JAGDEESH BASAVARAJU
ASU ID: 1213004713

Statistical Machine Learning Assignment #1

1. Given Prior, $P(\theta)$ is Beta(β_H, β_T), Prove that posterior also follows Beta $(\beta_H + \alpha_H, \beta_T + \alpha_T)$.

Sol^n:- Prior, $P(\theta) = \text{Beta}(\beta_H, \beta_T)$

$$P(\theta) = \frac{\theta^{\beta_H - 1} (1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \quad \text{— (1)}$$

Posterior $P(\theta/D) \propto P(D/\theta) P(\theta)$ — (2)

Consider $P(D/\theta) P(\theta)$

$$P(D/\theta) P(\theta) = \theta^{\alpha_H} (1-\theta)^{\alpha_T} \cdot \frac{\theta^{\beta_H - 1} (1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \quad [\because \text{From (1)}]$$

$$P(D/\theta) P(\theta) = \frac{1}{B(\beta_H, \beta_T)} \theta^{\alpha_H + \beta_H - 1} (1-\theta)^{\alpha_T + \beta_T - 1} \quad \text{— (3)}$$

From (1) & (2)

$$P(\theta/D) = \frac{1}{C} P(D/\theta) P(\theta)$$

$$P(\theta/D) \propto \theta^{\alpha_H + \beta_H - 1} (1-\theta)^{\alpha_T + \beta_T - 1}$$

$$\therefore P(\theta/D) \sim \text{Beta}(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

Mean of posterior = mean of Beta $(\alpha_H + \beta_H, \alpha_T + \beta_T)$

$$= \frac{\alpha_H + \beta_H}{\alpha_H + \beta_H + \alpha_T + \beta_T}$$

Mode of posterior (Map) $= \hat{\theta}_{MAP} = \arg\max_\theta P(\theta/D)$ = mode of beta

$$= \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

2) Parameter Estimation

i) Prove that $\hat{\mu}_{MLE}$ is unbiased.

$D = x_1, x_2, \ldots, x_N \in R$ and is $N(\mu, \sigma^2)$

$$\therefore P(D/\mu,\sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$$

$$\ln P(D/\mu,\sigma) = -N \ln(\sigma\sqrt{2\pi}) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}$$

To find $\hat{\mu}_{MLE}$, $\frac{d}{d\mu} \ln(P(D/\mu,\sigma)) = 0$

$$\frac{d}{d\mu}\left[\ln P(D/\mu,\sigma)\right] = \frac{d}{d\mu}\left[-N \ln(\sigma\sqrt{2\pi}) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{-1}{2\sigma^2} \sum_{i=1}^{N} 2(x_i-\mu).(-1) = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i-\mu) = 0$$

$$\therefore \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \mu = 0$$

$$\sum_{i=1}^{N} x_i = N\mu$$

$$\therefore \hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Expected value of $\hat{\mu}_{MLE} = E(\hat{\mu}_{MLE}) = \frac{1}{N} \sum_{i=1}^{N} E(x_i)$

$$= \frac{1}{N} \sum_{i=1}^{N} \mu = \frac{1}{N} . N\mu$$

$$= \mu$$

$$\text{Bias} = E(\hat{\mu}_{MLE}) - \mu = \mu - \mu = 0$$

$\therefore \hat{\mu}_{MLE}$ is unbiased.

2 ii) Prove that $\hat{\sigma}^2_{MLE}$ is biased.

$$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_{MLE})^2$$

Expected value of $\hat{\sigma}^2_{MLE} = E(\hat{\sigma}^2_{MLE})$

$$= \frac{1}{N}\sum_{i=1}^{N} E(x_i - \hat{\mu}_{MLE})^2$$

$E(x_i - \hat{\mu}_{MLE})^2$ can't be equal to σ^2 as $\hat{\mu}_{MLE}$ is just an estimate and we don't know true value.

$\therefore$ Bias $= E(\hat{\sigma}^2_{MLE}) - \sigma^2$

$$= \left\{\frac{1}{N}\sum_{i=1}^{N} E(x_i - \hat{\mu}_{MLE})^2\right\} - \sigma^2$$

$$= \left\{\frac{1}{N}\sum_{i=1}^{N} E\left(x_i^2 - 2x_i\hat{\mu}_{MLE} + \hat{\mu}^2_{MLE}\right)\right\} - \sigma^2$$

Bias $\neq 0$

Hence $\hat{\sigma}^2_{MLE}$ is biased.

$$= \left\{\frac{1}{N}\sum_{i=1}^{N}\left(E(x_i^2) - 2\hat{\mu}_{MLE}E(x_i) + E(\hat{\mu}^2_{MLE})\right)\right\} - \sigma^2$$

$$= \left\{\frac{1}{N}\sum_{i=1}^{N}\left(E(x_i^2) - 2\hat{\mu}_{MLE}\hat{\mu}_{MLE} + \hat{\mu}^2_{MLE}\right)\right\} - \sigma^2$$

$$= \left\{\frac{1}{N}\sum_{i=1}^{N} E(x_i^2) - 2\hat{\mu}^2_{MLE} + \mu^2_{MLE}\right\} - \sigma^2$$

$$= \left\{\frac{1}{N}\sum_{i=1}^{N} E(x_i^2) - \hat{\mu}^2_{MLE}\right\} - \left\{E(x_i^2) - \mu^2\right\}$$

Bias $\neq 0$

Hence $\hat{\sigma}^2_{MLE}$ is biased

3) Naive Bayes Classifier

i) 13 independent parameters

Consider age as X_1, income as X_2, student as X_3, credit_rating as X_4 and buys_computer as Y.

1) $P(Y = yes)$
2) $P(X_1 = youth / Y = yes)$
3) $P(X_1 = middle\_aged / Y = yes)$
4) $P(X_2 = high / Y = yes)$
5) $P(X_2 = medium / Y = yes)$
6) $P(X_3 = yes / Y = yes)$
7) $P(X_4 = fair / Y = yes$
8) $P(X_1 = youth / Y = no)$
9) $P(X_1 = middle\_aged / Y = no)$
10) $P(X_2 = high / Y = no)$
11) $P(X_2 = medium / Y = no)$
12) $P(X_3 = yes / Y = no)$
13) $P(X_4 = fair / Y = no)$

Since the class is binary, we need one parameter (1)
Age has 3 distinct values, so need 2 parameters ~~when~~ per class (4)
Income also has 3 distinct values, so need 2 per class (4)
Student has 2 distinct values, so 1 per class (2)
Credit_rating ~~also~~ has 2 distinct values, so 1 per class (2)
Hence totally 13 independent parameters

3 ii) Estimated value of all these parameters

Age : X_1, Income : X_2 Student : X_3 Credit_rating : X_4

Buys_Computer : Y

1) $P(Y = yes) = 9/14$ $\therefore P(Y = no) = 5/14$

2) $P(X_1 = youth / Y = yes) = 2/9$

3) $P(X_1 = middle\_aged / Y = yes) = 4/9$ $\therefore P(X_1 = senior / Y = yes) = 3/9$

4) $P(X_2 = high / Y = yes) = 2/9$

5) $P(X_2 = medium / Y = yes) = 4/9$ $\therefore P(X_2 = low / Y = yes) = 3/9$

6) $P(X_3 = yes / Y = yes) = 6/9$ $\therefore P(X_3 = no / Y = yes) = 3/9$

7) $P(X_4 = fair / Y = yes) = 6/9$ $\therefore P(X_4 = excellent / Y = yes) = 3/9$

8) $P(X_1 = youth / Y = no) = 3/5$

9) $P(X_1 = middle\_aged / Y = no) = 0/5 = 0$ $\therefore P(X_1 = senior / Y = no) = 2/5$

10) $P(X_2 = high / Y = no) = 2/5$

11) $P(X_2 = medium / Y = no) = 2/5$ $\therefore P(X_2 = low / Y = no) = 1/5$

12) $P(X_3 = yes / Y = no) = 1/5$ $\therefore P(X_3 = no / Y = no) = 4/5$

13) $P(X_4 = fair / Y = no) = 2/5$ $\therefore P(X_4 = excellent / Y = no) = 3/5$

3 iii) $X = (\text{youth}, \text{medium}, \text{yes}, \text{fair})$, find $P(y=\text{yes} \mid x)$

$$P(Y=1 \mid X_1 \ldots X_n) = \frac{P(Y=1)\, P(X_1 \ldots X_n / Y=1)}{P(Y=1)\, P(X_1 \ldots X_n \mid Y=1) + P(Y=0)\, P(X_1 \ldots X_n \mid Y=0)}$$

$$= \frac{P(Y=1)\; P(X_1 \mid Y=1) \ldots P(X_n \mid Y=1)}{P(Y=1)\, P(X_1 \mid Y=1) \ldots P(X_n \mid Y=1) + P(Y=0)\, P(X_1 \mid Y=0) \ldots P(X_n \mid Y=0)}$$

$$\therefore P(Y=\text{yes} \mid X) = P(Y=\text{yes})\, P(X_1=\text{youth} \mid Y=\text{yes})\, P(X_2=\text{medium} \mid Y=\text{yes})\, P(X_3=\text{yes} \mid Y=\text{yes}) \cdot P(X_4=\text{fair} \mid Y=\text{yes})$$

$$\therefore P(Y=\text{yes} \mid X) = \frac{\left(\frac{9}{14}\right)\left(\frac{2}{9}\right)\left(\frac{4}{9}\right)\left(\frac{6}{9}\right)\left(\frac{6}{9}\right)}{\left[\left(\frac{9}{14}\right)\left(\frac{2}{9}\right)\left(\frac{4}{9}\right)\left(\frac{6}{9}\right)\left(\frac{6}{9}\right) + \left(\frac{5}{14}\right)\left(\frac{3}{5}\right)\left(\frac{2}{5}\right)\left(\frac{1}{5}\right)\left(\frac{2}{5}\right)\right]}$$

$$= \frac{0.0282}{[0.0282 + 0.00685 7]} = \frac{0.0282}{0.035057}$$

$P(Y=\text{yes} \mid X) = 0.8044$

Since its greater than 0.5, the classifier has predicted $y=$ yes for this person.

i.e buys_computer = yes

4. Logistic Regression

$X_1 = (1,0)$ $X_2 = (0,-1)$ $X_3 = (0,1)$ $X_4 = (-1,0)$

$y_1 = 1$ $y_2 = 1$ $y_3 = 0$ $y_4 = 0$

i) Initial weight vector, $w^{(0)} = (0,0,0)^T$

Iteration 1: $\hat{P}(y^j = 1 / x^j, w^{(t)}) = \dfrac{\exp(w_0^{(t)} + \sum_{i=1}^{2} w_i^{(t)} x_i)}{1 + \exp(w_0^{(t)} + \sum_{i=1}^{2} w_i^{(t)} x_i)}$

$j = 1,2,3,4$

$$\hat{P}(Y^1 = 1 / X^1, w^{(0)}) = \frac{\exp(0 + 0\times 1 + 0\times 0)}{1 + \exp(0 + 0\times 1 + 0\times 0)} = \frac{1}{2} = 0.5$$

$$\hat{P}(Y^2 = 1 / X^2, w^{(0)}) = \frac{\exp(0 + 0\times 0 + 0\times(-1))}{1 + \exp(0 + 0\times 0 + 0\times(-1))} = \frac{1}{2} = 0.5$$

$$\hat{P}(Y^3 = 1 / X^3, w^{(0)}) = \frac{\exp(0 + 0\times 0 + 0\times 1)}{1 + \exp(0 + 0\times 0 + 0\times 1)} = \frac{1}{2} = 0.5$$

$$\hat{P}(Y^4 = 1 / X^4, w^{(0)}) = \frac{\exp(0 + 0\times(-1) + 0\times 0)}{1 + \exp(0 + 0\times(-1) + 0\times 0)} = \frac{1}{2} = 0.5$$

$$w_0^{(t+1)} = w_0^{(t)} + \eta \sum_j [y^j - \hat{P}(y^j = 1 / x^j, w^{(t)})]$$

$$\therefore w_0^{(1)} = w_0^{(0)} + \eta [(1-0.5) + (1-0.5) + (0-0.5) + (0-0.5)]$$

$$= 0 + \eta [0] = 0$$

$$w_i^{(t+1)} = w_i^{(t)} + \eta \sum_j x_i^j [y^j - \hat{P}(y^j = 1 / x^j, w^{(t)})]$$

$i = 1,2$

$$\therefore w_1^{(1)} = w_1^{(0)} + \eta [1(1-0.5) + 0(1-0.5) + 0(0-0.5) + (-1)(0-0.5)]$$

$$= 0 + \eta [0.5 + 0.5] = \eta$$

$$\therefore w_2^{(1)} = w_2^{(0)} + \eta [0(1-0.5) + (-1)(1-0.5) + 1(0-0.5) + 0(0-0.5)]$$

$$= 0 + \eta [-0.5 - 0.5] = -\eta$$

$\therefore w^{(1)} = (0, \eta, -\eta)'$

Iteration 2 :-

$$\hat{P}(Y^1=1 / X^1, w^{(1)}) = \frac{\exp(0+\eta(1)+(-\eta)(0))}{1+\exp(0+\eta(1)+(-\eta)(0))} = \frac{\exp(\eta)}{1+\exp(\eta)}$$

$$\hat{P}(Y^2=1 / X^2, w^{(1)}) = \frac{\exp(0+\eta(0)+(-\eta)(-1))}{1+\exp(0+\eta(0)+(-\eta)(-1))} = \frac{\exp(\eta)}{1+\exp(\eta)}$$

$$\hat{P}(Y^3=1 / X^3, w^{(1)}) = \frac{\exp(0+\eta(0)+(-\eta)(1))}{1+\exp(0+\eta(0)+(-\eta)(1))} = \frac{\exp(-\eta)}{1+\exp(-\eta)} = \frac{1}{1+\exp(\eta)}$$

$$\hat{P}(Y^4=1 / X^4, w^{(1)}) = \frac{\exp(0+\eta(-1)+(-\eta)(0))}{1+\exp(0+\eta(-1)+(-\eta)(0))} = \frac{\exp(-\eta)}{1+\exp(-\eta)} = \frac{1}{1+\exp(\eta)}$$

$$w_0^{(2)} = w_0^{(1)} + \eta\left[\left(1-\frac{\exp(\eta)}{1+\exp(\eta)}\right)+\left(1-\frac{\exp(\eta)}{1+\exp(\eta)}\right)+\left(0-\frac{1}{1+\exp(\eta)}\right)+\left(0-\frac{1}{1+\exp(\eta)}\right)\right]$$

$$= 0 + \eta\left[\frac{2}{1+\exp(\eta)} - \frac{2}{1+\exp(\eta)}\right] = 0$$

$$w_1^{(2)} = w_1^{(1)} + \eta\left[1\left(\frac{1}{1+\exp(\eta)}\right)+0\left(\frac{1}{1+\exp(\eta)}\right)+0\left(\frac{-1}{1+\exp(\eta)}\right)+(-1)\left(\frac{-1}{1+\exp(\eta)}\right)\right]$$

$$= \eta + \eta\left[\frac{2}{1+\exp(\eta)}\right] = \eta + \frac{2\eta}{1+\exp(\eta)}$$

$$w_2^{(2)} = w_2^{(1)} + \eta\left[0\left(\frac{1}{1+\exp(\eta)}\right)+(-1)\left(\frac{1}{1+\exp(\eta)}\right)+1\left(\frac{-1}{1+\exp(\eta)}\right)+0\left(\frac{-1}{1+\exp(\eta)}\right)\right]$$

$$= -\eta + \eta\left[\frac{-2}{1+\exp(\eta)}\right] = -\eta - \frac{2\eta}{1+\exp(\eta)}$$

$$\therefore w^{(2)} = \left(0, \eta + \frac{2\eta}{1+\exp(\eta)}, -\eta - \frac{2\eta}{1+\exp(\eta)}\right)'$$

As we continue the number of iterations like this, we get a final w, i.e $w^{(*)} = (0, \infty, -\infty)'$

Hence $w^{(*)} = (0, \infty, -\infty)'$ when $w^{(0)} = (0, 0, 0)'$

4 ii) Initial weight vector, $w^{(0)} = (0, 1, 0)^T$

Iteration 1:

$$\hat{P}(Y^1=1 / X^1, w^{(0)}) = \frac{\exp(0+1(1)+0(0))}{1+\exp(0+1(1)+0(0))} = \frac{\exp(1)}{1+\exp(1)} = \frac{e}{1+e}$$

$$\hat{P}(Y^2=1 / X^2, w^{(0)}) = \frac{\exp(0+1(0)+0(-1))}{1+\exp(0+1(0)+0(-1))} = \frac{1}{2} = 0.5$$

$$\hat{P}(Y^3=1 / X^3, w^{(0)}) = \frac{\exp(0+1(0)+0(1))}{1+\exp(0+1(0)+0(-1))} = \frac{1}{2} = 0.5$$

$$\hat{P}(Y^4=1 / X^4, w^{(0)}) = \frac{\exp(0+1(-1)+0(0))}{1+\exp(0+1(-1)+0(0))} = \frac{e^{-1}}{1+e^{-1}} = \frac{1}{1+e}$$

$$\therefore w_0^{(1)} = w_0^{(0)} + \eta\left[(1-\frac{e}{1+e}) + (1-0.5) + (0-0.5) + (0-\frac{1}{1+e})\right]$$

$$= 0 + \eta\left[(\frac{1}{1+e}) + (0.5) + (-0.5) + (\frac{-1}{1+e})\right] = 0 + \eta(0) = 0$$

$$\therefore w_1^{(1)} = w_1^{(0)} + \eta\left[1(\frac{1}{1+e}) + 0(0.5) + 0(-0.5) + (-1)(\frac{-1}{1+e})\right]$$

$$= 1 + \eta\left[\frac{2}{1+e}\right] = 1 + \frac{2\eta}{1+e}$$

$$\therefore w_2^{(1)} = w_2^{(0)} + \eta\left[0(\frac{1}{1+e}) + (-1)(0.5) + 1(-0.5) + 0(\frac{-1}{1+e})\right]$$

$$= 0 + \eta\left[-0.5 - 0.5\right] = -\eta$$

$$\therefore w^{(1)} = (0, 1+\frac{2\eta}{1+e}, -\eta)^T$$

Iteration 2:

$$\hat{P}(Y^1=1 / X^1, w^{(1)}) = \frac{\exp(0+(1+\frac{2\eta}{1+e})(1)+(-\eta)(0))}{1+\exp(0+(1+\frac{2\eta}{1+e})(1)+(-\eta)(0))}$$

$$= \frac{\exp(1+\frac{2\eta}{1+e})}{1+\exp(1+\frac{2\eta}{1+e})}$$

$$\hat{P}(Y^2=1 \mid X^2, w^{(1)}) = \frac{\exp\left(0 + \left(1+\frac{2\eta}{1+\epsilon}\right)(0) + (-\eta)(-1)\right)}{1+\exp\left(0+\left(1+\frac{2\eta}{1+\epsilon}\right)(0)+(-\eta)(-1)\right)}$$

$$= \frac{\exp(\eta)}{1+\exp(\eta)}$$

$$\hat{P}(Y^3=1 \mid X^3, w^{(1)}) = \frac{\exp\left(0 + \left(1+\frac{2\eta}{1+\epsilon}\right)(0) + (-\eta)(1)\right)}{1+\exp\left(0+\left(1+\frac{2\eta}{1+\epsilon}\right)(0)+(-\eta)(1)\right)}$$

$$= \frac{\exp(-\eta)}{1+\exp(-\eta)} = \frac{1}{1+\exp(\eta)}$$

$$\hat{P}(Y^4=1 \mid X^4, w^{(1)}) = \frac{\exp\left(0 + \left(1+\frac{2\eta}{1+\epsilon}\right)(-1) + (-\eta)(0)\right)}{1+\exp\left(0+\left(1+\frac{2\eta}{1+\epsilon}\right)(-1)+(-\eta)(0)\right)}$$

$$= \frac{\exp\left(-\left(1+\frac{2\eta}{1+\epsilon}\right)\right)}{1+\exp\left(-\left(1+\frac{2\eta}{1+\epsilon}\right)\right)} = \frac{1}{1+\exp\left(1+\frac{2\eta}{1+\epsilon}\right)}$$

$$\therefore w_0^{(2)} = w_0^{(1)} + \eta\left[\left(1 - \frac{\exp\left(1+\frac{2\eta}{1+\epsilon}\right)}{1+\exp\left(1+\frac{2\eta}{1+\epsilon}\right)}\right) + \left(1 - \frac{\exp(\eta)}{1+\exp(\eta)}\right) + \left(0 - \frac{1}{1+\exp(\eta)}\right) + \left(0 - \frac{1}{1+\exp\left(1+\frac{2\eta}{1+\epsilon}\right)}\right)\right]$$

$$= 0 + \eta\left[\frac{1}{1+\exp\left(1+\frac{2\eta}{1+\epsilon}\right)} + \frac{1}{1+\exp(\eta)} - \frac{1}{1+\exp(\eta)} - \frac{1}{1+\exp\left(1+\frac{2\eta}{1+\epsilon}\right)}\right]$$

$$= 0 + \eta[0] = 0$$

$$\therefore \quad w_1^{(2)} = w_1^{(1)} + \eta\left[1.\left(\frac{1}{1+\exp\left(1+\frac{2\eta}{1+e}\right)}\right) + 0(\,) + 0(\,) + (-1)\left(\frac{-1}{1+\exp\left(1+\frac{2\eta}{1+e}\right)}\right)\right]$$

$$= \left(1+\frac{2\eta}{1+e}\right) + \eta\left[\frac{2}{1+\exp\left(1+\frac{2\eta}{1+e}\right)}\right]$$

$$= 1 + \frac{2\eta}{1+e} + \frac{2\eta}{1+\exp\left(1+\frac{2\eta}{1+e}\right)}$$

$$w_2^{(2)} = w_2^{(1)} + \eta\left[0(\,) + (-1)\left(\frac{1}{1+\exp(\eta)}\right) + 1\left(\frac{-1}{1+\exp(\eta)}\right) + 0(\,)\right]$$

$$= -\eta + \eta\left[\frac{-2}{1+\exp(\eta)}\right] = -\eta - \frac{2\eta}{1+\exp(\eta)}$$

$$= -\left[\eta + \frac{2\eta}{1+\exp(\eta)}\right]$$

$$\therefore w^{(2)} = \left(0,\ 1+\frac{2\eta}{1+e} + \frac{2\eta}{1+\exp\left(1+\frac{2\eta}{1+e}\right)},\ -\left[\eta + \frac{2\eta}{1+\exp(\eta)}\right]\right)^T$$

As we continue the number of iterations, the pattern continues and we get a final w.

i.e $w^{(*)} = (0, \infty, -\infty)^T$

Hence the final vector, $w^{(*)}$ will be the same for the two different initial values. ∎

5) Naive Bayes Classifier and Logistic Regression

i) Gaussian Naive Bayes & Logistic Regression

– Number of independent parameters in GNB?

Sol^n :- $4d+1$, d being the number of features

1{ $P(Y=1)$

d { $\mu_{1,0}$, $\mu_{2,0}$, $\mu_{3,0}$, ⋮, $\mu_{d,0}$ } d { $\mu_{1,1}$, $\mu_{2,1}$, $\mu_{3,1}$, ⋮, $\mu_{d,1}$ } d { $\sigma_{1,0}$, $\sigma_{2,0}$, $\sigma_{3,0}$, ⋮, $\sigma_{d,0}$ } d { $\sigma_{1,1}$, $\sigma_{2,1}$, $\sigma_{3,1}$, ⋮, $\sigma_{d,1}$ }

Hence $4d+1$ independent parameters

– No, 'w' cannot be translated into the parameters of an equivalent GNB classifier without any extra assumption.

Extra assumption : Variance is the same across all the classes.

i.e $\sigma_{iK} = \sigma_i$ $\forall$ K (no. of classes)

With this assumption, w can be translated.

Explanation :

From Gaussian Naive Bayes,

$$P(Y=1/X) = \frac{P(Y=1)\,(P(X/Y=1))}{P(Y=1)\,P(X/Y=1) + P(Y=0)\,P(X/Y=0)}$$

$$P(Y=1/X) = \frac{1}{1+\exp\left(\ln\left(\frac{P(Y=0)}{P(Y=1)}\right) + \sum_i \ln \frac{P(X_i/Y=0)}{P(X_i/Y=1)}\right)}$$

Assuming $P(Y=1)=\theta$, $P(Y=0)=1-\theta$ & $\sigma_{ik}=\sigma_i$, i.e variance same across classes

we get,

$$P(Y=1/X) = \frac{1}{1+\exp\left(\left(\ln\frac{1-\theta}{\theta}\right)+\sum_i\left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}\right)X_i + \left(\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right)\right)}$$

This is of the form

$$P(Y=1/X) = \frac{1}{1+\exp\left(w_0+\sum_i w_i X_i\right)}$$

which is the form of Logistic Regression. Hence parameters of logistic regression W, can be translated into equivalent Gaussian Naive Bayes classifier.

Question 5.2: Implementation of Gaussian Naïve Bayes and Logistic Regression

i) Pseudocode

Load X(Samples), y(class) from the dataset

For fraction in fractions (0.01, 0.02, 0.05, 0.1, 0.625, 1)

Divide data into 3 folds making one of the folds test and other 2 folds training set

We will end up with 3 different combinations of train, test data

In each such combination, randomly pick a fraction of train data

Repeat the above step 5 times so that we will have 5 sets of data for each combination

Train the model for each of the set and learn the parameters

Predict the class for the test data from the learnt parameters and calculate the accuracy

Calculate the mean of accuracies which will be the accuracy for the fraction

Plot accuracy vs the fraction size curve

Gaussian Naïve Bayes

Function **Train**(X_train, y_train):

P(y=1) = number of positive samples / total number of samples

Find $Mean_{i,1}$ for training data of class y=1

Find $Mean_{i,0}$ for training data of class y=0

Find $Variance_{i,1}$ for training data of class y=1

Find $Variance_{i,0}$ for training data of class y=0

Learnt_parameters = P(y=1), $Mean_{i,1}$, $Mean_{i,0}$, $Variance_{i,1}$, $Variance_{i,0}$

Return Learnt_parameters

Function **Predict**(X_test, y_test, Learnt_parameters):

$P(X/Y=1) = (1/sqrt(2*pi* Variance_{i,1})) \exp(-((X\_test - Mean_{i,1})^2)/(2* Variance_{i,1}))$

$H\_pos = P(Y=1) * \prod_I P(X_i/Y=1)$

$P(X/Y=0) = (1/sqrt(2*pi* Variance_{i,0})) \exp(-((X\_test - Mean_{i,0})^2)/(2* Variance_{i,0}))$

$H\_zero = P(Y=0) * \prod_I P(X_i/Y=0)$

If H_pos > H-zero then

y_pred = 1.0

else

y_pred = 0.0

Accuracy = (# y_pred == y_test) / (# y_pred)

Return Accuracy

Logistic Regression

Function **train**(X_train, y_train, learning_rate)

W_0=0, W=(0, 0, 0, 0)

For 500 iterations

$Z = w_0 + \sum_i w_i X\_train_i$

P(Y=1/X_train,w) = exp(z)/(1+exp(z)) = 1/(1+exp(-z)) = sigmoid(-z) = a

$w\_0 = w\_0 + learning\_rate * (\sum_j (y\_train^j - P(Y=1/X\_train^j,w)))$

$w = w + learning\_rate * (\sum_j X^j(y\_train^j - P(Y=1/X\_train^j,w)))$

return w, w_0

Function **predict**(X_test, y_test, w, w_0)

$Z = w_0 + \sum_i w_i X\_test_i$

P(Y=1/X_test,w) = exp(z)/(1+exp(z)) = 1/(1+exp(-z)) = sigmoid(-z) = a

If P(Y=1/X_test,w) > 0.5 then

y_pred = 1.0

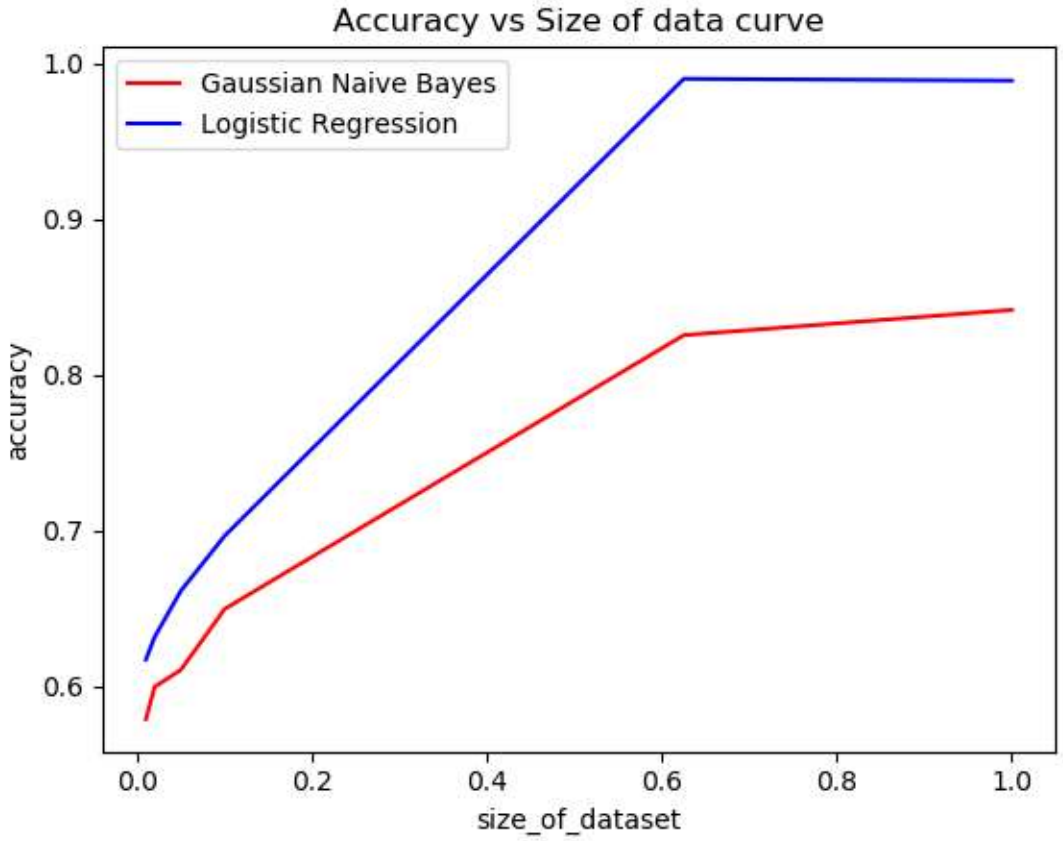
else

y_pred = 0.0

Accuracy = (# y_pred == y_test) / (# y_pred)

Return Accuracy

ii) Learning curve



iii) Generate examples and compare their mean and variance with the training data

Fold 1

Accuracy = 0.838074398249
Mean of training set: [-1.92548409 -1.22379259 2.43710485 -1.18223808]
Mean of generated set: [-1.87141318 -1.2079421 2.60292242 -1.20923457]
Percentage of Mean Deviated: [2.80817191 1.2951936 6.37043817 2.23252709]
Variance of training set: [3.70027822 30.50494584 28.38578698 4.1709176]
Variance of generated set: [3.44684103 28.88546879 29.63427956 3.83852848]
Percentage of Variance Deviated: [6.84913883 5.30889996 4.21300131 7.96920828]

Fold 2

Accuracy = 0.851203501094
Mean of training set: [-1.83526396 -0.80990072 1.90860182 -1.33436393]
Mean of generated set: [-1.8923333 -1.10565946 2.0417084 -1.37504932]
Percentage of Mean Deviated: [3.01581866 26.74953262 6.51937279 2.9588308]
Variance of training set: [3.51025097 29.10009686 27.13543712 4.59173298]
Variance of generated set: [3.31589324 31.0324415 29.20502293 4.53426649]
Percentage of Variance Deviated: [5.53686119 6.22685341 7.08640364 1.25152073]

Fold 3

Accuracy = 0.835886214442
Mean of training set: [-1.84221312 -0.93518316 2.0838102 -1.22831574]
Mean of generated set: [-1.95093491 -1.15531175 1.98556775 -1.19001075]
Percentage of Mean Deviated: [5.5728046 19.05360942 4.71455866 3.11849693]
Variance of training set: [3.38007627 27.77767508 27.22473826 4.08723726]
Variance of generated set: [3.35737334 31.72728121 26.58498813 3.94825671]
Percentage of Variance Deviated: [0.67166924 12.44861199 2.34988531 3.40035454]

As it can be observed, the mean and variance of the training samples is matching with the mean and variance of the generated samples. This is because both are generated from the same underlying gaussian distribution, N(mean,variance).