

# DS4400 HW3

Xin Guan

1. **MAP estimation.** Consider a Bernoulli random variable  $x$  with  $p(x = 1) = \theta$ . Given a dataset  $D = \{x_1, \dots, x_N\}$ , assume  $N_1$  is the number of trials where  $x_i = 1$ ,  $N_0$  is the number of trials where  $x_i = 0$  and  $N = N_0 + N_1$  is the total number of trials. Consider the following prior, that believes the experiments is biased:

$$p(\theta) = \begin{cases} 0.2 & \text{if } \theta = 0.6 \\ 0.8 & \text{if } \theta = 0.8 \\ 0 & \text{otherwise} \end{cases}$$

- (a) Write down the likelihood function, i.e.,  $p(D|\theta)$ . What is the maximum likelihood solution for  $\theta$  (we already have derived this in the class)?

**Solution:**

we write  $P(X = x) = \theta^x(1 - \theta)^{(1-x)}$

$$P(D|\theta) = P(X_1 = x_1, X_2 = x_2, \dots, X_N = x_N)$$

$$= \prod_{i=1}^N \theta^{x_i}(1 - \theta)^{(1-x_i)}$$

$$= \theta^{N_1}(1 - \theta)^{N_0}$$

$$\text{Let } J(\theta) = \log P(D|\theta) = N_1 \log(\theta) + N_0 \log(1 - \theta)$$

$$\frac{\partial J(\theta)}{\partial \theta} = \frac{N_1}{\theta} - \frac{N_0}{1-\theta}. \text{ Let } \frac{\partial J(\theta)}{\partial \theta} = 0.$$

$$\text{Then } \hat{\theta} = \frac{N_1}{N_0 + N_1} = \frac{N_1}{N}$$

Therefore, the maximum likelihood solution is  $\hat{\theta} = \frac{N_1}{N}$

- (b) Consider maximizing the posterior distribution,  $p(D|\theta) \times p(\theta)$ , that takes advantage of the prior. What is the MAP estimation?

**Solution:**

$$p(D|\theta) \times p(\theta) = \begin{cases} 0.2 \cdot 0.6^{N_1} \cdot 0.4^{N_0} & \theta = 0.6 \\ 0.8 \cdot 0.8^{N_1} \cdot 0.2^{N_0} & \theta = 0.8 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{P(D|0.6)P(0.6)}{P(D|0.8)P(0.8)} = \frac{1}{4} \left(\frac{3}{4}\right)^{N_1} (2)^{N_0} = 3^{N_1} 4^{-N_1-1} 2^{N_0} = 2^{N_1 \log_2 3} 2^{-2N_1-2} 2^{N_0} = 2^{N_0 - (2 - \log_2 3)N_1 - 2}$$

Therefore, when  $\frac{P(D|0.6)P(0.6)}{P(D|0.8)P(0.8)} \geq 1$ :

$$N_0 - (2 - \log_2 3)N_1 - 2 \geq 0 \Rightarrow N_0 \geq (2 - \log_2 3)N_1 + 2$$

$$\text{Therefore, } \hat{\theta} = \begin{cases} 0.6 & N_0 \geq (2 - \log_2 3)N_1 + 2 \\ 0.8 & N_0 < (2 - \log_2 3)N_1 + 2 \end{cases}$$

2. **Naive Bayes Classifier.** Assume you have the following training set with two binary features  $x_1$  and  $x_2$ , and a binary response/output  $y$ . Suppose you have to predict  $y$  using a naive Bayes classifier.

$x_1$	$x_2$	$y$
1	0	0
0	1	0
0	0	0
1	0	1
0	0	1
0	1	1
1	1	1

- (a) Compute the Maximum Likelihood Estimates (MLE) for  $\theta_j^y$  for  $j = 0, 1$  as well as  $\theta_{\tilde{x}_\ell|y}^{x_\ell|y}$  for  $j = 0, 1$  and for  $\ell = 1, 2$ .

**Solution:**

$$\theta_j^y = \begin{cases} \frac{3}{7} & j = 0 \\ \frac{4}{7} & j = 1 \end{cases}$$

$$\theta_{\tilde{x}_\ell|j}^{x_\ell|y} = \begin{cases} \frac{1}{3} & x_1 = 1, j = 0 \\ \frac{2}{3} & x_1 = 0, j = 0 \\ \frac{1}{3} & x_2 = 1, j = 0 \\ \frac{2}{3} & x_2 = 0, j = 0 \\ \frac{1}{2} & x_1 = 1, j = 1 \\ \frac{1}{2} & x_1 = 0, j = 1 \\ \frac{1}{2} & x_2 = 1, j = 1 \\ \frac{1}{2} & x_2 = 0, j = 1 \end{cases}$$

- (b) After learning via MLE is complete, what would be the estimate for  $P(y = 0|x_1 = 0, x_2 = 1)$ .

**Solution:**

$$\begin{aligned} & P(y = 0|x_1 = 0, x_2 = 1) \\ &= \frac{P(x_1=0, x_2=1|y=0)P(y=0)}{P(x_1=0, x_2=1|y=0)P(y=0) + P(x_1=0, x_2=1|y=1)P(y=1)} \\ &= \frac{P(x_1=0|y=0)P(x_2=1|y=0)P(y=0)}{P(x_1=0|y=0)P(x_2=1|y=0)P(y=0) + P(x_1=0|y=1)P(x_2=1|y=1)P(y=1)} \\ &= \frac{\theta_{0|0}^{x_1|y} \theta_{1|0}^{x_2|y} \theta_0^y}{\theta_{0|0}^{x_1|y} \theta_{1|0}^{x_2|y} \theta_0^y + \theta_{0|1}^{x_1|y} \theta_{1|1}^{x_2|y} \theta_1^y} \\ &= \frac{\frac{2}{3} \cdot \frac{1}{3} \cdot \frac{3}{7}}{\frac{2}{3} \cdot \frac{1}{3} \cdot \frac{3}{7} + \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{4}{7}} \\ &= \frac{2}{5} \end{aligned}$$

- (c) What would be the solution of the previous part without the naive Bayes assumption?

**Solution:**

Without Naive Bayes Assumption:

$$\begin{aligned} & P(y = 0|x_1 = 0, x_2 = 1) \\ &= P(y = 0, x_1 = 0, x_2 = 1)/P(x_1 = 0, x_2 = 1) \\ &= \frac{1}{2} \end{aligned}$$

3. Constrained Optimization. Consider the regression problem on a dataset  $\{(x_i, y_i)\}_{i=1}^N$ , where  $x_i \in \mathbb{R}^k$  denotes the input and  $y_i$  denotes the output/response. Let  $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_N]^T$  and  $\mathbf{y} = [y_1 \dots y_N]^T$ . Consider the following regression optimization.

$$\begin{aligned} & \min_{\theta} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\theta\|_2^2 \\ & \text{s.t. } \omega^T \theta = \mathbf{b} \end{aligned}$$

where  $\omega$  and  $b$  are given and indicate the parameters of a hyperplane on which the desired parameter vector,  $\theta$ , lies.

- (a) Assume that  $X^T X = I^k$ , where  $I_k$  denotes the identity matrix. Find the closed-form solution of the above regression problem.

**Solution:**

$$\text{Let } L(\theta, \alpha) = \frac{1}{2} \|\mathbf{y} - \mathbf{X}\theta\|_2^2 + \alpha(\omega^T \theta - b)$$

Then:

$$\frac{\partial L(\theta, \alpha)}{\partial \theta} = \frac{1}{2} \frac{\partial \|\mathbf{y} - \mathbf{X}\theta\|_2^2}{\partial \theta} + \alpha \frac{\partial \omega^T \theta - b}{\partial \theta}$$

$$\begin{aligned} & \frac{\partial}{\partial \theta} \left\| \begin{bmatrix} y_1 - x_1 \theta \\ y_2 - x_2 \theta \\ \vdots \\ y_N - x_N \theta \end{bmatrix} \right\|_2^2 \\ &= \frac{1}{2} \frac{\partial \sum_{i=1}^N (y_i - x_i \theta)^2}{\partial \theta} + \alpha \omega \\ &= \frac{1}{2} \frac{\partial \sum_{i=1}^N (y_i - x_i \theta)^2}{\partial \theta} + \alpha \omega \\ &= \frac{1}{2} \sum_{i=1}^N 2(y_i - x_i \theta)(-x_i) + \alpha \omega \\ &= \sum_{i=1}^N (-x_i y_i + (x_i)^T x_i \theta)(-x_i) + \alpha \omega \\ &= -X^T \mathbf{y} + X^T X \theta + \alpha \omega \\ &= -X^T \mathbf{y} + \theta + \alpha \omega \\ &\frac{\partial L(\theta, \alpha)}{\partial \alpha} = \omega^T \theta - b. \end{aligned}$$

$$\text{Let } \begin{cases} \frac{\partial L(\theta, \alpha)}{\partial \theta} = 0 \\ \frac{\partial L(\theta, \alpha)}{\partial \alpha} = 0 \end{cases} \quad \text{Then, } \begin{cases} -X^T \mathbf{y} + \theta + \alpha \omega = 0 \\ \omega^T \theta - b = 0 \end{cases}$$

$$\text{Since, } \theta = X^T \mathbf{y} - \alpha \omega,$$

$$\text{then } \omega^T (X^T \mathbf{y} - \alpha \omega) = b.$$

$$\Rightarrow \omega^T X^T \mathbf{y} - \omega^T \alpha \omega - b = 0.$$

$$\Rightarrow \omega^T X^T \mathbf{y} - b = \omega^T \alpha \omega$$

$$\Rightarrow \alpha = \frac{\omega^T X^T \mathbf{y} - b}{\omega^T \omega}.$$

$$\text{Therefore } \theta^* = X^T \mathbf{y} - \frac{\omega^T X^T \mathbf{y} - b}{\omega^T \omega} \omega$$

- (b) Verify if your obtained solution  $\theta^*$  satisfies the constraint  $\omega^T \theta^* = b$ .

**Solution:**

$$\begin{aligned} & \omega^T \theta^* \\ &= \omega^T \left( X^T \mathbf{y} - \frac{\omega^T X^T \mathbf{y} - b}{\omega^T \omega} \omega \right) \\ &= \omega^T X^T \mathbf{y} - \frac{\omega^T X^T \mathbf{y} - b}{\omega^T \omega} \omega^T \omega \\ &= \omega^T X^T \mathbf{y} - (\omega^T X^T \mathbf{y} - b) \\ &= b \end{aligned}$$

- (c) What you have been the solution of this optimization, if the constraint  $\omega^T \theta = b$  was not present?

**Solution:**

$$\text{If there is no constraint } \omega^T \theta = b, \text{ it is just } \min_{\theta} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\theta\|_2^2.$$

$$\text{Then } \frac{\partial \frac{1}{2} \|\mathbf{y} - \mathbf{X}\theta\|_2^2}{\partial \theta} = X^T X \theta - X^T \mathbf{y}. \text{ Let it to be 0.}$$

$$\text{Then } \theta^* = (X^T X)^{-1} X^T \mathbf{y}$$

$$\text{If we still preserve the assumption in problem a where } X^T X = I^k, \theta^* = X^T \mathbf{y}$$