CSE512 Fall 2018 Machine Learning - Homework 3

Your Name: Caitao Zhan

Solar ID: 111634527

NetID email address: caitao.zhan@stonybrook.edu

Names of people whom you discussed the homework with:

1 Naive Bayes and Logistic Regression

1.1

There are **7** parameters that must estimate.

1. For X_1, need to estimate $\theta_{10} = P(X_1 = 1|Y = 0)$ and $\theta_{11} = P(X_1 = 1|Y = 1)$
2. For X_2, need to estimate mean $\mu_{20} = E[X_2|Y = 0],\ \mu_{21} = E[X_2|Y = 1]$ and variance $\sigma_{20}^2 = E[(X_2 - \mu_{20})^2|Y = 0],\ \sigma_{21}^2 = E[(X_2 - \mu_{21})^2|Y = 1]$
3. For Y, need to estimate $\pi = P(Y = 1)$

For X_1, the Bernoulli distribution: $b(x|\theta) = \theta^x(1-\theta)^{1-x}$
For X_2, the Gaussian distribution: $g(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Assume the feature values are x_1 and x_2:

$$
\begin{aligned}
P(Y = 0|\mathbf{X}) &= \frac{P(\mathbf{X}|Y = 0)P(Y = 0)}{P(\mathbf{X})} \\
&= \frac{P(X_1 = x_1|Y = 0)P(X_2 = x_2|Y = 0)P(Y = 0)}{\sum_{j=0}^{1} P(X_1 = x_1|Y = j)P(X_2 = x_2|Y = j)P(Y = j)} \\
&= \frac{b(x_1|\theta_{10}) \times g(x_2|\mu_{20}, \sigma_{20}^2) \times (1-\pi)}{b(x_1|\theta_{10}) \times g(x_2|\mu_{20}, \sigma_{20}^2) \times (1-\pi) + b(x_1|\theta_{11}) \times g(x_2|\mu_{21}, \sigma_{21}^2) \times \pi}
\end{aligned}
$$

$$
P(Y = 1|\mathbf{X}) = \frac{b(x_1|\theta_{11}) \times g(x_2|\mu_{21}, \sigma_{21}^2) \times \pi}{b(x_1|\theta_{10}) \times g(x_2|\mu_{20}, \sigma_{20}^2) \times (1-\pi) + b(x_1|\theta_{11}) \times g(x_2|\mu_{21}, \sigma_{21}^2) \times \pi}
$$

1.2

For Y, assume $\pi = P(Y = 1)$ is estimated.
For X_i, assume $\theta_{i0} = P(X_i = 1|Y = 0)$ and $\theta_{i1} = P(X_i = 1|Y = 1)$ are estimated.
$P(X_i|Y = 0) = \theta_{i0}^{X_i}(1 - \theta_{i0})^{1-X_i}$
$P(X_i|Y = 1) = \theta_{i1}^{X_i}(1 - \theta_{i1})^{1-X_i}$

$$\begin{aligned}
P(Y = 1|X) &= \frac{P(X|Y = 1)P(Y = 1)}{P(X|Y = 0)P(Y = 0) + P(X|Y = 1)P(Y = 1)} \\
&= \frac{1}{1 + \frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}} \\
&= \frac{1}{1 + \exp[\ln \frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}]} \\
&= \frac{1}{1 + \exp[\sum_i \ln \frac{P(X_i|Y=0)P(Y=0)}{P(X_i|Y=1)P(Y=1)} + \ln \frac{P(Y=0)}{P(Y=1)}]} \\
&= \frac{1}{1 + \exp[\sum_i \ln \frac{\theta_{i0}^{X_i}(1-\theta_{i0})^{1-X_i}}{\theta_{i1}^{X_i}(1-\theta_{i1})^{1-X_i}} + \ln \frac{1-\pi}{\pi}]} \\
&= \frac{1}{1 + \exp[\sum_i (\ln \frac{\theta_{i0}}{\theta_{i1}} X_i + \ln \frac{1-\theta_{i1}}{1-\theta_{i0}}(1 - X_i)) + \ln \frac{1-\pi}{\pi}]} \\
&= \frac{1}{1 + \exp[\sum_i (\ln \frac{\theta_{i0}}{\theta_{i1}} - \ln \frac{1-\theta_{i1}}{1-\theta_{i0}})X_i + \ln \frac{1-\pi}{\pi} + \sum_i \ln \frac{1-\theta_{i0}}{1-\theta_{i1}}]} \\
&= \frac{1}{1 + \exp(-(\sum_{i=1}^{d} w_i X_i + w_{d+1}))}
\end{aligned}$$

where,
$w_{d+1} = -(\ln \frac{1-\pi}{\pi} + \sum_i \ln \frac{1-\theta_{i0}}{1-\theta_{i1}})$
$w_i = \ln \frac{\theta_{i0}}{\theta_{i1}} - \ln \frac{1-\theta_{i1}}{1-\theta_{i0}}$

2 Implementation of Logistic Regression

2.1

We assume $P(Y = 1|\bar{X}^i; \boldsymbol{\theta}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^T \bar{X}^i)} = \frac{\exp(\boldsymbol{\theta}^T \bar{X}^i)}{1+\exp(\boldsymbol{\theta}^T \bar{X}^i)}$

$$\begin{aligned}
log(P(Y^i|\bar{X}^i; \boldsymbol{\theta})) &= Y^i log(P(Y = 1|\bar{X}^i; \boldsymbol{\theta})) + (1 - Y^i) log(P(Y = 0|\bar{X}^i; \boldsymbol{\theta})) \\
&= Y^i log(\frac{\exp(\boldsymbol{\theta}^T \bar{X}^i)}{1 + \exp(\boldsymbol{\theta}^T \bar{X}^i)}) + (1 - Y^i) log(\frac{1}{1 + \exp(\boldsymbol{\theta}^T \bar{X}^i)}) \\
&= Y^i \boldsymbol{\theta}^T \bar{X}^i - log(1 + \exp(1 + \boldsymbol{\theta}^T \bar{X}^i))
\end{aligned}$$

$$\begin{aligned}
\frac{\partial log(P(Y^i|\bar{X}^i; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} &= (Y^i - \frac{\exp(\boldsymbol{\theta}^T \bar{X}^i)}{1 + \exp(\boldsymbol{\theta}^T \bar{X}^i)})\bar{X}^i \\
&= (Y^i - P(Y = 1|\bar{X}^i; \boldsymbol{\theta}))\bar{X}^i
\end{aligned}$$

2.2

2.3

1. (a)Number of epochs till termination = 369

   (b)$L(\boldsymbol{\theta})$ as a function of epoch

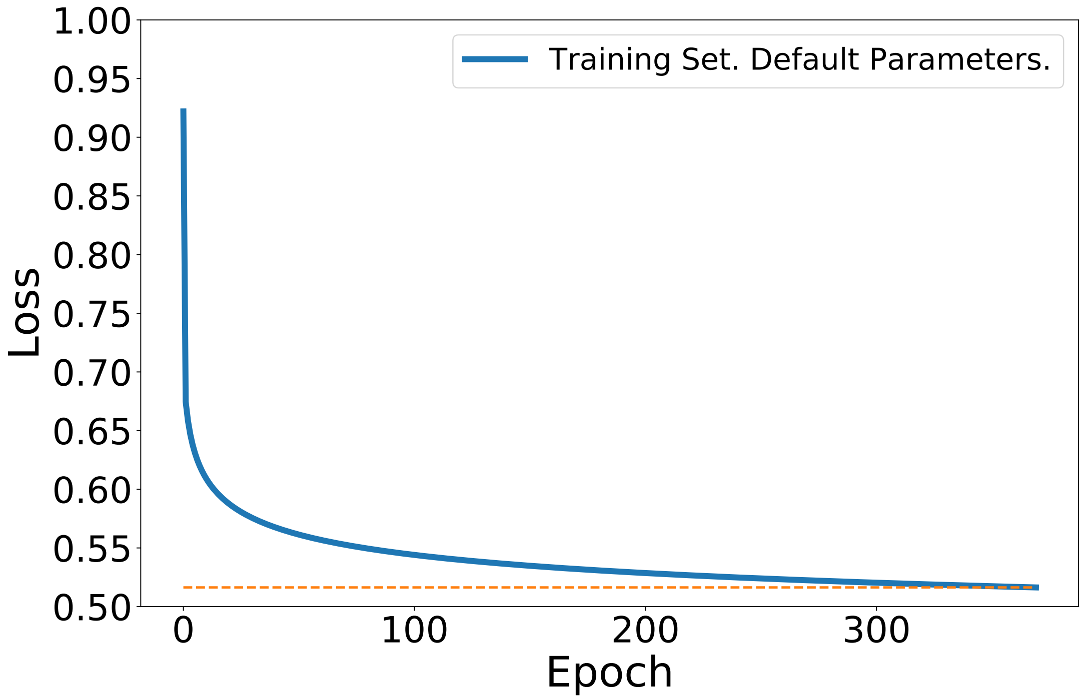

Figure 1: Default parameters

   (c) Final value of $L(\boldsymbol{\theta})$ after optimization = 0.516355139283

2. (a)Find a pair of η that leads to fast converge.

   Best value for, $\eta_0 = 3$, $\eta_1 = 1$

   Number of epochs for training = 13

   Final value of $L(\boldsymbol{\theta}) = 0.31952785732$

   (b)

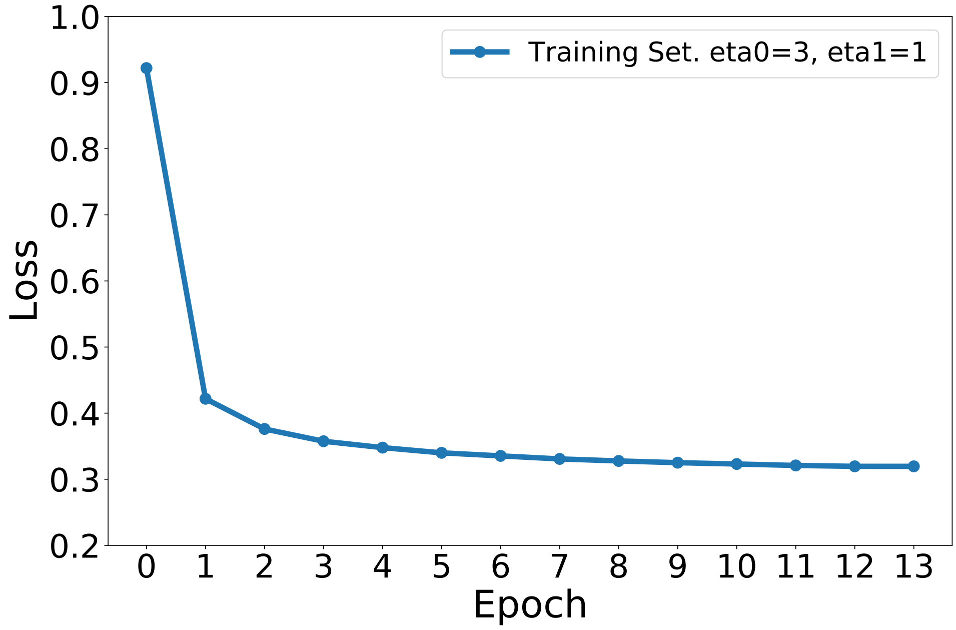

Figure 2: $\eta_0 = 3$, $\eta_1 = 1$

3. (a)$L(\boldsymbol{\theta})$ as a function of epoch

   When $\eta_0 = 500$, $\eta_1 = 1$, the loss function converge at around 0.223 on training data set, converge at around 0.261 on validation data set.

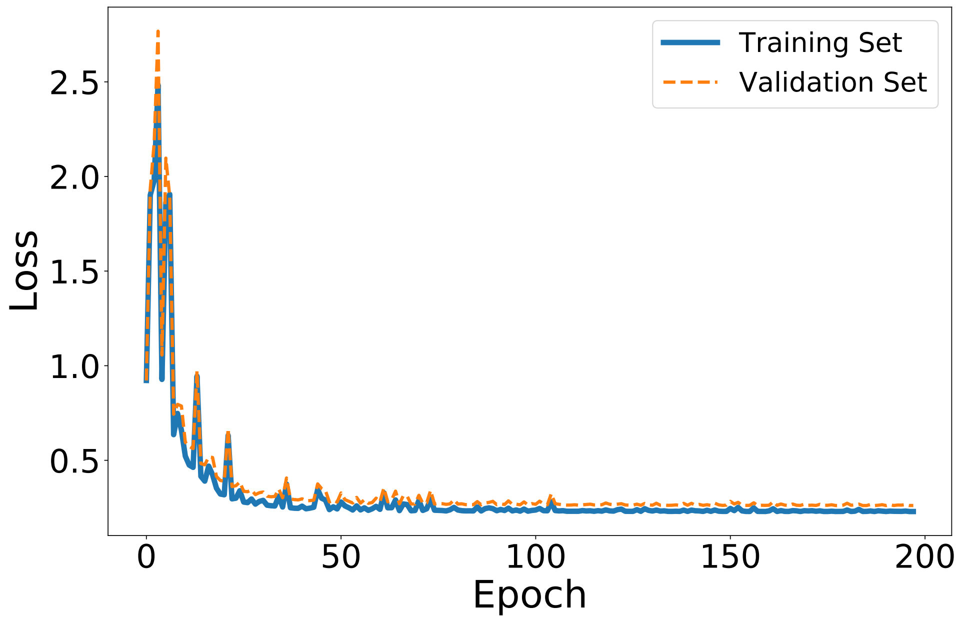

Figure 3: $\eta_0 = 500$, $\eta_1 = 1$

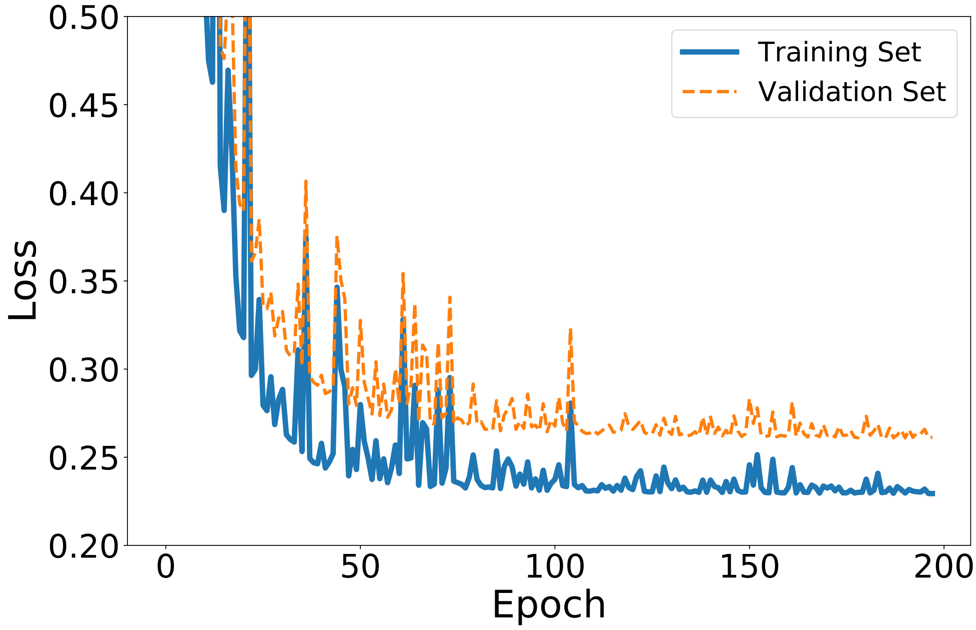

Figure 4: $\eta_0 = 500$, $\eta_1 = 1$. Different Y axis scale than figure 3

(b)Accuracy as function of epoch.

When $\eta_0 = 500$, $\eta_1 = 1$, the accuracy converge at around 0.91 on training data set, converge at around 0.89 on validation data set.

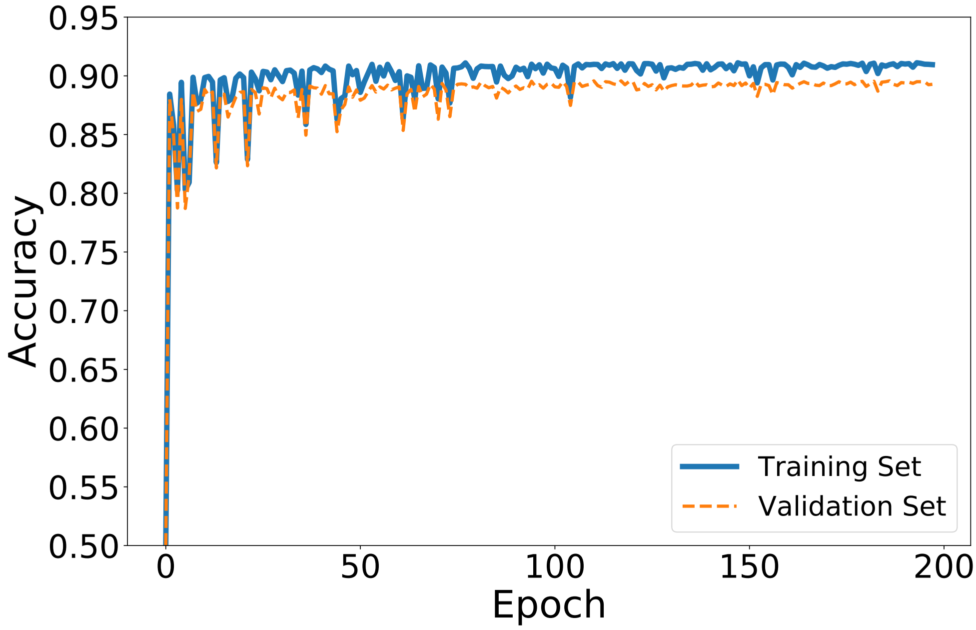

Figure 5: $\eta_0 = 500$, $\eta_1 = 1$.

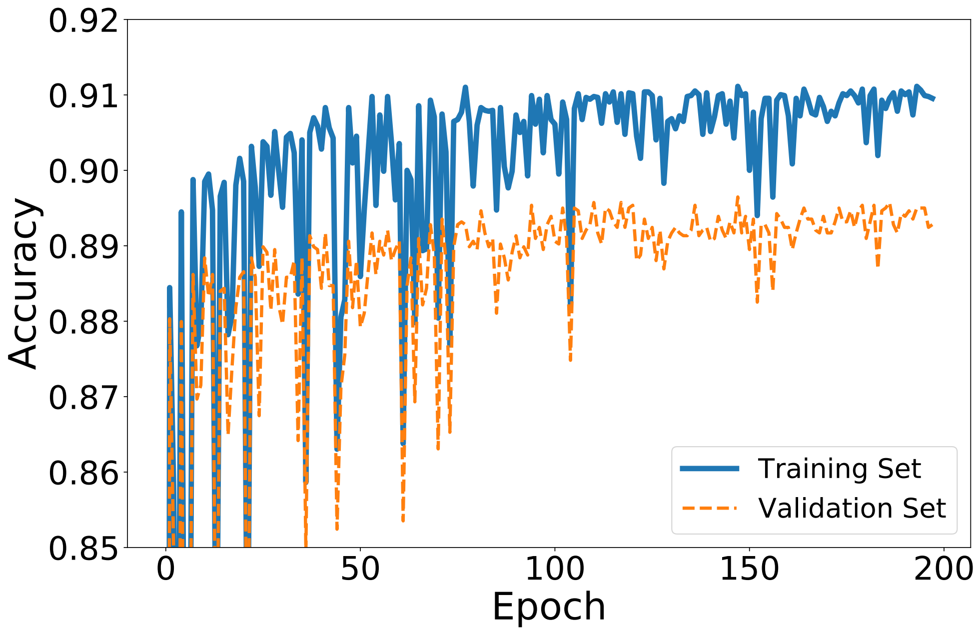

Figure 6: $\eta_0 = 500$, $\eta_1 = 1$. Different Y axis scale than figure 5

4. (a) Receiver operator curve on validation data set.

   Area under curve = 0.960097981503403 64

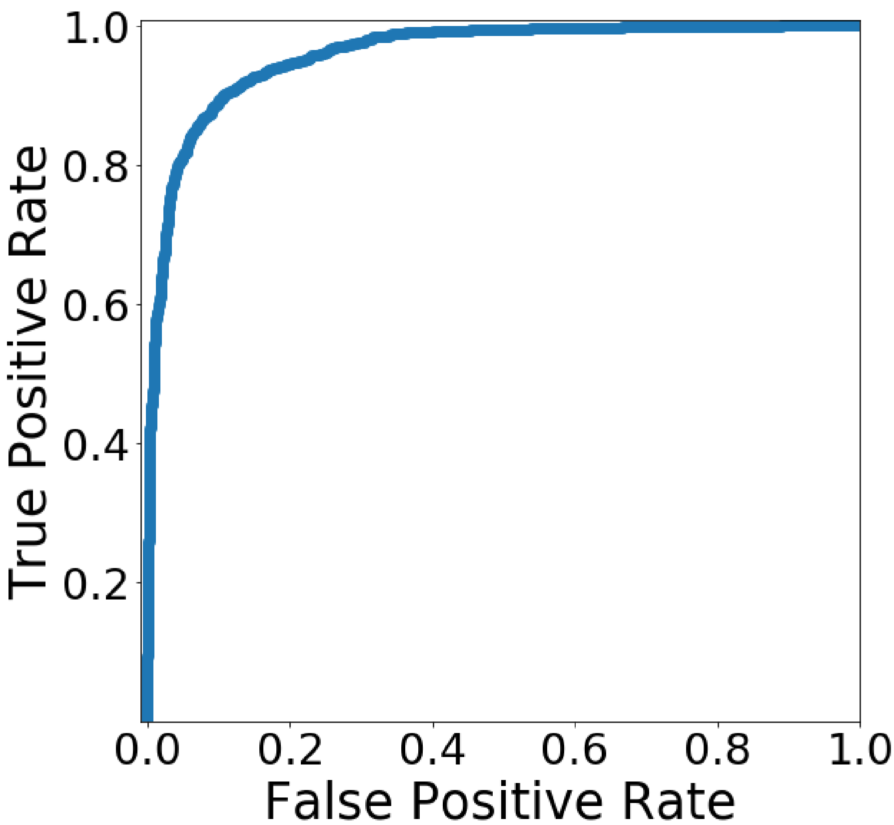

Figure 7: $\eta_0 = 500$, $\eta_1 = 1$, $m = 3$, δ=0.00005.

(b) Precision recall curve on validation data set.

Average Precision = 0.9571824249686488

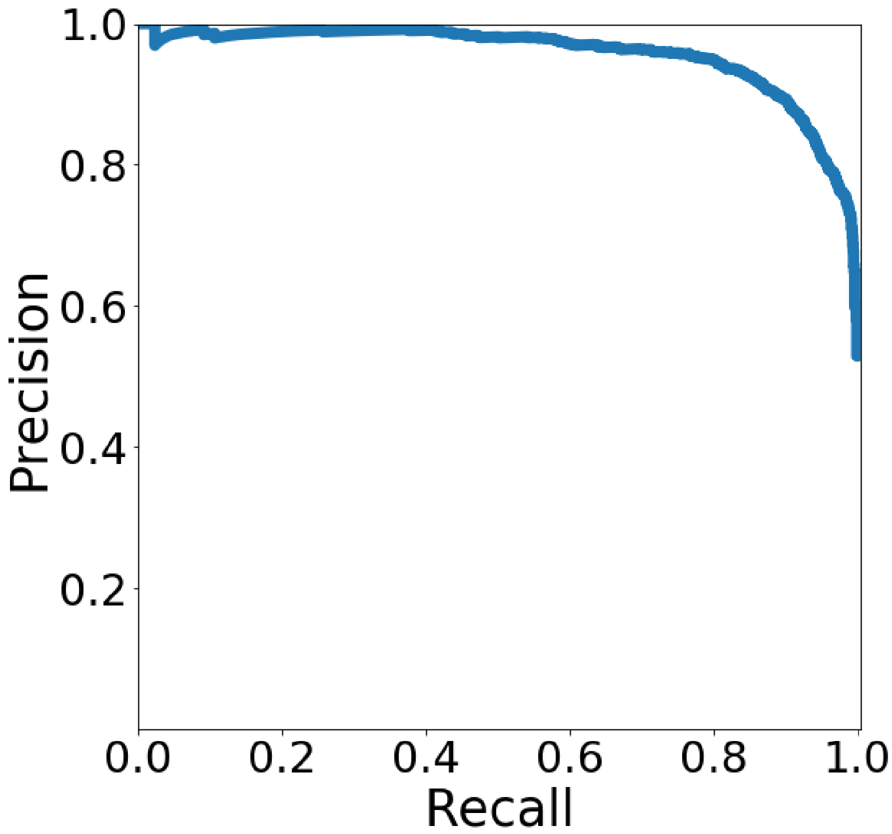

Figure 8: $\eta_0 = 500$, $\eta_1 = 1$, $m = 3$, δ=0.00005.

2.4

Kaggle submit.

Figure 9: Submit with name Caitao Zhan