Problem 1

(1) $P(X_1) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{\frac{1}{2}(X_1^T-\mu^T)(X_1-\mu)}{\sigma^2 I}\right)$

$\log P(X_1) = \text{constant} - \frac{1}{2\sigma^2}(X_1^T-\mu^T)(X_1-\mu) - \log \sigma I$

$\max P(X_1)\cdot P(X_2)\cdots P(X_n) \Longleftrightarrow \max \log P(X_1)\cdot P(X_2)\cdots P(X_n).$

$\Rightarrow L = \max \log \sim = \text{constant} + \sum_{i=1}^{n} -\frac{(X_i^T-\mu^T)(X_i-\mu)}{2\sigma^2 I} - \log\sigma I$

$\frac{\partial L}{\partial \mu} = \sum_{i=1}^{n} \frac{1}{\sigma^2}(X_i-\mu) = 0 \Rightarrow \underline{\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i}$

$\frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma I} + \frac{1}{\sigma^3 I}\sum (X_i^T-\mu^T)(X_i-\mu) = 0.$

$\Rightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^T(X_i-\hat{\mu})$

$\Rightarrow \underline{\sigma^2 = \left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^T(X_i-\hat{\mu})\right)}$

(2) $P(X_1) = \frac{1}{\det(2\pi\Lambda)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X^T-\mu^T)\Lambda^{-1}(X-\mu)\right)$

$L = \log P(X_1)\cdots P(X_n) = \text{constant} + \sum_{i=1}^{n} -\frac{1}{2}(X_i^T-\mu^T)\Lambda^{-1}(X_i-\mu) - \frac{1}{2}\log(\Lambda)$

$\frac{\partial L}{\partial \mu} = \Lambda^{-1}(X_i-\mu) = 0 \Rightarrow \underline{\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i}$

$\frac{\partial L}{\partial \sigma_1} = -\frac{1}{\sigma_1} + \frac{1}{\sigma_1^3}(X_1^T-\mu^T)(X_1-\mu) = 0 \Rightarrow \sigma_1^2 = \left((X_1-\hat{\mu})^T(X_1-\hat{\mu})\right)$

$\Rightarrow \underline{\Lambda = \begin{bmatrix} (X_1-\hat{\mu})^T(X_1-\hat{\mu}) & & 0 \\ & \ddots & \\ 0 & & ((X_N-\hat{\mu})^T(X_N-\hat{\mu})) \end{bmatrix}}$

(3) $P(X_1) = \frac{1}{\det(2\pi\Lambda)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X_1-A\mu)^T\Lambda^{-1}(X_1-A\mu)\right).$

$\Rightarrow L = \log P(X_1)\cdots P(X_n) = \text{constant} + \sum_{i=1}^{n} -\frac{1}{2}(X_i-A\mu)^T\Lambda(X_i-A\mu) + f(\Lambda)$

$\frac{\partial L}{\partial \mu} = \sum_{i=1}^{n} A(X_i-A\mu) = 0 \quad \Rightarrow \cancel{\mu = \sum A^{-1}X_i} \; \underline{\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} A^{-1}X_i}$

Problem 2

1. $\mu_i = 1/(1+\exp(-\beta x_i)) \Rightarrow \frac{d\mu_i}{d(\beta x_i)} = \frac{e^{-\beta x_i}}{(1+e^{-\beta x_i})^2} = \mu_i(1-\mu_i) \Rightarrow \nabla \mu_{i\beta} = x_i \frac{d\mu_i}{d\beta}$

$$\frac{d}{d\beta}\left(\sum_{i=1}^{n}\left[y_i \log\mu_i + (1-y_i)\log(1-\mu_i)\right]\right) = \sum \left(\frac{y_i}{\mu_i}\nabla\mu_i - \frac{1-y_i}{1-\mu_i}\nabla\mu_i\right) = \sum (y_i-\mu_i)X_i$$

$$= X^T(y-\mu)$$

$\Rightarrow$ gradient of a: $2\lambda\beta - X^T(y-\mu)$ $\quad \left(\mu = \begin{bmatrix}\mu_1\\ \mu_2\\ \vdots \\ \mu_N\end{bmatrix}\right)$

gradient of b: $2\lambda\beta + X^TX\beta - X^Ty$

2. Hessian of a: $2\lambda + \sum X_i^T \cdot \nabla\mu_{i\beta} = 2\lambda + \sum X_i^T X_i (1-\mu_i)\mu_i$

Hessian of a $\Rightarrow 2\lambda + X^T\Omega X$ $\quad \left(\Omega = \begin{bmatrix}(1-\mu_1)\mu_1 & & \\ & (1-\mu_2)(\mu_2) & \\ & & \ddots \\ & & & (1-\mu_N)\mu_N\end{bmatrix}\right)$

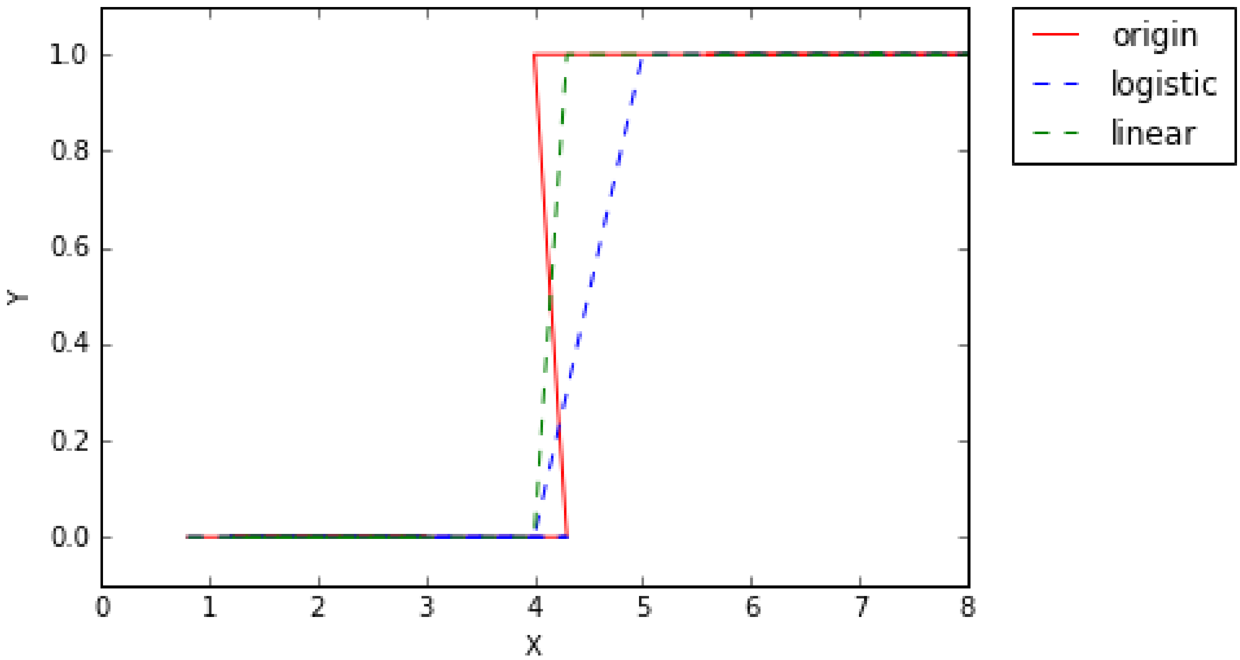
Hessian of b: $2\lambda + X^TX$

3. update $= (\nabla^2)^{-1} \cdot \nabla$

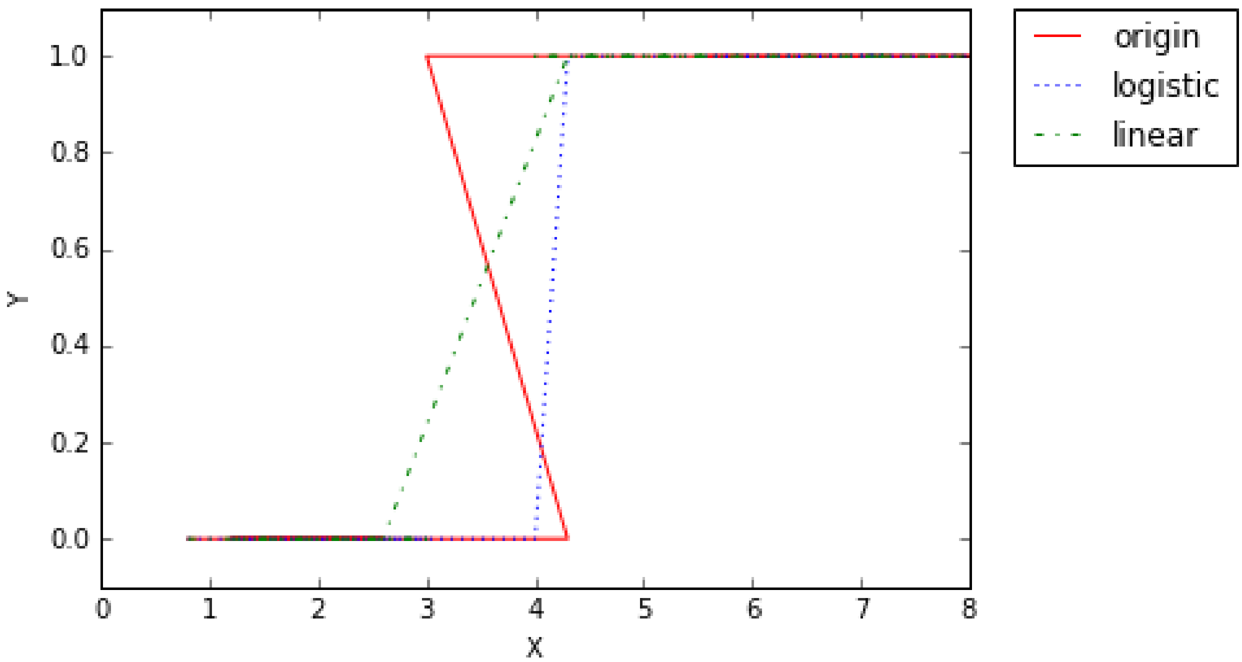
a: $(2\lambda + X^T\Omega X)^{-1} \cdot (2\lambda\beta - X^T(y-\mu))$ $\quad \left(\Omega = \begin{bmatrix}(1-\mu_1)\mu_1 & & \\ & \ddots & \\ & & (1-\mu_N)\mu_N\end{bmatrix} \; \mu = \begin{bmatrix}\mu_1\\ \vdots \\ \mu_N\end{bmatrix}\right.$

b: $(2\lambda + X^TX)^{-1} \cdot (2\lambda\beta + X^TX\beta - X^Ty)$

4.


origin
logistic
linear
1.0
0.8
0.6
0.4
0.2
0.0
Y
0
1
2
3
4
5
6
7
8
X

5.


origin
logistic
linear
1.0
0.8
0.6
0.4
0.2
0.0
Y
0
1
2
3
4
5
6
7
8
X

Problem 3

1. $J_\lambda(\beta) = \frac{1}{2} Tr\left[(y-X\beta)^T(y-X\beta)\right] + \cancel{\lambda Tr\|\beta\|} \lambda\|\beta\|_1$

$$= \frac{1}{2} y^T y - y^T X\beta + \frac{1}{2}\beta^T \underline{\underline{X^T X}} \beta + \lambda \sum_{i=1}^{d} |\beta_i|$$

$X^T X = nI$.

$$\Rightarrow J_\lambda(\beta) = \underbrace{\frac{1}{2} y^T y}_{g(y)} + \underbrace{\sum_{i=1}^{d} \lambda|\beta_i| - y^T X_i \cdot \beta_i + \frac{1}{2}\beta_i^2}_{\sum f(X_i, y, \beta_i, \lambda)}$$

So β_i only depends on i-th feature.

2. if $\beta_i^* > 0 \Rightarrow \frac{d}{d\beta_i} J_\lambda(\beta_i) = \lambda - y^T X_i \cancel{\beta_i} + \beta_i$

$$\Rightarrow \beta_i = y^T X_i - \lambda$$

3. if $\beta_i^* > 0 \Rightarrow \frac{d}{d\beta_i} J_\lambda(\beta_i) = -\lambda - y^T X_i + \beta_i$

$$\Rightarrow \beta_i = y^T X_i + \lambda$$

4. when $\beta_i^* > 0$ $\min f(X_i, y, \beta_i, \lambda) = -\frac{1}{2}(y^T X_i - \lambda)^2$

when $\beta_i^* < 0$, $\min f = -\frac{1}{2}(y^T X_i + \lambda)^2$

$\Rightarrow$ when $\beta_i^* = 0$, $\min f = 0$.

$\Rightarrow 0 \leq -\frac{1}{2}(y^T X_i - \lambda)$ and $-\frac{1}{2}(y^T X_i + \lambda)^2$

$\Rightarrow \lambda = 0$, and $y^T X_i = 0$

5. for $\beta^* = \arg\min_\beta \{ J_\lambda(\beta) = \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \}$

$$\Rightarrow \frac{\partial J(\lambda)(\beta_i)}{\partial \beta_i} = \lambda \beta_i^2 + \frac{1}{2}\beta_i^2 - y^T x_i \beta_i$$

$$\Rightarrow \beta_i^* = -\frac{y^T x_i}{2\lambda+1} == 0 \quad \Rightarrow y^T x_i = 0.$$

(not need $\lambda = 0$)

Problem 4

1. gradient: $2\lambda\beta - X^T(y-\mu)$.

2. stochastic gradient: $2\lambda\beta - X_i^T(y_i-\mu_i)$.

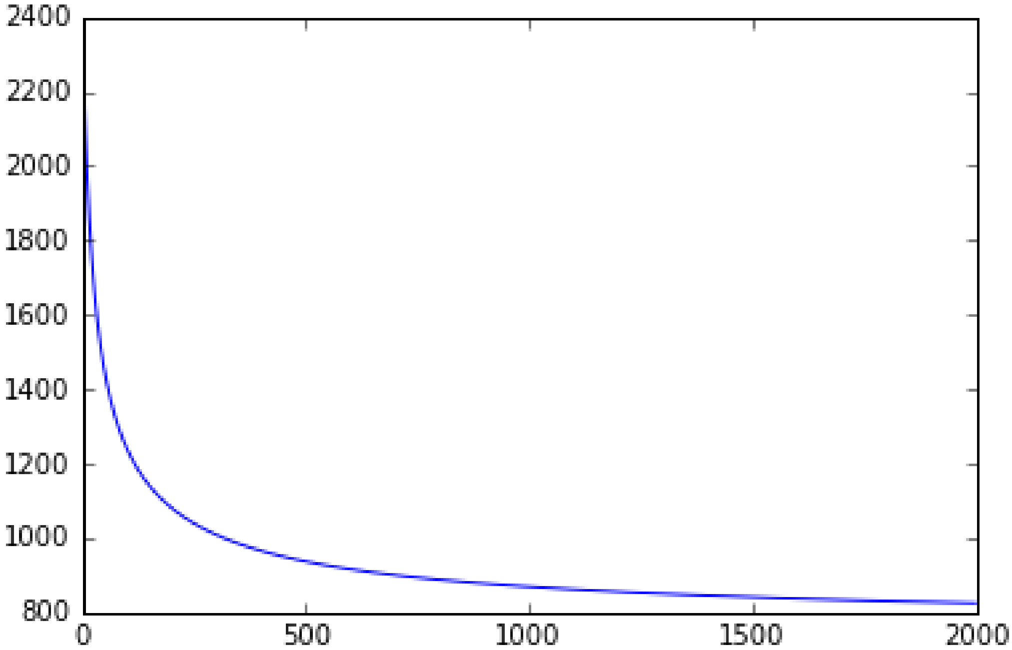
curve in (1) is smoother and has less "loss".

3.

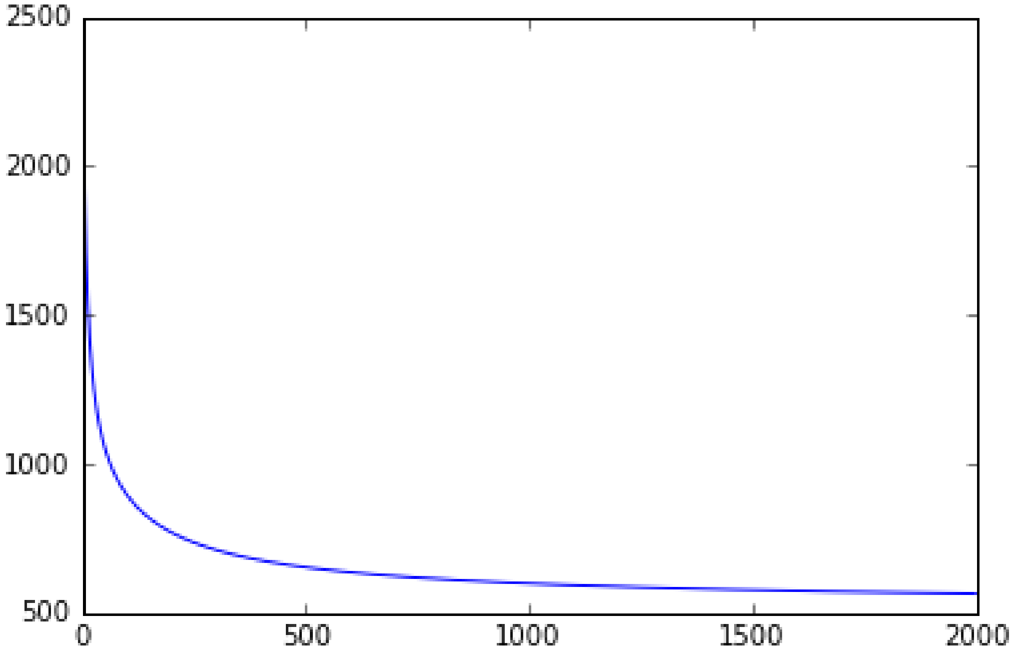
3. The learning rate decreased during the iteration. And the training accuracy dropped down.

1. **plot**

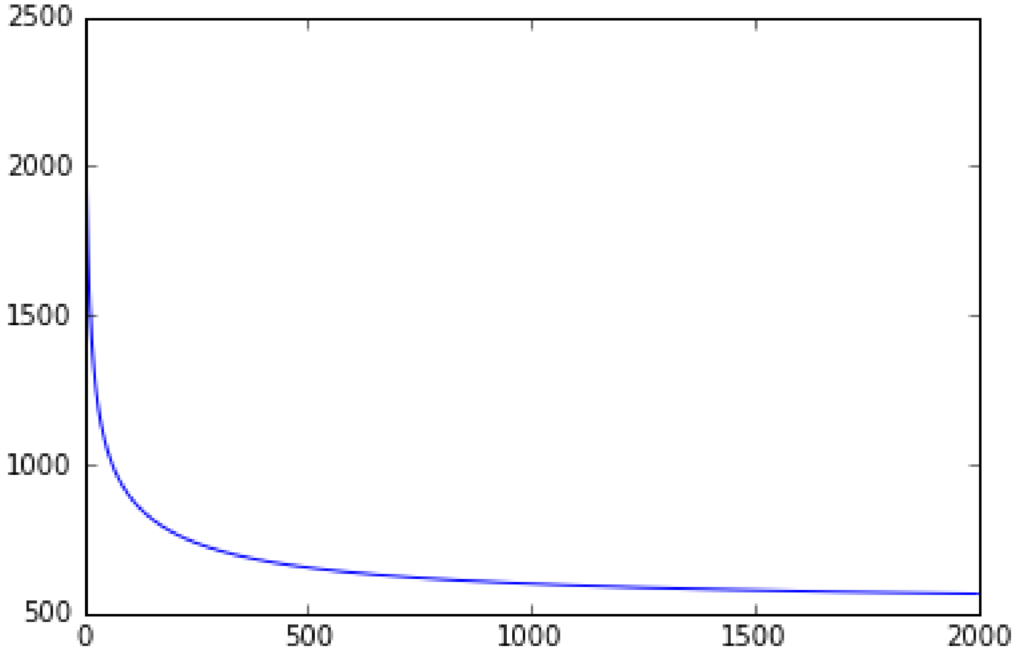
Standardize:



Logistic:

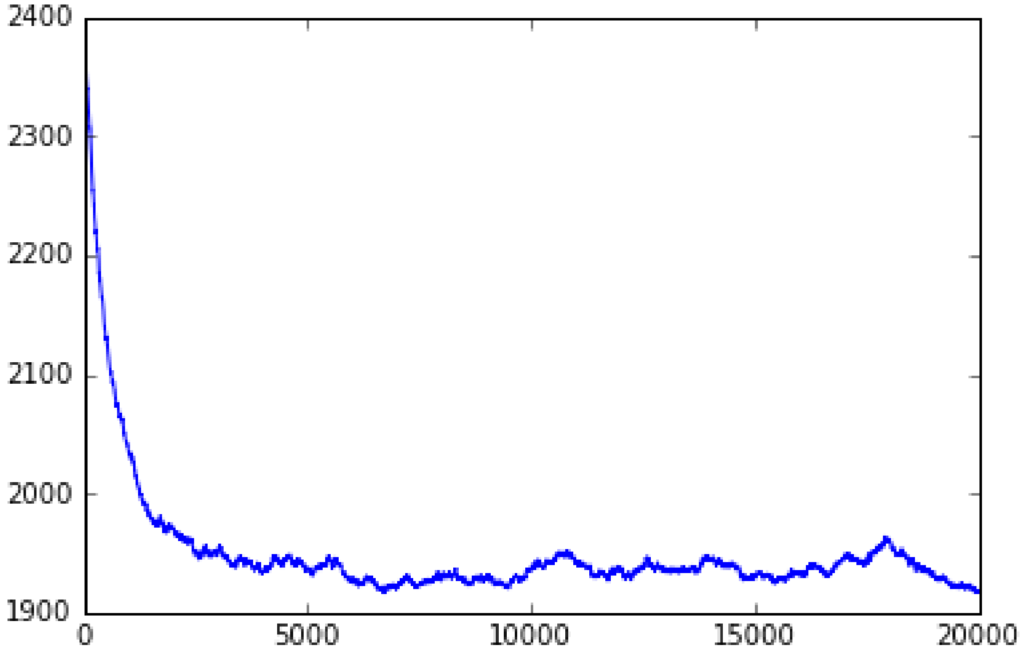


Binary:

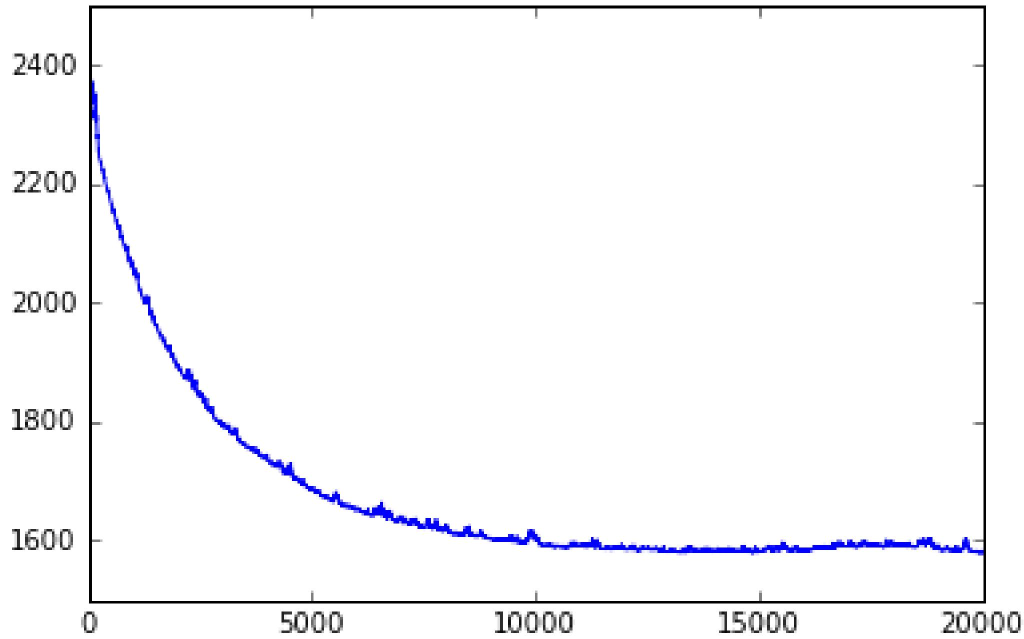


2. plot

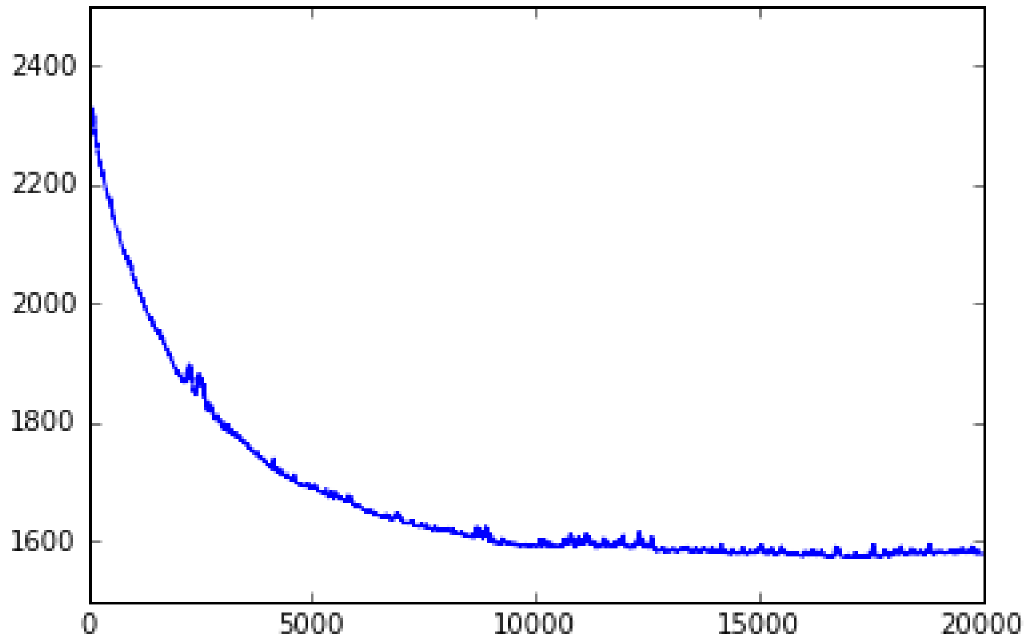
Standardize:



Logistic:

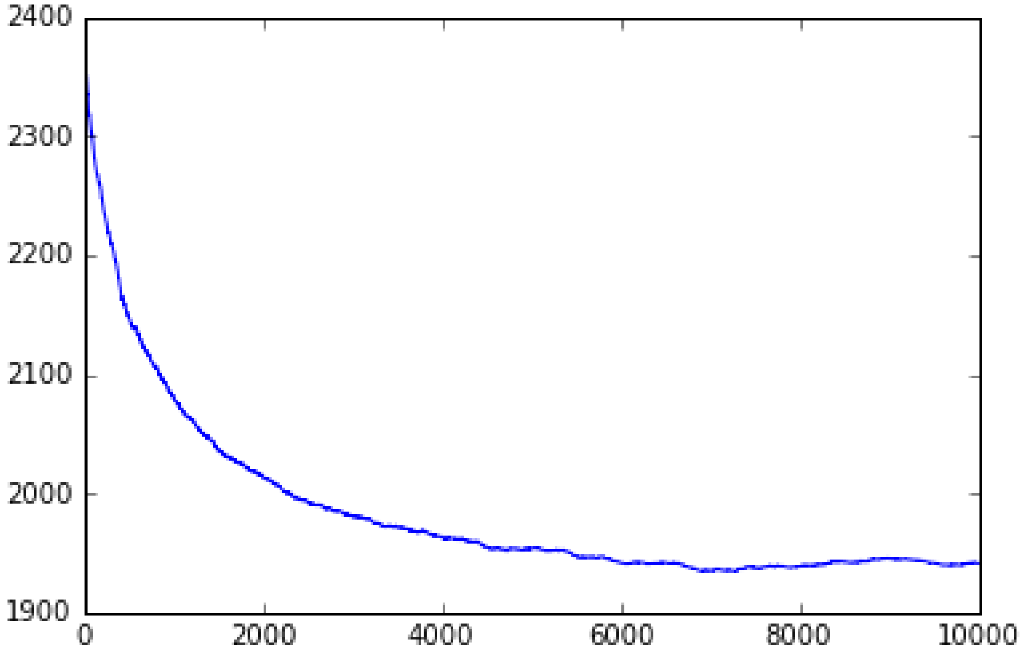


Binary:

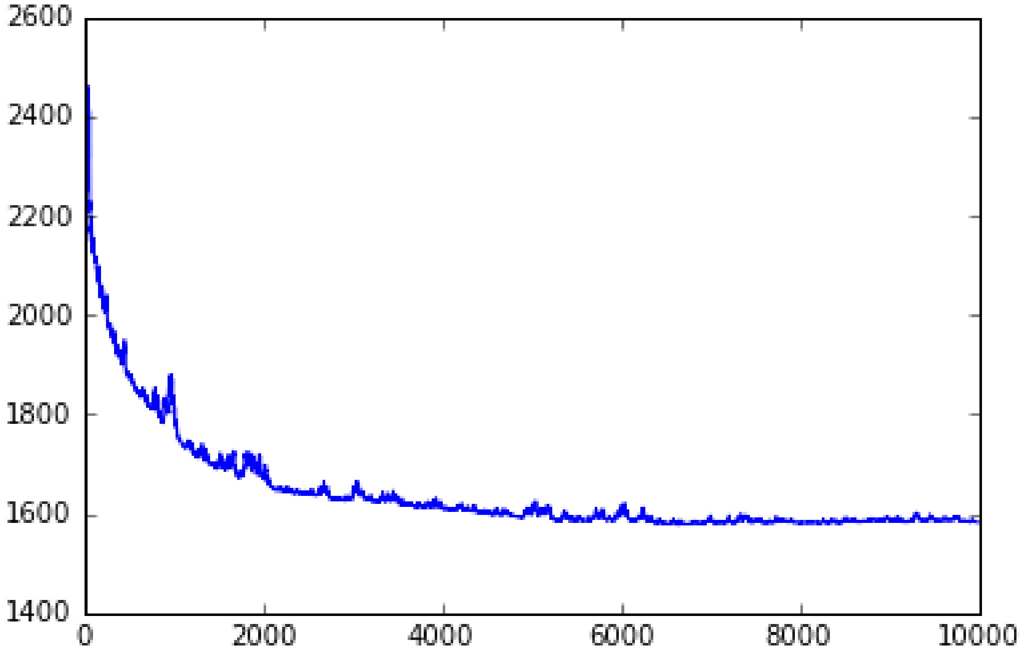


3. plot

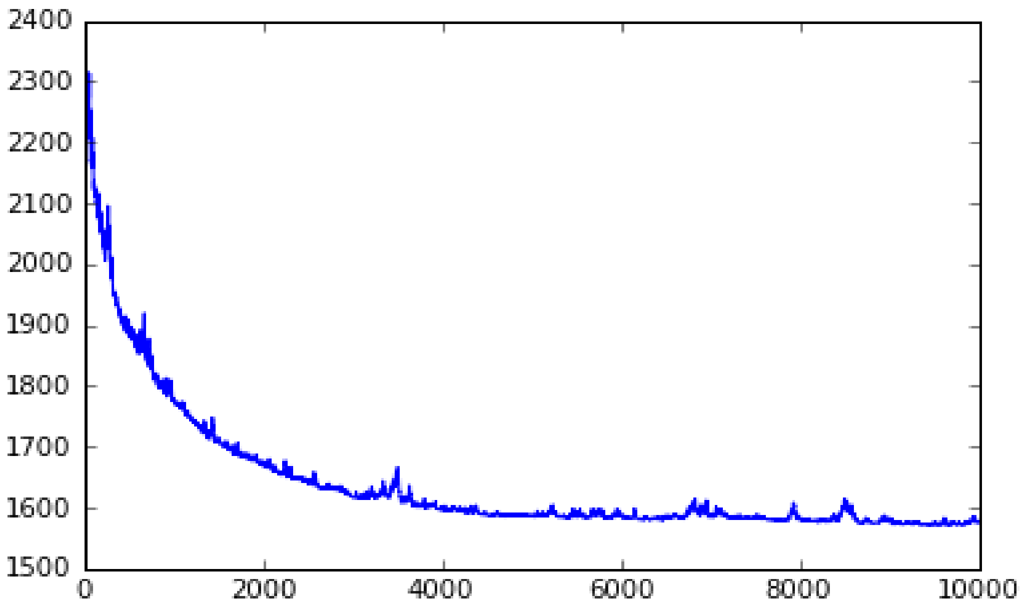
Standardize:



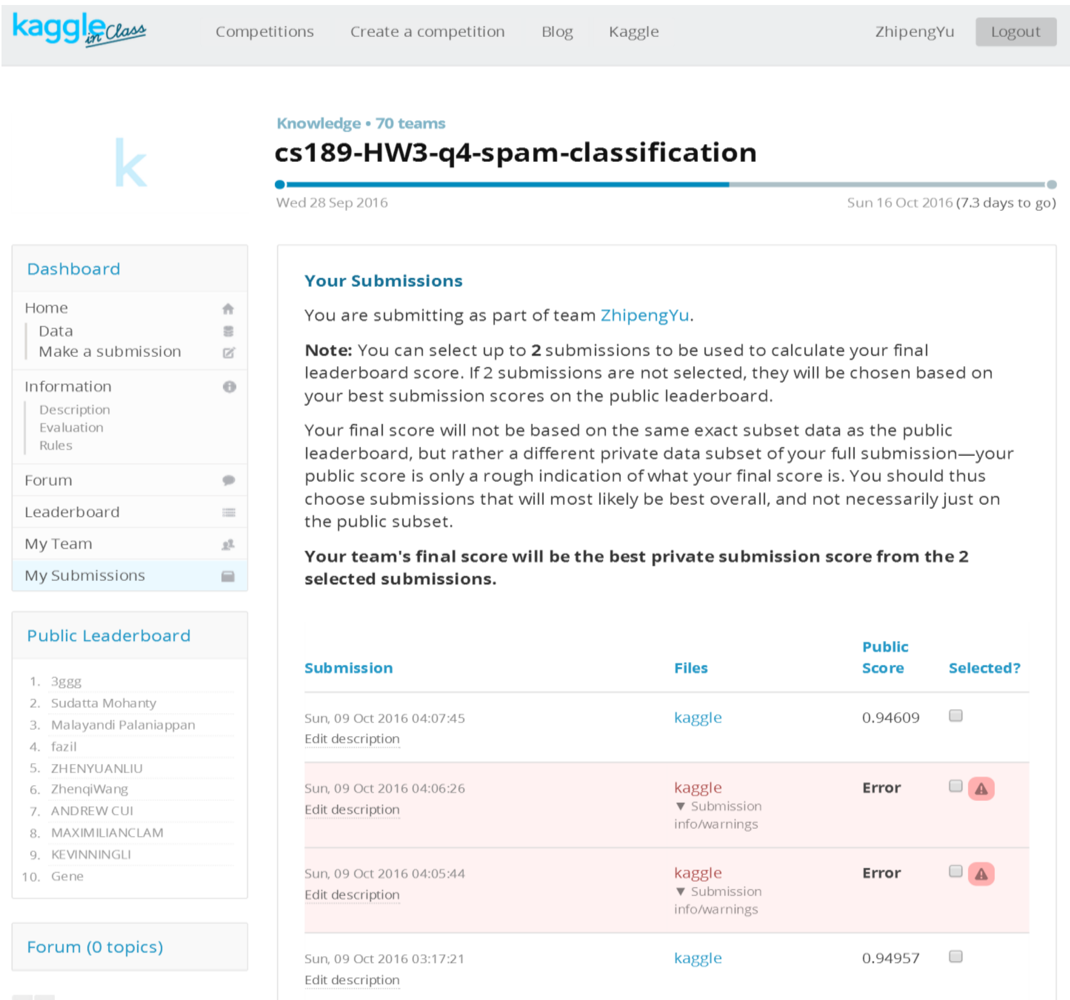
Logistic:



Binary:


2400
2300
2200
2100
2000
1900
1800
1700
1600
1500
0
2000
4000
6000
8000
10000

5.


kaggle in Class
Competitions
Create a competition
Blog
Kaggle
ZhipengYu
Logout
Knowledge • 70 teams
cs189-HW3-q4-spam-classification
Wed 28 Sep 2016
Sun 16 Oct 2016 (7.3 days to go)
Dashboard
Home
Data
Make a submission
Information
Description
Evaluation
Rules
Forum
Leaderboard
My Team
My Submissions
Public Leaderboard
1. 3ggg
2. Sudatta Mohanty
3. Malayandi Palaniappan
4. fazil
5. ZHENYUANLIU
6. ZhenqiWang
7. ANDREW CUI
8. MAXIMILIANCLAM
9. KEVINNINGLI
10. Gene
Forum (0 topics)
Your Submissions
You are submitting as part of team ZhipengYu.
Note: You can select up to 2 submissions to be used to calculate your final leaderboard score. If 2 submissions are not selected, they will be chosen based on your best submission scores on the public leaderboard.
Your final score will not be based on the same exact subset data as the public leaderboard, but rather a different private data subset of your full submission—your public score is only a rough indication of what your final score is. You should thus choose submissions that will most likely be best overall, and not necessarily just on the public subset.
Your team's final score will be the best private submission score from the 2 selected submissions.
Submission
Files
Public Score
Selected?
Sun, 09 Oct 2016 04:07:45
Edit description
kaggle
0.94609
Sun, 09 Oct 2016 04:06:26
Edit description
kaggle
▼ Submission info/warnings
Error
Sun, 09 Oct 2016 04:05:44
Edit description
kaggle
▼ Submission info/warnings
Error
Sun, 09 Oct 2016 03:17:21
Edit description
kaggle
0.94957