Hw2

2022年3月11日 上午 9:27

Solve: ∵ $P(y_i|x_i,\beta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\beta^T x_i)^2}{2\sigma^2}}$. $P(\beta) = \frac{\tau}{2} e^{-\tau|\beta|}$.

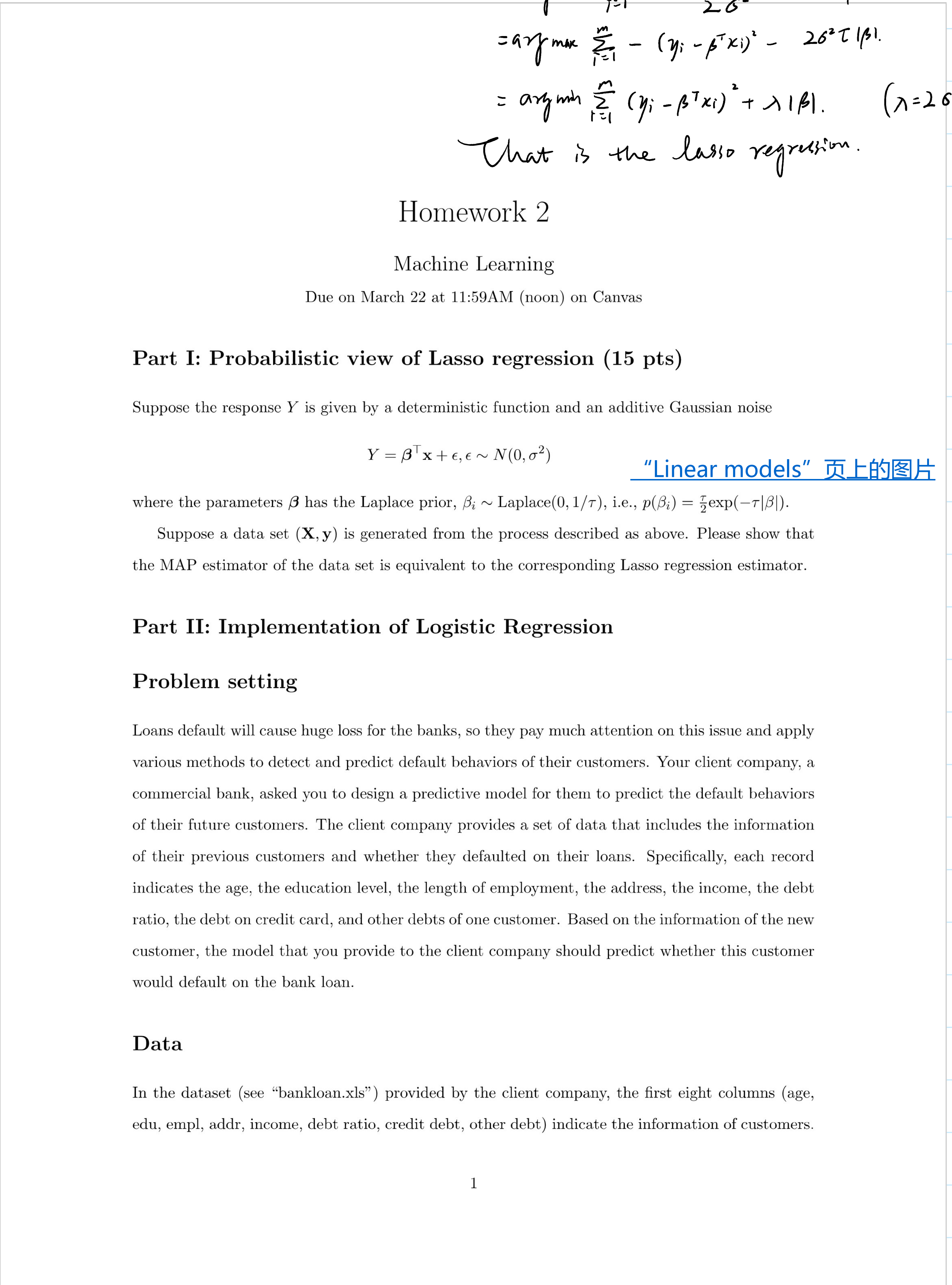
$$\hat{\beta}_{MAP} = \arg\max \sum_{i=1}^{m} \log p(y_i|x_i;\beta) + \log p(\beta).$$

$$= \arg\max \sum_{i=1}^{m} -\frac{(y_i-\beta^T x_i)^2}{2\sigma^2} - \tau|\beta|$$

$$= \arg\max \sum_{i=1}^{m} -(y_i-\beta^T x_i)^2 - 2\sigma^2\tau|\beta|.$$

$$= \arg\min \sum_{i=1}^{m} (y_i-\beta^T x_i)^2 + \lambda|\beta|. \quad (\lambda = 2\sigma^2\tau.)$$

That is the lasso regression.



Homework 2

Machine Learning

Due on March 22 at 11:59AM (noon) on Canvas

Part I: Probabilistic view of Lasso regression (15 pts)

Suppose the response Y is given by a deterministic function and an additive Gaussian noise

$$Y = \boldsymbol{\beta}^\top \mathbf{x} + \epsilon, \epsilon \sim N(0, \sigma^2)$$

"Linear models" 页上的图片

where the parameters $\boldsymbol{\beta}$ has the Laplace prior, $\beta_i \sim \text{Laplace}(0, 1/\tau)$, i.e., $p(\beta_i) = \frac{\tau}{2}\exp(-\tau|\beta|)$.

Suppose a data set $(\mathbf{X}, \mathbf{y})$ is generated from the process described as above. Please show that the MAP estimator of the data set is equivalent to the corresponding Lasso regression estimator.

Part II: Implementation of Logistic Regression

Problem setting

Loans default will cause huge loss for the banks, so they pay much attention on this issue and apply various methods to detect and predict default behaviors of their customers. Your client company, a commercial bank, asked you to design a predictive model for them to predict the default behaviors of their future customers. The client company provides a set of data that includes the information of their previous customers and whether they defaulted on their loans. Specifically, each record indicates the age, the education level, the length of employment, the address, the income, the debt ratio, the debt on credit card, and other debts of one customer. Based on the information of the new customer, the model that you provide to the client company should predict whether this customer would default on the bank loan.

Data

In the dataset (see "bankloan.xls") provided by the client company, the first eight columns (age, edu, empl, addr, income, debt ratio, credit debt, other debt) indicate the information of customers.