Q1 Given $P_r[y_i = c] = \pi_c$

$x_i / y_i \sim N(\mu_c, \Sigma_c)$.

Gaussian Naive Bayes classifier is similar to GDA. The difference is in the construction of $P(x_i/y_i)$. NB makes an (navie but strong) assumption that features are independent so that

$$P(x/y=k) = P(x_1/y=k) \cdot P(x_2/y=k) \cdot P(x_3/y=k) \cdots P(x_n/y=k)$$

$$= \prod_{j=1}^{n} p(x_j | y=k)$$

For multivarite normal distribution. $p(x_j | y=k) \sim N(\mu_j, \sigma_j^2)$

We define probability for any class given x.

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

a) maximum likelihood estimation for the model parameter.

$$L = -\sum_{i=1}^{n} \log P(y(i)) \prod_{j=1}^{n} (x_j | y_i).$$

$$= -\left(\sum_{i=1}^{n} \log P(y^{(i)} = c) + \sum_{i=1}^{n} \log\left(\prod_{j=1}^{d} P(x_j^{(i)} | y^{(i)})\right)\right) \quad (1)$$

If we contrain ~~Σ_c to be diagonal then~~ Σ_c (when $c = 1, \ldots k$) to be diagonal then. we can rewrite $P(x, y^i)$ as a product of $P(x_j | y^{(i)})$ for $j = 1 \cdots d$ features.

$$p(x/y^{(i)}) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_{y^{(i)}})}} \exp\left(-\frac{1}{2}(x_j - \mu_{j y^{(i)}})^T \Sigma_{y^{(i)}}^{-1} (x_j - \mu_{j y^{(i)}})\right).$$

$$= \prod_{j=1}^{d} \frac{1}{\sqrt{(2\pi)^d}\, \sigma_{j y^{(i)}}} \exp\left(-\frac{1}{2\sigma_{j y^{(i)}}} \|x_j - \mu_{j y^{(i)}}\|_2^2\right). = \prod_{j=1}^{D} p(x_j / t).$$

Using this fact we can modify our equation (1) as.

$$= -\left(\sum_{i=1}^{n} \log \pi_{y^{(i)}} + \sum_{i=1}^{n}\sum_{j=1}^{d} \log \frac{1}{\sqrt{2\pi}\,\sigma_{j y^{(i)}}} \exp\left(-\frac{(x_i - \mu_{i y^{(i)}})^2}{2\sigma^2_{j\, y^{(i)}}}\right)\right).$$

$$= -\left(\sum_{i=1}^{n} \log \pi_{y^{(i)}} + \sum_{i=1}^{n}\sum_{j=1}^{d} \log \frac{1}{\sqrt{2\pi}\,\sigma_{j y^{(i)}}} + \sum_{i=1}^{n}\sum_{j=1}^{d} \log \exp\left(\frac{-(x_i - \mu_{i y^{(i)}})^2}{2\sigma^2_{j\, y^{(i)}}}\right)\right).$$

$$= -\left(\sum_{i=1}^{n} \log \pi_{y^{(i)}} - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d} \log(2\pi\sigma_{j\, y^{(i)}}) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{d} \log \exp \frac{(x_j - \mu_{j y^{(i)}})^2}{2\sigma^2_{j\, y^{(i)}}}\right)$$

~~Derivative with respect to $\mu_{j y^{(i)}}$~~

$$\cancel{\mu_{j y^{(i)}} = \frac{\sum_{t=1}^{n} 1[y^{(t)} = y^{(i)}]\, x_j^{(t)}}{\sum_{t=1}^{n} 1[y^{(t)} = y^{(i)}]}}$$

taking derivative with respect to $\mu_j(y^{(i)} = c)$.

$$\mu_{jc} = \frac{\sum_{i=1}^{n} 1[y^{(i)} = c]\, x_j^{(i)}}{\sum_{i=1}^{n} 1[y^{(i)} = c]}$$

ith training example

1 if $y^{(i)} = c$ or 0.

jth feature.

c^{th} class.

taking derivative with respect to σ_{jc}^2

ith training example.

$$\sigma_{jc}^2 = \frac{\sum_{i=1}^{n} 1[y^{(i)} = c]\,(x_j^{(i)} - \mu_{jc})^2}{\sum_{i=1}^{n} 1[y^{(i)} = c]}$$

j^{th} feature c^{th} class.

taking derivative with respect to π_c.

$$\pi_c = \frac{\sum_{i=1}^{n} 1(y^{(i)} = c)}{n}$$

(b.) Now with estimated model parameters π_c, μ_c, σ_c; given any new data point we can predict $\hat{y}_0$ in following way.

$$\hat{y}_0 = \underset{y_c,\, c=(1,\dots K)}{\arg\max}\ P(y_c)\, P(x_0/y_c).$$

$$P(x/y_c) = \prod_{i=1}^{d} P(x_i/y_c) \qquad \because \text{ we know } \pi_c,\ \mu_{ic},\ \sigma_{ic}^2.$$

$$= \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\sigma_{ic}^2}} \exp\left[\frac{-(x_i-\mu_{ic})^2}{2\sigma_{ic}^2}\right]$$

$$\underset{y_c}{\arg\max}\ \pi_c \cdot \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\sigma_{ic}^2}} \exp\left[\frac{-(x_i-\mu_{ic})^2}{2\sigma_{ic}^2}\right]$$

new value of x_0 for which prediction has to be found.

plug the estimated values in the equation to get predicted $\hat{y}_0$

Q2

Given:- Lasso regularized K-class logistic regression problem.

$$\underset{\theta_1 \dots \theta_K}{\text{minimize}} \ \frac{1}{n}\sum_{i=1}^{n}\left(\log\sum_{c=1}^{K}\exp(\theta_c^T\phi_i) - \theta_{y_i}^T\phi_i\right) + \frac{\lambda}{2}\sum_{c=1}^{K}\|\theta_c\|_1$$

This is of the form $f(x) = g(x) + h(x)$ so using Proximal stochastic gradient descent algorithm to solve this will be our approach of the algorithm.

Pesudo code for solving this can be as follows:

→ Chose or initialize $\theta^{(0)}$ and repeat following steps for $K = 1,2,3 \dots$ max-iteration.

→ For a random sample i uniformily drawn from $\{1, \dots, n\}$. The stochastic step.

→ Calculate the gradient for i

if $i = y_i$; input class is equal to y_i

$$\nabla g(\theta^{(K-1)}) = \frac{\phi_i e^{\theta_c^T\phi_i}}{\sum_{c=1}^{K} e^{\theta_c^T}\phi_i} - \phi_i \quad \left\{ \begin{array}{l} \text{using } f(z) = \log\sum\exp(z) \\ \nabla f(z) = \frac{1}{\sum\exp(z)}\exp(z) \end{array} \right.$$

else

$$\nabla g(\theta^{(K-1)}) = \frac{\phi_i e^{\theta_c^T\phi_i}}{\sum_{c=1}^{K} e^{\theta_c^T}\phi_i}$$

→ update $\theta^{(K)}$

$$\theta^{(K)} = \theta^{(K-1)} - t_K\nabla g(\theta^{(K-1)})$$

where t_K is the diminized step size of our algorithm.

→ feed the new $\theta^{(k)}$ into proximal operation

where $\text{prox}_t(\hat{\theta}) = \underset{\theta}{\arg\min} \frac{1}{2} \|\theta - \tilde{\theta}\|^2 + \underbrace{h(\theta)}$ → lasso regularization in our case.

prox. operator tries to find a point that makes $h(\theta)$ small but is approximately close to θ.

For lasso regularization proximal operator can be explicitly expressed as soft thresholding.

$$\text{Prox}_{t_K \frac{\lambda}{2}\|\cdot\|_1}(\tilde{\theta})_k = \begin{cases} \tilde{\theta}^{(k)} - t_K \frac{\lambda}{2} & \tilde{\theta}^{(k)} > t_K \lambda/2 \\ 0 & |\tilde{\theta}^{(k)}| \leq t_K \lambda/2 \\ \tilde{\theta}^{(k)} + t_K \frac{\lambda}{2} & \tilde{\theta}^{(k)} < -t_K \frac{\lambda}{2} \end{cases}$$

→ repeat for given number of iteration.

→ End of pseudo code.

Q3

Given

$$\Pr(y_i = c) = \pi_c \qquad \Pr(x_i/y_i = c) \sim \text{Multi}(P_c, L_i)$$

$$P(x_i/y_i = c) = \frac{L_i!}{\prod_{j=1}^{d} x_{ij}!} \prod_{j=1}^{d} P_{cj}^{x_{ij}}$$

(a) For maximum likelihood formulation for estimating $P_1, P_2 \dots P_K$ and π we find the log likelihood of probability mass function.

$$\Rightarrow \sum_{i=1}^{n} \log P(x_i, y_i ; \pi, P_c) \quad \text{where } c = 1, 2, \dots k.$$

$$\Rightarrow \sum_{i=1}^{n} \log p(y_i = c) + \sum_{i=1}^{n} \log P(x_i / y_i ; P_c).$$

$$\Rightarrow \sum_{i=1}^{n} \log p(y_i = c) + \sum_{i=1}^{n} \sum_{j=1}^{d} x_{ij} \log P_{cj} + \underbrace{\text{constant}}_{\text{parameter-free term}}.$$

$$\Rightarrow \sum_{i=1}^{n} \sum_{c=1}^{K} \mathbb{I}\{y_i = c\} \log \pi_c + \sum_{i=1}^{n} \sum_{c=1}^{K} \sum_{j=1}^{d} x_{ij} \, \mathbb{I}\{y_i = c\} \log P_{cj} + \text{constant} \qquad (1)$$

where this is either 1 or 0

(b) The E-step or expectation step requires that we compute the expected complete log likelihood under multinomial distribution.

$$E_{p(\cdot | x_i, \pi_c, P_c)} \left[\sum_{i=1}^{n} \log p(x_i, y_i ; \pi, P_c) \right].$$

$$= \sum_{i=1}^{n} \sum_{c=1}^{K} \phi_{i,c} \log \pi_c + \sum_{i=1}^{n} \sum_{c=1}^{K} \sum_{j=1}^{d} x_{ij} \, \phi_{i,c} \log P_{cj}$$

where we define $\phi_{i,c} = P(y_i = c | x_i ; \pi, P_c)$ As in the GMM case. studies in class., it suffices to compute ϕ_{ic} to complete this step.

By Bayes' Rule.

$$P(y_i = c | x_i ; \pi, P_c) \propto P(y_i = c ; \pi) \, P(x_i / y_i ; P_c)$$

$$P(y_i = c \mid x_i ; \pi, P_c) = \frac{\pi_c f(x_i, P_c)}{\sum_{l=1}^{k} \pi_l f(x_i, P_l)}$$

where $f(x_i, P_c) \propto \prod_{j=1}^{d} P_{cj}^{x_{ij}}$ using multimonial distributed.

<u>The M-step</u>. maximize the expected log likelihood we have a closed form solution for the parameter. updates.

$$\pi_c = \frac{1}{n} \sum_{i=1}^{n} \phi_{i,c}$$ taking derivative of equ (1) function w.r.t to π_c

$$P_{cj} = \frac{\sum_{i=1}^{n} x_{ij} \phi_{i,c}}{\sum_{i=1}^{n} L_i \phi_{i,c}}.$$

(c)

```
Top 10 key words for cluster: 0
['ooo\n', 'colour\n', 'moslems\n', 'italian\n', 'subjects\n', 'atom\n',
'explaining\n', 'league\n', 'crucifixion\n', 'honor\n']
Top 10 key words for cluster: 1
['koran\n', 'innermost\n', 'geoffrey\n', 'wesleyan\n', 'nrp\n',
'league\n', 'coincidence\n', 'disobeying\n', 'creatures\n', 'exist\n']
Top 10 key words for cluster: 2
['koran\n', 'innermost\n', 'disobeying\n', 'nrp\n', 'creatures\n',
'exist\n', 'coincidence\n', 'necessity\n', 'shits\n', 'fri\n']
Top 10 key words for cluster: 3
['koran\n', 'innermost\n', 'disobeying\n', 'nrp\n', 'geoffrey\n',
'creatures\n', 'coincidence\n', 'exist\n', 'wesleyan\n', 'shits\n']
Top 10 key words for cluster: 4
```

```
['ooo\n', 'colour\n', 'moslems\n', 'subjects\n', 'league\n',
'italian\n', 'geoffrey\n', 'atom\n', 'crucifixion\n', 'explaining\n']
Top 10 key words for cluster: 5
['moslems\n', 'explaining\n', 'italian\n', 'colour\n', 'ooo\n',
'gaul\n', 'atom\n', 'climbed\n', 'honor\n', 'wherein\n']
Top 10 key words for cluster: 6
['koran\n', 'geoffrey\n', 'league\n', 'wesleyan\n', 'blood\n',
'husbands\n', 'ooo\n', 'subjects\n', 'crucifixion\n', 'innermost\n']
Top 10 key words for cluster: 7
['disobeying\n', 'innermost\n', 'koran\n', 'creatures\n', 'exist\n',
'nrp\n', 'coincidence\n', 'necessity\n', 'shits\n', 'fri\n']
Top 10 key words for cluster: 8
['disobeying\n', 'innermost\n', 'koran\n', 'creatures\n', 'nrp\n',
'exist\n', 'coincidence\n', 'necessity\n', 'shits\n', 'fri\n']
Top 10 key words for cluster: 9
['disobeying\n', 'koran\n', 'innermost\n', 'creatures\n', 'nrp\n',
'exist\n', 'coincidence\n', 'necessity\n', 'shits\n', 'fri\n']
Top 10 key words for cluster: 10
['koran\n', 'geoffrey\n', 'league\n', 'wesleyan\n', 'innermost\n',
'blood\n', 'husbands\n', 'ooo\n', 'subjects\n', 'omnipotent\n']
Top 10 key words for cluster: 11
['ooo\n', 'moslems\n', 'colour\n', 'italian\n', 'explaining\n',
'atom\n', 'subjects\n', 'gaul\n', 'honor\n', 'climbed\n']
Top 10 key words for cluster: 12
['disobeying\n', 'innermost\n', 'koran\n', 'creatures\n', 'nrp\n',
'exist\n', 'coincidence\n', 'necessity\n', 'shits\n', 'fri\n']
Top 10 key words for cluster: 13
['koran\n', 'innermost\n', 'disobeying\n', 'nrp\n', 'creatures\n',
'coincidence\n', 'exist\n', 'shits\n', 'necessity\n', 'fri\n']
Top 10 key words for cluster: 14
['moslems\n', 'explaining\n', 'colour\n', 'italian\n', 'ooo\n',
'gaul\n', 'atom\n', 'climbed\n', 'honor\n', 'subjects\n']
Top 10 key words for cluster: 15
['moslems\n', 'ooo\n', 'colour\n', 'italian\n', 'explaining\n',
'atom\n', 'gaul\n', 'subjects\n', 'climbed\n', 'honor\n']
Top 10 key words for cluster: 16
['ooo\n', 'league\n', 'colour\n', 'moslems\n', 'subjects\n',
'geoffrey\n', 'italian\n', 'crucifixion\n', 'blood\n', 'atom\n']
Top 10 key words for cluster: 17
['explaining\n', 'moslems\n', 'italian\n', 'colour\n', 'ooo\n',
'gaul\n', 'atom\n', 'climbed\n', 'honor\n', 'fondly\n']
Top 10 key words for cluster: 18
['ooo\n', 'moslems\n', 'colour\n', 'italian\n', 'explaining\n',
'subjects\n', 'atom\n', 'gaul\n', 'honor\n', 'climbed\n']
Top 10 key words for cluster: 19
['koran\n', 'innermost\n', 'disobeying\n', 'nrp\n', 'creatures\n',
'exist\n', 'coincidence\n', 'necessity\n', 'shits\n', 'fri\n']
```

Q4 ~~The given~~.

(a) Given

$$\text{minimize} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\tilde{x}_i^T \tilde{x}_j - \tilde{y}_i^T y_j \right)^2.$$

$$\Rightarrow \text{minimize } \| \tilde{X}^T \tilde{X} - \tilde{Y}^T \tilde{Y} \| \quad \text{—— (1)}$$

Using Eckart-Young theorem.

$\|A - B\| \geq \|A - A_K\|$ where rank$(B) \leq K$.

and A_K is the best approximation of A to rank K, as taking largest k eigenvalues in SVD.

Using this theorem we can state that.

$$\| \tilde{X}^T \tilde{X} - Y^T Y \| \geq \| \tilde{X}^T \tilde{X} - (\tilde{X}^T \tilde{X})_K \|.$$

At optimality.

$$Y^T Y = (\tilde{X}^T \tilde{X})_K \quad \text{—— (2)}$$

→ Take SVD and largest k eigenvalues.

Now

$$\Rightarrow \tilde{X}^T \tilde{X} = U \Lambda U^T$$

$$\Rightarrow (\tilde{X}^T \tilde{X})_K = U_K \Lambda_K U_K^T \quad \text{—— (3)}$$

from (2) and (3)

$$Y^T Y = U_K \Lambda_K U_K^T$$

$Y = \Lambda^{1/2} U^T$ as Y is positive definite

$Y^T = U \Lambda^{1/2}$

Sine ~~we~~ only keep k eigenvalues. the corresponding columns in ~~D~~ Y using $Y = \Lambda^{1/2} U^T$ are ~~ith~~ y values that is y_i is i^{th} column of Y which same as PCA.

(b)

Let $K = X^T X$

with $k = diag(K_{ii}) \in R^n$

$D = d_{ij}^2$ given in question ~~does~~ becomes.

$$= k \cdot 1^T + 1 \cdot k^T - 2K \quad \text{— (1)}$$

Also given in question.

$$\tilde{x}_i = x_i - \frac{1}{n}\sum_{j=1}^{n} x_j \} \longrightarrow \text{nothing but mean.}$$

$$\tilde{X} = X - \frac{1}{n} X \cdot 1 \cdot 1^T \quad \text{— (2)}$$

Given $B = -\frac{1}{2}\left(1 - \frac{1}{n} 1 \cdot 1^T\right) D \left(1 - \frac{1}{n} 1 \cdot 1^T\right)^T$

we have to prove $B = \tilde{X}^T \tilde{X}$.

using (1)

$$B = -\frac{1}{2}\left(1 - \frac{1}{n} 1 \cdot 1^T\right)\left(k \cdot 1^T + 1 \cdot k^T - 2K\right)\left(1 - \frac{1}{n} \cdot 1 \cdot 1^T\right)^T$$

expanding this we have.

$$B = HKH^T - \frac{1}{2} H \cdot 1 \cdot \cancel{K}^T \cdot H^T - \frac{1}{2} H \cdot \cancel{K} \cdot 1^T H^T$$

where $H = \left(1 - \frac{1}{n} \cdot 1 \cdot 1^T\right)$.

$$H \cdot 1 \cdot K^T \cdot H^T = H \cdot 1 \cdot K^T \cdot H^T = K \cdot 1 \left(1 - \frac{1}{n} \cdot 1 \cdot 1^T\right)$$

$$= K \cdot 1 - K\left(\frac{1^T \cdot 1}{n}\right) \cdot 1$$

$$= 0.$$

$$B = HKH^T - \frac{1}{2} \underbrace{H \cdot 1 \cdot K^T H^T}_{0} - \frac{1}{2} \underbrace{H K 1^T H^T}_{0}$$

$$B = HKH^T = \left(1 - \frac{1}{n} \cdot 1 \cdot 1^T\right) K \left(1 - \frac{1}{n} \cdot 1 \cdot 1^T\right)^T$$

$$= K - \frac{1}{n} \cdot 1 \cdot 1 \cdot K - \frac{1}{n} \cdot K \cdot 1 \cdot 1^T + \frac{1}{n^2} \cdot 1 (1^T K \cdot 1) 1^T$$

$$= K = X^T X$$

Hence proved.