Introduction to Machine Learning: Assignment #2

Shadi Albarqouni
shadi.albarqouni@ukbonn.de

University of Bonn — November 7, 2022

1 Logistic Regression

Question 1

_____ is a widely used discriminative classification model $p(y|\boldsymbol{x};\boldsymbol{\theta})$, where $\boldsymbol{x} \in \mathbb{R}^D$ is a fixed-dimensional input vector, $\boldsymbol{y} \in \{0, 1\}$ is the class label, and θ are the parameters.

(a) Conditional Probability

(b) Linear Regression

(c) Multinominal Logistic Regression

(d) Binary Logistic Regression

Question 2

The sigmoid function $\sigma(a) = \frac{1}{1+e^{-a}}$ is typically used in the logistic regression because (Check all that apply)

- ☐ it squeezes the logits a to a value between 0 and 1
- ☐ it is differentiable
- ☐ it is a linear function
- ☐ it has a value of 0.5 for any $a > 0$

Question 3

In logistic regression, the plane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ is often called the _____ seperating the 3d space into two halfs.

(a) decision boundary

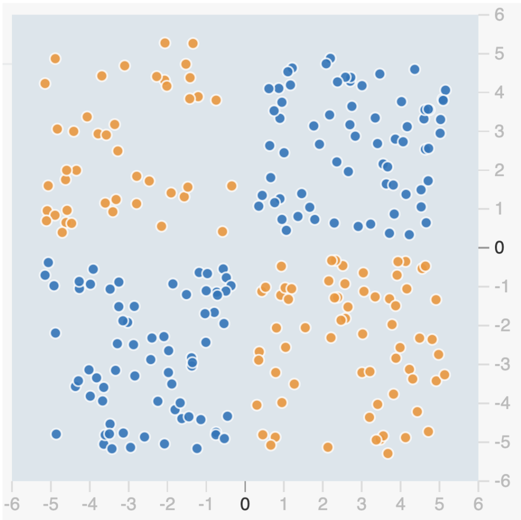
(b) lineraly seperable

(c) perceptron

(d) prediction

Question 4

In logistic regression, we can often make a problem inearly separable by preprocessing the inputs in a suitable way. Let $w = [0; 0; 1]$, which of the following non-linear functions $\phi(x_1, x_2)$ is the most suitable one for the given data points:



(a) $\phi(x_1, x_2) = [1; x_1^2; x_2^2]$

(b) $\phi(x_1, x_2) = [1; x_1 x_2; x_2]$

(c) $\phi(x_1, x_2) = [1; \cos(x_1); \sin(x_2)]$

(d) $\phi(x_1, x_2) = [1; x_1; x_1 x_2]$

Question 5

A non linearly-separable data can always be made linearly-separable in another feature space

(a) True

(b) False

Question 6

To ensure the objective function is convex, we must prove the hessian is negative semi-definite

(a) True

(b) False

Question 7

Which of the following solutions/estimates avoids overfitting:

(a) Maximum Likelihood Estimator (MLE)

(b) Maximum A Posterior (MAP)

(c) Iteratively Reweighted Least Squares (IRLS)

(d) Ordinary Least Squares (OLS)

Question 8

The *Negative Log Likelihood (NLL)* for the multi-label logistic regression $\prod_{n=1}^{N}\prod_{c=1}^{C} \text{Ber}(y_c|\sigma(\boldsymbol{w}_c^T\boldsymbol{x_n}))$ with $\text{Ber}(y|\theta) \triangleq \theta^y(1-\theta)^{1-y}$:

(a) $-\frac{1}{N}\sum_{n=1}^{N} y_n \log\sigma(\boldsymbol{w}_c^T\boldsymbol{x_n}) + (1-y_n)\log(1-\sigma(\boldsymbol{w}_c^T\boldsymbol{x_n}))$

(b) $-\frac{1}{N}\sum_{n=1}^{N} \sigma(\boldsymbol{w}_c^T\boldsymbol{x_n})\log y_n + (1-\sigma(\boldsymbol{w}_c^T\boldsymbol{x_n}))\log(1-y_n)$

(c) $-\frac{1}{N}\sum_{n=1}^{N}\left[\sum_{c=1}^{C} y_{nc}\log\sigma(\boldsymbol{w}_c^T\boldsymbol{x_n}) + (1-y_{nc})\log(1-\sigma(\boldsymbol{w}_c^T\boldsymbol{x_n}))\right]$

(d) $-\frac{1}{N}\sum_{n=1}^{N}\left[\sum_{c=1}^{C} \sigma(\boldsymbol{w}_c^T\boldsymbol{x_n})\log y_{nc} + (1-\sigma(\boldsymbol{w}_c^T\boldsymbol{x_n}))\log(1-y_{nc})\right]$

Question 9

In Multinominal Logistic Regression $p(y|\boldsymbol{x};\boldsymbol{\theta}) = Cat(y|\psi(W\boldsymbol{x}+\boldsymbol{b}))$, we commonly use the following activation function $\psi(\cdot)$:

(a) Softmax

(b) Sigmoid

(c) Heaviside step function

(d) Rectified Linear Unit

Question 10

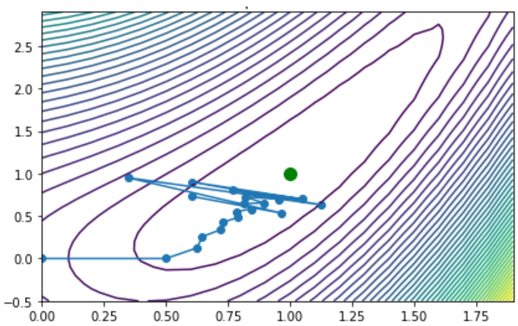
The model on the right hand side suffers from convergence. This could be attributed to (Check all that apply)

☐ low learning rate

☐ high learning rate

☐ low weight decay

☐ high weight decay



Question 11

Consider the following dataset for a binary classification problem with input of $D = 3$ features and binary output $y \in 0, 1$. Then, it is possible to achieve 100% accuracy on this dataset.

(a) True

(b) False

x_1	x_2	x_2	y
3	4	5	1
2	4	3	1
2	3	1	1
2	4	3	0
1	3	5	0

Question 12

The vector $\boldsymbol{w}$ defines the _____ of the decision boundary, and its magnitude, $\|w\|_2 = \sqrt{\sum_{d=1}^{D} w_d^2}$ controls the _____ of the sigmoid, and hence the confidence of the predictions.

(a) steepness, orientation

(b) weights, prediction

(c) orientation, steepness

(d) prediction, weights

Question 13

The vector $\boldsymbol{w}$ defines the _____ of the decision boundary, and its magnitude, $\|w\|_2 = \sqrt{\sum_{d=1}^{D} w_d^2}$ controls the _____ of the sigmoid, and hence the confidence of the predictions.

(a) steepness, orientation

(b) weights, prediction

(c) orientation, steepness

(d) prediction, weights

Question 14 From ML Course – Andrew Ng

Consider training a logistic regression classifier by stochastic gradient descent. You observe that the average cost over the last 100 examples, plotted as a function of the number of iterations, is slowly increasing. Which of the following changes is likely to have the greatest impact?

(a) Attempt to reduce the learning rate by half, and see if the cost drops consistently. If not, reduce the learning rate by half again until it does.

(b) Train with fewer examples.

(c) Consider averaging the cost over a smaller number of examples.

(d) Using stochastic gradient descent, this is not possible, because theta converges to the optimum.

Question 15

Consider a classficiation model with NLL as an objective function. Let $\theta_0 \triangleq (w, b) = (4, 5)$ with a gradient $g_0 = (4, 10)$. What is the suitable learning rate η to reach the optimal parameter $\theta_{opt} = (1, -1)$ given the gradient at the second iteration is $g_1 = (2, 2)$:

(a) 1

(b) 0.5

(c) -1

(d) -0.5