SF1935 Probability Theory and Statistics with Application to Machine Learning

Project in Bayesian linear regression

Pawel Herman

Lab review deadline: May 23, 2023

You will present your group project both orally and in writing. The oral presentation in the lab will be evaluated by one of teaching assistants. Please prepare a few slides introducing the project and illustrating your key findings as well as your observations and reflections that you find worth sharing as part of the discussion of your results. Importantly, please avoid showing your code, implementation etc. and focus instead on communicating the fruits of your intellectual effort. The presentation should not be counted for more than 10 min. After your presentation, be prepared to answer questions formulated by the teaching assistant and try to get all group members equally involved in the entire presentation and discussion.

The written part of project work consists in a short report (one per group) not exceeding 4-5 A4 pages. Your report should include a very brief introduction, a short section on methods and the main part should be devoted to presenting key results and discussing your findings (again, sharing your observations, reflections). The introduction in essence amounts to stating the purpose and aims of the project (ca. 0.5 page). The method section covers some fundamental aspects of your implementation, e.g. which programming/scripting language and accompanying libraries you have used (0.3 page). The results section should be written as a narrative story where you first remind the purpose of your simulation, then a short description of your experimental setup (e.g. how you use data, evaluate, find parameters) followed by the actual results that are accompanied by figures and/or tables (altogether 2-3 pages). Finally, the discussion section should mainly account for your reflections, observations, lessons learnt and conclusions (1 page). The project report should be submitted through Canvas no later than the actual oral presentation.

The reason why the project is evaluated (P/F) based on both oral and written account is to help you develop competence in communicating your work in a concise manner. Please keep in mind that an ability to communicate your results and conclusions is a key aspect of any data science / machine learning practitioner. It is up to you as an author to make sure that the report and presentation clearly shows what you have done, how you prioritise your findings and what observations you have made (and consider worth discussing).

As far as the implementation is concerned, please use any programming/scripting language/environment that you feel comfortable with. I can just recommend Python due to its popularity among data scientists. Matlab, R are other good options.

1 Background - theory

1.1 Theory

Regression is the task of estimating a continuous target variable $\mathbf{t}$ from an observed variate $\mathbf{X}$. The target and the observed variates are related to each other through a mapping,

$$f : \mathbf{X} \rightarrow \mathbf{t}, \tag{1}$$

where f indicates the mapping. Given input output pairs $\{\mathbf{x}_i, t_i\}_1^N$ our task is to estimate the mapping f such that we can infer the associated t_i from previously unseen $\mathbf{x}_i$. In this task we will work with real vectorial data such that $\mathbf{x}_i \in \mathbf{X}$ where $\mathbf{x}_i \in \mathbb{R}^q$ and $t_i \in \mathbb{R}^1$. Being probabilistic means that we need to consider the uncertainty in both the observations as well as the relationship between the variates. Starting with the relationship between two *single* points $\mathbf{x}_i$ and t_i we can assume the following form of the likelihood,

$$p(t_i|f, \mathbf{x}_i) \sim \mathcal{N}(f(\mathbf{x}_i), \beta^{-1}). \tag{2}$$

Assuming that each output point is conditionally independent given the input and the mapping we can write the likelihood of the data as follows:

$$p(\mathbf{t}|f, \mathbf{X}) = \prod_{i=1}^{N} p(t_i|f, \mathbf{x}_i). \tag{3}$$

The task of regression means that we wish to infer t_i from its corresponding variate $\mathbf{x}_i$. These two variates are related to each other by the mapping f so from a probabilistic viewpoint we wish to find the mapping from the observed data. More specifically, taking uncertainty into account, what we wish to reach is the posterior distribution over the mapping given the observations,

$$p(f|\mathbf{X}, \mathbf{t}). \tag{4}$$

1.1.1 Linear Regression

Let's make an assumption about the mapping and model the relationship between the variates. Next, let's assume that the structure of the noise in the observations follows additive Gaussian distribution ($i = 1, .., N$):

$$t_i = \mathbf{w}^T \phi(\mathbf{x}_i) + \epsilon, \tag{5}$$

where $\epsilon \sim \mathcal{N}(0, \beta^{-1})$. From this we can formulate the likelihood of the data,

$$p(\mathbf{t}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^{N} p(t_i|f, \mathbf{x}_i) = \prod_{i=1}^{N} \mathcal{N}(t_i|\mathbf{w}^T \phi(\mathbf{x}_i), \beta^{-1}). \tag{6}$$

The posterior is the object that integrates our prior beliefs with the data

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}) = \frac{1}{Z} p(\mathbf{t}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}). \tag{7}$$

In the above equation we can see that we need to formulate our belief of the model parameters $\mathbf{W}$ in a prior $p(\mathbf{W})$. We can make many different choices of priors, but a sensible choice would be to pick the conjugate prior i.e. a Gaussian prior over the parameters:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \tag{8}$$

2 Project assignment

2.1 Warm-up task with a single input variable

Now we will implement the approach we studied in the previous part. Remember to save images and figures to support your claims in part 1 as this will make the presentation much easier to examine. There are a couple of packages in Python that are really useful (similar functionality can also be found in Matlab, R and other languages):

```
import pylab as pb
import numpy as np
from math import pi
from scipy.spatial.distance import cdist

# To sample from a multivariate Gaussian
f = np.random.multivariate_normal(mu,K);
# To compute a distance matrix between two sets of vectors
D = cdist(x1,x2)
# To compute the exponetial of all elements in a matrix
E = np.exp(D)
```

In this task we will implement the linear regression that we looked at in the previous task. We will examine both the prior and the posterior over the parameters $\mathbf{W} = [w_0, w_1]$ and evaluate the effect this will have on the model. To do so we will need to have some data to experiment with. What we want to show is that the methodology that we have learned is capable of recovering the true underlying mapping from the observed data. Therefore let's generate some data $(\mathbf{x}, \mathbf{t})$ and then simply throw the generating parameters away.

$$t_i = w_0 x_i + w_1 + \epsilon = 0.5 x_i - 1.5 + \epsilon \tag{9}$$

$$\mathbf{x} = [-1, -0.99, \ldots, 0.99, 1] \tag{10}$$

$$\epsilon \sim \mathcal{N}(\mu, \sigma^2) \text{ , where } \mu = 0 \text{ and } \sigma^2 = 0.2 \tag{11}$$

$$\tag{12}$$

Task 1:

1. *Set the prior distribution over* $\mathbf{W}$ *and visualise it.*
2. *Pick a single data point (x,t) and visualise the posterior distribution over* $\mathbf{W}$*.*
3. *Draw 5 samples from the posterior and plot the resulting functions.*
4. *Repeat* $2-3$ *by adding additional data points up to 7.*
5. *Given the plots explain the effect of adding more data on the posterior as well as the functions. How would you interpret this effect?*
6. *Finally, test the exercise for different values of* σ^2*, e.g. 0.1, 0.4 and 0.8. How does your model account for data with varying noise levels? What is the effect on the posterior?*

2.2 Warm-up task with linear regression in multidimensional input space

In the second part of the project, your aim is to perform a regression task where x is a vector in two-dimensional space so that

$$t_i = \mathbf{w}^T(\mathbf{x}_i) + \epsilon \text{ , where } \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{13}$$

Please assume some concrete values for the weight vector, just as you did for the first task. Now you can generate 1000 data points over the two-dimensional domain of the input space. Your objective is to develop a linear regression model on a training set and then make predictions on a test set, and

this evaluate the test performance of your model. For this purpose please randomly choose 80% of the data as the training (design) set and the remaining 20% will serve as your test set. We follow a classical methodology where you fit the model based on the available training data, then you apply the model for making predictions on the unseen test data and finally evaluate the model's predictive performance. With this basic methodology, please do the following tasks

Task 2:

1. *Fit the model using the maximum likelihood principle for different fixed values of σ (say, 0.1, 0.3, 0.5) and evaluate the predictive performance of the model*
2. *For different fixed values of σ (say, 0.1, 0.3, 0.5) choose also different values of the uncertainty parameter of the Gaussian prior over the weight parameters (α) and perform Bayesian linear regression.*
3. *Make comparison between the two modelling approaches in terms of their predictive performance on the test data.*
4. *When you make predictions with the Bayesian model examine also their uncertainties (Bayesian model outputs the Gaussian distribution with the mean and standard deviation), i.e. vary α and observe the effect on the estimated uncertainty of your predictions with the Bayesian model.*

Good Luck!

References

[1] C.M. Bishop. *Pattern recognition and machine learning.* 2006