LUMS
A Not-for-Profit University

Department of Electrical Engineering
School of Science and Engineering

EE514/CS535 Machine Learning

ASSIGNMENT 2 – SOLUTIONS

Due Date: 4:00 pm, Tuesday, April 11, 2023.
Format: 6 problems, for a total of 100 marks
Instructions:

- You are allowed to collaborate with your peers but copying your colleague's solution is strictly prohibited. This is not a group assignment. Each student must submit his/her own assignment.
- Solve the assignment on blank A4 sheets and staple them before submitting.
- Submit in the dropbox labeled EE-514 outside the instructor's office.
- Write your name and roll no. on the first page.
- Feel free to contact the instructor or the teaching assistants if you have any concerns.

- You represent the most competent individuals in the country, do not let plagiarism come in between your learning. In case any instance of plagiarism is detected, the disciplinary case will be dealt with according to the university's rules and regulations.

Problem 1 (15 marks)

(a) [**5 marks**] Show that the cross-entropy function is a convex function.

Solution:

$$f(w) = -[y * ln(\hat{y}) + (1-y) * ln(1-\hat{y})]$$
$$\hat{y} = \sigma(z), z = w^T x$$

We find the 2nd derivative of $f(w)$ and show that it is convex.

$$-f(w) = y * ln(\frac{e^{w^T x}}{1+e^{w^T x}}) + (1-y) * ln(\frac{1}{1+e^{w^T x}})$$

simplifying this we get,

$$-f(w) = y * ln(e^{w^T x}) - ln(1+e^{w^T x})$$
$$f(w) = ln(1+e^{w^T x}) - wxy$$
$$\frac{df(w)}{dw} = \frac{1}{1+e^{w^T x}} e^{w^T x} x - xy = \frac{x}{1+e^{-w^T x}} - xy$$
$$\frac{d^2 f(w)}{dw^2} = x \frac{d}{dw}(\frac{1}{1+e^{-w^T x}}) = x \frac{d}{dw}(\sigma(x^T w))$$
$$= x(\sigma(x^T w)((1-\sigma(x^T w)))$$
$$= x^2(\frac{1}{1+e^{-x^T w}})(\frac{e^{-w^T x}}{1+e^{-w^T x}})$$
$$\frac{x^2 e^{-w^T x}}{(1+e^{-w^T x})^2}$$

All three terms are always ≥ 0. So $\frac{d^2 f(w)}{dw^2} \geq 0$, and function is convex.

(b) [**5 marks**] Prove that softmax with 2 classes is the same as sigmoid

Solution: Let,

$$softmax(z_i) = \frac{e^{z_i}}{\Sigma e^{z_j}}$$
$$P(Y_i = 1) = 1 - P(Y_i = 0) = 1 - softmax(z_i) = 1 - \frac{e^{z_0}}{e^{z_0} + e^{z_1}} = \frac{e^{z_1}}{e^{z_0} + e^{z_1}}$$

Can be written as,

$$P(Y_i = 1) = \frac{1}{1 + e^{z_0 - z_1}}$$

Let $-z = z_0 - z_1$

$$P(Y_i = 1) = \frac{1}{1 + e^{-z}}$$

Which is same as sigmoid.

Problem 2 (15 marks)

You have been given miniature training and test documents from the actual dataset of movie reviews. The documents belong to either the positive, negative, or neutral class.

Dataset	Sentiment	Text
Training	Positive	great acting by everyone and amazing movie
Training	Positive	superb plot and cinematography
Training	Neutral	average acting but the storyline is good.
Training	Negative	lacks proper plot.
Training	Negative	the movie is an utter disaster.
Testing	?	great acting by Leonardo and amazing storyline.

You need to develop a multinomial Naive Bayes classifier for this problem by following the steps below.

(a) [**2 marks**] A list of stop words is given to you.
Stop words = [as, if, at, by, and, the, an, but, is]

Apply preprocessing to the training and test data by removing stop words from them and showing the documents after preprocessing.

Solution:

Dataset	Sentiment	Text
Training	Positive	great acting everyone amazing movie
Training	Positive	superb plot cinematography
Training	Neutral	average acting storyline good.
Training	Negative	lacks proper plot.
Training	Negative	movie utter disaster.
Testing	?	great acting Leonardo amazing storyline.

(b) [**3 marks**] You need to work with the preprocessed documents from now onwards. Construct vocabulary from the data and tell its size.

Solution: V = {great, acting, everyone, amazing, movie, superb, plot, cinematography, average, storyline, good, lacks, proper, utter, disaster}
$|V| = 15$

(c) [**3 marks**] Construct prior probabilities.

Solution:

$$P(Positive) = \frac{2}{5}$$

$$P(Negative) = \frac{2}{5}$$

$$P(Neutral) = \frac{1}{5}$$

(d) [**4 marks**] Compute the likelihoods of all the words in the training data using Laplace add-one smoothing.

Solution:

$$P(great|Positive) = \frac{1+1}{8}, P(great|Negative) = \frac{0+1}{6}, P(great|Neutral) = \frac{0+1}{4}$$

$$P(plot|Positive) = \frac{1+1}{8}, P(plot|Negative) = \frac{1+1}{6}, P(plot|Neutral) = \frac{0+1}{4}$$

$$P(acting|Positive) = \frac{1+1}{8}, P(acting|Negative) = \frac{0+1}{6}, P(acting|Neutral) = \frac{1+1}{4}$$

...

(e) [**3 marks**] Now, predict the sentiment of the test data and show your working.

Solution:

$$P(class|X) = \frac{P(X|class)P(class)}{P(X)}$$

$P(X)$ is same always so we can ignore. Here

$$X = [w_1, w_2, ..., w_n] = [great, acting, Leonardo, amazing, storyline]$$

$$P(class = Positive|X) = P(X|class = Positive)P(class = Positive)$$

Using Naive assumption (indepence of words) Posterior becomes,

$$P(class = Positive|X) =$$

$$P(great|Positive) * P(acting|Positive) * P(Leonardo|Positive) *$$
$$P(amazing|Positive) * P(storyline|Positive) * P(Positive)$$

Similarly for Negative class,

$$P(class = Negative|X) =$$
$$P(great|Negative) * P(acting|Negative) * P(Leonardo|Negative) *$$
$$P(amazing|Negative) * P(storyline|Negative) * P(Negative)$$

Plug in and you will find that

$$P(class = Positive|X) > P(class = Neutral|X) > P(class = Negative|X)$$

Hence we classify the test review as Positive class.

Problem 3 (25 marks)

We often use regularization to reduce overfitting. In the case of ridge regression, we added the following (square of Euclidean norm of the weights $\mathbf{w}$)

$$\lambda||\mathbf{w}||_2^2$$

as the regularization term in the objective function to be minimized. In this question, we extend this to logistical regression.

(a) **[10 marks]** Formulate a loss function for the logistic regression and add this regularization term in the objective function. Compute the gradient of the new loss function with respect to weights $\mathbf{w}$.

Solution:

$$L(w) = \sum_{i=1}^{n} log(1 + exp(-y_i \mathbf{w^T x_i})) + \lambda||w||_2^2,$$

where,

$$y_i \in \{1, -1\}$$

$$\begin{aligned}\frac{dL(w)}{dw} &= \sum_{i=1}^{n} \frac{1}{1 + exp(-y_i \mathbf{w^T x_i})} exp(-y_i \mathbf{w^T x_i})(-y_i \mathbf{x_i}) + 2\lambda \mathbf{w} \\ &= -\sum_{i=1}^{n} y_i(1 - \hat{y}_i)\mathbf{x_i} + 2\lambda \mathbf{w}\end{aligned}$$

where,

$$\hat{y}_i = \frac{1}{1 + exp(-y_i \mathbf{w^T x_i})}$$

(b) **[15 marks]** Another way to minimize the objective function with regularization term is to obtain Maximum A Posteriori (MAP) estimate given by

$$\mathbf{w}_{\text{MAP}} = \max \prod_{i=1}^{n} P(y_i \mid \mathbf{x}_i, \mathbf{w}) P(\mathbf{w}),$$

where y_i is i-th label associated with the i-th input $\mathbf{x}_i$.

We make the following assumptions:

$P(y_i \mid x_i, \mathbf{w}) = 1/1 + \exp(-y_i \mathbf{w}^T x_i)$ for all $i \in 1, 2, ...N,$. $\mathbf{w}$ (Prior on $\mathbf{w}$) is normally distributed with zero mean and the covariance matrix is a multiple of the identity matrix.

$$P(w) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}\sigma} \exp(-w_j^2/2\sigma^2).$$

We require you to show that for a particular value of λ and σ, the MAP estimate is the same as the $\mathbf{w}$ obtained by minimizing the objective function formulated for regularized logistic regression.

Solution:

$$\begin{aligned}\mathbf{w_{MAP}} &= \arg\max \prod_{i=1}^{n} P(y_i|\mathbf{x_i}, \mathbf{w})P(\mathbf{w}) \\ &= \arg\max \log \prod_{i=1}^{n} P(y_i|\mathbf{x_i}, \mathbf{w})P(\mathbf{w}) \\ &= \arg\max \left(\sum_{i=1}^{n} \log P(y_i|\mathbf{x_i}, \mathbf{w}) + \log P(\mathbf{w}) \right)\end{aligned}$$

$$= \arg\max \left(\sum_{i=1}^{n} \log \frac{1}{1 + exp(-y_i \mathbf{x_i^T w})} + \log \frac{1}{\sqrt{2\pi\sigma}} \prod_{i=1}^{d} e^{-\frac{w_j^2}{2\sigma^2}} \right)$$

$$= \arg\max \left(\sum_{i=1}^{n} \log \frac{1}{1 + \exp(-y_i \mathbf{x_i^T w})} + \sum_{j=1}^{d} \log \left(\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{w_j^2}{2\sigma^2}} \right) \right)$$

$$= \arg\max \left(\sum_{i=1}^{n} \log \frac{1}{1 + \exp(-y_i \mathbf{x_i^T w})} + \sum_{j=1}^{d} \left(-\frac{w_j^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right) \right)$$

$$= \arg\max \left(-\sum_{i=1}^{n} \log(1 + \exp(-y_i \mathbf{x_i^T w})) - \sum_{j=1}^{d} \frac{w_j^2}{2\sigma^2} \right)$$

$$= \arg\max \left(-\sum_{i=1}^{n} \log(1 + \exp(-y_i \mathbf{x_i^T w})) - \frac{||w||_2^2}{2\sigma^2} \right)$$

$$= \arg\min \left(\sum_{i=1}^{n} \log(1 + \exp(-y_i \mathbf{x_i^T w})) + \frac{||w||_2^2}{2\sigma^2} \right)$$

This is same as the loss function from part (a) with $\lambda = \frac{1}{2\sigma^2}$

Problem 4 (10 marks)

You are in a noisy bar diligently studying for your midterm, and your friend is trying to get your attention, using only a two-word vocabulary. She has said a sentence but you can't hear one of the words.

$$w_1 = \text{hi}, \quad w_2 = \text{yo}, \quad w_3 = ?, \quad w_4 = \text{hi}$$

Assume that your friend was generating words from this first-order Markov model:

$$\begin{array}{ll} p(hi \mid hi) = 0.7, & p(yo \mid hi) = 0.3 \\ p(hi \mid yo) = 0.5, & p(yo \mid yo) = 0.5 \end{array}$$

Given these parameters, what is the posterior probability of whether the missing word is "hi" or "yo"?

Solution: By the markov assumption,

$$P(w_3|w_1, w_2, w_4) = P(w_3|w_2, w_4)$$

Using Bayes rule,

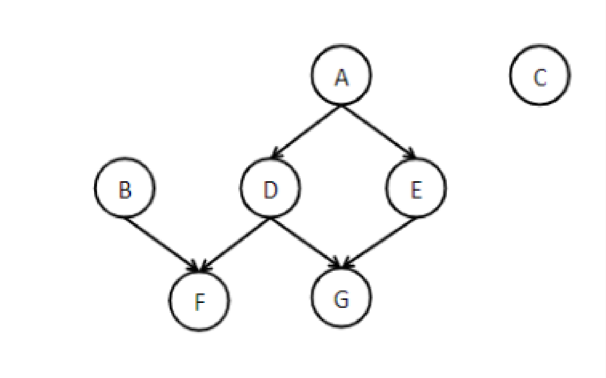
$$\begin{aligned} P(w_3|w_2, w_4) = P(w_3|w_2)P(w_4|w_2, w_3) &= P(w_3|w_2)P(w_4|w_3) \\ P(w_3 = hi|w_2 = yo, w_4 = hi) &= P(w_3 = hi|w_2 = yo)P(w_4 = hi|w_3 = hi) \\ &= (0.5)(0.7) = 0.35 \\ P(w_3 = yo|w_2 = yo, w_4 = hi) &= P(w_3 = yo|w_2 = yo)P(w_4 = hi|w_3 = yo) \\ &= (0.5)(0.5) = 0.25 \end{aligned}$$

Problem 5 (20 marks)

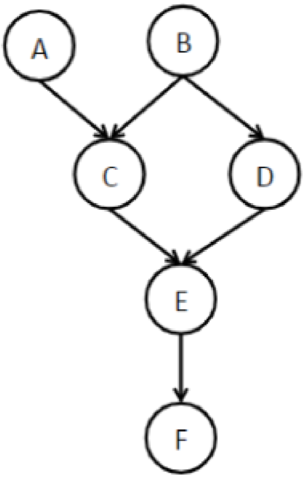
(a) [**5 marks**] Please draw the directed graph corresponding to the following distribution:

$$P(A,B,C,D,E,F,G) = P(A)P(B)P(C)P(D \mid A)P(E \mid A)P(F \mid B,D)P(G \mid D,E)$$

Solution:



(b) [**5 marks**] Please write down the factorial joint distribution represented by the graph below:



Solution: $P(A,B,C,D,E,F,G) = P(A)P(B)P(C \mid A,B)P(D \mid B)P(E \mid C,D)P(F \mid E)$

(c) [**4 marks**] Assume the random variables in the graph shown above are Boolean. How many parameters are needed in total to fully specify the Bayesian network? Justify your answer.

Solution: We need just 1 parameter for P(x), because P(x = True) = 1 - P(x = False)

- 1 for P(A)
- 1 for P(B)
- 4 for P(C | A, B) - 1 for each combination of values of A and B
- 2 for P(D | B)
- 4 for P(E | C, D)
- 2 for P(F | E)

Total parameters = 14

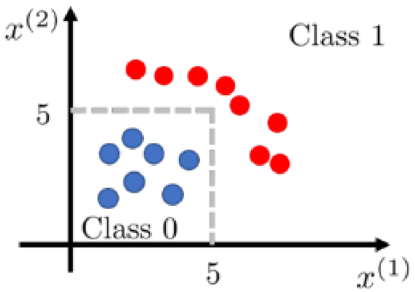
(d) [**6 marks**] Based on the graph shown in part (b), state whether the following are true or false: Where $A \perp\!\!\!\perp B$ represents A is independent of B.

	$True/False$
$A \perp\!\!\!\perp B$	
$A \perp\!\!\!\perp B \mid C$	
$C \perp\!\!\!\perp D$	
$C \perp\!\!\!\perp D \mid E$	
$C \perp\!\!\!\perp D \mid B, F$	
$F \perp\!\!\!\perp B$	
$F \perp\!\!\!\perp B \mid C$	
$F \perp\!\!\!\perp B \mid C, D$	
$F \perp\!\!\!\perp B \mid E$	
$A \perp\!\!\!\perp F$	
$A \perp\!\!\!\perp F \mid C$	
$A \perp\!\!\!\perp F \mid D$	

Solution:

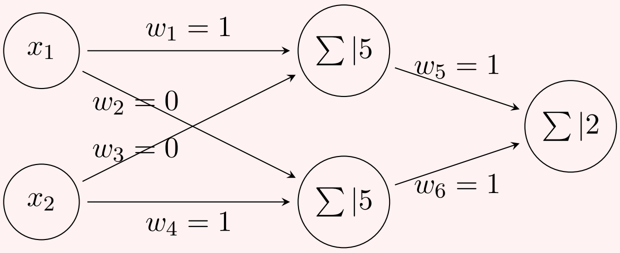
	$True/False$
$A \perp\!\!\!\perp B$	True
$A \perp\!\!\!\perp B \mid C$	False
$C \perp\!\!\!\perp D$	False
$C \perp\!\!\!\perp D \mid E$	False
$C \perp\!\!\!\perp D \mid B, F$	False
$F \perp\!\!\!\perp B$	False
$F \perp\!\!\!\perp B \mid C$	False
$F \perp\!\!\!\perp B \mid C, D$	True
$F \perp\!\!\!\perp B \mid E$	True
$A \perp\!\!\!\perp F$	False
$A \perp\!\!\!\perp F \mid C$	False
$A \perp\!\!\!\perp F \mid D$	False

Problem 6 (15 marks)

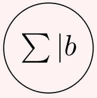


Design a perceptron with the dashed line indicated in the figure as its (approximate) decision boundary. You must draw a perceptron indication of inputs, output, weights, and biases and provide a brief explanation of your design.
Hint: You might want to start with a perceptron that determines on which side of the dashed boundary the data lies.

Solution:



Where,



$$= \begin{cases} 1 & \text{if, } \sum_i x^{(i)} w^{(i)} \geq b \\ 0 & \text{else.} \end{cases}$$

— End of Assignment —