CSCI-6522, Fall 2023
Advanced Machine Learning II
Study Guide for Test#1

(Please don't distribute this study guide.
The guide is for your study purpose only)

1. What are the properties of a (univariate) Gaussian / Normal distribution *N* $(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$ **?**

Ans: Here, variable $x \in (-\infty, \infty)$

Mean $\mu \in (-\infty, \infty)$ and

Variance $\sigma^2 > 0$

Mean = Median = Mode

And, $\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx = 1$

The plotted equation generates a bell-shaped and symmetric curve.

2. Using maximum likelihood estimation, for a given data D{ $x_1, x_2, \ldots x_N$ **} where iid is followed and pdf is Gaussian [*P*(*x*) =** $\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$ **], estimate the best mean (** μ **).(see Chapter 01 (ANN), page 5-7).**

3. Using maximum likelihood estimation, for a given data D{ $x_1, x_2, \ldots x_N$ **} where iid is followed and pdf is Gaussian [*P*(*x*) =** $\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$ **], justify the least-square approach regression method (see Chapter 01 (ANN), page 7-8).**

4. Explain Bernoulli distribution?
Ans: (see Chapter #1 (ANN), page # 8-9)

5. Assume for a binary class classification modeling, the posterior class probabilities are given by:

$$\textbf{Pr}(G = 1|X = x) = \frac{\exp(\beta_0 + \beta^T x)}{1+\exp(\beta_0 + \beta^T x)},$$

$$\textbf{Pr}(G = 2|X = x) = \frac{1}{1+\exp(\beta_0 + \beta^T x)}.$$

How will you derive their corresponding *logit* transformation, explain?

Ans: The *logit* or *log-odds* of a probability p is given by, $logit(p) = \log\left(\frac{p}{1-p}\right)$. Therefore, given $p = \Pr(G = 1|X = x) = \frac{\exp(\beta_0 + \beta^T x)}{1+\exp(\beta_0 + \beta^T x)}$, we can have,

(1-p) = $\left(1-\frac{\exp(\beta_0 + \beta^T x)}{1+\exp(\beta_0 + \beta^T x)}\right)=\frac{1}{1+\exp(\beta_0 + \beta^T x)}$, which is basically, Pr ($G = 2|X = x$).

Thus, we can compute the log $[p/(1-p)]$ as follows:

$$\log\frac{\Pr(G=1 \mid X=x)}{\Pr(G=2 \mid X=x)} = \beta_0 + \beta^T x$$

6. Write down the steps of the IRLS Logistic regression (classification) algorithm.

Ans:

1. Initialize β with zero(s).
2. Load **X**
3. Load **y** matrix {y = 0 if Benign else y = 1 for Malignant class}
4. Compute $\boldsymbol{\eta}$ as: $\eta_i = p(x_i;\beta) = \frac{e^{X_i^T\beta}}{1+e^{X_i^T\beta}}$, where, i =1 to N.
5. Compute $\mathbf{W}_{N\times N}$ (the diagonal matrix) and the i^{th} diagonal element $p(x_i;\beta)(1-p(x_i;\beta)) = (\frac{e^{X_i^T\beta}}{1+e^{X_i^T\beta}})(\frac{1}{1+e^{X_i^T\beta}})$, where, i =1 to N.
6. Compute, **z** as: $\mathbf{z} = \left(\mathbf{X}\beta(t) + \mathbf{W}^{-1}(\mathbf{y}-\boldsymbol{\eta})\right)$
7. Compute next β as: $\beta(t+1) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{W}\mathbf{z}$
8. Check exit condition, else goto step 4.

7. How do we get the term to be maximized with regularization, that is, $\max_{\beta_o,\beta}\left\{\sum_{i=1}^{N}\left[y_i(\beta_0+\beta^T x_i)-\log(1+e^{\beta_0+\beta^T x_i})\right]-\lambda\sum_{j=1}^{p}|\beta_j|\right\}$, for logistic regression when the derived equation, $l(\beta)=\log L(\beta)=\sum_{i=1}^{N}\{y_i\log\eta_i+(1-y_i)\log(1-\eta_i)\}$ is given and $\eta=\frac{1}{1+e^{-x^T\beta}}=\frac{e^{x^T\beta}}{1+e^{x^T\beta}}$ and $(1-\eta)=\frac{1}{1+e^{x^T\beta}}$ are available?

Ans:

The derived equation $l(\beta)=\log L(\beta)=\sum_{i=1}^{N}\{y_i\log\eta_i+(1-y_i)\log(1-\eta_i)\}$ …… (A)

is given where, $\eta=\frac{1}{1+e^{-x^T\beta}}=\frac{e^{x^T\beta}}{1+e^{x^T\beta}}$ and $(1-\eta)=\frac{1}{1+e^{x^T\beta}}$.

Continuing from (A), we can write:

$$\sum_{i=1}^{N}\{y_i\log\eta_i+(1-y_i)\log(1-\eta_i)\}$$

$$=\sum_{i=1}^{N}\left\{y_i\log(\frac{e^{x^T\beta}}{1+e^{x^T\beta}})+(1-y_i)\log(\frac{1}{1+e^{x^T\beta}})\right\}$$

$$=\sum_{i=1}^{N}\left\{y_i[\log e^{x^T\beta}-\log(1+e^{x^T\beta})]+(1-y_i)[\log(1)-\log(1+e^{x^T\beta})]\right\}$$

$$=\sum_{i=1}^{N}\left\{y_i[x^T\beta-\log(1+e^{x^T\beta})]+(1-y_i)[\log(1)-\log(1+e^{x^T\beta})]\right\}$$

$$=\sum_{i=1}^{N}\left\{y_i\,x^T\beta-y_i\log(1+e^{x^T\beta})-\log(1+e^{x^T\beta})+y_i\log(1+e^{x^T\beta})\right\}$$

$$=\sum_{i=1}^{N}\left\{y_i\,x^T\beta-\log(1+e^{x^T\beta})\right\}$$

$$=\sum_{i=1}^{N}\left[y_i(\beta_0+\beta^T x_i)-\log(1+e^{\beta_0+\beta^T x_i})\right] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(B)}$$

$$where,\ \beta_0+\beta^T x_i=x^T\beta$$

Since, the problem term falls under maximization problem we subtract regularization part "$\lambda\sum_{j=1}^{p}|\beta_j|$" from (B). Therefore, finally we can write the target with regularization is $\max_{\beta_o,\beta}\left\{\sum_{i=1}^{N}\left[y_i(\beta_0+\beta^T x_i)-\log(1+e^{\beta_0+\beta^T x_i})\right]-\lambda\sum_{j=1}^{p}|\beta_j|\right\}$.

8. How can we implement multiclass classification using logistic regression – explain (see Chapter 01 (ANN), page 14-15).

9. Draw the characteristic curve of (i) sigmoid function and (ii) hyperbolic tangent function (see class note or, find it by yourself)

10. For the given Artificial Neural Network below, write the vector-equations involved in forward propagation:

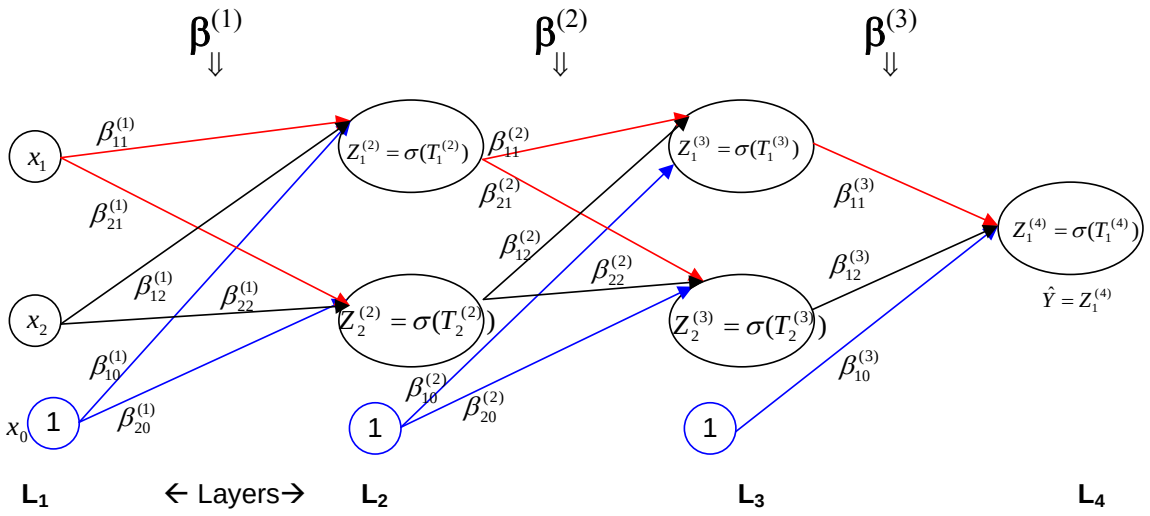


Figure: A multilayer neural network demonstrating the notations.

Assume that the transformation function σ = f_{sig} and bias unit (input "1") in each layer is presented as $Z_0^{(l)}$, where *l* is the layer number.

Ans:

$$Z^{(1)} = X$$

$$T^{(2)} = \beta^{(1)^T} Z^{(1)}$$

$$Z^{(2)} = f_{sig} \bullet (T^{(2)}) \quad \text{and add } Z_0^{(2)}$$

$$T^{(3)} = \beta^{(2)^T} Z^{(2)}$$

$$Z^{(3)} = f_{sig} \bullet \left(T^{(3)}\right) \text{ and add } Z_0^{(3)}$$

$$T^{(4)} = \beta^{(3)^T} Z^{(3)}$$

$$Z^{(4)} = \hat{Y} = f_{sig} \bullet (T^{(4)})$$

[Note: Don't forget the '.'s (dots) in the above equations, they indicate element-wise operations**]**

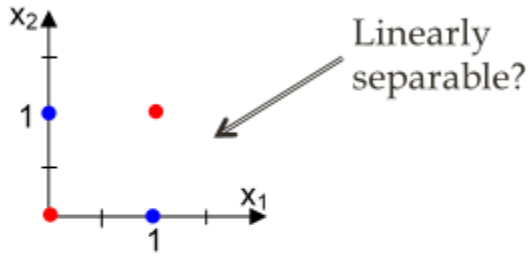
11. Show why logic function XOR is not linearly separable. Design an ANN for a two-input XOR function (See chapter # 1 (ANN), pages 18-19 and class notes).

Ans:

Hidden Layer, When? ...

XOR

x_1	x_2	$\hat{Y}$
0	0	0
0	1	1
1	0	1
1	1	0



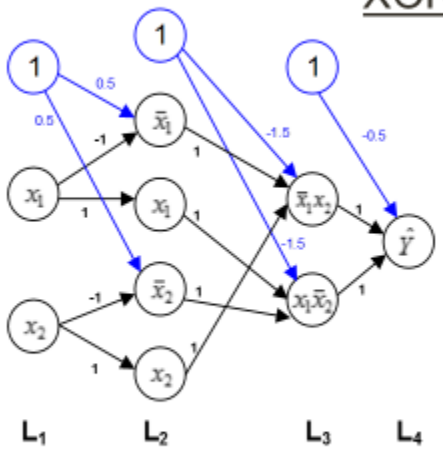
We can extend the truth-table to use the linearly separable logic functions, right?

x_1	x_2	$(\bar{x}_1 x_2 + x_1 \bar{x}_2)$	$\hat{Y}$
0	0	0 + 0	0
0	1	1 + 0	1
1	0	0 + 1	1
1	1	0 + 0	0

33

Hidden Layer, When? ...

XOR



x_1	x_2	$(\bar{x}_1 x_2 + x_1 \bar{x}_2)$	$\hat{Y}$
0	0	0 + 0	0
0	1	1 + 0	1
1	0	0 + 1	1
1	1	0 + 0	0

Answer: 'Hidden layer', when the classes are **NOT** linearly separable.

34

12. Explain what should be the starting or initial values of the weight of an ANN. (See chapter #1 (ANN), page 28 …)

Ans: Starting values:

- If the weights are near zero, then the operative part of the sigmoid is roughly linear, and hence the neural network collapses into an approximately linear model. The use of exact zero weights leads to zero derivatives and perfect symmetry, and the algorithm never moves. Starting instead with large weights often leads to poor solutions.

- Usually, starting values for weights are chosen to be random values near zero. Hence the model starts out nearly linear and becomes nonlinear as the weights increase.

- With standardized inputs, it is typical to take random uniform weights over the range [−0.7, +0.7].

13. Explain what should be the number of hidden units and layers in an ANN (see Chapter #1 (ANN), page 29 …).

14. Write down the delta-rule or, error back-propagation algorithm.
Ans:

Algorithm: Error Back-propagation

BEGIN

1. From a data point (x_i, y_i), apply an input vector x_i to the network and forward propagate through the network and find the output error E. Then to perform the following steps to computation the rate of change of error w.r.t the network weights βs to compute the next value of βs.

2. For each of the output node/unit compute the error term δ^L :

$$\delta^{(L)} = (Z_k^{(L)} - Y_k)\, Z_k^{(L)}(1 - Z_k^{(L)})$$

3. For each of the hidden layer node /unit compute the error term $\delta^{(L-1)}$:

$$\delta^{(L-1)} = Z^{(L-1)}(1 - Z^{(L-1)}) \times \sum_{k=1}^{K} \delta^{(L)} \times \beta^{(L-1)}$$

And Compute: $\delta^{(L-2)}, \delta^{(L-3)}, ..., \delta^{(2)}$, except $\delta^{(1)}$ because it is the input layer and we do not want to change the original input data.

4. Update the weights (β) of the network:

$$\beta^{(i)}(t+1) = \beta^{(i)}(t) - \alpha\,\delta^{(i+1)}\,Z^{(i)}$$
$$\beta_0^{(i)}(t+1) = \beta_0^{(i)}(t) - \alpha\,\delta^{(i+1)} \qquad \text{[For bias terms]}$$
where, i=*1*, *2*, …, (*L*-1)

5. Go to Step 1 to loop or, exit if the exit-condition is met.

END

--- x ---