Deep Learning Interview

Kindergarten

Logistic Regression

Key Concepts

- odds, odds ratio, and probability

$$odds(p) = (\frac{p}{1-p})$$

$$odds\ ratio = \frac{\frac{X_A}{X_B}}{\frac{Y_A}{Y_B}}$$

$$relative\_risk = \frac{\frac{X_A}{X_A+Y_A}}{\frac{X_B}{X_B+Y_B}}$$

where X is treated, Y is control, A is impacted, B is not impacted

$$probability = \frac{odds}{1+odds} = \frac{4}{1+4} = 0.8$$

- Distribution of logistic regression predictor and outcome variables

$$Z = logit(P) = log(odds) = log(\frac{P}{1-P}) = \theta^T x = \theta_0 + \theta_1$$

$$e^Z = \frac{P}{1-P}$$

$$P = \frac{e^Z}{1+e^Z} = \frac{1}{1+e^{-Z}}$$

- Sigmoid function (logistic function for binary classification and a neuron activation function)

$$\sigma(x) = \frac{1}{1+e^{-\theta x}}$$

Derivative of sigmoid funtion (we can expand this to softmax)

$$\frac{d}{dx}\sigma(x) = \frac{e^{-x}}{(1+e^{-x})^2}$$

or

$$\sigma'(x) = \sigma(x)(1-\sigma(x))$$

- Logistic Regression Definition (put the above concept together) Hypothesis function $h_\theta(x)$ Logit: $Z = \theta^T x$

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

Decision Boundry:

$$h_\theta(x) \geq 0.5 \rightarrow y = 1$$
$$h_\theta(x) < 0.5 \rightarrow y = 0$$

or

$$\theta^T \geq 0 \rightarrow y = 1$$
$$\theta^T < 0 \rightarrow y = 0$$

Cost Function (Measure the goodness of our hypothesis with respect to all data samples)

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} Cost(h_\theta(x^{(i)}), y^(i))$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (-y^i log(h_\theta(x^i)) - (1 - y^i) log(1 - h_\theta(x^i)))$$

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} (y^i log(h_\theta(x^i)) + (1 - y^i) log(1 - h_\theta(x^i)))$$

Questions and Answers

prb-4

True or False: For a fixed number of observations in a data set, introducing more variables normally generates a model that has a better fit to the data. What may be the drawback of such a model fitting strategy?

AN: True. Overfitting

prb-5

Define the term "odds of success" both qualitatively and formally. Give a numerical example that stresses the relation between probability and odds of an event occurring.

AN: Odds of success = probability of success / probability of failure whereas probability is freq of success / total Using the event of rolling a dice: odds of rolling a 6 is 1/5, prob of rolling a 6 is 1/6

$$Odds(p) = (\frac{p}{1 - p})$$

prb-6

Define what is meant by the term "interaction", in the context of a logistic regression predictor variable.

AN: 1. An interaction is the product of two single predictor variables implying a non-additive effect.

2. What is the simplest form of an interaction? Write its formula? AN: The simplest interaction model includes a predictor variable formed by multiplying two ordinary predictors. Let us assume two variables X and Z. Then, the logistic regression model that employs the simplest form of interaction follows:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

β_3 is the coefficient of the interaction term

3. What statistical tests can be used to attest the significance of an interaction term?

For testing the contribution of an interaction, two principal methods are commonly employed; the Wald chi-squared test or a likelihood ratio test between the model with and without the interaction term.

prb-7

True or False: In machine learning terminology, unsupervised learning refers to the mapping of input covariates to a target response variable that is attempted at being predicted when the labels are known

AN: false, labels are unknown for unsupervised learning, the description describe the supervised learning model

prob-8

Complete the following sentence: In the case of logistic regression, the response variable is the log of the odds of being classified in [. . .]

AN: in a group of binary or multi-class responses

prb-9

Describe how in a logistic regression model, a transformation to the response variable is applied to yield a probability distribution. Why is it considered a more informative representation of the response?

AN: When a transformation to the response variable is applied, it yields a probability distribution over the output classes, which is bounded between 0 and 1; this transformation can be employed in several ways, e.g., a softmax layer, the sigmoid function or classic normalization. This representation facilitates a soft-decision by the logistic regression model, which permits construction of probability-based processes over the predictions of the model.

- Softmax is for multi-classs problem
- Sigmoid is for binary class problem

prb-10

Minimizing the negative log likelihood also means maximizing the [. . .] of selecting the [. . .] class

AN: probability/likelihood of selecing the correct class

prb-11

Assume the probability of an event occurring is p = 0.1.

1. What are the odds of the event occurring?.

$$odds = \frac{p}{1-p} = \frac{0.1}{0.9} = 0.111$$

2. What are the log-odds of the event occurring?.

$$log\ odds = \ln(\frac{p}{1-p}) = \ln(0.1/0.9) = -2.197$$

3. Construct the probability of the event as a ratio that equals 0.1.

$$odds = \frac{p}{1-p}$$

$$p = \frac{odds}{1+odds} = \frac{0.11}{1+0.11} = 0.1$$

prb-12

True or False: If the odds of success in a binary response is 4, the corresponding probability of success is 0.8.

An: True

$$p = \frac{odds}{1 + odds} = \frac{4}{1 + 4} = 0.8$$

prb-13

Draw a graph of odds to probabilities, mapping the entire range of probabilities to their respective odds.

$$odds = \frac{P}{1 - P}$$

If we plot $y(x) = \frac{x}{1-x}$

- Assume x axis is prob, y axis is odds
- Wnen probability is 0, odds is 0
- When probability is 1. odds is infinity
- It is concav up and asymptote to infinity when prob at 1

prob-14

The logistic regression model is a subset of a broader range of machine learning models known as generalized linear models (GLMs), which also include analysis of variance (ANOVA), vanilla linear regression, etc. There are three components to a GLM; identify these three components for binary logistic regression.

3 components of GLM: * Random component: refers to the probability distribution of the response variable (Y) * Systematic component: describes the explanatory variables * Link function: specifies the link between random and systematic components

For binary logistic regression: * The Random cmoponent is binormial distribution * Systematic component is $\sum \theta_i x_i$ * Link function: how the expected value of the response relates to the linear predictor of explanatory variables.

prb-15

Let us consider the logit transformation, i.e., log-odds. Assume a scenario in which the logit forms the linear decision boundary:

$$\log(\frac{Pr(Y = 1|X)}{Pr(Y = 0|X)}) = \theta_0 + \theta^T X$$

for a given vector of systematic components X and predictor variables θ. Write the mathematical expression for the hyperplane that describes the decision boundary

AN: Decision boundry of logistic regression gives:

$$h(\theta) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} = 0.5$$

Therefore

$$e^{-\theta^T x} = 1$$

or

$$\theta_0 + \theta^T X = 0$$

prb-16

True or False: The logit function and the natural logistic (sigmoid) function are inverses of each other.

AN: True

Proof:

We know logit function is defined as:

$$Z = logit(P) = log(\frac{P}{1-P})$$

taking the inverse of logit:

$$e^Z = \frac{P}{1-P}$$

we get

$$P = \frac{e^Z}{1+e^Z}$$

or

$$P = \frac{1}{1+e^{-Z}}$$

Which is the sigmoid function

prb-17

Compute the derivative of the natural sigmoid function: $\sigma(x) = \frac{1}{1+e^{-x}} \in (0,1)$

AN:

$$\frac{d}{dx}\sigma(x) = \frac{e^{-x}}{(1+e^{-x})^2} \in (\frac{1}{4}, \frac{e}{(1+e)^2})$$

The weakness of signmoid function as activiation function is the when the input arguement is very large or very small the signmoid function is very flat, its derivative becomes very small therefore the training becomes very slow

prb-18

Remember that in logistic regression, the hypothesis function for some parameter vector β and measurement vector x is defined as:

$$h_\beta(x) = g(\beta^T x) = \frac{1}{1+e^{-\beta^T x}} = P(y=1|x;\beta)$$

where y holds the hypothesis value. Suppose the coefficients of a logistic regression model with independent variables are as follows: $\beta_0 = -1.5, \beta_1 = 3, \beta_2 = -0.5$, As a result, the logit equestion becomes:

AN:

$$logit = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

1. What is the value of the logit for this observation?

$$logit = -1.5 + 3 * 1 - 0.5 * 5 = -1$$

2. What is the value of the odds for this observation?

$$odds = e^{logit} = e^{-1} = \frac{1}{e}$$

3. What is the value of $P(y = 1)$ for this observation?

$$p = \frac{e^{-1}}{1 + e^{-1}} = \frac{1}{1 + e}$$

prb-19

Proton therapy questions

Tumour eradication table:

Cancer Type	Yes	No
Breast	560	260
lung	69	36

1. What is the explanatory variable and what is the response variable?

explanatory variable (X): cancer type response variable (Y): tumour eradication

2. Explain the use of relative risk and odds ratio for measuring association.

$$odds\_ratio = \frac{\frac{Y_1}{Y_2}}{\frac{N_1}{N_2}}$$

$$relative\_risk = \frac{\frac{Y_1}{Y_1+N_1}}{\frac{Y_2}{Y_2+N_2}}$$

Relative risk (RR) is the ratio of risk of an event in one group (e.g., exposed group) versus the risk of the event in the other group (e.g., non-exposed group). The odds ratio (OR) is the ratio of odds of an event in one group versus the odds of the event in the other group.

$$odds\_ratio = \frac{\frac{560}{69}}{\frac{260}{36}} = 1.123$$

$$relative\_risk = \frac{\frac{560}{820}}{\frac{69}{105}} = 1.039$$

3. Are the two variables positively or negatively associated? Find the direction and strength of the association using both relative risk and odds ratio.

Yes, they are positively correlated The odds ratio is larger than one, indicating that the odds for a breast cancer is more than the odds for a lung cancer to be eradicated.

4. Compute a 95% confidence interval (CI) for the measure of association. To test association, we will use chi-square test

Cancer Type	Yes	No	total
Breast	560	260	820
lung	69	36	105
total	629	296	925

Breast OR: (560/820)/(1-(560/820)) = 2.15 lung OR: (69/105)/(1-(69/105)) = 1.91 OR = 2.15/1.91 = 1.13

95% CI of odd ratio

$$\log(OR) \pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}})$$

$$95\%\ CI\ log\ odds = \log(1.123) \pm 1.96\sqrt{\frac{1}{560}+\frac{1}{260}+\frac{1}{69}+\frac{1}{36}})$$

$$95\%\ CI\ odds = e^{log(1.123)\pm 1.96\sqrt{\frac{1}{560}+\frac{1}{260}+\frac{1}{69}+\frac{1}{36}})}$$

$$CI = (0.810, 1.909)$$

5. Interpret the results and explain their significance. The CI (0.810, 1.909) contains 1, which indicates that the true odds ratio is not significantly different from 1 and there is not enough evidence that tumour eradication is dependent on cancer type.

prb-20

1. Estimate the probability that, given a patient who undergoes the treatment for 40 milliseconds and who is presented with a tumour sized 3.5 centimetres, the system eradicates the tumor.

$e^{(-6+0.05*40+1*3.5)} = 0.61$ $p = \frac{0.61}{1+0.61} = 0.38$

2. How many milliseconds the patient in part (a) would need to be radiated with to have exactly a 50% chance of eradicating the tumor?

50 millisecond

prb-21

1. Using X1 and X2, express the odds of a patient having a migraine for a second time.

$$P = \frac{1}{1+e^{-\beta^T x}}$$

where

$$\beta^T x = -6.36 - 1.02x_1 + 0.12x_2$$

$$odds = e^{-6.36-1.02x_1+0.12x_2}$$

2. Calculate the probability of a second migraine for a patient that has at least four amalgams and drank 100 cups per month?

   We plug in 1 for x_1 and 100 for x_2, we get p = 0.99 or 99%

3. For users that have at least four amalgams, is high coffee intake associated with an increased probability of a second migraine?

   Yes, the coefficient for X2 (0.119) is a positive number and P-value is 0.0304< 0.05

4. Is there statistical evidence that having more than four amalgams is directly associated with a reduction in the probability of a second migraine?

   No, since the P-value for the coefficient is 0.3818 > 0.05 and is not statistically significant

prb-22

1. Estimate the probability of improvement when the count of gum bacteria of a patient is 33.

$$P = \frac{1}{1 + e^{-\beta^T x}}$$

where

$$\beta^T x = -4.88 + 0.0258x$$

$$p = \frac{1}{1 + e^{-(-4.88+0.0258*33)}} = 0.017$$

2. Find out the gum bacteria count at which the estimated probability of improvement is 0.5.

$$P = \frac{1}{1 + e^{-4.88+0.0258x}} = 0.5$$
$$e^{-(-4.88+0.0258x)} = 1$$
$$-4.88 + 0.0258x = 0$$
$$x = 189$$

The bacteria count is 189

3. Find out the estimated odds ratio of improvement for an increase of 1 in the total gum bacteria count.

$$odds\ ratio = odds_{(x+1)}/odds_{(x)}$$
$$log(odds\ ratio) = log(odds_{(x+1)}) - log(odds_{(x)})$$
$$log(odds\ ratio) = -4.8792 + 0.0258(x+1) - (-4.8792 + 0.0258x) = 0.0258$$
$$odds\ ratio = e^{0.0258} = 1.0261$$

4. Obtain a 99% confidence interval for the true odds ratio of improvement increase of 1 in the total gum bacteria count. Remember that the most common confidence levels are 90%, 95%, 99%, and 99.9%. Table 9.1 lists the z values for these levels.

$$99\%\ CI = 0.0258 \pm 2.576 \times 0.0194$$
$$99\%\ True\ CI = e^{0.0258 \pm 2.576 \times 0.0194}$$

prb-23

1. Find the sample odd ratio

$$odd\ ratio = \frac{\frac{60}{130}}{\frac{6833}{6778}} = 0.458$$

2. Find the sample log-odd ratio

$$log\ odds\ ratio = log(0.458)$$

3. Compute a 95% confidence interval (z0.95 = 1.645; z0.975 = 1.96) for the true log odds ratio and true odds ratio.

99% CI of odd ratio = $e^{(-0.783 \pm 1.96\sqrt{\frac{1}{60}+\frac{1}{130}+\frac{1}{6833}+\frac{1}{6778}})}$

prb-24

Entropy loss of a single binary outcome with probability p

$$H(p) = -plog(p) - (1-p)log(1-p)$$

1. At what p does H(p) attain its maximum value? when p = 0.5, H(p) = 0.693
2. What is the relationship between the entropy H(p) and the logit function, given p?

$$\frac{dH(p)}{dp} = -logit(p)$$

prb-25

```
#include <iostream>
#include <cmath>
#include <vector>
#include <numeric>

std::vector<double> theta {-6,0.05,1.0};
double sigmoid(double x) {
  double tmp =1.0 / (1.0 + exp(-x));
  std::cout << "prob=" << tmp<<std::endl;
  return tmp;
}

double hypothesis(std::vector<double> x){
  double z;
  z=std::inner_product(std::begin(x), std::end(x), std::begin(theta), 0.0);
  std::cout << "inner_product=" << z <<std::endl;
  return sigmoid(z);
}

int classify(std::vector<double> x){
  int hypo=hypothesis(x) > 0.5f;
  std::cout << "hypo=" << hypo<<std::endl;
  return hypo;
}
int main() {
  std::vector<double> x1 {1,40,3.5};
  classify(x1);
}
```

1. Explain the purpose of line 10, i.e., inner_product

Calculate the logit function, or probability of binary class

2. Explain the purpose of line 15, i.e., hypo(x) > 0.5f

make binary classification of prob > 0.5 return true otherwise return false

3. What does θ stand for in line 2? coefficient of logistic regression
4. Compile and run the code, you can use: https://repl.it/languages/cpp11 to evaluate the code. What is the output?

inner_product=-0.5 prob=0.377541 hypo=0

prob-26

```
import torch
import torch.nn as nn

lin = nn.Linear(5, 7)
data = (torch.randn(3, 5))
print(lin(data).shape)
```

shape is (3, 7)

prob-27

```
from scipy.special import expit
import numpy as np
import math

def Func001(x):
    e_x = np.exp(x - np.max(x))
    return e_x / e_x.sum()
def Func002(x):
    return 1 / (1 + math.exp(-x))

def Func003(x):
    return x * (1-x)
```

Func001 is a softmax function Func002 is a sigmoid function Func003 is the derivative of a sigmoid function

Note:

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

prob-28

```
from scipy.special import expit
import numpy as np
import math

def Func006(y_hat, y):
    if y == 1:
        return -np.log(y_hat)
    else:
        return -np.log(1 - y_hat)
```

What important concept in machinelearning does it implement?

AN: It implement the binary cross-entropy function (negative log-loss)

prob-29

```
from scipy.special import expit
import numpy as np
import math

def Ver001(x):
    return 1 / (1 + math.exp(-x))
```

```python
def Ver002(x):
    return 1 / (1 + (np.exp(-x)))

WHO_AM_I = 709
def Ver003(x):
    return 1 / (1 + np.exp(-(np.clip(x, -WHO_AM_I, None))))
```

1. Which mathematical function do these methods implement? The sigmoid objective function of logistic regression (probability)
2. What is significant about the number 709 in line 11? exceeding pyhton floating point number boundry
3. Given a choice, which method would you use? Ver003 is the best to ensure numerical stability