b = 0, \mathcal{V})$
5. $p(c = 1|f = 0, b = 1, \mathcal{V})$
6. $p(c = 1|f = 0, b = 0, \mathcal{V})$

HINT: Uniform distribution corresponds to beta distribution with $\alpha_h = \alpha_t = 1$.

Task 2 – Bayesian Linear Regression [15 points]

a) Prove or disprove the following :

> Training a linear regression model by minimizing the squared loss without regularization is the same as maximizing the likelihood function for the probabilistic linear regression.

b) Prove or disprove the following :

> Training a ridge regression model, that is, minimizing the squared loss plus an L2-regularizer, is the same as maximizing the posterior distribution for probabilistic linear regression (using a normal prior as given in lecture)

By "is the same" we mean that for suitably chosen hyperparameters (regularization weight λ, variances σ^2, σ_p^2) the solution of the optimization problem is the same. For this task, please do not use the result for the posterior in Bayesian logistic regression (Slides 32 and 33). Rather, start by writing the posterior distribution according to Bayes rule and plugging in the prior and likelihood function. Then you have to check if the argmax over this posterior is the same as the argmin over the objective in the non-probabilistic version (loss or loss plus regularizer). Hint: look at the logarithm of the posterior.

c) Now assume that the prior distribution is given by a product of Laplace distributions on the individual elements of the parameter vector, $p(\boldsymbol{\theta}) = \prod_{m=1}^{M} p(\theta_m)$ where

$$p(\theta_m) = \frac{1}{2b} e^{-\frac{|\theta_m - \mu|}{b}}. \tag{1}$$

We assume a location parameter $\mu = 0$ and the scale b is treated as a hyperparameter. What regularization term would we have to choose for regularized linear regression in order to get the same result as maximizing the posterior distribution in Bayesian linear regression under this prior?

Task 3 – Visualizing Posterior Distribution (Programming) [10 points]

In this task, we will train a Bayesian polynomial regression model on the toy sine data set and visualize the resulting posterior distribution.

Use the notebook *Exercise06_Task3.ipynb* to generate a toy sine data set with $N = 20$ instances. The notebook also contains code for a polynomial feature representation for this data set.

Based on the formulae in the lecture, compute the model $\boldsymbol{\theta}_{MAP}$ for the toy data set. Use the hyperparameters $\sigma = 0.1$ and $\sigma_p = 100$. Plot the MAP model (for example code for plotting a model, you can look at Task 2 of the last exercise sheet if you like). Based on the formulae in the lecture, compute the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})$. Draw a sample of 10 models $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{10}$ from the posterior distribution and plot them (all in one plot). Also run your code for data sets of size $N = 3$, $N = 10$, and $N = 100$ instances.