1] Show that a kernel function $K(x_1, x_2)$ satisfies the Cauchy-Schwarz generalization $K(x_1, x_2)^2 \leq K(x_1, x_1) K(x_2, x_2)$.

↳ Hint: $|u^T v|^2 \leq \|u\|^2 \|v\|^2 \quad \forall \; u, v \in \mathbb{R}^n$

Consider a feature mapping $\phi: \mathbb{R}^m \to \mathbb{R}^n$.

Then, the corresponding kernel is defined as $K(x_1, x_2) = \phi(x_1)^T \phi(x_2) \quad \forall x_1, x_2 \in \mathbb{R}^m$.

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

$= |\phi(x_1)^T \phi(x_2)|^2$ b/c $x^2 = |x|^2 \quad \forall \; x \in \mathbb{R}$

$\leq \|\phi(x_1)\|^2 \|\phi(x_2)\|^2$ by cauchy-schwarz

$= \phi(x_1)^T \phi(x_1) \; \phi(x_2)^T \phi(x_2)$ b/c $u^T u = \|u\|^2$

$= K(x_1, x_1) \; K(x_2, x_2)$ by definition.

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

□

Zack Beyer

2] Given $K_1(x,x')$ and $K_2(x,x')$ are valid, Prove the following is valid.

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

$K_1(x,x')$ and $K_2(x,x')$ are valid Kernels, and so are symmetric.

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

$\therefore K(x,x')$ is symmetric.

The kernel matrix K for $K(x,x')$ is defined as $K = K_1 + K_2$.
(Kernel matrices for $K_1(x,x')$ and $K_2(x,x')$.)

Consider some arbitrary vector y. Then,

$$\begin{aligned} y^T K y &= y^T (K_1 + K_2) y \\ &= y^T K_1 y + y^T K_2 y \\ &\geq 0 + 0 \\ &= 0 \end{aligned}$$

$\therefore K$ is positive semidefinite.

Hence, $K(x,x')$ is a valid kernel.

Generally, to prove K is valid, must show it is symmetric and its matrix is positive semidefinite.

□

b) $K(x,x') = K_1(x,x')\,K_2(x,x')$. The kernel matrix K is defined as $K = K_1 \circ K_2$. (Pointwise multiplication)

Positive semi-definite:

Consider some arbitrary vector y.

$$\begin{aligned} y^T K y &= y^T (K_1 \circ K_2) y \\ &= y^T K_1 y \circ y^T K_2 y \\ &\geq 0 \circ 0 \\ &= 0 \end{aligned}$$

$\therefore K$ is positive semidefinite.

Symmetry:

Also, since $K_1(x,x')$ and $K_2(x,x')$ are symmetric,

$$\begin{aligned} K(x,x') &= K_1(x,x')\,K_2(x,x') \\ &= K_1(x',x)\,K_2(x',x) \\ &= K(x',x) \end{aligned}$$

$\therefore K$ is symmetric.

Hence, $K(x,x')$ is a valid kernel.

□

c) $K(x,x') = \exp(K_1(x,x'))$.

Suppose K_1, K_2 are valid kernels.

Claim ①: If $\alpha, \beta \geq 0$, then $K(x,x') := \alpha K_1(x,x') + \beta K_2(x,x')$ is a valid kernel.

Claim ②: If f is a polynomial with positive coefficients, then $K(x,x') := f(K_1(x,x'))$ is a valid kernel.

By definition, $\exp(x) = \sum_{i=0}^{\infty} \frac{x^i}{i!}$ is a polynomial with positive coefficients

Then, $\exp(K_1(x,x'))$ is equivalent to writing an infinite polynomial with positive coefficients as a function of $K_1(x,x')$.

By claim ②, $K = \exp(K_1(x,x'))$ is a valid kernel.

□

Proof of claim ①: Similar to part a, except $K = \alpha K_1 + \beta K_2$ and we note $y^T K y = \alpha y^T K_1 y + \beta y^T K_2 y \geq 0$.

□

Proof of claim ②: Each polynomial term is a product of kernels multiplied by some positive coefficient.

By ①, each term is a valid kernel.

By ①, the sum of the terms is a valid kernel multiplied by a coefficient

□

2] Show that the parameter b can be determined using

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

M = index set for data s.t. $0 < \alpha_n < C$
S = index set for data s.t. $\alpha_n \neq 0$.

Consider Primal Problem of Soft SVM.

The Lagrangian is $\mathcal{L} = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i + \sum_{i=1}^{m} \alpha_i (1 - \xi_i - y_i(w^T x_i + b)) - \sum_{i=1}^{m} \beta_i \xi_i$

→ Here, α_i and β_i denote Lagrange multipliers. $\alpha_i, \beta_i \geq 0$.

Taking derivatives, $\frac{\partial \mathcal{L}}{\partial w} = \frac{1}{2} \cdot 2w - \sum_{i=1}^{m} \alpha_i y_i x_i$.

Taking $\frac{\partial \mathcal{L}}{\partial w} = 0$, obtain $w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \leftrightarrow w = \sum_{i=1}^{m} \alpha_i y_i x_i$
at the desired optimum.

Let S denote the set of indices of data points having $\alpha_i \neq 0$.

Then, $w = \sum_{i=1}^{m} \alpha_i y_i x_i = \sum_{m \in S} \alpha_m y_m x_m = w$

Now, let M denote the index set of points x_i for which $0 < \alpha_i < C$.

↳ Recall the complementary slackness conditions at the optimum,

- $\alpha_i (y_i(w^T x_i + b) - 1 + \xi_i) = 0$
- $\beta_i \xi_i = 0$

Also note that $\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \beta_i = 0$ at the optimum (KKT conditions).

Because $0 < \alpha_i < C$ and $C - \alpha_i - \beta_i = 0$, $0 < \beta_i < C$.

Because $0 < \beta_i$ and $\beta_i \xi_i = 0$, $\xi_i = 0$.

Because $\alpha_i > 0$ and $\alpha_i (y_i (w^T x_i + b) - 1 + \xi_i) = 0$, $y_i(w^T x_i + b) - 1 + \xi_i = 0$.

Because $\xi_i = 0$ and $y_i(w^T x_i + b) - 1 + \xi_i = 0$, $y_i(w^T x_i + b) - 1 = 0$.

$\Rightarrow y_i(w^T x_i + b) = 1$

By the above argument, for points x_i whose index is in M (i.e. $0 < \alpha_i < C$), $y_i(w^T x_i + b) = 1$.

→

We can algebraically solve for b for any point whose index $\in M$:

$$y_i(w^T x_i + b) = 1$$

$$\updownarrow$$

$$w^T x_i + b = \frac{1}{y_i}$$

$$\updownarrow$$

$$w^T x_i + b = y_i \quad (\text{b/c } y_i \in \{+1, -1\},\ \frac{1}{y_i} = y_i)$$

$$\updownarrow$$

$$\underline{b = y_i - w^T x_i}$$

Using $w = \sum_{m \in S} a_m y_m x_m$ as before, $b = y_i - \left(\sum_{m \in S} a_m y_m x_m\right)^T x_i$

$$\Rightarrow b = y_i - \sum_{m \in S} a_m y_m \langle x_i, x_m \rangle$$

If we average b over all vectors whose index is in the set M, then we obtain the desired result, assuming we found optimal a:

$$b = \frac{1}{N_M} \sum_{n \in M} \left(y_n - \sum_{m \in S} a_m y_m \langle x_n, x_m \rangle \right)$$

□

4] Consider RVs A, B, C with listed joint $P(A,B,C)$. Marginals computed by summing rows / columns...

	$C=0$		$C=1$	
	$B=0$	$B=1$	$B=0$	$B=1$
$A=0$	0.096	0.024	0.24	0.03
$A=1$	0.224	0.056	0.27	0.03

$P(A=0) = 0.42$

$P(A=1) = 0.58$

$P(C=0) = 0.4 \quad P(C=1) = 0.6$

$P(B=0) = 0.86 \quad P(B=1) = .14$

a) $P(A \mid C=0) = \dfrac{P(A, C=0)}{P(C=0)} = \begin{cases} A=0: \dfrac{(.096+.024)}{0.40} = 0.3 \\ A=1: \dfrac{(0.224+0.056)}{0.40} = 0.7 \end{cases}$

$P(B \mid C=0) = \dfrac{P(B, C=0)}{P(C=0)} = \begin{cases} B=0: \dfrac{(0.096+0.224)}{0.40} = 0.8 \\ B=1: \dfrac{(0.024+0.056)}{0.40} = 0.2 \end{cases}$

$P(A, B \mid C=0) = \begin{cases} A=0, B=0: \dfrac{0.096}{0.40} = 0.24 \\ A=0, B=1: \dfrac{0.024}{0.40} = 0.06 \end{cases} \quad \begin{cases} A=1, B=0: \dfrac{0.224}{0.40} = 0.56 \\ A=1, B=1: \dfrac{.056}{.40} = .14 \end{cases}$

$P(A \mid C=0) = 0.3$ if $A=0$, 0.7 if $A=1$

$P(B \mid C=0) = 0.8$ if $B=0$, 0.2 if $B=1$

$P(A,B \mid C=0) = 0.24$ if $A,B=0$, 0.14 if $A,B=1$,
0.06 if $A=0, B=1$
0.56 if $A=1, B=0$

b) $P(A|C=1) = \begin{cases} A=0: & \frac{(0.27+0.03)}{0.6} = \frac{1}{2} \\ A=1: & \frac{(0.27+0.03)}{0.6} = \frac{1}{2} \end{cases}$

$P(B|C=1) = \begin{cases} B=0: & \frac{(0.27+0.27)}{0.6} = 0.9 \\ B=1: & \frac{(0.03+0.03)}{0.6} \quad 0.1 \end{cases}$

$P(A,B|C=1) = \begin{cases} A=0,B=0: & \frac{0.27}{0.60} = 0.45 \\ A=0,B=1: & \frac{0.03}{0.60} = 0.05 \end{cases} \quad \begin{cases} A=1,B=0: & \frac{0.27}{0.60} = 0.45 \\ A=1,B=1: & \frac{0.03}{0.60} = 0.05 \end{cases}$

$P(A|C=1) = 0.5$ if $A=0$, 0.5 if $A=1$

$P(B|C=1) = 0.9$ if $B=0$, 0.1 if $B=1$

$P(A,B|C=1) = 0.45$ if $A,B=0$, 0.05 if $A,B=1$,
0.05 if $A=0, B=1$
0.45 if $A=1, B=0$

c) 8 cases to check to see if A conditionally independent of B given C.

- $P(A=0,B=0|C=0) = 0.24 = 0.24 = P(A=0|C=0)P(B=0|C=0)$
- $P(A=0,B=1|C=0) = 0.06 = 0.06 = P(A=0|C=0)P(B=1|C=0)$
- $P(A=1,B=0|C=0) = 0.56 = 0.56 = P(A=1|C=0)P(B=0|C=0)$
- $P(A=1,B=1|C=0) = 0.14 = 0.14 = P(A=1|C=0)P(B=1|C=0)$
- $P(A=0,B=0|C=1) = 0.45 = 0.45 = P(A=0|C=1)P(B=0|C=1)$
- $P(A=0,B=1|C=1) = 0.05 = 0.05 = P(A=0|C=1)P(B=1|C=1)$
- $P(A=1,B=0|C=1) = 0.45 = 0.45 = P(A=1|C=1)P(B=0|C=1)$
- $P(A=1,B=1|C=1) = 0.05 = 0.05 = P(A=1|C=1)P(B=1|C=1)$

$\Rightarrow$ A is conditionally independent of B given C **yes**

d) $P(A) = \begin{cases} A=0: 0.42 \\ A=1: 0.58 \end{cases}$ $P(B) = \begin{cases} B=0: 0.86 \\ B=1: 0.14 \end{cases}$

$P(A,B) = \begin{cases} A=0, B=0: 0.27 + 0.096 = 0.366 \\ A=1, B=1: 0.056 + 0.03 = 0.086 \\ A=0, B=1: 0.024 + 0.03 = .054 \\ A=1, B=0: 0.224 + 0.27 = 0.494 \end{cases}$

$P(A) = 0.42$ if $A=0$, 0.58 if $A=1$

$P(B) = 0.86$ if $B=0$, 0.14 if $B=1$

$P(A,B) = 0.366$ if $A,B=0$, 0.086 if $A,B=1$
0.054 if $A=0, B=1$
0.494 if $A=1, B=0$

e) 4 cases to consider to see if A is independent of B.

- $P(A=0, B=0) = 0.366 \neq 0.362 = P(A=0)P(B=0)$

$\Rightarrow$ Since $P(A=0, B=0) \neq P(A=0)P(B=0)$, A is NOT independent of B. **NO**

5] Use Naive Bayes to decide if 9, 10 good or not.

a) Calculate MLE of $P(G)$, $P(x|G)$ for $x \in \{O, B, C, A\}$.

↳ $P(G=1) = \frac{6}{8}$

$P(O=1|G=1) = \frac{P(O=1, G=1)}{P(G=1)} = \frac{3/8}{6/8} = \frac{3}{6}$ $\quad P(O=1|G=0) = \frac{2}{2}$

$P(B=1|G=1) = 2/6$ $\quad P(B=1|G=0) = 2/2$

$P(C=1|G=1) = 3/6$ $\quad P(C=1|G=0) = 1/2$

$P(A=1|G=1) = 5/6$ $\quad P(A=1|G=0) = 0/2$

b) Decide on #9, #10 with $G_i = \underset{G \in \{0,1\}}{\text{argmax}}\ P(G_i) P(O_i, B_i, C_i, A_i | G_i)$.

↳ use conditional independence to write

$P(O_i, B_i, C_i, A_i | G_i) = P(O_i|G_i) P(B_i|G_i) P(C_i|G_i) P(A_i|G_i)$

- For sample #9, $(O=0, B=1, C=0, A=1)$

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

$= (\frac{2}{8})(\frac{0}{2})(\frac{2}{2})(\frac{1}{2})(\frac{0}{2}) = 0$

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

$= (\frac{6}{8})(\frac{3}{6})(\frac{2}{6})(\frac{3}{6})(5/6) = \frac{5}{96}$

Since $\frac{5}{96} > 0$, $G = 1$

- For sample #10, $(O=1, B=1, C=1, A=1)$

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

$= (\frac{2}{8})(\frac{2}{2})(\frac{2}{2})(1/2)(0/2) = 0$

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

$= (\frac{6}{8})(\frac{3}{6})(\frac{2}{6})(\frac{3}{6})(5/6) = \frac{5}{96}$

Since $5/96 > 0$, $G = 1$

Sample #9: $G = 1$
Sample #10: $G = 1$

c) Redo part (a) using Laplace Smoothing for $P(X|G)$.

- $P(O=1|G=1) = \frac{3+1}{6+2} = \frac{4}{8}$
- $P(B=1|G=1) = \frac{2+1}{6+2} = \frac{3}{8}$
- $P(C=1|G=1) = \frac{3+1}{6+2} = \frac{4}{8}$
- $P(A=1|G=1) = \frac{5+1}{6+2} = \frac{6}{8}$
- $P(O=1|G=0) = \frac{2+1}{2+2} = \frac{3}{4}$
- $P(B=1|G=0) = \frac{2+1}{2+2} = \frac{3}{4}$
- $P(C=1|G=0) = \frac{1+1}{2+2} = \frac{2}{4}$
- $P(A=1|G=0) = \frac{0+1}{2+2} = \frac{1}{4}$

d) Repeat (b) with data from (c).

- For sample #9...
  - → $G=0$: $P(G=0)P(O=0|G=0)P(B=1|G=0)P(C=0|G=0)P(A=1|G=0)$
    $= (\frac{2}{8})(\frac{1}{4})(\frac{3}{4})(\frac{2}{4})(\frac{1}{4}) = \frac{3}{512}$
  - → $G=1$: $P(G=1)P(O=0|G=1)P(B=1|G=1)P(C=0|G=1)P(A=1|G=1)$
    $= (\frac{6}{8})(\frac{4}{8})(\frac{3}{8})(\frac{4}{8})(\frac{6}{8}) = \frac{27}{512}$

  Since $\frac{27}{512} > \frac{3}{512}$, $G = 1$

- For sample #10...
  - → $G=0$: $P(G=0)P(O=1|G=0)P(B=1|G=0)P(C=1|G=0)P(A=1|G=0)$
    $= (\frac{3}{8})(\frac{3}{4})(\frac{3}{4})(\frac{2}{4})(\frac{1}{4}) = \frac{9}{512}$
  - → $G=1$: $P(G=1)P(O=1|G=1)P(B=1|G=1)P(C=1|G=1)P(A=1|G=1)$
    $= (\frac{6}{8})(\frac{4}{8})(\frac{3}{8})(\frac{4}{8})(\frac{6}{8}) = \frac{27}{512}$

  Since $\frac{27}{512} > \frac{9}{512}$, $G = 1$

Sample #9: $G = 1$
Sample #10: $G = 1$

6] Extend Naive Bayes to binary feature values.

a) Use Naive Bayes Assumption to write joint probability of data, $P(x^1, \ldots, x^m, y^1, \ldots, y^m)$ in terms of θs.

↳ Naive Bayes Assumption: class conditional independence

The likelihood of one example, $P(x_i, y_i)$ is given by

$$P(y_i)P(x_i|y_i) = P(y_i)P(x_{i1}, \ldots, x_{in}|y_i) = P(y^i)\prod_{j=1}^{n} P(x^{ij}|y^i)$$

Then, extending to m examples,

$$P(x^1, \ldots, x^m, y^1, \ldots, y^m) = \prod_{i=1}^{m} P(x^i, y^i) \quad \text{by independence of data points.}$$

$$= \prod_{i=1}^{m} P(y^i)\prod_{j=1}^{n} P(x^{ij}|y^i)$$

Using the θ notation and indicator on $P(x^i, y^i)$, we find

$$P(x^i, y^i) = P(y^i)\prod_{j=1}^{n} P(x^{ij}|y^i)$$

$$= \theta_0^{1[y=0]}(1-\theta_0)^{1[y=1]} \prod_{j=1}^{n}\prod_{k=1}^{s} \theta_{j,k|y=0}^{1[y=0, x_j=k]} \theta_{j,k|y=1}^{1[y=1, x_j=k]}$$

$$\therefore \boxed{P(x^1, \ldots, x^m, y^1, \ldots, y^m) = \prod_{i=1}^{m} \theta_0^{1[y_i=0]}(1-\theta_0)^{1[y_i=1]} \prod_{j=1}^{n}\prod_{k=1}^{s} \theta_{j,k|y=0}^{1[y_i=0, x_{ij}=k]} \theta_{j,k|y=1}^{1[y_i=1, x_{ij}=k]}}$$

Note, $\log(P(x^1, \ldots, x^m, y^1, \ldots, y^m))$

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

Let the boxed expression be denoted as $\star$.

b) Maximize the joint from (a) wrt θ_0, $\theta_{j,k|y=0}$, $\theta_{j,k|y=1}$. Explain the meaning of the results.

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

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

$$= 0$$

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

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

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

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

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

$$= 0$$

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

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

$$\Rightarrow \theta_{j',k'|y=0} = \frac{\theta_{j',s|y=0}\sum_{i=1}^{m} 1[y_i=0, x_{ij'}=k']}{\sum_{i=1}^{m} 1[y_i=0, x_{ij'}=s]} \quad (1)$$

→ Now, must find $\theta_{j's|y=0}$.

Take sum of both sides of equation iterating k from 1 to S-1.

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

$$\updownarrow$$

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

$$\updownarrow$$

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

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

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

$$\Longrightarrow \theta_{j's|y=0} = \frac{\sum_{i=1}^{m} 1[y_i=0, x_{ij'}=S]}{\sum_{i=1}^{m} 1[y_i=0]} \quad (2)$$

→

Combining (1) and (2),

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

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

We can find the estimation of the remaining parameters by symmetry, with a summary and explanation of the results as follows:

- $\theta_0 = \frac{1}{m} \sum_{i=1}^{m} 1[y_i = 0]$
  - ↳ Relative frequency of data points class $y=0$, i.e. the ratio of points with class $y=0$ divided by total # of data points.
- $\theta_{j,k|y=0} = \frac{\sum_{i=1}^{m} 1[y_i = 0, x_{ij} = k]}{\sum_{i=1}^{m} 1[y_i = 0]}$
  - ↳ Relative frequency of datapoints with class $y=0$ whose j^{th} feature equals the k^{th} feature value.
- $\theta_{j,k|y=1} = \frac{\sum_{i=1}^{m} 1[y_i = 1, x_{ij} = k]}{\sum_{i=1}^{m} 1[y_i = 1]}$
  - ↳ Relative frequency of datapoints with class $y=1$ whose j^{th} feature equals the k^{th} feature value.