ocinar1@remote04:~/Desktop/ML2$ python3 NaiveBayes.py train/spam train/ham test/spam test/ham
Accuracy for training: 98.0561555075594
Accuracy with stop words: 94.97907949790795
Accuracy without stop words: 94.35146443514644

① Derivative maximum likelihood estimates

a) parameter: p Bernoulli (p), sample of size n

Binomial → $f(x,p) = \begin{cases} p^x(1-p)^{1-x} & ; x=0,1,2 \\ 0 & ; \text{else} \end{cases}$

$x_i \to \text{Bern}(p)$ and x_i independent then f

log likelihood $L \to L(p) = \ln \prod_{i=1}^{n} f(x_i,p)$

$$L(p) = \sum_{i=1}^{n} \ln\left[p^{x_i}(1-p)^{1-x_i}\right]$$

$$L(p) = \sum_{i=1}^{n} \left[x_i \ln p + (1-x_i)\ln(1-p)\right]$$

we are going to maximize this

$$\arg\max_p L(p) \Rightarrow \frac{\partial L(p)}{\partial p} = 0$$

$$\frac{\partial}{\partial p} = \left[\sum_{i=1}^{n} x_i \ln(p) + (1-x_i)\ln(1-p)\right] = 0$$

$$\sum_{i=1}^{n} x_i \cdot \frac{1}{p} + \sum_{i=1}^{n} (1-x_i) \cdot \frac{1}{1-p} \cdot (-1) = 0$$

$$\left(\sum_{i=1}^{n} x_i\right)\left(\frac{1}{p} + \frac{1}{1-p}\right) + \sum_{i=1}^{n} \frac{-1}{1-p} = 0$$

$$\frac{1}{p(1-p)} \sum_{i=1}^{n} x_i - \frac{n}{1-p} = 0$$

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

b) parameter p, Binomial(N,p) sample size n

Binomial dist $\Rightarrow f(x_i, N, p) = \binom{N}{x_i} p^{x_i} (1-p)^{N-x_i}$

$$\prod_{i=1}^{n} f(x_i) = \left(\prod_{i=1}^{n} \binom{N}{x_i}\right) p^{\sum x_i} \cdot (1-p)^{nN - \sum_{i=1}^{n} x_i}$$

$$L(p) = \ln\left(\prod_{i=1}^{n} f(x_i)\right) = \ln\left(\prod_{i=1}^{n} \binom{N}{x_i}\right) + \left(\sum_{i=1}^{n} x_i\right) \ln(p) + \left(nN - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$

$$\Rightarrow \frac{dL(p)}{d(p)} = 0 \Rightarrow \left(\sum_{i=1}^{n} x_i\right)\left(\frac{1}{p}\right) + \left(nN - \sum_{i=1}^{n} x_i\right)\frac{1}{1-p}(-1) = 0$$

$$\sum_{i=1}^{n} x_i \left[\frac{1}{p} + \frac{1}{1-p}\right] - \frac{nN}{1-p} = 0 \Rightarrow \left(\sum_{i=1}^{n} x_i\right)\frac{1}{p(1-p)} - \frac{nN}{1-p} = 0$$

$$\Rightarrow \left(\sum_{i=1}^{n} x_i\right)\frac{1}{p} - nN = 0 \Rightarrow \hat{p} = \frac{1}{nN}\sum_{i=1}^{n} x_i$$

sample $\Rightarrow (3, 6, 2, 0, 0, 3)$ $\quad N = 10 \quad n = 6$

$$\sum_{i=1}^{n} x_i = 3+6+2+0+0+3 = 14$$

$$\hat{p} = \frac{1}{6 \cdot 10} \cdot 14 = \boxed{0.2333}$$

c) parameters a, b on a Uniform (a,b) sample size n

Uniform dist $\rightarrow f(x_i, a, b) = \begin{cases} \frac{1}{b-a} & a \le x_i \le b \\ 0 & \text{else} \end{cases}$

$L(p) = \ln \prod_{i=1}^{n} f(x_i, a, b) \Rightarrow \ln\left(\prod_{i=1}^{n} \frac{1}{(b-a)}\right) = \ln \frac{1}{(b-a)^n} = -n \ln(b-a)$

$\frac{\partial L(p)}{\partial a} = \frac{n}{b-a}$ $\rightarrow$ it is monotonically increasing

$\frac{\partial L(p)}{\partial b} = \frac{-n}{b-a}$ $\rightarrow$ it is monotonically decreasing

$\Rightarrow \boxed{\hat{a} = \min(x_1, x_2, x_3 \dots x_n) \quad \hat{b} = \max(x_1, x_2, x_3 \dots x_n)}$

d) $f(x_i, \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$

$\Rightarrow \prod_{i=1}^{n} f(x_i, \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x_i-\mu)^2}{2\sigma^2}} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2}}$

$\Rightarrow L(p) = \ln\left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2\sigma^2}}\right] = \sum_{i=1}^{n}\left[-\ln\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}(x_i-\mu)^2\right]$

$\frac{\partial L(p)}{\partial \mu} = 0 \Rightarrow \sum_{i=1}^{n} 2(-1)(x_i-\mu) \cdot \frac{(-1)}{2\sigma^2} = 0 \Rightarrow \sum_{i=1}^{n} \frac{(x_i-\mu)}{\sigma^2} = 0$

$\sum_{i=1}^{n} (x_i - \mu) = 0 \Rightarrow \boxed{\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i}$

e) $L(\theta) = \sum_{i=1}^{n} \left[-\ln(\sqrt{2\pi}\,\sigma) - \frac{1}{2\sigma^2}(x_i - \mu)^2 \right] \Rightarrow \frac{\partial L(\theta)}{\partial \sigma} = 0$

$\Rightarrow \sum_{i=1}^{n} \left[\frac{-1}{\sigma} + \frac{1}{\sigma^3}(x_i - \mu)^2 \right] = 0 \Rightarrow \frac{-n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$

$\Rightarrow \sum_{i=1}^{n} (x_i - \mu)^2 = n\sigma^2 \Rightarrow \boxed{\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}}$

f) $f(x_i, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i - \mu)^2}{2\sigma^2}}$

$L(\theta) = \sum_{i=1}^{n} \left[-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_i - \mu)^2 \right]$

$$\begin{cases} \frac{\partial L(\theta)}{\partial \mu} = 0 \Rightarrow \sum_{i=1}^{n} \frac{2}{2\sigma^2} \cdot (-1)(x_i - \mu) = 0 \\ \frac{\partial L(\theta)}{\partial \sigma^2} \Rightarrow \sum_{i=1}^{n} \left[-\frac{1}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(x_i - \mu)^2 = 0 \right] \end{cases}$$

$$\Rightarrow \begin{cases} \sum_{i=1}^{n} (x_i - \mu) = 0 \\ \frac{-n\mu}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \end{cases}$$

$$\Rightarrow \boxed{\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}$$

(5)

(2)

a/ MLE → $P(H)=\theta \quad P(T)=1-\theta \Rightarrow \theta \in [0,1]$

$P(D|\theta)=\theta^{n_H}(1-\theta)^{n_T}$

$\Rightarrow L(\theta)=\ln(P(D|\theta))=\ln[\theta^{n_H}(1-\theta)^{n_T}]=n_H\ln\theta+n_T\ln(1-\theta)$

$\frac{dL(\theta)}{d\theta}=0 \Rightarrow \frac{n_H}{\theta}-\frac{n_T}{1-\theta}=0 \Rightarrow n_H(1-\theta)=n_T\theta$

$$\hat{\theta}_{MLE}=\frac{n_H}{n_H+n_T}$$

MAP → Bayes → $\frac{P(D|\theta)P(\theta)}{P(D)}=P(\theta|D)$

$\theta_{map}=\arg\max_\theta P(\theta|D)=\arg\max_\theta \frac{P(D|\theta)P(\theta)}{P(D)}=\arg\max_\theta P(D|\theta)P(\theta)$

$=\arg\max_\theta \prod_{i=1}^{n} P(x_i|\theta)P(\theta)$

$\arg\max P(\theta|D)=\arg\max_\theta \log(P(\theta|D))=\arg\max_\theta \log\prod_{i=1}^{n} P(x_i|\theta)P(\theta)$

$=\arg\max_\theta \sum_{i=1}^{n}[\ln(P(x_i|\theta))+\ln P(\theta)]$

$\Rightarrow L(\theta)=\sum_{i=1}^{n}[\ln Bernoulli(x_i|\theta)+\ln Beta(\theta|\alpha,\beta)]$

$\frac{dL}{d\theta}=\sum_{i=1}^{n}\left[\frac{\partial}{\partial\theta}\ln(Bernoulli(x_i|\theta)+\frac{\partial}{\partial\theta}\ln Beta(\theta,\alpha,\beta)\right]$

first → $\sum_i \frac{\partial Bernoulli(x_i|\theta)}{\partial\theta}=\frac{1}{\theta}\sum x_i-\frac{1}{1-\theta}\sum(1-x_i)$

second → $\frac{\partial}{\partial\theta}\ln Beta(\theta|\alpha,\beta)=\frac{d}{d\theta}=\frac{\partial}{\partial\theta}\ln\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\theta^{\alpha-1}(1-\theta)^{\beta-1}\right]$

$=\frac{d}{d\theta}\ln\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}+\frac{d}{d\theta}\ln\theta^{\alpha-1}(1-\theta)^{\beta-1}=0+\frac{\partial}{\partial\theta}\ln\theta^{\alpha-1}(1-\theta)^{\beta-1}$

$=\frac{\partial(\alpha-1)}{\partial\theta}\frac{d}{d\theta}\ln\theta+(\beta-1)\frac{d}{\theta}(1-\theta)=\frac{\alpha-1}{\theta}\frac{\beta-1}{1-\theta}$

(b)

$$\hat{\theta}_{MAP} \Rightarrow \frac{d(L\theta)P}{d\theta} = 0$$

$$\Rightarrow \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{1}{1-\theta}\sum_{i=1}^{n}(1-x_i) + \frac{\alpha-1}{\theta} - \frac{\beta-1}{1-\theta} = 0$$

$$\Rightarrow \theta\left[\sum_{i=1}^{n}(1-x_i)+\beta-1\right] = (1-\theta)\left[\sum_{i=1}^{n} x_i + \alpha - 1\right]$$

$$\Rightarrow \theta\left[\sum_{i=1}^{n}(1-x_i) + \sum_{i=1}^{n} x_i + \beta - 1 + \alpha - 1\right] = \sum_{i=1}^{n} x_i + \alpha - 1$$

$$\Rightarrow \theta\left[\sum_{i=1}^{n} 1 + \beta + \alpha - 2\right] = \sum_{i=1}^{n} x_i + \alpha - 1$$

$\Rightarrow$ let head = 1 and tails = 0 $\Rightarrow \sum_{i=1}^{n} x_i = n_H$

$$\theta\left[\sum_{i=1}^{n} 1 + \beta + \alpha - 2\right] = \sum_{i=1}^{n} x_i + \alpha - 1$$

$$\theta\left[n + \beta + \alpha - 2\right] = n_H + \alpha - 1$$

$$\Rightarrow \boxed{\hat{\theta}_{MAP} = \frac{n_H + \alpha - 1}{n + \beta + \alpha - 2}}$$

b) thumbtack →

n	n_H	α	β	$\hat{\theta}_{MLE}$	$\hat{\theta}_{MAP}$
100	70	100	100	0.7	0.52
100	70	1000	1000	0.7	0.51
100	70	10000	10000	0.7	0.50

coin →

n	n_H	α	β	$\hat{\theta}_{MLE}$	$\hat{\theta}_{MAP}$
100	60	1	1	0.6	0.6
100	60	40	60	0.6	0.5
100	60	30	70	0.6	0.54

$$\hat{\theta}_{MLE} = \frac{n_H}{n} \text{ and } \hat{\theta}_{MAP} = \frac{n_H + \alpha - 1}{n + \beta + \alpha - 2}$$

In coin → if $\beta = 1$ & $\alpha = 1 \Rightarrow \hat{\theta}_{MLE} = \hat{\theta}_{MAP}$

★ So ⇒ MLE is a special case of MAP when the prior is uniform.

(2)

c) False ⇒ MLE approaches MAP in just one case which is $\alpha = \beta = 1$

d) True ⇒ example ⇒ coin ⇒ $n_H = 60$; $n = 100$ $\alpha = \beta = 100.000$

$\hat{\theta}_{MLE} = 0.6$ and $\hat{\theta}_{MAP} = 0.5$

thumbtack ⇒ $n_H = 40$; $n = 100$ $\alpha = \beta = 100.000$

$\hat{\theta}_{MLE} = 0.4$ and $\hat{\theta}_{MAP} = 0.5$

when we use large prior for both of them, they will have the same MAP