Author: Joseph D Ciaravino
ISYE 6740 CDA Fall 2022 - Homework 4

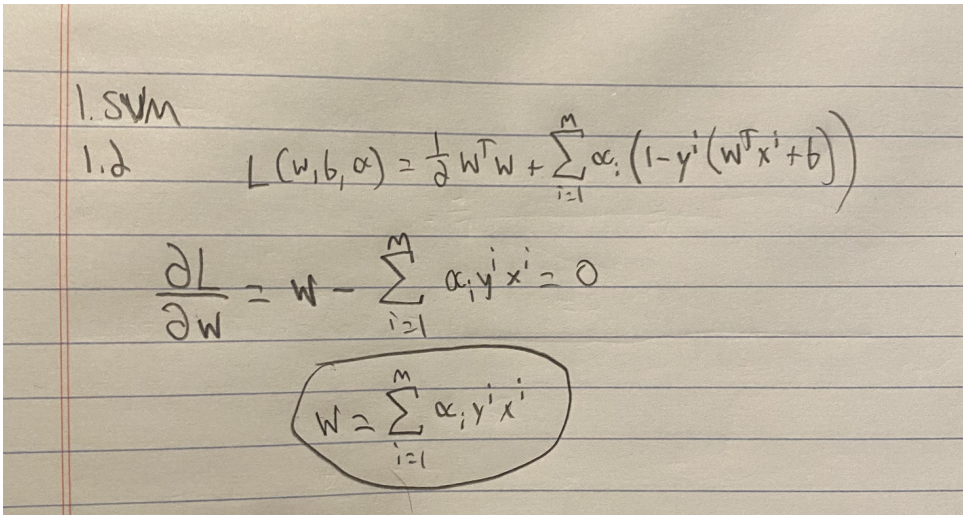
1. SVM

1.1. Explain why we can set the margin to c =1 to derive the SVM formulation?

(a)C is just a constant that scales w (orthogonal vector to decision boundary) and b (offset scalar). Setting c = 1 merely modifies the original mathematical problem into a cleaner form (2c/w l2 norm)

1.2. Using Langrangian dual formulation, show that the weight vector can be represented as:

(a)The optimal value of w is a linear combination, projection of the other variables



1.3. Explain why only the data points on the “margin” will contribute to the sum

(a)KKT condition aigi(w) = 0, not every constraint (data point) is relevant. Sparsity structure in solution

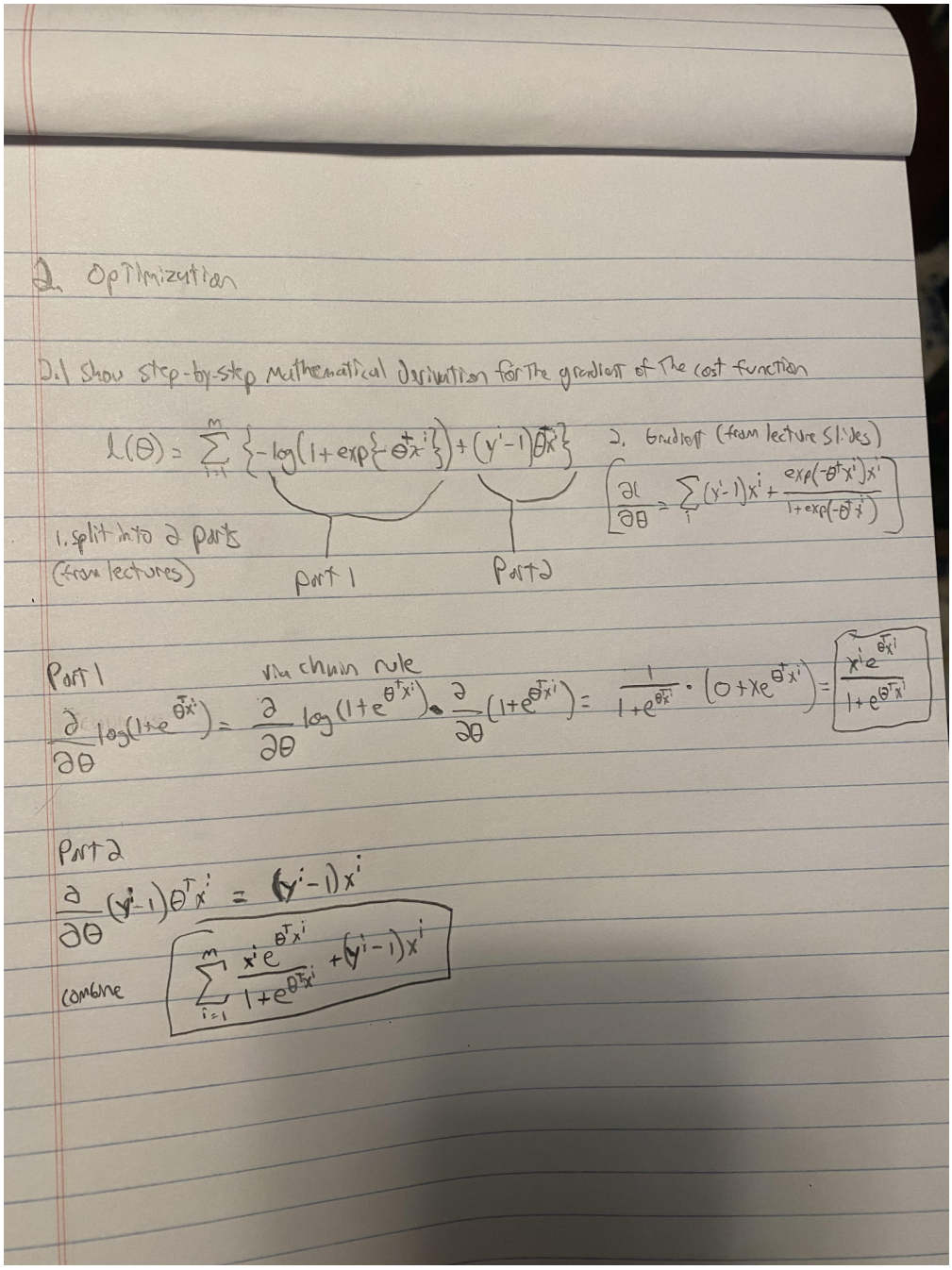
1.4. Simple SVM by hand

(a)Any value below 1 allows a line to linearly separate the data points

(b) Yes, the margin from the decision boundary will increase or decrease depending on the location of the support vectors from the boundary

2. Optimization

2.1. See attachment



2.2. Write pseudo code for gradient descent

From lecture slides:

Update rule

$$x^{t+1} = x^t - \gamma_t \nabla f(x^t)$$

$\gamma_t > 0$ is called the step size or learning rate

1. Set random initialization of parameter theta
2. Set step size γ_t
3. Set epsilon tolerance
4. While theta + 1 - theta is greater than epsilon, do theta + step*(gradient)
5. Initialize parameter θ^0

   Do

   gradient of $\nabla l(\theta)$ since we are solving maximization

   $$\theta^{t+1} \leftarrow \theta^t + \gamma_t \sum_i (y^i - 1)\, x^i + \frac{\exp\left(-\theta^{t\top} x^i\right) x^i}{1 + \exp\left(-\theta^{t\top} x^i\right)}$$

   While the $||\theta^{t+1} - \theta^t|| > \epsilon$
6. Return Theta when complete.

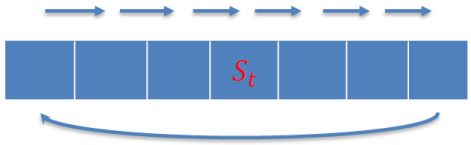
2.3. Write pseudo code for Stochastic Gradient Descent

From the lecture slides

- At each iteration, we randomly sample a small subset S_t data point $(x_i, y_i), i \in S_t$ (in the extreme case, there is just one sample in S_t)
- Use gradient estimated using a small subset of data (instead of full data)

$$\nabla \hat{l}(\theta) = \sum_{i \in S_t} (y^i - 1) x^i + \frac{\exp(-\theta^\top x^i) x^i}{1 + \exp(-\theta^\top x^i)}$$

- Each iteration use a different subset $S_t, t = 1, 2, \ldots.$
- Eventually loop through the entire training data, and may loop through the data multiple times



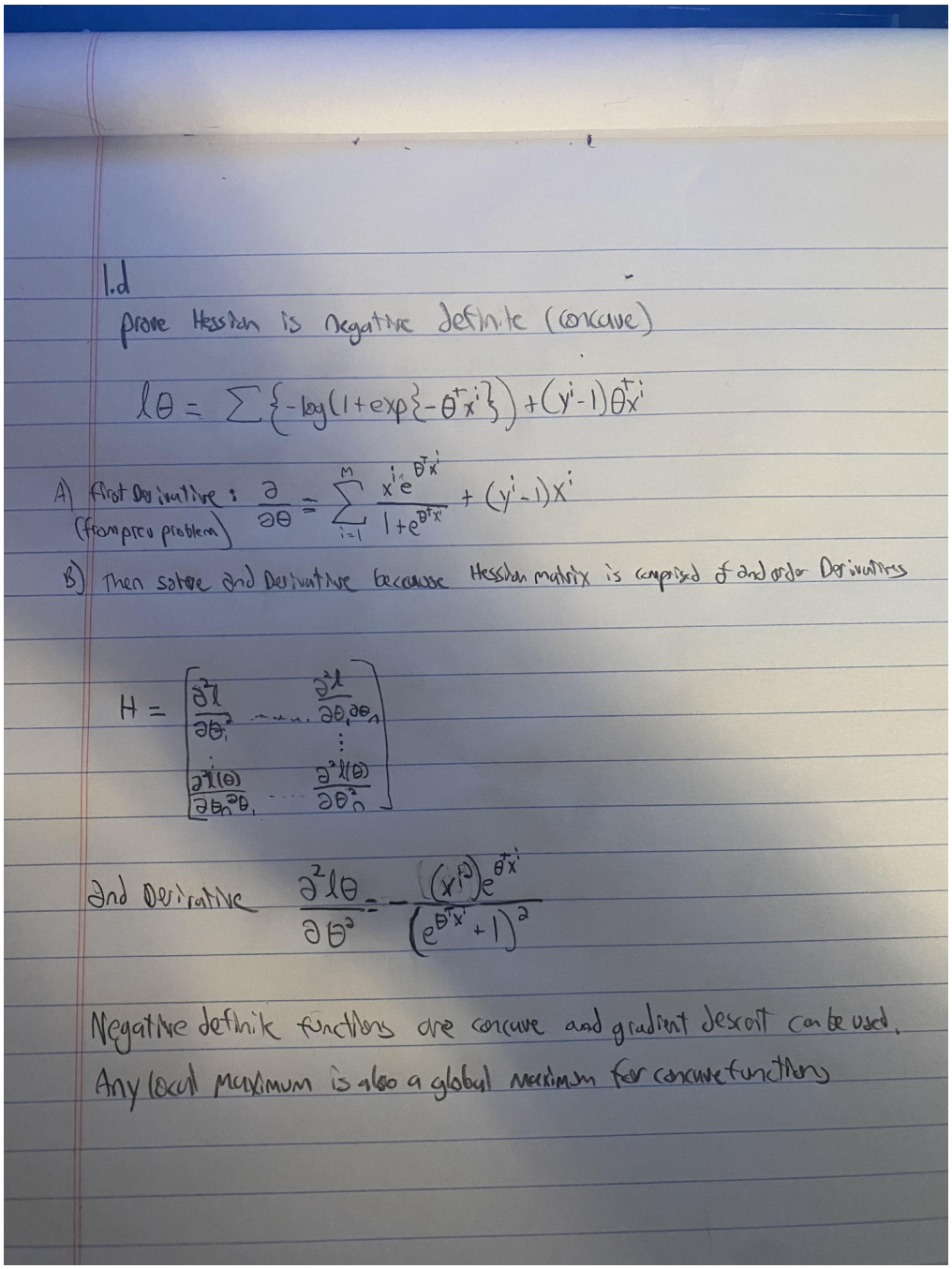
Georgia Tech

3. Initialize random theta
4. Set tolerance + maximum iteration limit
5. While less than maximum iterations + New_theta is greater than tolerance:
   5.1. Randomly sample a subset of data (St)
   5.2. New_Theta = Prev_Theta + step(1/t)(decreasing) * gradient(using subset)
6. End if max reached or New_theta less than tolerance
7. Return New_Theta

2.4 Show Training problem in basic logistic problem is concave

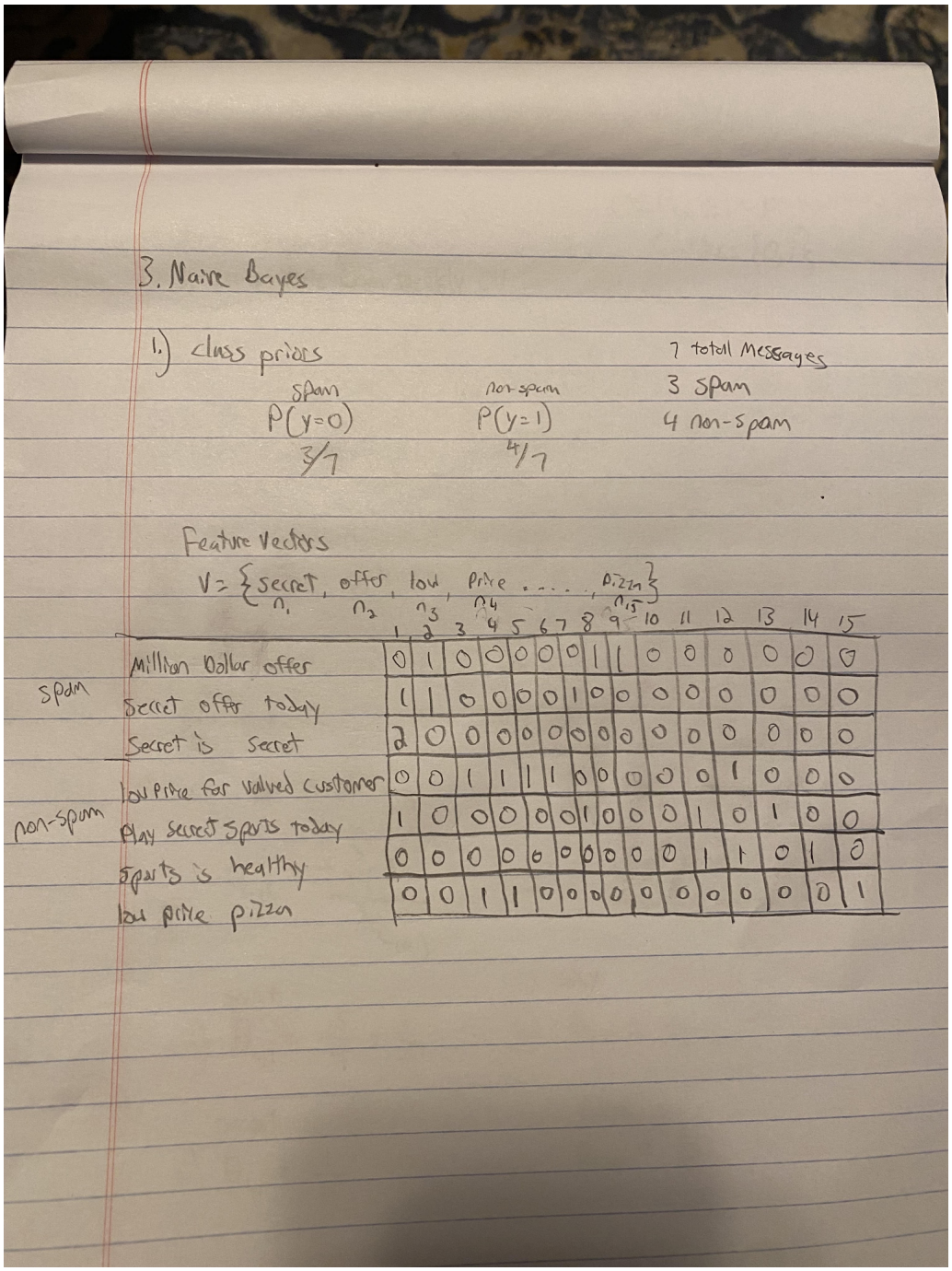
Concave = negative definite

See attachment



3. Naive Bayes Classifier

3.1



3. Naive Bayes

1.) class priors

spam: $P(y=0)$ = 3/7

non-spam: $P(y=1)$ = 4/7

7 totall Messages

3 spam

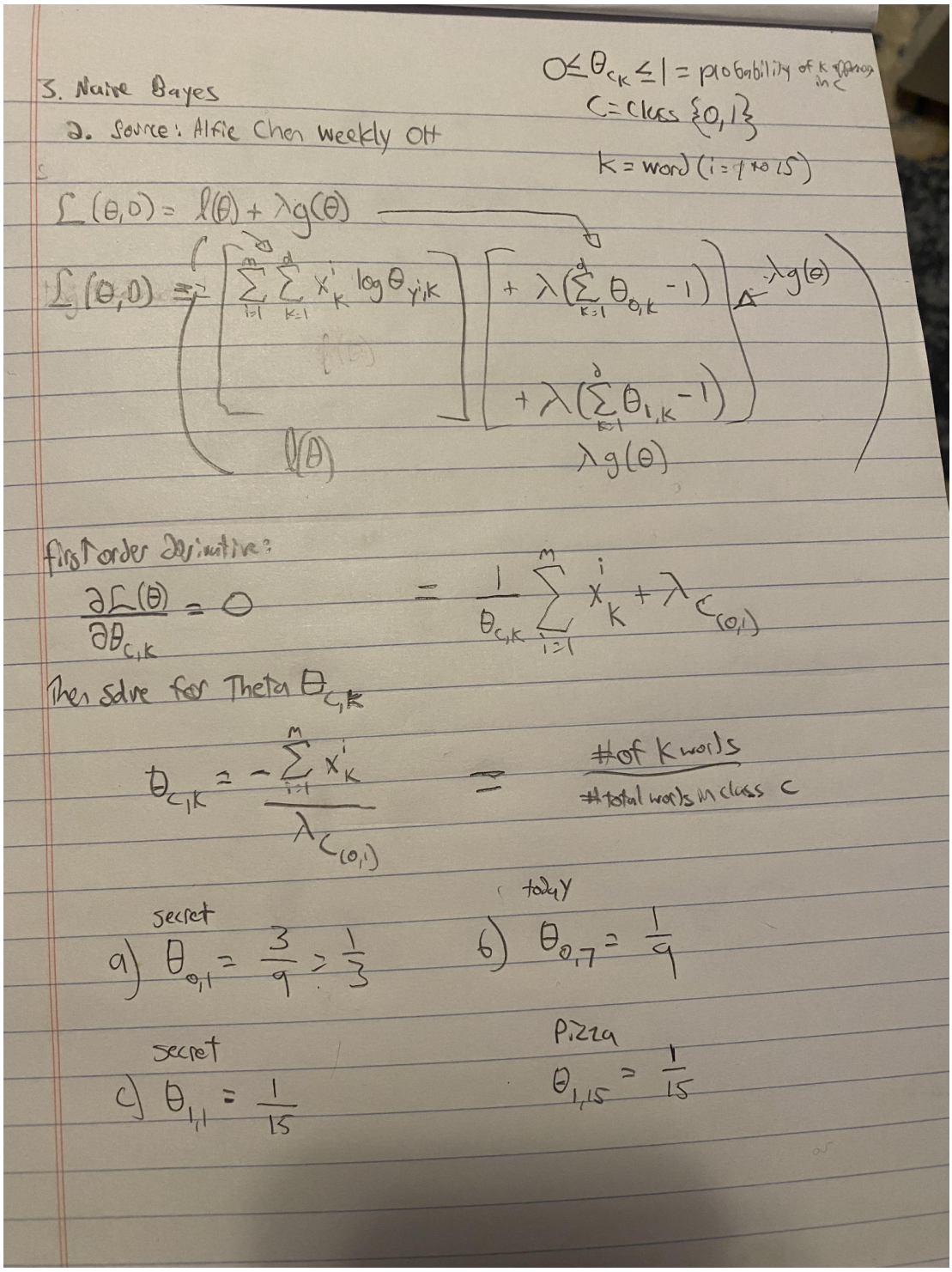
4 non-spam

Feature Vectors

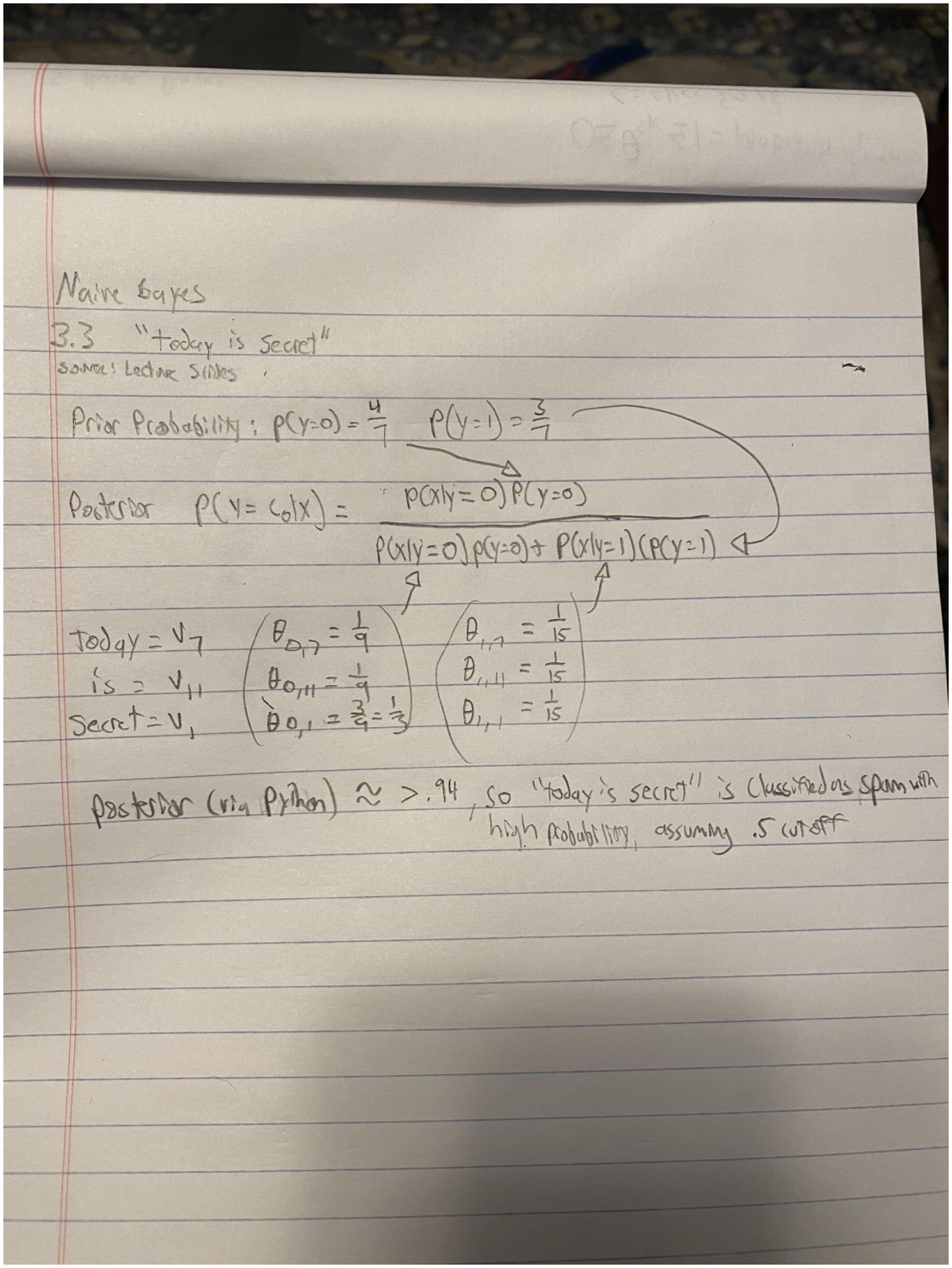
$V = \{\underset{n_1}{\text{secret}}, \underset{n_2}{\text{offer}}, \underset{n_3}{\text{low}}, \underset{n_4}{\text{price}}, \ldots, \underset{n_{15}}{\text{pizza}}\}$

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
spam	Million Dollar offer	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0
	Secret offer today	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0
	Secret is Secret	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
non-spam	low Price for valued customer	0	0	1	1	1	1	0	0	0	0	0	1	0	0	0
	Play secret sports today	1	0	0	0	0	0	1	0	0	0	1	0	1	0	0
	Sports is healthy	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0
	low price pizza	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1

3.2


3. Naive Bayes
2. Source: Alfie Chen weekly OH
$0 \le \theta_{c,k} \le 1$ = probability of k appearing in C
C = class {0,1}
k = word (i = 1 to 15)
$\mathcal{L}(\theta, 0) = l(\theta) + \lambda g(\theta)$
$\mathcal{L}(\theta, 0) = \left[\sum_{i=1}^{m}\sum_{k=1}^{d} x_k^i \log \theta_{y^i,k}\right] + \lambda\left(\sum_{k=1}^{d} \theta_{0,k} - 1\right) + \lambda\left(\sum_{k=1}^{d} \theta_{1,k} - 1\right)$
$l(\theta)$
$\lambda g(\theta)$
first order derivative:
$\frac{\partial \mathcal{L}(\theta)}{\partial \theta_{c,k}} = 0 = \frac{1}{\theta_{c,k}} \sum_{i=1}^{m} x_k^i + \lambda_{c_{(0,1)}}$
Then solve for Theta $\theta_{c,k}$
$\theta_{c,k} = \frac{-\sum_{i=1}^{m} x_k^i}{\lambda_{c_{(0,1)}}} = \frac{\#\text{of k words}}{\#\text{total words in class c}}$
secret
a) $\theta_{0,1} = \frac{3}{9} > \frac{1}{3}$
today
b) $\theta_{0,7} = \frac{1}{9}$
secret
c) $\theta_{1,1} = \frac{1}{15}$
Pizza
$\theta_{1,15} = \frac{1}{15}$

3.3



Naive Bayes

3.3 "today is secret"

source: Lecture Slides

Prior Probability: $P(y=0) = \frac{4}{7}$ $P(y=1) = \frac{3}{7}$

Posterior $P(y=c_0|x) = \frac{P(x|y=0)P(y=0)}{P(x|y=0)P(y=0) + P(x|y=1)(P(y=1)}$

today $= v_7$ $\theta_{0,7} = \frac{1}{9}$ $\theta_{1,7} = \frac{1}{15}$

is $= v_{11}$ $\theta_{0,11} = \frac{1}{9}$ $\theta_{1,11} = \frac{1}{15}$

secret $= v_1$ $\theta_{0,1} = \frac{3}{9} = \frac{1}{3}$ $\theta_{1,1} = \frac{1}{15}$

posterior (via Python) $\approx > .94$, so "today is secret" is classified as spam with high probability, assuming .5 cutoff