Introduction to ML (CS771), Autumn 2023

Indian Institute of Technology Kanpur

Homework Assignment Number 1

[redacted]

QUESTION

[redacted]

[redacted]

First, we will establish the expectation values for various vectors and matrices involved in the problem, since we are dealing with probabilities.

$\mathbb{E}[m_{nd}]$ = p, as m_{nd} is a Bernoulli random variable, $\mathbb{E}[v] = 1 \cdot p + 0 \cdot (1-p)$ for any bernoulli random variable v.

$\mathbb{E}[\widetilde{x}_{nd}] = \mathbb{E}[x_{nd} \cdot m_{nd}]$

$\mathbb{E}[\widetilde{x}_{nd}] = x_{nd} \cdot \mathbb{E}[m_{nd}]$

$\mathbb{E}[\widetilde{x}_{nd}] = x_{nd} \cdot p$

Thus, we also get:

$\mathbb{E}[\widetilde{x}_n] = p \cdot x_n$

Now, consider the matrix $\widetilde{x}_n \widetilde{x}_n^T$. Its (i,j)th entry is $\widetilde{x}_{ni} \cdot \widetilde{x}_{nj}$. Now, if i != j, $\widetilde{x}_{ni}$ and $\widetilde{x}_{nj}$ are two independent random variables, hence we will have

$$\mathbb{E}[\widetilde{x}_{ni} \cdot \widetilde{x}_{nj}] = \mathbb{E}[\widetilde{x}_{ni}] \cdot \mathbb{E}[\widetilde{x}_{nj}]$$

$$\mathbb{E}[\widetilde{x}_{ni} \cdot \widetilde{x}_{nj}] = p^2 \cdot x_{ni} \cdot x_{nj}$$

And for the diagonal elements, that is, i = j, it becomes $\mathbb{E}[\widetilde{x}_{nd}^2]$.

$$\mathbb{E}[\widetilde{x}_{nd}^2] = x_{ni}^2 \cdot \mathbb{E}[m_{ni}^2]$$

where m_{ni} is a Bernoulli random variable. For any Bernoulli random variable v with probability p of getting 1, $\mathbb{E}[v^2] = 1^2 \cdot p + 0^2 \cdot (1-p)$, thus the expected value of a square of Bernoulli random variable is p.

$$\mathbb{E}[\widetilde{x}_{nd}^2] = x_{ni}^2 \cdot p$$

Therefore, $\mathbb{E}[\widetilde{x_n}\widetilde{x_n}^T]$ is the matrix $x_n x_n^T$, with p^2 term multiplied in the nondiagonal elements and p term multiplied in the diagonal elements. We can add and subtract p^2 to get the following expression, where $diag(x_n)$ represents a $D \times D$ diagonal matrix with i'th diagonal entry as x_{ni}^2 :

$$\mathbb{E}[\widetilde{x_n}\widetilde{x_n}^T] = p^2 \cdot x_n x_n^T + (p - p^2) \cdot diag(x_n) \tag{2}$$

Now, consider the matrix $\widetilde{X}^T \widetilde{X}$.

$$\widetilde{X}^T \widetilde{X} = \sum_{n=1}^{N} \widetilde{x_n}\widetilde{x_n}^T$$

$$\mathbb{E}[\widetilde{X}^T \widetilde{X}] = \mathbb{E}[\sum_{n=1}^{N} \widetilde{x_n}\widetilde{x_n}^T]$$

Using linearity of expectation, we have:

$$\mathbb{E}[\widetilde{X}^T \widetilde{X}] = \sum_{n=1}^{N} \mathbb{E}[\widetilde{x_n}\widetilde{x_n}^T]$$

Now, from equation 2, we can write this as :

$$\mathbb{E}[\widetilde{X}^T \widetilde{X}] = \sum_{n=1}^{N} (p^2 \cdot x_n x_n^T + (p - p^2) \cdot diag(x_n))$$

$$\mathbb{E}[\widetilde{X}^T \widetilde{X}] = \sum_{n=1}^{N} p^2 \cdot x_n x_n^T + \sum_{n=1}^{N} (p - p^2) \cdot diag(x_n)$$

$$\mathbb{E}[\widetilde{X}^T \widetilde{X}] = (p^2) X^T X + (p - p^2) \sum_{n=1}^{N} diag(x_n)$$

Define $diag(X)$ as the $D \times D$ diagonal matrix with i'th entry as $\sum_{j=1}^{N} x_{ni}^2$. It's evident that $\sum_{n=1}^{N} diag(x_n) = diag(X)$. Hence, we have :

$$\mathbb{E}[\widetilde{X}^T \widetilde{X}] = (p^2) X^T X + (p - p^2) diag(X)$$

Now, the loss function for the problem is

$$L(w) = \sum_{n=1}^{N} (y_n - w^T \widetilde{x_n})^2$$

$$L(w) = (Y - \widetilde{X} w)^T (Y - \widetilde{X} w)$$

$$\mathbb{E}[L(w)] = \mathbb{E}[(Y - \widetilde{X} w)^T (Y - \widetilde{X} w)]$$

Again, we will expand the terms and use linearity of expectation

$$\mathbb{E}[L(w)] = \mathbb{E}[Y^T Y] - \mathbb{E}[Y^T \widetilde{X} w] - \mathbb{E}[w^T \widetilde{X}^T Y] + \mathbb{E}[w^T \widetilde{X}^T \widetilde{X} w]$$

$$\mathbb{E}[L(w)] = \mathbb{E}[Y^T Y] - 2\mathbb{E}[Y^T \widetilde{X} w] + \mathbb{E}[w^T \widetilde{X}^T \widetilde{X} w]$$

The terms not involving random variables can be taken out of expectation.

$$\mathbb{E}[L(w)] = Y^T Y - 2Y^T \mathbb{E}[\widetilde{X}] w + w^T \mathbb{E}[\widetilde{X}^T \widetilde{X}] w$$

Now, using these expectation values that we have already calculated,

$$\mathbb{E}[L(w)] = Y^T Y - 2pY^T X w + w^T (p^2 X^T X + (p - p^2) diag(X)) w$$

$$\mathbb{E}[L(w)] = Y^T Y - 2pY^T X w + p^2 w^T X^T X w + (p - p^2) w^T diag(X) w$$

Now, minimizing with respect to w is similar to minimizing with respect to pw, so define $w' = pw$, the equation becomes:

$$\mathbb{E}[L(w')] = Y^T Y - 2Y^T X w' + w'^T X^T X w' + (\frac{1-p}{p}) w'^T diag(X) w'$$

$$\mathbb{E}[L(w')] = (Y - Xw')^T (Y - Xw') + (\frac{1-p}{p}) w'^T diag(X) w'$$

$$\mathbb{E}[L(w')] = \sum_{n=1}^{N} (y_n - w'^T x_n)^2 + (\frac{1-p}{p}) w'^T diag(X) w'$$

Here, we can see that the second term is similar to regularization, with a different value of lambda for each dimension of the weight.

Also, on minimizing this with respect to w', we will get

$$\hat{w'} = (X^T X + (\frac{1-p}{p}) diag(X))^{-1} X^T Y$$

which is similar to solution of L2 regularized square loss function, with the λI term replaced by $(\frac{1-p}{p}) diag(X)$.