

mlbd2022fall-logistic-regression

Machine Learning & Big Data 2022 Fall homework 2: logistic regression

<https://github.com/keyork/mlbd2022fall-logistic-regression>

Task

See [task.md](#).

Usage

```
1 pip install matplotlib numpy pandas colorlog tqdm
2 python train.py -h
3 python train.py --args ARGS
```

Model

Log-likelihood

$$\mathcal{L}(\beta) = \sum_{i=1}^m y_i \log \left(\frac{1}{1 + e^{-\beta^T \mathbf{x}}} \right) + (1 - y_i) \log \left(1 - \frac{1}{1 + e^{-\beta^T \mathbf{x}}} \right)$$

$$\beta = [\beta_0, \beta_1, \beta_2]^T$$

$$\mathbf{x} = [x_0, x_1, x_2]^T \quad (x_0 = 1)$$

Regression

$$\beta^{(k+1)} = \beta^{(k)} + \alpha \left[y_i - g(\beta^{(k)^T}) \mathbf{x}_i \right] \mathbf{x}_i$$

$$f_\beta(\mathbf{x}) = \frac{1}{1 + e^{-\beta^T \mathbf{x}}}$$

Find boundary pseudo-code

```
1 procedure LOGISTICREGRESSION(dataset, beta)
2     for data, label in dataset
3         beta += alpha * (label - data / (1 + exp(-beta * data))) * data
4     end for
5 end procedure
```

5-fold

Split all data into 5 parts, each part is used as val-set, and the rest is used as train-set, compare the MSE-Loss of 5 times train and select the smallest as the best model.

Norm

$$x = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Back Line Search

See [hw1: minibatch-sgd](#).

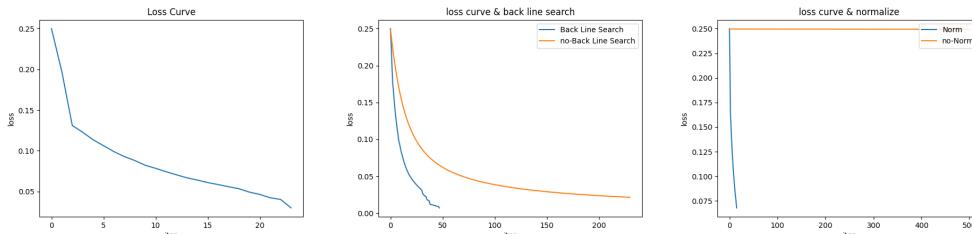
Experiments

logistic regression

Using normalize, 5-fold and back line search, we get the boundary by logistic regression and plot it in 3D, plot its loss curve(Fig. Left).

Ablation Experiment

Remove back line search(Fig. Middle), normalize(Fig. Right) and compare them with initial experiment.



Result

```
ckeyork@CdeMacBook-Pro:~/code/mlbd2022fall-logistic-regression
(mlbd) ckeyork@CdeMacBook-Pro ~ /code/mlbd2022fall-logistic-regression > python train.py
[WARNING]: Start (2022-10-18 16:47:56,236)
[INFO]: config list (2022-10-18 16:47:56,237)
fold: 5
    learning rate: 0.1
    iter: 500
    back line search: True
    normalize: True
    compare normalize: False
    compare back line search: False
    img dir: ./img/
    data path: ./data/l6_data.CSV
[INFO]: Start Train (2022-10-18 16:47:56,278)
[INFO]: Valid Fold ID: 0 (2022-10-18 16:47:56,278)
[INFO]: Valid Fold ID: 1 (2022-10-18 16:47:56,278)
100% | 500/500 [00:00:00:00, 2391.04it/s]
[INFO]: Valid Fold ID: 2 (2022-10-18 16:47:56,445)
100% | 500/500 [00:00:00:00, 3405.60it/s]
[INFO]: Valid Fold ID: 3 (2022-10-18 16:47:56,579)
100% | 500/500 [00:00:00:00, 3770.05it/s]
[INFO]: Valid Fold ID: 4 (2022-10-18 16:47:56,677)
100% | 500/500 [00:00:00:00, 3840.87it/s]
[INFO]: Result (2022-10-18 16:47:56,889)
Valid Pred # [1. 1. 0. 0. 0. 0.]
Valid Labels # [1. 1. 0. 0. 0. 0.]
Valid Acc: 100.0%
beta = [-0.98596654 5.76723161 6.04668562]
[INFO]: Draw 3D IMG (2022-10-18 16:47:56,888)
Path -> ./img/Img-fold-5-iter_500-bols_True-lr_0.1-norm_True.png
[INFO]: Draw Loss Curve (2022-10-18 16:48:00,589)
Path -> ./img/Loss-fold_5-iter_500-bols_True-lr_0.1-norm_True.png
[WARNING]: Done (2022-10-18 16:48:01,750)
(mlbd) ckeyork@CdeMacBook-Pro ~ /code/mlbd2022fall-logistic-regression > dev +
```

$$\beta = [\beta_0, \beta_1, \beta_2]^T = [-7.98596654, 5.76723161, 6.04668562]^T$$

