1. Suppose we would like to classify input x given by

$$x = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$$

Consider parameter W is given by

$$W = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

and bias b given by

$$b = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$$

We would like to use linear score $s = Wx + b$, where the 1st, 2nd and 3rd element of s represents the score for "cat" , "dog" and "ship", respectively.

(a) Calculate the score for "cat" category

(b) Calculate the score for "dog" category

(c) Calculate the score for "ship" category

SOL:

(a) 3

(b) 6

(c) 7

2. **Naive Bayes classification.** This example will guide you through building a Naive Bayes (NB) classifier for discrete-valued input data. Note that this model is different from logistic regression, and does not use parameterized estimates of probabilities, but uses simple estimates directly from data.

Consider the data in Table 1 of data regarding some illness in a hospital. Note $+$ and $-$ of three symptoms, Fever, Cough, Nose (runny nose) means the presence of symptom is positive and negative respectively. These $+$ or $-$ values of symptoms are input data. The output data is the presence of Illness, again $+$ or $-$ as positive or negative.

Suppose a doctor received new patient with symptoms: Fever:$-$, Cough:$-$, Nose: $+$. We would like to decide whether this patient is ill using NB classifier.

In this binary decision problem, we would like to decide which one is greater of

$$P(I_-|F_-, C_-, N_+) \text{ vs } P(I_+|F_-, C_-, N_+)$$

Fever	Cough	Nose	Illness
+	+	+	+
+	-	-	+
-	+	-	+
-	+	+	-
+	-	+	-
-	-	-	-

Table 1:

Note I_+ and I_- denotes the event of illness positive and negative respectively. Similarly F_+ and F_- denote the event of Fever positive or negative: other notations are defined similarly for C_- (cough) and N_+ (nose). In other words, which one is greater: probability of being ill or not, given that the patient's symptom is Fever:$-$, Cough:$-$, Nose: $+$? From Bayes' rule, this is equivalent to asking

$$\frac{P(F_-, C_-, N_+|I_+)P(I_+)}{P(F_-, C_-, N_+)} \text{ vs } \frac{P(F_-, C_-, N_+|I_-)P(I_-)}{P(F_-, C_-, N_+)}$$

or, because the denominators are the same,

$$P(F_-, C_-, N_+|I_+)P(I_+) \text{ vs } P(F_-, C_-, N_+|I_-)P(I_-)$$

From conditional independence assumption of NB,

$$P(F_-, C_-, N_+|I_+) = P(F_-|I_+)P(C_-|I_+)P(N_+|I_+)$$
$$P(F_-, C_-, N_+|I_-) = P(F_-|I_-)P(C_-|I_-)P(N_+|I_-)$$

There are several probabilities that needs to be estimated from the data:

- prior probabilities $P(I_+), P(I_-)$
- posterior probabilities $P(F_-|I_+), P(C_-|I_+), P(N_+|I_+), P(F_-|I_-), P(C_-|I_-), P(N_+|I_-)$.

(a) We want to estimate the probabilities $P(I_+)$ and $P(I_-)$. Make a maximum likelihood estimate (MLE) of $P(I_+)$. Note $P(I_-)$ is simply $1 - P(I_+)$.

(b) We want to estimate the probability $P(F_-|I_+)$. Make a MLE of $P(F_-|I_+)$.

(c) Use similar approach to make MLE of $P(C_-|I_+), P(N_+|I_+), P(F_-|I_-), P(C_-|I_-)$ and $P(N_+|I_-)$.

(d) Based on your estimates, what is your decision on whether ill(+) or not(-) on the patient?

SOL: Let us define the following:

$$p_{i+} = P(I_+),\ p_{F+} = P(F_+|I_+), p_{F-} = P(F_+|I_-),$$
$$p_{N+} = P(N_+|I_+), p_{N-} = P(N_+|I_-),\ p_{C+} = P(C_+|I_+), p_{C-} = P(C_+|I_-),$$

The likelihood function for the data in Table 1 can be represented using these probabilities. For example, for the likelihood of the first sample

$$P(I_+|F_+, C_+, N_+) \propto P(F_+, C_+, N_+|I_+)P(I_+) = P(F_+|I_+)P(C_+|I_+)P(N_+|I_+)P(I_+) = p_{F+}p_{C+}p_{N+}p_{i+}$$

Similarly, the likelihood for the second sample is $p_{F+}(1-p_{C+})(1-p_{N+})p_{i+}$, etc. Thus the overall likelihood Λ of data can be represented to be proportional to the product of these probabilities. Now, given the likelihood (or logarithm of it), we know how to maximize it: (partial) differentiate it and set to zero. For example, we know that the likelihood will have the factor

$$p_{i+}^3(1-p_{i+})^3$$

because there are three + and three - in the illness data. By taking the partial differentiation of λ with respect to p_{i+} and setting it to zero, we see that the optimal $p_{i+} = 0.5$. We can find similar maximum likelihood estimates for other probabilities $p_{i+},\ p_{F+}, p_{F-},\ p_{N+}, p_{N-}\ p_{C+}, p_{C-}$

(a) This is the optimal value of p_{i+} which is the maximum likelihood estimate. $P(I_+) = 3/6 = 0.5$. $P(I_-) = 1 - P(I_+) = 0.5$

(b) The maximum likelihood estimate for $p_{F+} = 2/3$, because there are two cases of $+$ and 1 case of $-$ Fever, given that Illness is $+$. Thus $P(F_-|I_+) = 1 - p_{F+} = 1/3$,

(c) $P(C_-|I_+) = 1/3, P(N_+|I_+) = 1/3$ and $P(F_-|I_-) = 2/3, P(C_-|I_-) = 2/3$ and $P(N_+|I_-) = 2/3$

(d) Using our estimates, we have that

$$P(F_-, C_-, N_+|I_+)P(I_+) = P(F_-|I_+)P(C_-|I_+)P(N_+|I_+)P(I_+) = 1/54$$

where

$$P(F_-, C_-, N_+|I_-)P(I_-) = P(F_-|I_-)P(C_-|I_-)P(N_+|I_-)P(I_-) = 8/54$$

So our decision is, the new patient is not ill or negative $(-)$

3. Suppose we have the score

$$s = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

What is the softmax applied to s?
SOL: TO BE GRADED:

$$\begin{bmatrix} \frac{\exp(2)}{\exp(2)+\exp(1)+\exp(0)} \\ \frac{\exp(1)}{\exp(2)+\exp(1)+\exp(0)} \\ \frac{\exp(0)}{\exp(2)+\exp(1)+\exp(0)} \end{bmatrix}$$

4. Write down the cross-entropy loss L for the linear score for class 1, 2 and 3

$$s = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

where the ground truth label for this data sample was class 2.
SOL:

$$L = -\log\left(\frac{\exp(1)}{\exp(2)+\exp(1)+\exp(0)}\right)$$

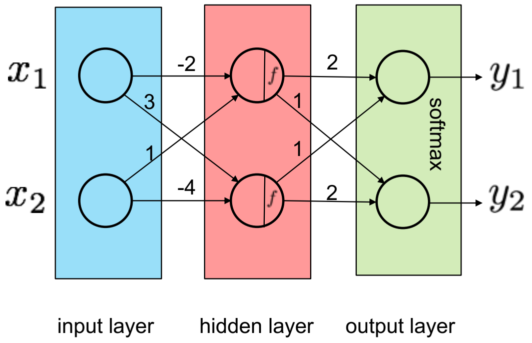

Figure 1: Neural Network

5. Consider the neural network in Fig. 1. The hidden layer uses ReLU activation. The output layer is a linear layer, and outputs the softmax of the linear score. Suppose the input is $(x_1, x_2) = (1, 3)$. What is the output (y_1, y_2)? Assume all the biases are 0.

   SOL: TO BE GRADED: The output of the first neuron of the hidden layer is

$$\max(1 \times (-2) + 3 \times 1, 0) = 1$$

   The output of the second neuron of the hidden layer is

$$\max(1 \times (3) + 3 \times (-4), 0) = 0$$

   So the output of the hidden layer is $(1, 0)$. Then the output layer is

$$\text{softmax}(1 \times 2 + 0 \times 1, 1 \times 1 + 0 \times 2) = \text{softmax}(2, 1) = \left(\frac{\exp(2)}{\exp(2) + \exp(1)}, \frac{\exp(1)}{\exp(2) + \exp(1)}\right)$$

6. Consider loss function $L(x, y) = 2x + 3xy$ with learning rate α. We would like to minimize the loss using gradient descent. What is the step for gradient descent at point (x, y)?
   SOL: TO BE GRADED:

$$-\alpha \nabla L = -\alpha \begin{bmatrix} \frac{\partial L}{\partial x} \\ \frac{\partial L}{\partial y} \end{bmatrix} = -\alpha \begin{bmatrix} 2 + 3y \\ 3x \end{bmatrix}$$

7. Consider a simple linear classifier with m classes. The input to the classifier is vector $x \in \mathbb{R}^n$. The first layer is linear layer whose output is Wx, and $W \in \mathbb{R}^{m \times n}$ is a parameter matrix. Then the cross-entropy loss function denoted by L is applied to the output (that is, first applying softmax to the output and applying negative log-likelihood). Suppose that the current input is x, and the ground truth label is given by $y \in \{1, \ldots, m\}$. Find the expression for

$$\frac{dL}{dW}$$

   Note that the answer should have the same shape as W.

   SOL: TO BE GRADED:

$$\frac{dL}{dW} = (\text{softmax}(Wx) - \mathbf{e}_y)x^T$$

8. Consider a network with a convolutional layer followed by ReLU activation. The input map is denoted by x given by

$$x = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 0 \\ 0 & -1 & -2 \end{bmatrix}$$

The parameter for convolutional filter is denoted by W which has spatial dimension of height 2 and width 2 (no bias), with no padding and stride of 1. The current value of W is given by

$$W = \begin{bmatrix} -1 & 1 \\ 1 & 2 \end{bmatrix}$$

Let the output of ReLU activation be y. Suppose that the upstream gradient of loss L with respect to y is given by

$$\frac{dL}{dy} = \begin{bmatrix} -3 & 2 \\ 0 & 1 \end{bmatrix}$$

Find the derivative of L with respect to parameter W, that is

$$\frac{dL}{dW}$$

Note that the answer should have the same shape as W.

SOL: Let the output of convolutional layer be z.

$$z = \begin{bmatrix} 4 & 2 \\ -2 & -6 \end{bmatrix}$$

Using chain rule, we have

$$\frac{dL}{dW} = \frac{dL}{dy}\frac{dy}{dz}\frac{dz}{dW}$$

Using properties of derivative of ReLU, we have that

$$\frac{dL}{dz} = \begin{bmatrix} -3 \cdot 1(4 \geq 0) & 2 \cdot 1(2 \geq 0) \\ 0 \cdot 1(-2 \geq 0) & 1 \cdot 1(-6 \geq 0) \end{bmatrix} = \begin{bmatrix} -3 & 2 \\ 0 & 0 \end{bmatrix}$$

Thus, we have

$$\frac{dL}{dW} = -3\begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} + 2\begin{bmatrix} 2 & 3 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & -3 \end{bmatrix}$$