**Q1**

The KL divergence is:

The cross entropy is:

The relationship is showed as formula below:

=

In the machine learning process, we wish to minimize the KL divergence (1). From (3), we know that minimize the is same as minimize the cross entropy since is unchanged. Approximating the cross entropy using data, we have:

Which is same as maximizing the log likelihood.

**Q2**

**=**

Therefore,

**Q3**

1.c. There is no penalty on w0, thus w0 = 1

2.b. The strong penalty on the weight will force the function have small slope.

3.a. The penalty is given to w0 either, resulting w0 < 1.

4. d. The regularization scheme put strong penalty on both the intercept and weight, results the intercept in figure d) relatively low and the model underfit.

**Q4**

The prediction of before iteration is:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X1 | 0 | 0 | 1 | 1 |
| X2 | 0 | 1 | 0 | 1 |
| Y | 1 | 0 | 0 | 1 |
|  | 0.1192 | 0.2689 | 0.2689 | 0.5 |

According to the batch gradient descent formula:

W0 = -2 + 0.1 \* 0.25 \* [(1 – 0.1192)\*1 + (0 – 0.2689)\*1 + (0 – 0.2689)\*1 + (1 – 0.5)\*1]

= -2 + 0.025 \* 0.843

= -1.978925

W1 = 1 + 0.1 \* 0.25 \* [(1 – 0.1192)\*0 + (0 – 0.2689)\*0 + (0 – 0.2689)\*1 + (1 – 0.5)\*1]

= 1 + 0.025 \* 0.2311

= 1.0058

W2 = 1 + 0.1 \* 0.25 \* [(1 – 0.1192)\*0 + (0 – 0.2689)\*1 + (0 – 0.2689)\*0 + (1 – 0.5)\*1]

= 1 + 0.025 \* 0.2311

= 1.0058

The prediction of After iteration is:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X1 | 0 | 0 | 1 | 1 |
| X2 | 0 | 1 | 0 | 1 |
| Y | 0 | 0 | 0 | 1 |
| Ypredict | 0 | 0 | 0 | 1 |

The training error is 0.

**Q5**

**1.**

The minimum achievable training error is 0.25, since a straight line cannot separate the dataset and classify the data correctly.

A possible weight is w0 = -1.1, w1 = w2 = 1

2.

The minimum achievable training error is 0

A possible weight is w0 = 1, w1 = w2 = -1, w3 = 2

**Q6**

1.

For the first updating rule, since x1 = 0 in all examples with y = 0 and large fraction of examples with y = 1, the update term will be 0.

For x1 = 1 and y = 1, we can infer that .

Generally, examples with x1 = 1 and y = 1 count a small fraction of training set, which means will be very small or even 0.

Therefore, w1 will increase very slowly or even keep constant with iterations.

2.

From , we can infer that

, which means the regularization will forces the weights to be smaller with

1 > .

Combine with conclusion drawn in q1, the w1 will increase slower than the speed in q1. When the iteration is terminated, the final value of w1 will be smaller than the final value of w1 in q1.