1.

a. The random variables for this problem are whether the table is free or not (Y={0 ,1}) and whether it is sunny or not (X={0 ,1}) . Since x and y can only take two outcomes each, we have model it using Bernoulli's . Therefore, our parameters are w=(α_0 , α_1) where α_0 , $\alpha_1 \in \{0,1\}$ and we can write their probabilities as the following:
$P(Y=y|\ \alpha_0) = \alpha_0^y(1-p)^{1-y}$ and $P(X=x|\ \alpha_1) = \alpha_1^x(1-p)^{1-x}$
D = {(x_i,y_i)} from i to n

Formalizing the problem we get the following:
Argmax(P(D|w)) = $\sum_{i=1}^{n} \ln p\big((\alpha_1 y_i)\big|(\alpha_1, \alpha_0)\big)$
= $\sum_{i=1}^{n} \ln \alpha_1^{x_i} (1-\alpha_1)^{1-x_i} + \sum_{i=1}^{n} \ln \alpha_0^{y_i} (1-\alpha_0)^{1-y_i}$

b. Once we get our value for w* (i.e. α_0^* , α_1^*) we can evaluate P(D|w) for each of the values of w that we found and find the value of w which gives us our maximum value and that would be the probability of our table being free given that it is sunny. Therefore P(Y=1|X=1)=

c. We would have one extra random variable Z whose values can either be {Morning, Afternoon, Evening} and as it will only have 3 values, it will be a Uniform Distribution and as such will have no extra parameters.

2.

Q2a $-\ln p(x|w) = \ln\left(0.5N(x|\mu_1,\sigma_1^2) + 0.5N(x|\mu_2,\sigma^2)\right)$

$D = \{x_i\}_{i=1}^{n}$

$w = (\mu_1, \mu_2, \sigma_1, \sigma_2)$

$$c(w) = -\ln p(D|w)$$
$$= -\sum_{i=1}^{n} \ln p(x_i|w)$$
$$= -\sum_{i=1}^{n} \ln\left(0.5N(x_i|\mu_1,\sigma_1^2) + 0.5N(x_i|\mu_2,\sigma^2)\right)$$
$$= -\sum_{i=1}^{n} \ln\left(\frac{1}{2\sigma_1\sqrt{2\pi}} e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}} + \frac{1}{2\sigma_2\sqrt{2\pi}} e^{\frac{-(x_i-\mu_2)^2}{2\sigma_2^2}}\right)$$

$$\frac{\delta c(w)}{\delta \sigma_1} = -\sum_{i=1}^{n} \frac{\delta}{\delta\sigma_1} \ln\left(\frac{1}{2\sigma_1\sqrt{2\pi}} e^{\frac{-(x_i-\mu_1)^2}{2\sigma_1^2}} + \frac{1}{2\sigma_2\sqrt{2\pi}} e^{\frac{-(x_i-\mu_2)^2}{2\sigma_2^2}}\right)$$
$$= -\sum_{i=1}^{n} \frac{\delta}{\delta\sigma_1} \ln\left(\frac{e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}}}{2\sigma_1\sqrt{2\pi}} + \frac{e^{-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}}}{2\sigma_2\sqrt{2\pi}}\right)$$
$$= -\sum_{i=1}^{n} \frac{\delta}{\delta\sigma_1} \ln\left(\frac{2\sigma_2\sqrt{2\pi}\, e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}} + e^{-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}} \cdot 2\sigma_1\sqrt{2\pi}}{8\pi\sigma_1\sigma_2}\right)$$
$$= -\sum_{i=1}^{n} \frac{\delta}{\delta\sigma_1}\left(\ln\left(2\sigma_2\sqrt{2\pi}\, e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}} + 2\sigma_1\sqrt{2\pi}\, e^{\frac{-(x_i-\mu_2)^2}{2\sigma_2^2}}\right)\right) - \frac{1}{\sigma_1}$$
$$= -\sum_{i=1}^{n} \frac{\left(\frac{(x_i-\mu_1)^2 2\sigma_2\sqrt{2\pi}}{\sigma_1^3} e^{\frac{(-x_i-\mu_1)^2}{2\sigma_1^2}} + \sqrt{8\pi}\, e^{\frac{-(x_i-\mu_2)^2}{2\sigma_2^2}}\right)}{2\sigma_2\sqrt{2\pi} \times e^{\frac{-(x_i-\mu_1)^2}{2\sigma_1^2}} + 2\sigma_1\sqrt{2\pi} \times e^{\frac{-(x_i-\mu_2)^2}{2\sigma_2^2}}} - \frac{1}{\sigma_1}$$

$$\frac{\delta c(w)}{\delta \sigma_2} = -\sum_{i=1}^{n} \left(\frac{e^{\frac{(x_i-\mu_1)^2}{2\sigma_1^2}} \times \sqrt{8\pi} + \frac{(x_i-\mu_2)^2 \times 2\sigma_1 \times \sqrt{2\pi}}{\sigma_2^3} e^{\frac{(-x_i-\mu_2)^2}{2\sigma_2^2}}}{2\sigma_2 \times \sqrt{2\pi} \times e^{\frac{-(x_i-\mu_1)^2}{2\sigma_1^2}} + 2\sigma_1\sqrt{2\pi} \times e^{\frac{-(x_i-\mu_2)^2}{2\sigma_2^2}}}\right) - \frac{1}{\sigma_2}$$

$$\frac{\delta C(w)}{\delta \mu_1} = -\sum_{i=1}^{n} \frac{\delta \ln}{\delta \mu_1}\left(2\sigma_2\sqrt{2\pi}\, e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}} + 2\sigma_1\sqrt{2\pi} \times e^{-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}}\right)$$

$$= -\sum_{i=1}^{n}\left(\frac{2\sigma_2\sqrt{2\pi}\, e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}} \times (x_i-\mu_1)}{2\sigma_2\sqrt{2\pi}\, e^{-\frac{(x_i-\mu_1)^2}{2\sigma_2}} + 2\sigma_1\sqrt{2\pi} \times e^{\frac{(-x_i-\mu_2)^2}{2\sigma_2^2}}}\right)$$

$$\frac{\delta C(W)}{\delta M_2} = -\sum_{i=1}^{n}\left(\frac{2\sigma_1\sqrt{2\pi}\,(x_i-M_2)\, e^{-\frac{(x_i-M_2)^2}{2\sigma_2^2}}}{2\sigma_2\sqrt{2\pi}\, e^{-\frac{(x_i-M_1)^2}{2\sigma_2}} + 2\sigma_1\sqrt{2\pi}\, e^{\frac{-(x_i-M_1)^2}{2\sigma_2^2}}}\right)$$

$\therefore$ Our gradient is $\left(\frac{\delta C(w)}{\delta \mu_1}, \frac{\delta C(w)}{\delta M_2}, \frac{\delta C(w)}{\delta \sigma_1}, \frac{\delta C(w)}{\delta \sigma_2}\right)$

b. $w_t - \eta t\left(\frac{\delta C(w)}{\delta \mu_1}, \frac{\delta C(w)}{\delta M_2}, \frac{\delta C(w)}{\delta \sigma_1}, \frac{\delta C(W)}{\delta \sigma_2}\right) = w_{t+1}$

3.

a. Done in Julia
b. Random Regressor

    Standard Deviation of Error: 6.89002963564

    Average: 38.366856218335

  Mean Regressor

    Standard Deviation of Error: 0.149530885382

    Average: 31.651414098074

  Stochastic Regressor

    Standard Deviation of Error: 0.0892452780084

    Average: 11.981075314047

c. Done in Julia
d. Done in Julia
e. Done in Julia
f. Mini Batch Approach:

    Standard Deviation of Error: 0.0891813440887 0.0910378224164

    Average: 11.976117848765 11.987213547431

  Stochastic Approach

    Standard Deviation of Error: 0.27268064502

    Average: 13.199427112118

  From this, we can see that that while we get a larger standard error (especially for the stochastic approach as it is larger by about 0.2) when using adaptive steps, we get a larger average (Not a notable change for Mini batch while for Stochastic, we get a larger difference). This mean we can cover more data within the distribution despite a larger sample error(indicated not a more closely distributed data set).