732A91 - Lab 2

Joris van Doorn || Weng Hang Wong

29 April 2020

1. Linear and polynomial regression

The dataset TempLinkoping.txt contains daily average temperatures (in Celcius degrees) at Malmslätt, Linköping over the course of the year 2018. The response variable is temp and the covariate is

$$time = \frac{the\,number\,of\,days\,since\,beginning\,of\,year}{365}$$

The task is to perform a Bayesian analysis of a quadratic regression

$$temp = \beta_0 + \beta_1 * time + \beta_2 * time^2 + \epsilon, \epsilon \sim^{iid} N(0, \sigma^2)$$

a.

Determining the prior distribution of the model parameters. Use the conjugate prior for the linear regression model. Your task is to set the prior hyperparameters μ_0, Ω_0, ν_0 and σ_0^2 to sensible values. Start with $\mu_0 = (-10, 100, -100)^T, \Omega_0 = 0.01 \cdot I_3, \nu_0 = 4$ and σ_0^2. Check if this prior agrees with your prior opinions by simulating draws from the joint prior of all parameters and for every draw compute the regression curve. This gives a collection of regression curves, one for each draw from the prior. Do the collection of curves look reasonable? If not, change the prior hyperparameters until the collection of prior regression curves agrees with your prior beliefs about the regression curve.

d.

To prevent overfitting we suggest adding a regularization term. The proposed prio would like as follows:

$$\beta_i | \sigma^2 \sim^{iid} N(0, \frac{\sigma^2}{\lambda})$$

where λ will be the smoothness/shrinkage/regularization term.

Variable	Data type	Meaning	Role
Work	Binary	Whether or not the woman works	Response
Constant	1	Constant to the intercept	Feature
HusbandInc	Numeric	Husband's income	Feature
EducYears	Counts	Years of education	Feature
ExpYears	Counts	Years of experience	Feature
ExpYears2	Numeric	(Years of experience)/10)^2	Feature
Age	Counts	Age	Feature
NSmallChild	Counts	Number of child <7 years in household	Feature
NBigChild	Counts	Number of child >6 years in household	Feature

2. Posterior approximation for cassification with logistic regression

The dataset WomenWork.dat contains n = 200 observations (i.e. women) on the following nine variables:

a.

Consider the logistic regression

$$Pr(y = 1|x) = \frac{e^{x^T \beta}}{1 + e^{x^T \beta}}$$

where y is the binary variable with y = 1 if the woman works and y = 0 if she does not. x is a 8-dimensional vector containing the eight features (including a one for the constant term that models the intercept). The goal is to approximate the posterior distribution of the 8-dim parameter vector β with a multivariate normal distribution

$$\beta|y, X \sim N(\hat{\beta}, J_y^{-1}(\hat{\beta}))$$

where $\hat{\beta}$ is the posterior mode and $J(\hat{\beta}) = -\frac{\delta^2 lnp(\beta|y)}{\delta\beta\delta\beta^T}|_{\beta=\hat{\beta}}$

Appendix

```r
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(fig.width=9, fig.height = 4.1)
library(tidyverse)
library(dplyr)
library(knitr)
set.seed(12345)
data0 <- read.table("TempLinkoping.txt", header = TRUE)
intercept <- rep(1,365)
data1 <- cbind(data0, "intercept"=intercept)
time2 <- data1$time^2
data1 <- cbind(data1, "time2"=time2)
```