1 a)

p(open)=1/9 p(close)=8/9

p(0|open)=1/6 p(0|close)=1/6

p(1|open)=1/6 p(1|close)=0.01

p(2|open)=2/6 p(2|close)=1/6

p(3|open)=1/6 p(3|close)=2/6

4(4|open)=1/6 p(4|close)=2/6

P(open|031)=p(0|open)p(3|open)p(1|open)p(open)=

1/6*1/6*1/6*1/9=0.005

P(close|031)=p(0|close)p(3|close)p(1|close)p(close)=

1/6*2/6*0.01*8/9=0.004

So open.

1 b)

So first we can write the Naïve Bayes class distribution in the form of logistic class distribution.

$$P(Y=1 \mid X_1,\dots,X_p) = \frac{P(Y=1)\prod_{j=1}^{p} P(X_j = x \mid Y=1)}{P(Y=1)\prod_{j=1}^{p} P(X_j = x \mid Y=1) + P(Y=0)\prod_{j=1}^{p} P(X_j = x \mid Y=0)}$$

$$= \frac{\pi \prod_{j=1}^{p} \theta_j^x (1-\theta_j)^{1-x}}{\pi \prod_{j=1}^{p} \theta_j^x (1-\theta_j)^{1-x} + (1-\pi)\prod_{j=1}^{p} \alpha_j^x (1-\alpha_j)^{1-x}}$$

Divide both numerator and denominator by the numerator

$$= \frac{1}{1 + \frac{(1-\pi)\prod_{j=1}^{p} \alpha_j^x (1-\alpha_j)^{1-x}}{\pi \prod_{j=1}^{p} \theta_j^x (1-\theta_j)^{1-x}}}$$

We can apply log and exp to the non-constant term

$$= \frac{1}{1+\exp\left(\log\frac{1-\pi}{\pi} + x\log\sum_{j=1}^{p}\frac{\alpha_j}{\theta_j} + (1-x)\log\sum_{j=1}^{p}\frac{1-\alpha_j}{1-\theta_j}\right)}$$

$$= \frac{1}{1+\exp\left(\log\frac{1-\pi}{\pi} + \sum_{j=1}^{p}\log\frac{1-\alpha_j}{1-\theta_j} + \sum_{j=1}^{p}\left(\log\frac{\alpha_j}{\theta_j} - \log\frac{1-\alpha_j}{1-\theta_j}\right)x\right)}$$

X is indexed by j so we can write

$$w_0 = \log\frac{1-\pi}{\pi} + \sum_{j=1}^{p}\log\frac{1-\alpha_j}{1-\theta_j}$$

$$w_j = \log\frac{\alpha_j}{\theta_j} - \log\frac{1-\alpha_j}{1-\theta_j}$$

So we have

$$P(Y=1 \mid X_1,\ldots,X_p) = \frac{1}{1+\exp(w_0 + \sum_{j=1}^{p} w_j x_j)}$$

If w_j for logistic regression is selected using naïve Bayes estimate for

θ_j α_j π and using the formula we derive above, it is the same as we use Naïve Bayes from the beginning. On the other hand, if we take the formula for P(Y = 1 | X_j) in Naive Bayes and optimize the maximum conditional likelihood, we will get logistic regression.

c) The conditional independence assumption. Features are conditionally independent makes this classifier less generic than a logistic classifier learned directly from the data.

2 1) the likelihood of the Gaussian Mixture Model

$\ln p(x|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)\}$

responsibility

$$\gamma(z_k) = \frac{\pi_k N(x|\mu_k, \Sigma_k \quad)}{\sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma_j \quad)}$$

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

$$\sum\nolimits_k = \frac{1}{\sum_{n=1}^{N} \gamma(z_{nk})} \sum_{n=1}^{N} \gamma(z_{nk})(x_k - \mu_k)(x_k - \mu_k)^T$$

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma(z_n k)}{N}$$

In this case

$$p(\vec{t}, \vec{X}|\pi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = \prod_{n=0}^{N-1} (\pi N(\vec{x}_n|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}))^{t_n}((1-\pi)N(\vec{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}))^{1-t_n}$$

$$\begin{aligned}
p(\vec{t}, \vec{X}|\pi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) &= \prod_{n=0}^{N-1} (\pi N(\vec{x}_n|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}))^{t_n}((1-\pi)N(\vec{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}))^{1-t_n} \\
\ln p(\vec{t}, \vec{X}|\pi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) &= \ln \prod_{n=0}^{N-1} (\pi N(\vec{x}_n|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}))^{t_n}((1-\pi)N(\vec{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}))^{1-t_n} \\
&= \sum_{n=0}^{N-1} t_n \ln(\pi N(\vec{x}_n|\boldsymbol{\mu}_0, \boldsymbol{\Sigma})) + (1-t_n)\ln((1-\pi)N(\vec{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})) \\
&= \sum_{n=0}^{N-1} t_n \ln(\pi N(\vec{x}_n|\boldsymbol{\mu}_0, \boldsymbol{\Sigma})) + (1-t_n)\ln((1-\pi)N(\vec{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})) \\
&= \sum_{n=0}^{N-1} t_n \ln(\pi) + (1-t_n)\ln(1-\pi) + const
\end{aligned}$$

$$\frac{N_0}{N_0 + N_1} = \pi$$

2)

$$P(y=1|\mathrm{x}) = \frac{P(\mathrm{x}|\mu_1, \Sigma_1)P(y=1)}{P(\mathrm{x}|\mu_1, \Sigma_1)P(y=1) + P(X|\mu_0, \Sigma_0)P(y=0)}$$

$$\log \frac{P(\mathrm{x}|\mu_1, \Sigma_1)}{P(\mathrm{x}|\mu_0, \Sigma_0)} = 0 \quad \text{or} \quad P(y=1|\mathrm{x}) = 0.5$$

$$P(y=1|x,\theta) = \frac{P(x|\mu_1,\sigma_1^2)}{P(x|\mu_1,\sigma_1^2) + P(x|\mu_0,\sigma_0^2)} = \frac{1}{1 + \exp\left\{-\log \frac{P(x|\mu_1,\sigma_1^2)}{P(x|\mu_0,\sigma_0^2)}\right\}}$$

$$\log \frac{P(x|\mu_1,\sigma_1^2)}{P(x|\mu_0,\sigma_0^2)} = \begin{cases} w_0 + w_1 x, & \text{when } \sigma_1^2 = \sigma_0^2 \\ w'_0 + w'_1 x + w'_2 x^2, & \text{otherwise} \end{cases}$$

$$P(y=1|x,\mathrm{w}) = \frac{1}{1 + \exp\{-(w_0 + w_1 x)\}} = g\,(\,w_0 + w_1 x\,)$$

$$g(z) = (1 + \exp\{-z\})^{-1}.$$

3 we need to minimize

$$\frac{\lambda}{2} \sum_{k=1}^{K} ||w_k||^2 - \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} w_k^T x_i + \sum_{i=1}^{N} log\left(\sum_{j=1}^{K} e^{w_j^T x_i}\right)$$

We do partial derivative to each k and sum it up

$$\sum_{k=1}^{K} \omega_k - (\sum_{k=1}^{K} \sum_{i=1}^{N} y_{ik}\, x_i - \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{e^{\omega_i^T x_i} x_i}{\sum_{k=1}^{K} e^{\omega_i^T x_i}})$$

The part in the parentheses will equal to zero, so $\sum_{k=1}^{K} \omega_k$=0

4 b) training accuracy

accuracy	raw data	standardized data
Batch GD	0.9074	0.9248
Newton	0.9270	0.9270
Glmfit	0.9270	0.9270

Test accuracy

accuracy	raw data	standardized data
Batch GD	0.9057	0.9226
Newton	0.9235	0.9248
Glmfit	0.9235	0.9257

c)

mutual information

81.0183	62.3229	153.9113	0.1429	179.9197	121.2667
160.7466	84.7307	71.8753			
87.9017	94.6529	68.3537	36.0170	3.4097	11.4636
231.3557	143.3183	105.4160			
111.1017	23.4632	265.9074	0.8627	139.7089	139.4310
68.1706	28.3794	29.8853			
3.3050	12.1973	6.1919	7.8922	3.6153	7.5089
3.3582	2.6534	-2.1636			
13.4738	0.0024	6.1358	-0.1202	1.8214	8.7830
3.3277	4.8160	-1.7745			
3.1158	0.0272	2.4972	-5.3677	18.8069	-5.1670
347.7491	349.8579	49.4948			
82.0853	85.8735	63.3480			

Pearson

0.1262 -0.0302 0.1970 0.0574 0.2419 0.2326

0.3321 0.2068 0.2316 0.1390 0.2345 0.0077

0.1329 0.0600 0.1959 0.2632 0.2632 0.2042

0.2737 0.1898 0.3832 0.0919 0.3348 0.2161 -

0.2567 -0.2330 -0.1834

-0.1588 -0.1335 -0.1711 -0.1269 -0.1142 -0.1199 -

0.1128 -0.1492 -0.1361

-0.1780 -0.0310 -0.1228 -0.0648 -0.0974 -0.1366 -

0.1357 -0.0946 -0.1404

-0.1461 -0.0447 -0.0840 -0.0596 -0.0897 -0.0647

0.2419 0.3236 0.0651 0.1100 0.2161 0.2492

New feature

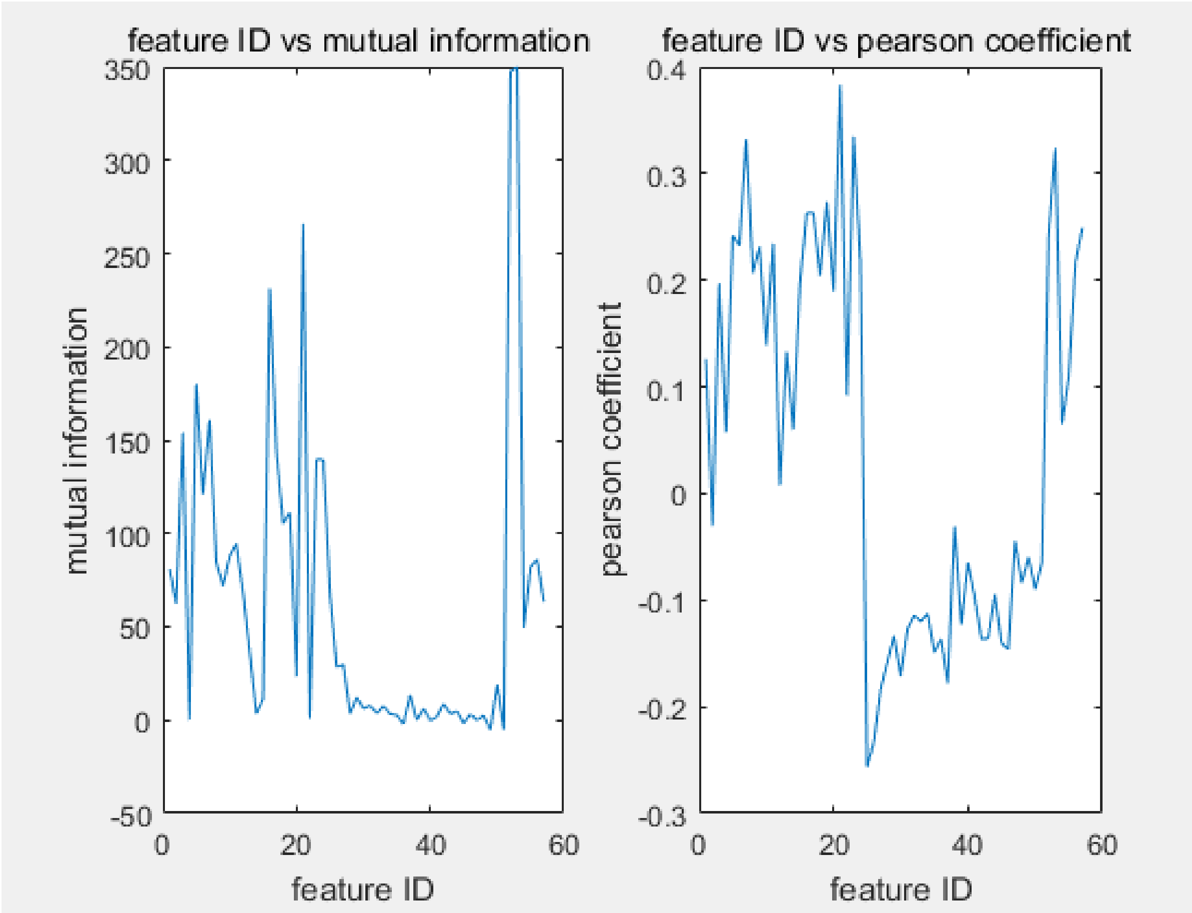
53 52 21 16 5 7 3 17 23 24

6 19 18 11 10 56

8 55 1 9

Accurary

Train 0.8900



4.2 1) model1

prob 0.2884

0.2861

0.4255

M1

1.4567 0.0234

M2

-1.5082 1.5186

M3

-1.5104 -1.5292

S1

0.9647	0.0093
0.0093	4.0493

S2

1.0422	-0.5505
-0.5505	1.1639

S3

1.0254	0.4562
0.4562	1.0007

accurary

0.9037

2)

prob 0.4469

0.2721

0.2810

M1

-1.4518 -1.4646

M2

-1.5591 1.6005

M3

1.4908 0.0345

S1

1.0842	0.5073
0.5073	1.0807

S2

1.0013	-0.4900
-0.4900	1.0680

S3

0.9298	-0.0093
-0.0093	4.0996

accurary

0.9021

w

3.2481	0.8281	0.1720
-0.1340	1.8552	-0.4990
0	0	0

accurary

0.8862