FinTech HW1

R08942025 徐瑋辰

1. Linear Regression

(1) Obtain training set and test set

Step 1: drop out other columns and keep the required ones

Step 2: apply one-hot encoder to discrete data

Step 3: shuffle data by pandas.DataFrame.sample

Step 4: separate training set and test set as follow:

Training set: first 80% data

Test set: last 20% data

(2) For regularization (c), find the optimal weights with maximum likelihood criterion

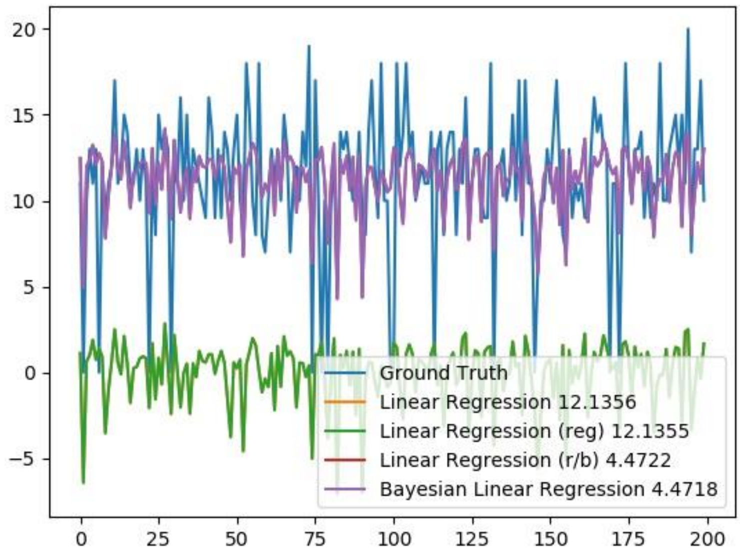
$$
\begin{aligned}
J(w) &= MSE_{train} + \frac{\lambda}{2} w^T w \\
&= (y - Xw)^T (y - Xw) + \frac{\lambda}{2} w^T w \\
&= (y^T - w^T X^T)(y - Xw) + \frac{\lambda}{2} w^T w \\
&= y^T y - y^T Xw - w^T X^T y + w^T X^T X w + \frac{\lambda}{2} w^T w \\
&= y^T y - 2 w^T X^T y + w^T X^T X w + \frac{\lambda}{2} w^T w
\end{aligned}
$$

$$
\frac{\partial}{\partial w} J(w) = -2X^T y + 2w X^T X + 2 \cdot \frac{\lambda}{2} \cdot w = 0
$$

$$
\Rightarrow w = \left(X^T X + \frac{\lambda}{2} I\right)^{-1} X^T y
$$

#

(3) compare the RMSEs and predicted G3 values



pseudo inverse:	RMSE = 12.1356
regularization without bias, lambda = 1:	RMSE = 12.1355
regularization with bias, lambda = 1:	RMSE = 4.4722
Bayesian Linear Regression (with bias), alpha = 1:	RMSE = 4.4718

Bayesian Linear Regression has the lowest RMSE. We already know that regularization with L2 norm has the same closed-form solution as Bayesian Linear Regression, but with different alpha (0.5 and 1, respectively), the result of (d) and (e) is slightly different.

(4) explain why predicted G3 values are closer to the ground truth for (d) and (e)

$$Y = b_0 + b_1 X_1$$

$$b_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$\begin{aligned} MSE &= \frac{1}{n}\sum(y-\hat{y})^2 \\ &= \frac{1}{n}\sum(y-(b_0+b_1x))^2 \\ &= \frac{1}{n}\sum(y-(\bar{y}-b_1\bar{x}+b_1x))^2 \\ &= \frac{1}{n}\sum(y-\bar{y}+b_1\bar{x}-b_1x)^2 \\ &= \frac{1}{n}\sum(\underline{y-b_1x}+b_1\bar{x}-\bar{y})^2 \\ &= \frac{1}{n}\sum[(y-\bar{y})-b_1(x-\bar{x})]^2 \end{aligned}$$

$$Y = b_1^* X_1$$

$$b_1^* = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\begin{aligned} MSE^* &= \frac{1}{n}\sum(y-\hat{y})^2 \\ &= \frac{1}{n}\sum(\underline{y-b_1^*x})^2 \end{aligned}$$

with $b_1^* > b_1$

$MSE < MSE^*$

2. Census Income Data Set

(1) Approach

- For target (the last column), I define “>50K” as label 1, “<=50K” as label 0.
- I use all the features, and for discrete items, I use one-hot encoder.
- The approach to obtain training set and test set is same as 1., but for each data, after calculating RMSE, I classify all the data into two groups. If the output > 0.5, it is seen as “>50K” (label 1), otherwise “<=50K” (label 0).
- After classification, the output is only 0 or 1 for each input, so I use “accuracy” to estimate the goodness of a model.

(2) Result

pseudo inverse:	RMSE = 0.56724069, acc = 77.8%
regularization without bias, lambda = 1:	RMSE = 0.56723977, acc = 77.8%
regularization with bias, lambda = 1:	RMSE = 0.50926444, acc = 83.5%
Bayesian Linear Regression (with bias), alpha = 1:	RMSE = 0.50926342, acc = 83.5%

(3) Compare

- For RMSE, the model with bias is better. This is the same as Prob. 1.
- For accuracy, the classification shows the same situation as RMSE.
- After tuning alpha, we get the lowest RMSE when alpha = 100000. However, we get the highest accuracy when alpha = 1.