1 CSE 512 Machine Learning - Homework 3

Name: Rohit Bhal
Solar ID: 112073893
NetID email address: rbhal@cs.stonybrook.edu / rohit.bhal@.stonybrook.edu

1. Question 1

1.1 Given $X \longrightarrow Y$, where Y is boolean

where $X = (X_1, X_2)$

X_1 is boolean, X_2 is continuous

$$P(X_1|Y_k) = \prod_{i=1}^{4} (P_{ki})^{x_i} (1-P_{ki})^{1-x_i}$$

$$P(X_2=v|Y_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

$$P(Y|k) = \frac{P(k|Y)\; P(Y)}{P(k)}$$

$$\therefore P(Y_k|(X_1, X_2)) = \frac{P(X_1|Y_k)\; P(X_2|Y_k)\; P(Y)}{\sum_{k=0}^{1} P(X_1|Y_k)\; P(X_2|Y_k)\; P(Y_k)}$$

$$= \frac{P(X_1|Y_k)\; P(X_2|Y_k)\; P(Y)}{P(Y_0)\; P(X_1|Y_0)\; P(X_2|Y_0) + P(Y_1)\; P(X_1|Y_1) P(X_2|Y_1)}$$

Now, we need the following parameters to estimate $P(Y|X)$:

$\sigma_k, \mu_k, P(Y_k), P_{ki}$ where $k \in \{0,1\}$

P_{ki} is the probability of class Y_k generating the term x_i.

$$P(Y|X) = \frac{\left(\prod_{i=1}^{n} (P_{ki})^{x_i} (1-P_{ki})^{1-x_i}\right) \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}} P(Y_k)}{\left(P(Y_0)\left(\prod_{i=1}^{n} (P_{0i})^{x_i} (1-P_{0i})^{1-x_i}\right) \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(v_0-\mu_0)^2}{2\sigma_0^2}} + P(Y_1)\left(\prod_{i=1}^{n} (P_{1i})^{x_i} (1-P_{1i})^{1-x_i}\right) \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(v-\mu_1)^2}{2\sigma_1^2}}\right)}$$

1.2

Given Y is boolean & $X = (X_1 \ldots X_d)$ is a vector of boolean variables.

$$P(Y=1 \mid X) = \frac{P(X \mid Y=1)\, P(Y=1)}{P(Y=0, X) + P(Y=1, X)}$$

$$= \frac{P(X \mid Y=1)\, P(Y=1)}{P(Y=1)\, P(X \mid Y=1) + P(Y=0)\, P(X \mid Y=0)}$$

$$P(Y=1|X) = \frac{1}{1+\frac{P(Y=0)\,P(X|Y=0)}{P(Y=1)\,P(X|Y=1)}}$$

Now, $P(Y)$ follows the binomial distribution.

$$P(X|Y=k) = \prod_{i=1}^{n} (P_{ki})^{x_i} (1-P_{ki})^{1-x_i}$$

$$\therefore\ P(Y=1|X) = \frac{1}{1+\frac{P(Y=0)\,P(X|Y=0)}{P(Y=1)\,P(X|Y=1)}}$$

$$= \frac{1}{1+\exp\log\left(\frac{P(Y=0)\,P(X|Y=0)}{P(Y=1)\,P(X|Y=1)}\right)}$$

$$= \frac{1}{1+\exp\log\frac{P(Y=0)}{P(Y=1)} + \exp\log\frac{P(X|Y=0)}{P(X|Y=1)}}$$

$$= \frac{1}{1+\exp\log\left(\frac{1-\gamma}{\gamma}\right) + \Sigma\log\frac{P_{0i}^{x_i}(1-P_{0i})^{1-x_i}}{P_{1i}^{x_i}(1-P_{1i})^{1-x_i}}}$$

$$= \frac{1}{1 + \exp \log\left(\frac{1-\gamma}{\gamma}\right) + \Sigma\left[X_i \log \frac{P_{0i}}{P_{1i}} + (1-X_i)\log\left(\frac{1-P_{0i}}{1-P_{1i}}\right)\right]}$$

$$= \frac{1}{1 + \exp \underbrace{\ln\left(\frac{1-\gamma}{\gamma}\right) + \left(\Sigma \ln \frac{1-P_{0i}}{1-P_{1i}}\right)}_{\text{constant}} + \left[\Sigma X_i \left(\ln \frac{P_{0i}}{P_{1i}} - \ln \frac{1-P_{0i}}{1-P_{1i}}\right)\right]}$$

$$\theta_{d+1} = \ln\left(\frac{1-\gamma}{\gamma}\right) + \Sigma \ln\left(\frac{1-P_{0i}}{1-P_{1i}}\right)$$

$$\theta_i = \left(\ln \frac{P_{0i}}{P_{1i}} - \ln \frac{1-P_{0i}}{1-P_{1i}}\right)$$

$\therefore$ This also follows the same form as logistic regression.

2. $$P(Y=1 | X; \theta) = \frac{1}{1 + \exp(-\theta^T X)}$$

To Prove

$$\frac{\partial \log(P(Y^i | X^i; \theta))}{\partial \theta} = \left(Y^i - P(Y=1 | X^i; \theta)\right) X^i$$

$$\frac{\partial \log P(Y^i|X^i;\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\Sigma\left(\frac{\left(\ln\left(1+e^{-\theta^T x_i}\right)^{-1}\right)Y^i + (1-Y^i)\ln 1-\frac{1}{1+e^{-\theta^T x^i}}}{}\right) -$$

$$= \frac{\partial \Sigma\left(\left(\ln 1 - \ln 1+e^{-\theta^T x^i}\right)(Y^i) + (1-Y^i)\ln \frac{e^{-\theta^T x^i}}{1+e^{-\theta^T x^i}}\right)}{\partial \theta}$$

$$= \frac{\partial \Sigma\left(-Y^i\left(\ln 1+e^{-\theta^T x^i}\right) + (1-Y^i)\left[\ln e^{-\theta^T x^i} + -\ln 1+e^{-\theta^T x^i}\right]\right)}{\partial \theta}$$

$$= \frac{\partial \Sigma\left(-Y^i \cancel{\ln 1+e^{-\theta^T x^i}} + -\theta^T x^i - \ln 1+e^{-\theta^T x^i} - Y^i(-\theta^T x^i) + Y^i \cancel{\ln 1+e^{-\theta^T x^i}}\right)}{\partial \theta}$$

$$= -x^i - \frac{e^{-\theta^T x^i}}{1+e^{-\theta^T x^i}}(-x^i) + Y^i x^i$$

$$= x^i\left[Y^i - \left(1 - \frac{e^{-\theta^T x^i}}{1+e^{-\theta^T x^i}}\right)\right]$$

$$= \left[Y^i - \frac{1}{1+e^{-\theta^T x^i}}\right]x^i = \left[Y^i - P(Y=1|X^i;\theta)\right]X^i$$

Hence Proved.

2.3

1. (a) Number of epochs till termination = 330

   (c) $L(\theta) = 0.508056883686 2912$

2. (a) eta0 = 1, eta1 = 0.001

   Number of epochs for training = 253

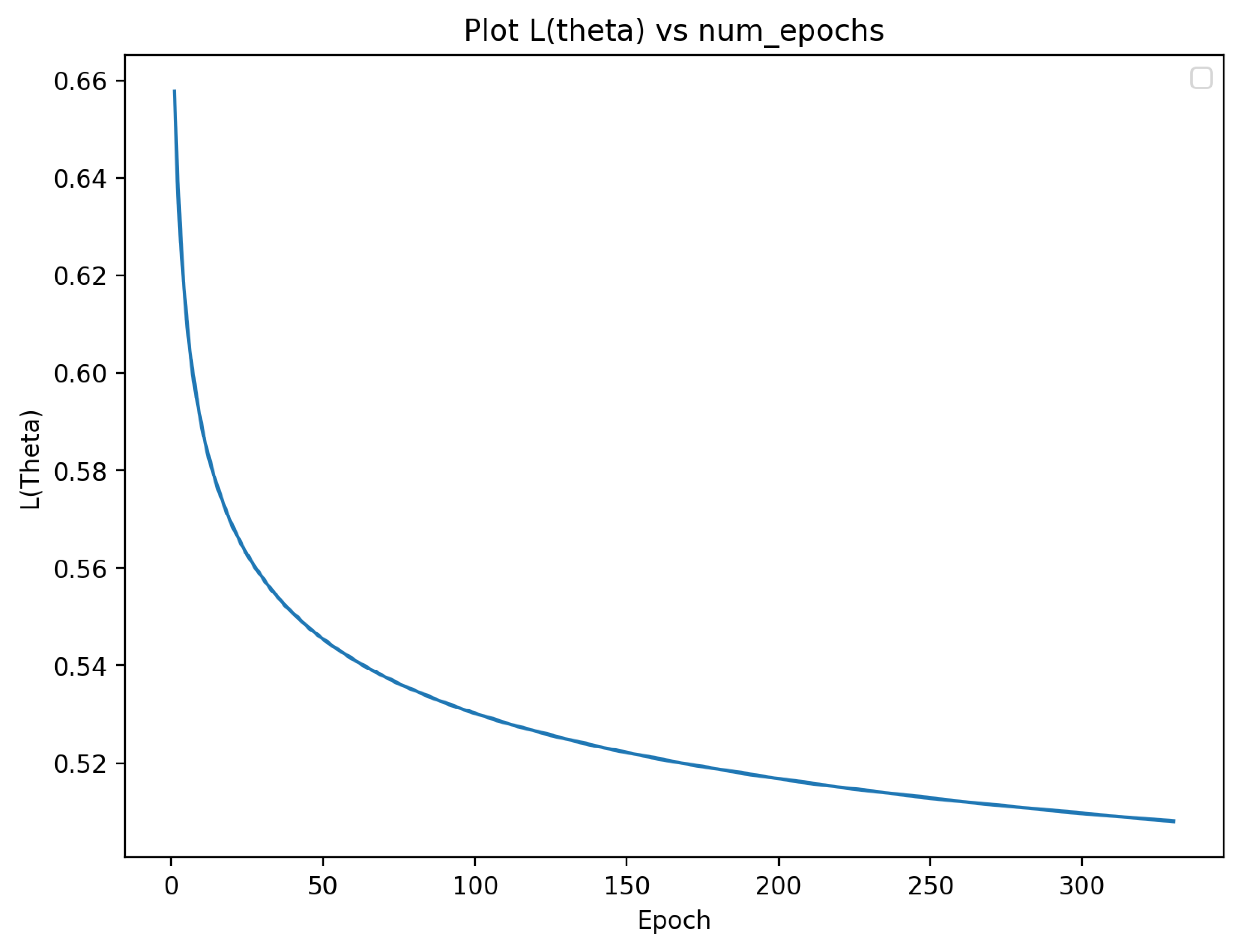
   $L(\theta) = 0.32257825069 56712$

4. (a) Area Under Curve = 0.89

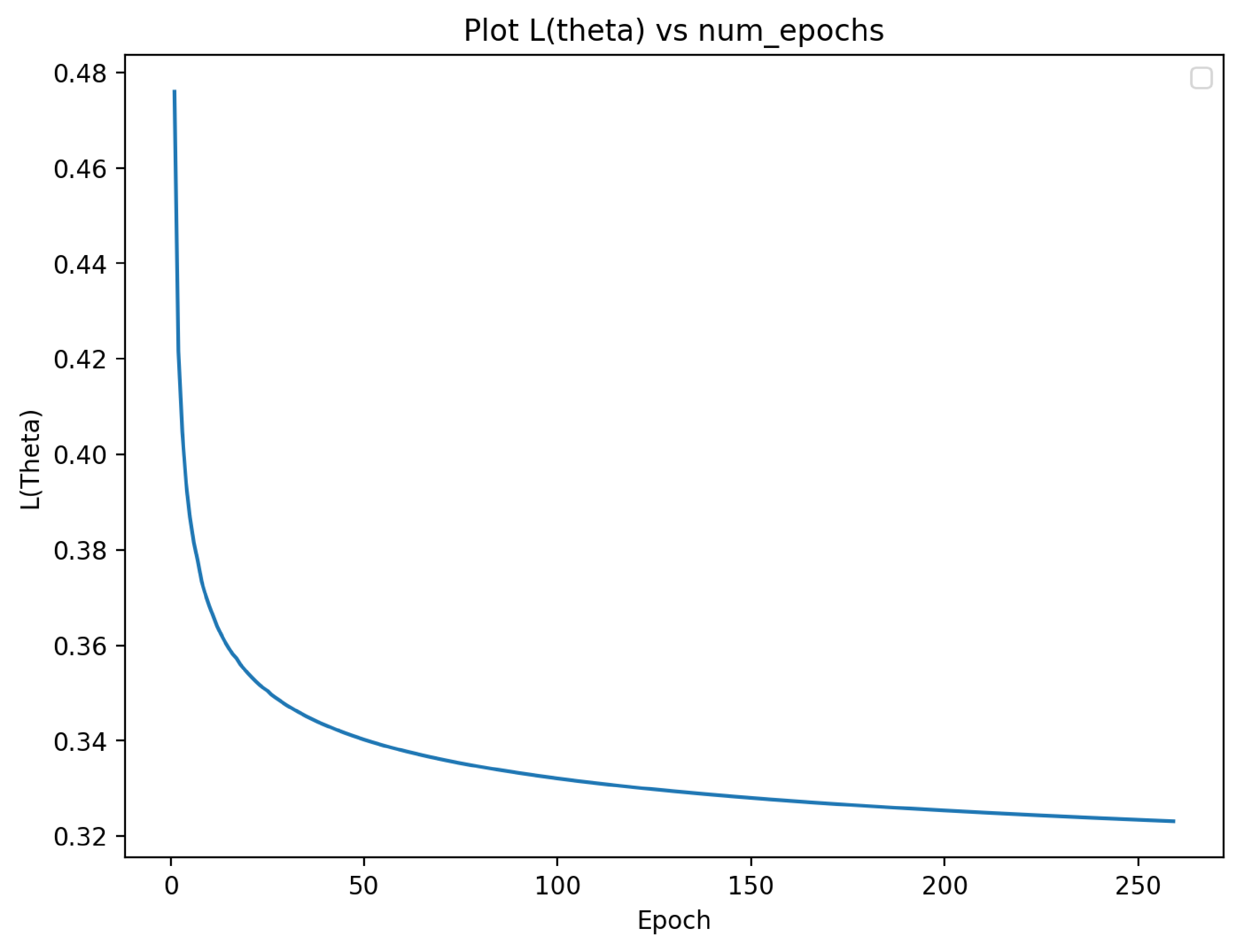
   (b) Average Precision = 0.86

2.4

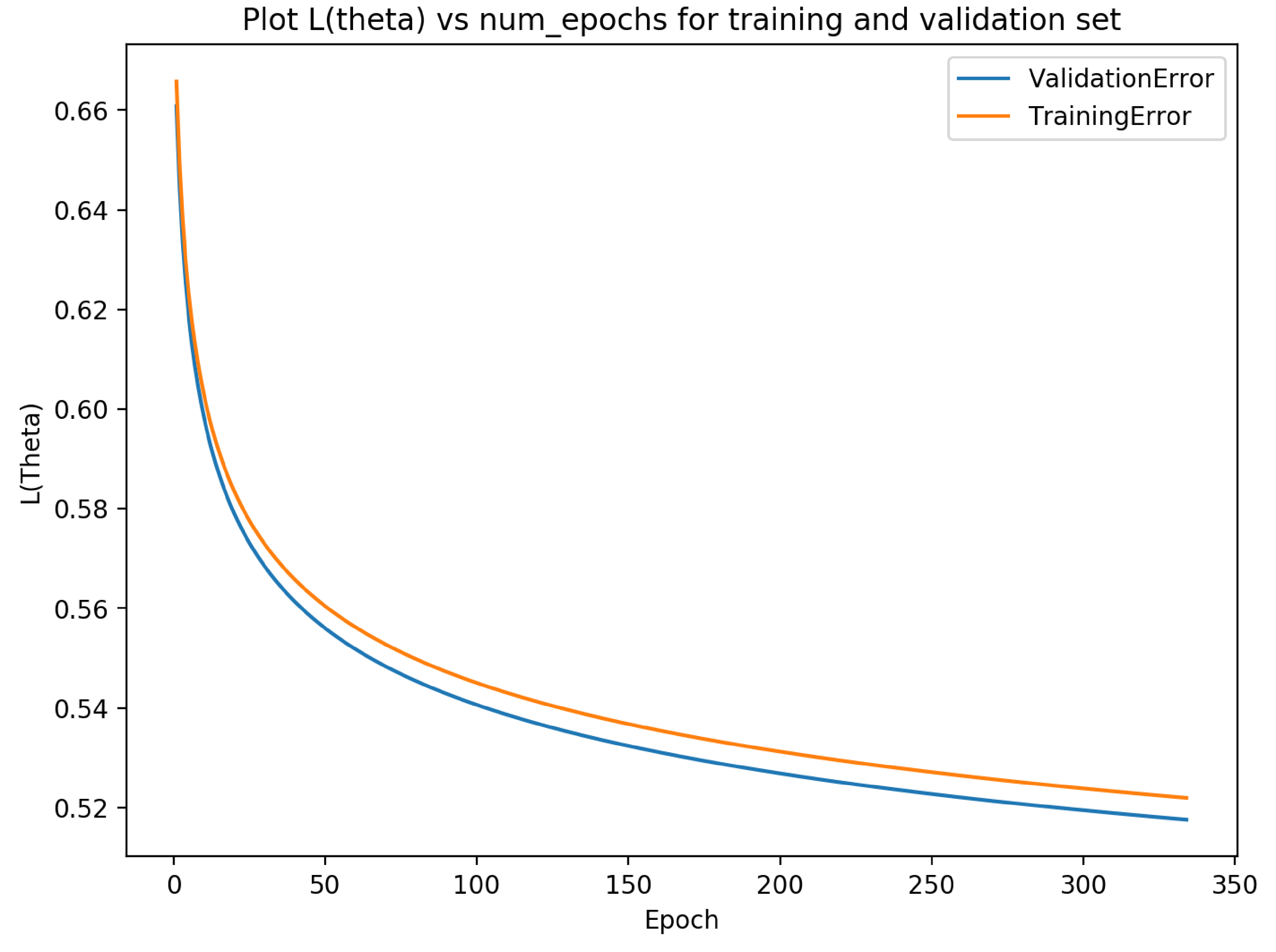
1. Best Accuracy = 89.029%.

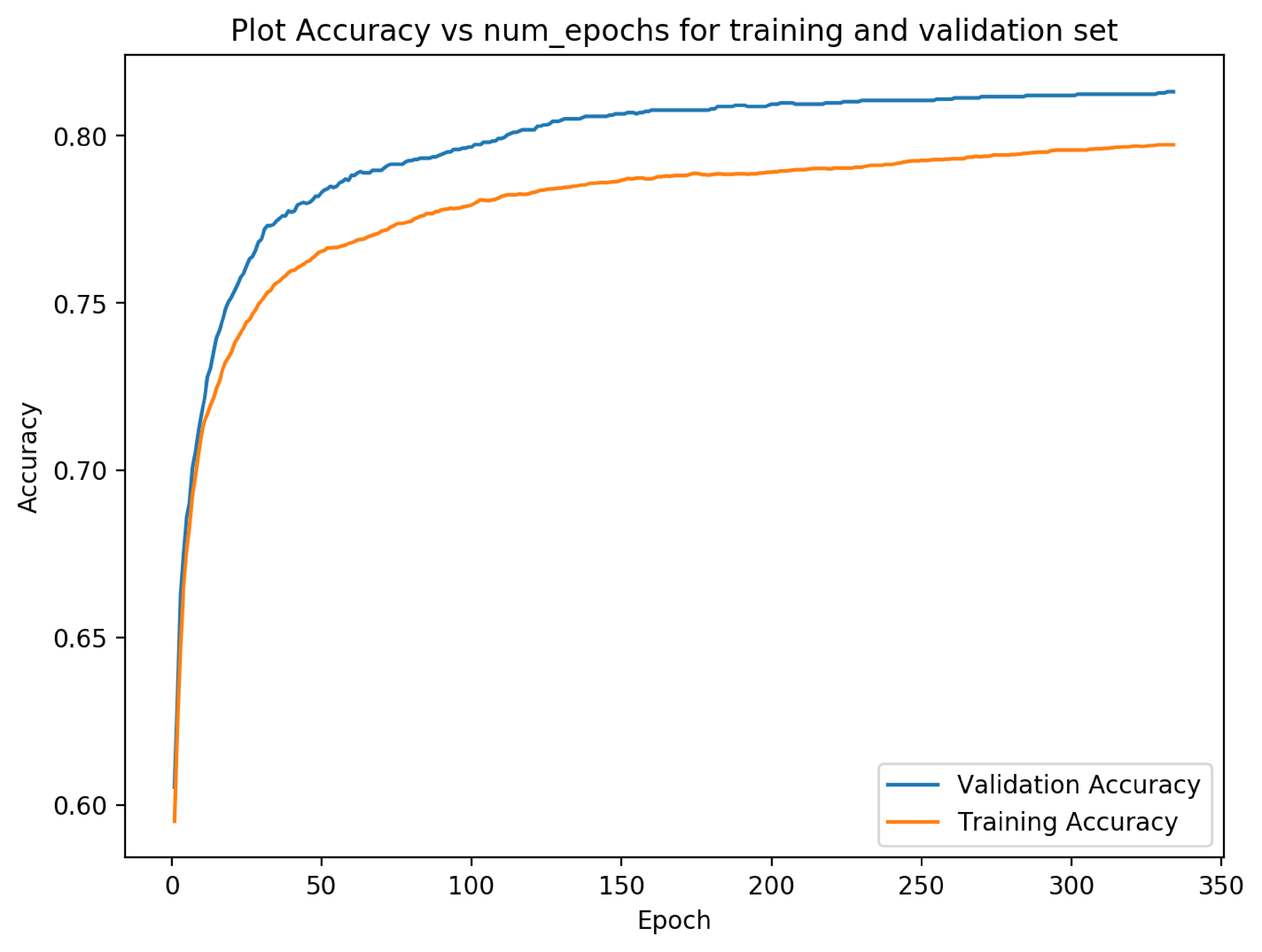


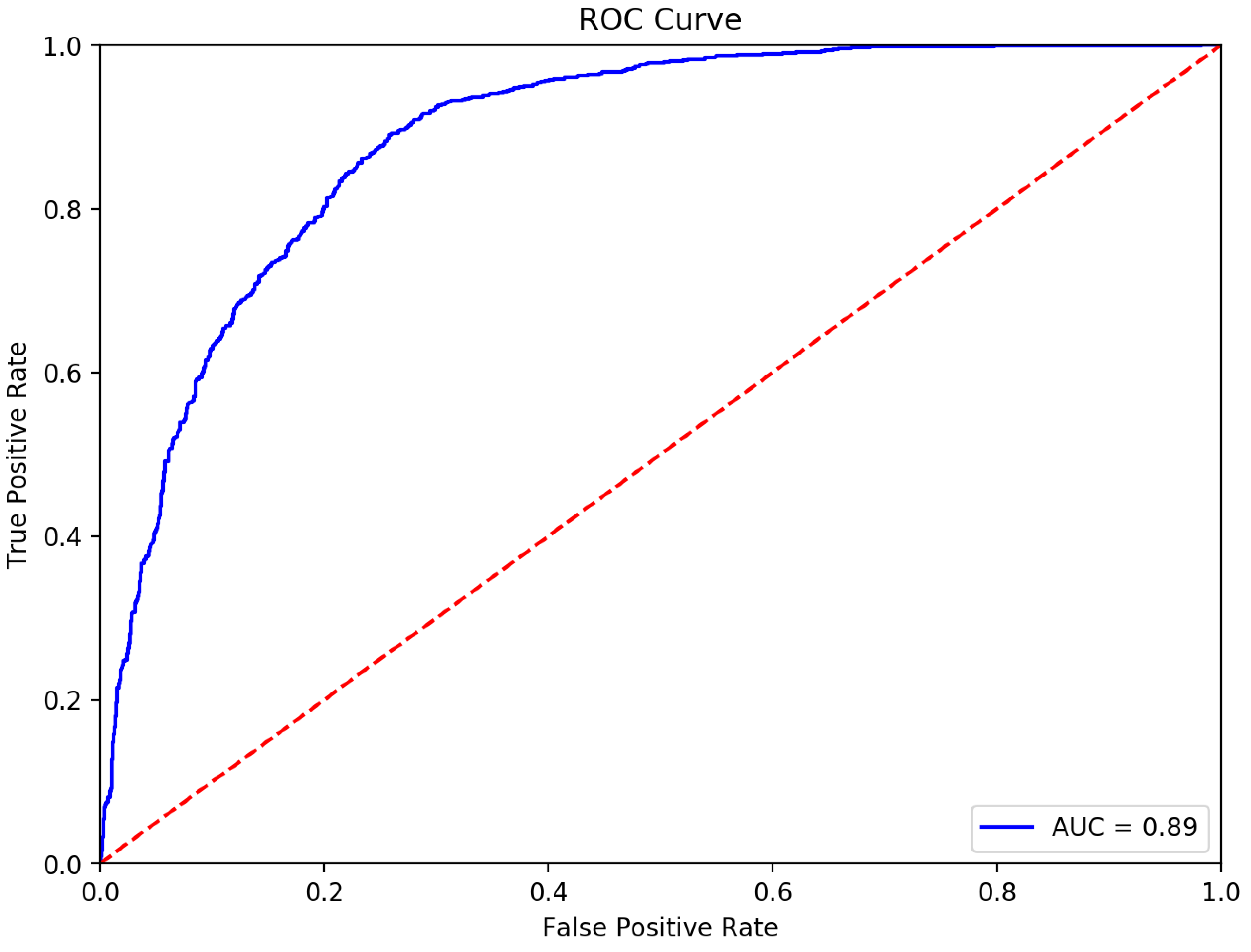
Plot L(theta) vs num_epochs
L(Theta)
Epoch
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
0
50
100
150
200
250
300

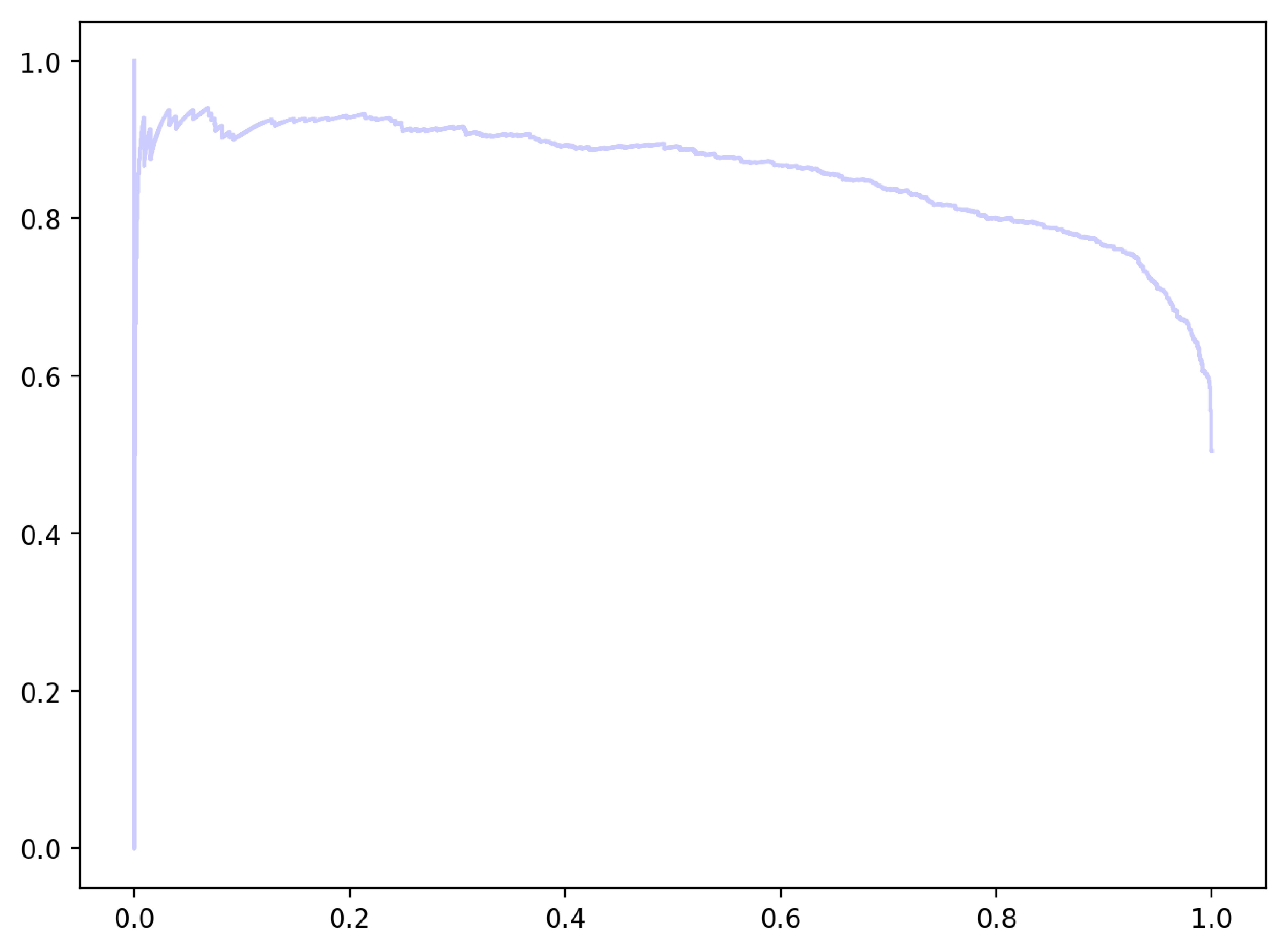


Plot L(theta) vs num_epochs
0.48
0.46
0.44
0.42
0.40
0.38
0.36
0.34
0.32
L(Theta)
0
50
100
150
200
250
Epoch

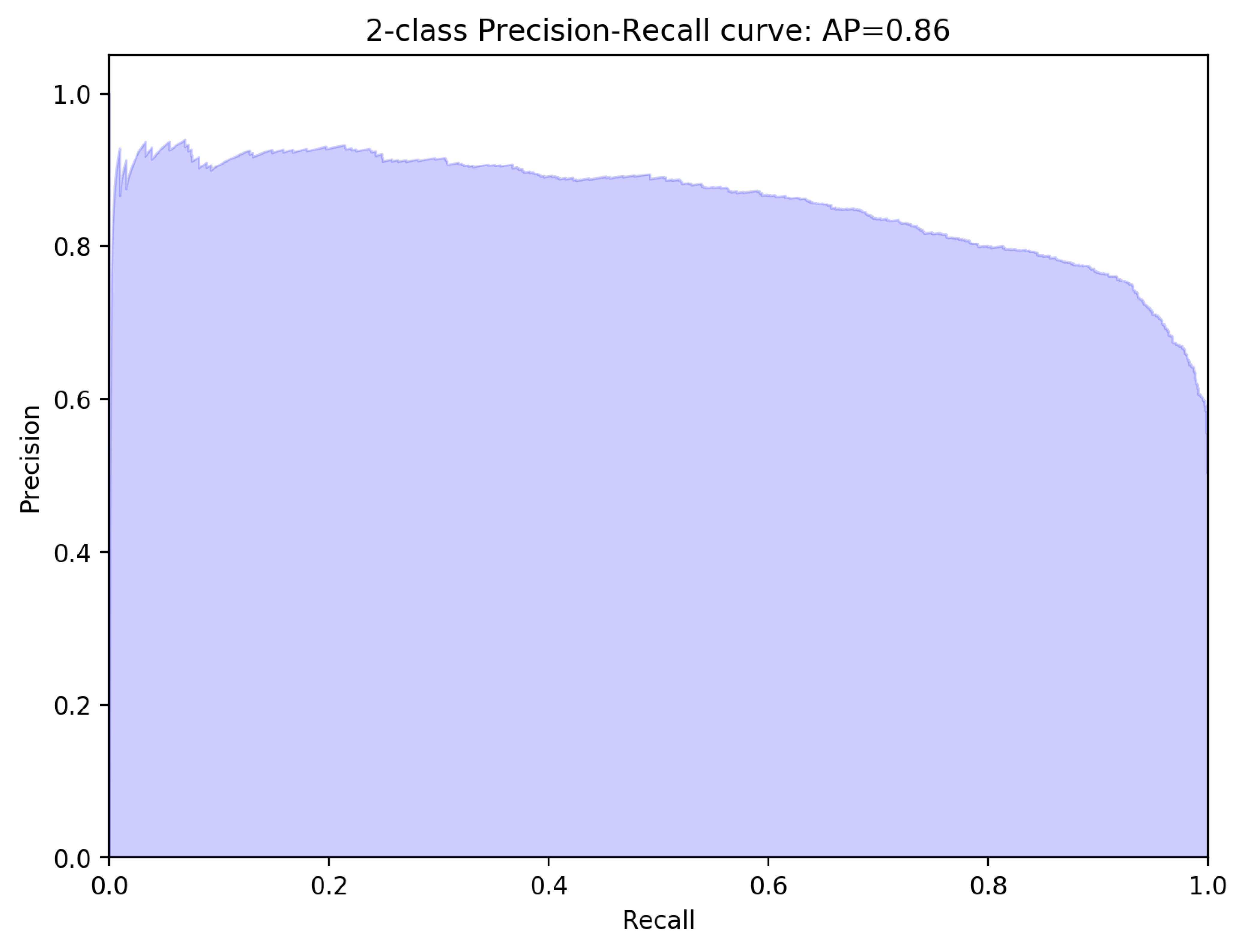

Plot L(theta) vs num_epochs for training and validation set
ValidationError
TrainingError
L(Theta)
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.52
Epoch
0
50
100
150
200
250
300
350


Plot Accuracy vs num_epochs for training and validation set
Accuracy
Epoch
0.60
0.65
0.70
0.75
0.80
0
50
100
150
200
250
300
350
Validation Accuracy
Training Accuracy


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.89



1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


2-class Precision-Recall curve: AP=0.86
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0