ECE M146 HW #5
Melody Chew
#705120273

1. Show that a kernel function $K(x_1, x_2)$ satisfies the following generalization of the Cauchy-Schwartz inequality:

$$K(x_1, x_2)^2 \leq K(x_1, x_1)K(x_2, x_2).$$

Hint: The Cauchy-Schwartz inequality states that: for two vectors u and v, $|u^T v|^2 \leq \|u\|^2\|v\|^2$.

Definition of Kernel: $K(x_1, x_2) = (\phi(x_1)^T \phi(x_2))$

$$(K(x_1, x_2))^2 = (\phi(x_1)^T \phi(x_2))^2$$

$$K(x_1, x_1)\, K(x_2, x_2) = (\phi(x_1)^T \phi(x_1))(\phi(x_2)^T \phi(x_2))$$

$$= \|\phi(x_1)\|^2 \cdot \|\phi(x_2)\|^2$$

Using the Cauchy-Schwartz inequality $|u^T v|^2 \leq \|u\|^2\|v\|^2$,

$$(\phi(x_1)^T \phi(x_2)) \leq \|\phi(x_1)\|^2 \|\phi(x_2)\|^2$$

$$\Rightarrow (K(x_1, x_2))^2 \leq K(x_1, x_1)\, k(x_2, x_2)$$

2. Given valid kernels $K_1(x, x')$ and $K_2(x, x')$, show that the following kernels are also valid:

(a) $K(x, x') = K_1(x, x') + K_2(x, x')$.

(b) $K(x, x') = K_1(x, x')K_2(x, x')$.

(c) $K(x, x') = \exp(K_1(x, x'))$. Hint: use your results in (a) and (b).

a) want to show $y^T \begin{pmatrix} k(x,x) & k(x,x') \\ k(x',x) & k(x',x') \end{pmatrix} y \geq 0$

$\begin{pmatrix} k(x,x) & k(x,x') \\ k(x',x) & k(x',x') \end{pmatrix} = \begin{pmatrix} K_1(x,x)+K_2(x,x) & K_1(x,x')+K_2(x,x') \\ K_1(x',x)+K_2(x',x) & K_1(x',x')+K_2(x',x') \end{pmatrix}$ ← given.

$$[y_1, y_2] \begin{pmatrix} K_1(x,x)+K_2(x,x) & K_1(x,x')+K_2(x,x') \\ K_1(x',x)+K_2(x',x) & K_1(x',x')+K_2(x',x') \end{pmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$$= \begin{bmatrix} y_1(K_1(x,x)+K_2(x,x)) + y_2(K_1(x',x)+K_2(x',x)) & y_1(K_1(x,x')+K_2(x,x')) + y_2(K_1(x',x')+K_2(x',x')) \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$$= y_1\left(y_1(K_1(x,x)+K_2(x,x)) + y_2(K_1(x',x)+K_2(x',x))\right) + y_2\left(y_1(K_1(x,x')+K_2(x,x')) + y_2(K_1(x',x')+K_2(x',x'))\right)$$

$$= y_1^2 K_1(x,x) + y_1 y_2 K_1(x',x) + y_1 y_2 K_1(x,x') + y_2^2 K_1(x',x')$$
$$+ y_1^2 K_2(x,x) + y_1 y_2 K_2(x',x) + y_1 y_2 K_2(x,x') + y_2^2 K_2(x',x')$$

$$= \underbrace{[y_1, y_2] K_1(x,x') \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}}_{\text{we know that this is PSD}} + \underbrace{[y_1, y_2] K_2(x,x') \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}}_{\text{also PSD.}} \geq 0$$

b) Let $x_1, \ldots, x_n \in S$, let $u \in \mathbb{R}^n$ want to show that there exist $u^T E u \geq 0$. then $K(x, x')$ is PSD.

Gram matrix representation of our kernels.

$C = (C_{ij})$ $C_{ij} = K_1(x_i, x_j)$ $C = A^T A$ by kernel property. $A = (a_1 \ldots a_n)$

$D = (d_{ij})$ $d_{ij} = K_2(x_i, x_j)$ $D = B^T B$ $B = (b_1 \ldots b_n)$

$E = (e_{ij})$ $e_{ij} = C_{ij} d_{ij}$ $C_{ij} = a_i^T a_j = \sum_k a_{ik} a_{jk}$ $d_{ij} = b_i^T b_j = \sum_\ell b_{i\ell} b_{j\ell}$

$$u^T E u = \sum_{ij} u_i u_j C_{ij} d_{ij} = \sum_{ij} u_i u_j \sum_k a_{ik} a_{jk} \sum_\ell b_{i\ell} b_{j\ell}$$

$$= \sum_{ij} \sum_{k\ell} u_i u_j a_{ik} a_{jk} b_{i\ell} b_{j\ell} = \sum_{k\ell} \left(\sum_i u_i a_{ik} b_{i\ell}\right)\left(\sum_j u_j a_{jk} b_{j\ell}\right)$$

$$= \sum_{k\ell} \left(\sum_i u_i a_{ik} b_{i\ell}\right)^2 \geq 0.$$

c) We rewrite $K(x, x')$ with its Taylor Series form:

$$K(x, x') = \sum_{n=0}^{\infty} \frac{K_1(x, x')^n}{n!}$$

Since we have proved from (a) that sum of kernel is still valid kernel and that product of valid kernel is still valid kernel, we know that the above summation of product of kernel is still a PSD kernel.

3. In class, we learned that the soft margin SVM has the primal problem:

$$\begin{aligned} \min_{\xi,w,b} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ s.t. \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \cdots, m \\ & \xi_i \geq 0, \quad i = 1, \cdots, m, \end{aligned}$$

and the dual problem:

$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)}, x^{(j)}\rangle \\ s.t. \quad & 0 \leq \alpha_i \leq C, i = 1, \cdots, m, \\ & \sum_{i=1}^{m}\alpha_i y^{(i)} = 0. \end{aligned}$$

Note that $\langle z, s\rangle$ is an alternative expression for the inner product $z^T s$. As usual, $y^{(i)} \in \{+1, -1\}$.

Now suppose we have solved the dual problem and have the optimal α. Show that the parameter b can be determined using the following equation:

$$b = \frac{1}{N_{\mathcal{M}}}\sum_{n\in\mathcal{M}}\left(y^{(n)} - \sum_{m\in\mathcal{S}}\alpha_m y^{(m)}\langle x^{(n)}, x^{(m)}\rangle\right). \tag{1}$$

In (1), $\mathcal{M}$ denotes the set of indices of data points having $0 < \alpha_n < C$, parameter $N_{\mathcal{M}}$ denotes the size of the set $\mathcal{M}$, and $\mathcal{S}$ denotes the set of indices of data points having $\alpha_n \neq 0$.

For all points in $\mathcal{M}$, we know that:

$$y^{(n)}(w^T x_n + b) = 1$$

We know that $y^{(n)} = w^T x^{(n)} + b$

$$w = \sum_{m\in\mathcal{S}} \alpha_m y^{(m)} x^{(m)}$$

$$y^{(n)} = \left(\sum_{m\in\mathcal{S}} \alpha_m y^{(m)} x^{(m)}\right)^T x_n + b$$

$$b = y^{(n)} - \left(\sum_{m\in\mathcal{S}} \alpha_m y^{(m)} x^{(m)}\right)^T x_n$$

We want to find stable b, so we average b over all points in $\mathcal{M}$.

$$b = \frac{1}{N_{\mathcal{M}}}\sum_{n\in\mathcal{M}}\left(y^{(n)} - \left(\sum_{m\in\mathcal{S}} \alpha_m y^{(m)}\langle x^{(n)}, x^{(m)}\rangle\right)\right)$$

4. Consider 3 random variables A,B and C with joint probabilities $P(A, B, C)$ listed in the following table.

	C=0		C=1	
	B=0	B=1	B=0	B=1
A=0	0.096	0.024	0.27	0.03
A=1	0.224	0.056	0.27	0.03

(a) Calculate $P(A|C=0)$, $P(B|C=0)$, and $P(A, B|C=0)$.

(b) Calculate $P(A|C=1)$, $P(B|C=1)$, and $P(A, B|C=1)$.

(c) Is A conditionally independent of B given C?

(d) Calculate $P(A)$, $P(B)$, and $P(A, B)$.

(e) Is A independent of B?

a) $P(A|C=0) = \begin{cases} \frac{P(A=1,C=0)}{P(C=0)} = \frac{0.28}{0.4} = 0.7, & A=1 \\ \frac{P(A=0,C=0)}{P(C=0)} = 0.3, & A=0 \end{cases}$

$P(B|C=0) = \begin{cases} \frac{P(B=1,C=0)}{P(C=0)} = \frac{0.08}{0.4} = 0.2, & B=1 \\ 0.8, & B=0 \end{cases}$

$P(A,B|C=0) = \begin{cases} A=1, B=1, & \frac{P(A=1,B=1,C=0)}{P(C=0)} = \frac{0.056}{0.4} = 0.14 \\ A=0, B=1, & \frac{P(A=0,B=1,C=0)}{P(C=0)} = \frac{0.024}{0.4} = 0.06 \\ A=1, B=0, & \frac{P(A=1,B=0,C=0)}{P(C=0)} = \frac{0.224}{0.4} = 0.56 \\ A=0, B=0, & 0.24 \end{cases}$

b) $P(A|C=1) = \begin{cases} A=0, & \frac{P(A=0,C=1)}{P(C=1)} = \frac{0.3}{0.6} = 0.5 \\ A=1, & 0.5 \end{cases}$

$P(B|C=1) = \begin{cases} B=0, & \frac{P(B=0,C=1)}{P(C=1)} = \frac{0.54}{0.6} = 0.9 \\ B=1, & 0.1 \end{cases}$

$P(A,B|C=1) = \begin{cases} A=0, B=0 = \frac{P(A=0,B=0,C=1)}{P(C=1)} = \frac{0.27}{0.6} = 0.45 \\ A=0, B=1 = \frac{P(A=0,B=1,C=1)}{P(C=1)} = \frac{0.03}{0.6} = 0.05 \\ A=1, B=0 = \frac{P(A=1,B=0,C=1)}{P(C=1)} = \frac{0.27}{0.6} = 0.45 \\ A=1, B=1, & 0.05 \end{cases}$

c) Conditionally independent if $P(A \cap B|C) = P(A|C)P(B|C)$.

yes, $P(A=0|C=0) \cdot P(B=0|C=0) = P(A=0, B=0|C=0)$

$P(A=1|C=0) \cdot P(B=1|C=0) = P(A=1, B=1|C=0)$

$\vdots$

true for all combinations of A, B, C.

d)

$$p(A) = \begin{cases} A=0, & 0.42 \\ A=1, & 0.58 \end{cases} \qquad p(B) = \begin{cases} B=0, & 0.86 \\ B=1, & 0.14 \end{cases}$$

$$p(A,B) = \begin{cases} A=0, B=0, & 0.366 \\ A=0, B=1, & 0.054 \\ A=1, B=0, & 0.494 \\ A=1, B=1, & 0.086 \end{cases}$$

e) A is not independent of B.

Ex. $p(A=0)p(B=0) \neq p(A=0, B=0)$.

5. Let us revisit the restaurant selection problem in HW3. You are trying to choose between two restaurants (sample 9 and sample 10) to eat at. To do this, you will train a classifier based on your past experiences (sample 1-8). The features for each restaurants and your judgment on the goodness of sample 1-8 are summarized by the following chart. In this exercise, instead of a decision tree, you will use the Naïve

Sample #	HasOutdoorSeating	HasBar	IsClean	HasGoodAtmosphere	IsGoodRestaurant
1	0	0	0	1	1
2	1	1	0	0	0
3	0	1	1	1	1
4	1	0	0	1	1
5	1	1	1	0	0
6	1	0	1	0	1
7	1	1	0	1	1
8	0	0	1	1	1
9	0	1	0	1	?
10	1	1	1	1	?

Bayes classifier to decide whether restaurant 9 and 10 are good or not. For clarity, we abbreviate the names of the features and label as follows: HasOutdoorSeating $\to O$, HasBar $\to B$, IsClean $\to C$, HasGoodAtmosphere $\to A$, and IsGoodRestaurant $\to G$.

(a) Train the Naïve Bayes classifier by calculating the maximum likelihood estimate of class priors and class conditional distributions. Namely, calculate the maximum likelihood estimate of the following: $P(G)$, and $P(X|G), X \in \{O, B, C, A\}$.

(b) For Sample #9 and #10, make the decision using

$$\hat{G}_i = \underset{G_i \in \{0,1\}}{\text{argmax}} \quad P(G_i)P(O_i, B_i, C_i, A_i|G_i),$$

where O_i, B_i, C_i, and A_i are the feature values for the i-th sample.

(c) We use Laplace smoothing to avoid having class conditional probabilities that are strictly 0. To use Laplace smoothing for a binary classifier, add 1 to the numerator and add 2 to the denominator when calculating the class conditional distributions. Let us re-calculate the class conditional distributions with Laplace smoothing. Namely, calculate the maximum likelihood estimate of $P(X|G), X \in \{O, B, C, A\}$.

(d) Repeat (b) with the class conditional distributions you get from (c).

a) $P(G=0)=\frac{2}{8}$ $P(G=1)=\frac{6}{8}$

$P(O=0|G=0)=0$
$P(O=0|G=1)=\frac{3}{6}=\frac{1}{2}$
$P(O=1|G=0)=1$
$P(O=1|G=1)=\frac{1}{2}$
$P(B=0|G=0)=0$
$P(B=0|G=1)=\frac{4}{6}=\frac{2}{3}$
$P(B=1|G=0)=1$
$P(B=1|G=1)=\frac{2}{6}=\frac{1}{3}$
$P(C=0|G=0)=\frac{1}{2}$
$P(C=0|G=1)=\frac{1}{2}$
$P(C=1|G=0)=\frac{1}{2}$
$P(C=1|G=1)=\frac{1}{2}$
$P(A=0|G=0)=1$
$P(A=0|G=1)=\frac{1}{6}$
$P(A=1|G=0)=0$
$P(A=1|G=1)=\frac{5}{6}$

b) $i=9$

$P(G=1)P(O=0|G=1)P(B=1|G=1)P(C=0|G=1)P(A=1|G=1)$
$=\frac{3}{4}\cdot\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{2}\cdot\frac{5}{6}=0.05208$

$P(G=0)P(O=0|G=0)P(B=1|G=0)P(C=0|G=0)P(A=1|G=0)$
$=\frac{1}{4}\cdot 0\cdot 1\cdot\frac{1}{2}\cdot 0=0$

$G_9=1$

$i=10$

$P(G=1)P(O=1|G=1)P(B=1|G=1)P(C=1|G=1)P(A=1|G=1)$
$=\frac{3}{4}\cdot\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{2}\cdot\frac{5}{6}=0.05208$

$P(G=0)P(O=1|G=0)P(B=1|G=0)P(C=1|G=0)P(A=1|G=0)$
$=\frac{1}{4}\cdot 1\cdot 1\cdot\frac{1}{2}\cdot 0=0$

$G_{10}=1$

c) $P(G=0)=\frac{2}{8}=\frac{1}{4}$ $P(G=1)=\frac{6}{8}=\frac{3}{4}$

$P(O=0|G=0)=\frac{1}{4}$
$P(O=0|G=1)=\frac{4}{8}=\frac{1}{2}$
$P(O=1|G=0)=\frac{3}{4}$
$P(O=1|G=1)=\frac{5}{10}=\frac{1}{2}$

$P(B=0|G=0)=\frac{1}{4}$
$P(B=0|G=1)=\frac{5}{8}$
$P(B=1|G=0)=\frac{3}{4}$
$P(B=1|G=1)=\frac{3}{8}$

$P(C=0|G=0)=\frac{1}{2}$
$P(C=0|G=1)=\frac{1}{2}$
$P(C=1|G=0)=\frac{1}{2}$
$P(C=1|G=1)=\frac{1}{2}$

$P(A=0|G=0)=\frac{3}{4}$
$P(A=0|G=1)=\frac{1}{4}$
$P(A=1|G=0)=\frac{1}{4}$
$P(A=1|G=1)=\frac{3}{4}$

d) $i = 9$

$P(G=1)P(O=0|G=1)P(B=1|G=1)P(C=0|G=1)P(A=1|G=1)$

$= \frac{3}{4} \cdot \frac{1}{2} \cdot \frac{3}{8} \cdot \frac{1}{2} \cdot \frac{3}{4} = 0.0527$

$P(G=0)P(O=0|G=0)P(B=1|G=0)P(C=0|G=0)P(A=1|G=0)$

$= \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{3}{4} \cdot \frac{1}{2} \cdot \frac{1}{4} = 0.00586$

$G_9 = 1.$

$i = 10$

$P(G=1)P(O=1|G=1)P(B=1|G=1)P(C=1|G=1)P(A=1|G=1)$

$= \frac{3}{4} \cdot \frac{1}{2} \cdot \frac{3}{8} \cdot \frac{1}{2} \cdot \frac{3}{4} = 0.0527$

$P(G=0)P(O=1|G=0)P(B=1|G=0)P(C=1|G=0)P(A=1|G=0)$

$= \frac{1}{4} \cdot \frac{3}{4} \cdot \frac{3}{4} \cdot \frac{1}{2} \cdot \frac{1}{4} = 0.01757$

$G_{10} = 1.$

6. In class, we learned a Naïve Bayes classifier for binary feature values, i.e., $x_j \in \{0,1\}$ where we model the class conditional distribution to be Bernoulli. In this exercise, you are going to extend the result to the case where features that are non-binary.

We are given a training set $\{(x^{(i)}, y^{(i)}); i = \{1, \cdots, m\}\}$, where $x^{(i)} \in \{1, 2, \cdots, s\}^n$ and $y^{(i)} \in \{0, 1\}$. Again, we model the label as a biased coin with $\theta_0 = P(y^{(i)} = 0)$ and $1 - \theta_0 = P(y^{(i)} = 1)$. We model each non-binary feature value $x_j^{(i)}$ (an element of $x^{(i)}$) as a biased dice for each class. This is parameterized by:

$$P(x_j = k|y = 0) = \theta_{j,k|y=0},\ k = 1, \cdots, s-1;$$
$$P(x_j = s|y = 0) = \theta_{j,s|y=0} = 1 - \sum_{k=1}^{s-1} \theta_{j,k|y=0};$$
$$P(x_j = k|y = 1) = \theta_{j,k|y=1},\ k = 1, \cdots, s-1;$$
$$P(x_j = s|y = 1) = \theta_{j,s|y=1} = 1 - \sum_{k=1}^{s-1} \theta_{j,k|y=1};$$

Notice that we do not model $P(x_j = s|y = 0)$ and $P(x_j = s|y = 1)$ directly. Instead we use the above equations to guarantee all probabilities for each class sum to 1.

(a) Using the **Naïve Bayes (NB) assumption**, write down the joint probability of the data:

$$P(x^{(1)}, \cdots, x^{(m)}, y^{(1)}, \cdots, y^{(m)})$$

in terms of the parameters θ_0, $\theta_{j,k|y=0}$ and $\theta_{j,k|y=1}$. You may find the indicator function $\mathbf{1}(\cdot)$ useful.

(b) Now, maximize the joint probability you get in (a) with respect to each of θ_0, $\theta_{j,k|y=0}$, and $\theta_{j,k|y=1}$. Write down your resulting θ_0, $\theta_{j,k|y=0}$ and $\theta_{j,k|y=1}$ and show intermediate steps. Explain in words the meaning of your results.

a) $\prod_{i=1}^{m} p(x_i, y_i) = \prod_{i=1}^{m} p(y_i) \prod_{j=1}^{n} \prod_{k=1}^{s} p(x_{j,k,i} \mid y_i)$

$= \prod_{i=1}^{m} \theta_0^{1[y_i=0]} (1-\theta_0)^{1[y_i=1]} \cdot \prod_{j=1}^{n} \prod_{k=1}^{s} \theta_{j,k|y=0}^{1\cdot[x_{j,i}=k, y_i=0]} \theta_{j,k|y=1}^{1\cdot[x_{j,i}=k, y_i=1]}$

b) We take the negative log of equation above: (also works if we take positive log).

$$J = -\sum_{i=1}^{m} \{1[y_i=0] \log \theta_0 + 1[y_i=1] \log(1-\theta_0) + \sum_{j=1}^{n} \sum_{k=1}^{s} 1[x_{j,i}=k, y_i=0] \log(\theta_{j,k|y=0}) + 1[x_{j,i}=k, y_i=1] \log(\theta_{j,k|y=1})\}$$

$$\frac{\partial J}{\partial \theta_0} = -\sum_{i=1}^{m} \left(\frac{1[y_i=0]}{\theta_0} - \frac{1[y_i=1]}{1-\theta_0} \right) = 0$$

$$= -\sum_{i=1}^{m} \frac{1[y_i=0]}{\theta_0} + \sum_{i=1}^{m} \frac{1[y_i=1]}{1-\theta_0} = 0$$

$$\frac{\sum_{i=1}^{m} 1[y_i=0]}{\theta_0} = \frac{\sum_{i=1}^{m} 1[y_i=1]}{1-\theta_0}$$

$$(1-\theta_0) \sum_{i=1}^{m} 1[y_i=0] = \theta_0 \sum_{i=1}^{m} 1[y_i=1]$$

$$\sum_{i=1}^{m} 1[y_i=0] = \theta_0 \sum_{i=1}^{m} 1[y_i=1] + \theta_0 \sum_{i=1}^{m} 1[y_i=0]$$

$$\theta_0 = \frac{\sum_{i=1}^{m} 1[y_i=0]}{m}$$

$\theta_0 = p(y_i=0)$ means that probability of $y_i=0$ for a data point w/o other information is equal to frequency of $[y_i=0]$ for all data points from $i = 1 \ldots m$.

We take derivative w.r.t a particular $\theta_{j,k|y=0}$, since $\theta_{j,s|y=0}$ is not included in $\theta_{j,k|y=0}$, we need to exclude it from derivative below.

$$\frac{\partial J}{\partial \theta_{j,k|y=0}} = -\sum_{i=1}^{m} \frac{1[x_{j,i}=k, y_i=0]}{\theta_{j,k|y=0}} + \sum_{i=1}^{m} \frac{1[x_{j,i}=s, y_i=0]}{\theta_{j,s|y=0}} = 0$$

$$\sum_{i=1}^{m} 1[x_{j,i}=s, y_i=0] \cdot \theta_{j,k|y=0} = \sum_{i=1}^{m} 1[x_{j,i}=k, y_i=0] \cdot \theta_{j,s|y=0}$$

$$\theta_{j,k|y=0} = \frac{\sum_{i=1}^{m} 1[x_{j,i}=k, y_i=0]}{\sum_{i=1}^{m} 1[x_{j,i}=s, y_i=0]} \cdot \theta_{j,s|y=0}$$

$$\theta_{j,s|y=0} = 1 - \sum_{k=1}^{s-1} \theta_{j,k|y=0} = \frac{\sum_{i=1}^{m} 1[x_{ij}=s, y_i=0]}{\sum_{i=1}^{m} 1[y_i=0]}$$ ← by definition of $\theta_{j,s|y=0}$.

$$\theta_{j,k|y=0} = \frac{\sum_{i=1}^{m} 1[x_{j,i}=k, y_i=0]}{\sum_{i=1}^{m} 1[x_{j,i}=s, y_i=0]} \left(1 - \sum_{k=1}^{s-1} \theta_{j,k|y=0}\right)$$

$$= \frac{\sum_{i=1}^{m} 1[x_{j,i}=k, y_i=0]}{\sum_{i=1}^{m} 1[x_{j,i}=s, y_i=0]} \cdot \frac{\sum_{i=1}^{m} 1[x_{ij}=s, y_i=0]}{\sum_{i=1}^{m} 1[y_i=0]} = \frac{\sum_{i=1}^{m} 1[x_{ij}=k, y_i=0]}{\sum_{i=1}^{m} 1[y_i=0]}$$

This means that the probability of $p(x_j=k|y=0)$ $k=1,\dots,s-1$, depends on the frequency i.e. the total # of times event $x_{i,j}=k$ and $y_i=0$ occurs in our dataset out of the total number of times event $y_i=0$ occurs. This clearly relates to Bayes theorem of conditional probability: $P(A|B) = \frac{P(A,B)}{P(B)}$.

For $\theta_{j,k|y=1}$, we follow the same steps above, setting $\frac{\partial J}{\partial \theta_{j,k|y=1}} = 0$, excluding $\theta_{j,s|y=1}$, doing the same algebra, substituting in

$$\theta_{j,s|y=1} = 1 - \sum_{k=1}^{s-1} \theta_{j,k|y=1} = \frac{\sum_{i=1}^{m} 1[x_{ij}=s, y_i=1]}{\sum_{i=1}^{m} 1[y_i=1]},$$

and finally getting $\theta_{j,k|y=1} = \frac{\sum_{i=1}^{m} 1[x_{ij}=k, y_i=1]}{\sum_{i=1}^{m} 1[y_i=1]}$.

This means that $\theta_{j,k|y=1} = p(x_{ij}=k|y=1)$ is the relative frequency of the event $x_{ij}=k \cap y_i=1$ occurring out of all the event of $y_i=1$ in our data set. This makes sense as conditional probability base on bayes theorem of $p(A|B) = \frac{P(A,B)}{P(B)}$.