Gaussian Naïve Bayes

Nazerfard, Ehsan
nazerfard@aut.ac.ir

Gaussian/Normal Distribution

- A lot of things in nature follow the normal distribution, such as people's height, etc.

Gaussian Distribution

- A type of continuous distribution
- (Univariate) probability density function:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

  - where μ and σ^2 are the mean and variance of the distribution.

Gaussian Distribution (cont.)

- (Multivariate) probability density function:

$$f_{\mathbf{X}}(x_1, \dots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X}-\boldsymbol{\mu})}$$

$$\mathbf{X} = (X_1, \dots, X_{\mathrm{k}})^T$$

- where $\mu_{k\times 1}$ and $\Sigma_{\mathrm{k\times k}}$ are the mean vector and covariance matrix of the distribution:

- $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}] = \left[\mathbb{E}[X_1], \dots, \mathbb{E}[X_k]\right]^T$

- $\Sigma = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}}]$

Gaussian naïve Bayes (GNB)

- How to distinguish <u>c</u>hildren from <u>a</u>dults, using a Gaussian naïve Bayes classifier?
  - Classes: {a,c}
  - Features: <u>h</u>eight and <u>w</u>eight (continuous inputs)

- Training data $\{h_i, w_i, y_i\}$
  - 12 <u>c</u>hildren (black)
  - 4 <u>a</u>dults (red)

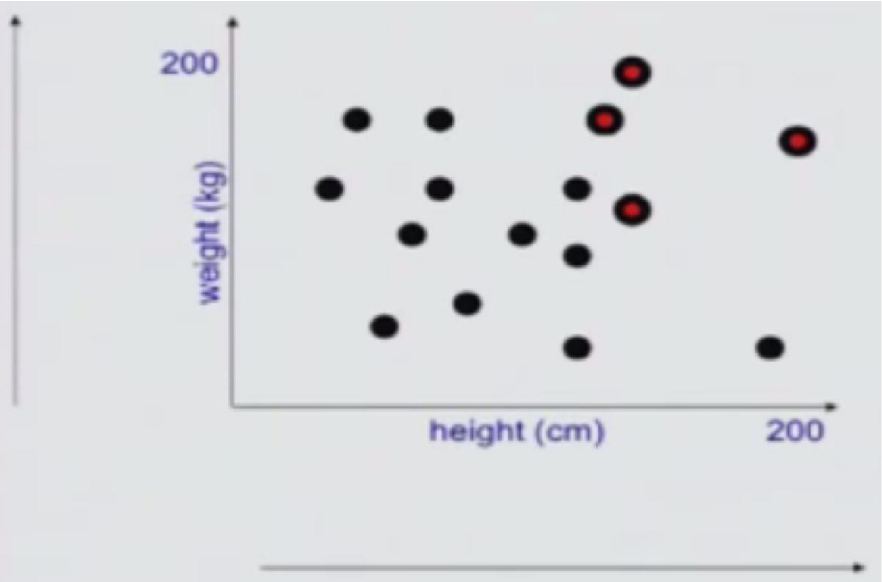


Figure © Victor Lavrenko, Lecture on GNB, University of Edinburgh

Prior Probabilities

- Prior class probabilities:
  - $p(Y = a) = 4/16 = 0.25$
  - $p(Y = c) = 12/16 = 0.75$

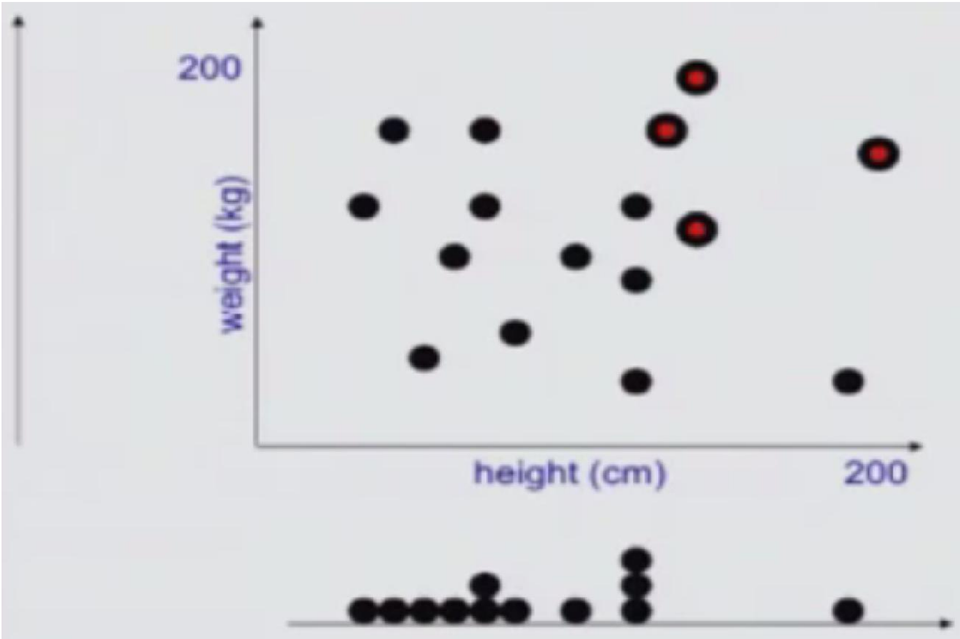


Figure © Victor Lavrenko, Lecture on GNB, University of Edinburgh

Children’s Features

- Assuming that children’s heights and weights follow the Gaussian distribution:

  - $\mu_{h,c} = \frac{1}{12}\sum_{i:y_i=c} h^i$
  - $\sigma^2_{h,c} = \frac{1}{12-1}\sum_{i:y_i=c}\left(h^i - \mu_{h,c}\right)^2$
  - $\mu_{w,c} = \frac{1}{12}\sum_{i:y_i=c} w^i$
  - $\sigma^2_{w,c} = \frac{1}{12-1}\sum_{i:y_i=c}\left(w^i - \mu_{w,c}\right)^2$

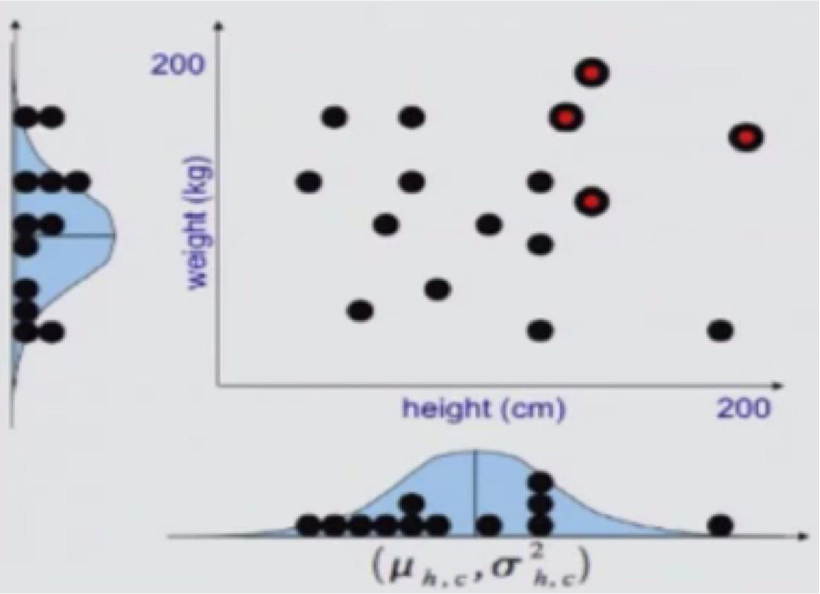


Figure © Victor Lavrenko, Lecture on GNB, University of Edinburgh

2-Dimensional Gaussian Distribution

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \; \Sigma = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

where $\rho = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$ is the correlation coefficient between X and Y.

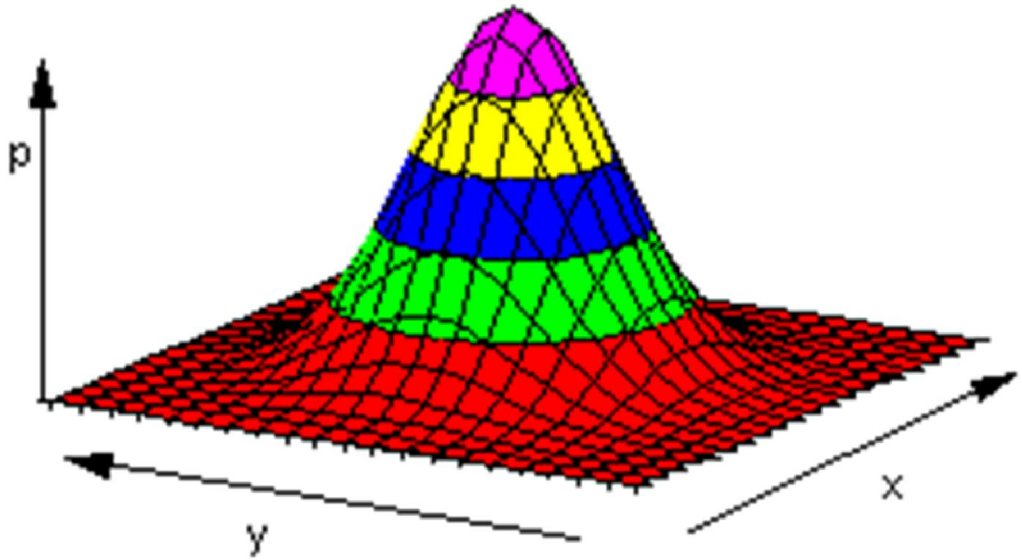
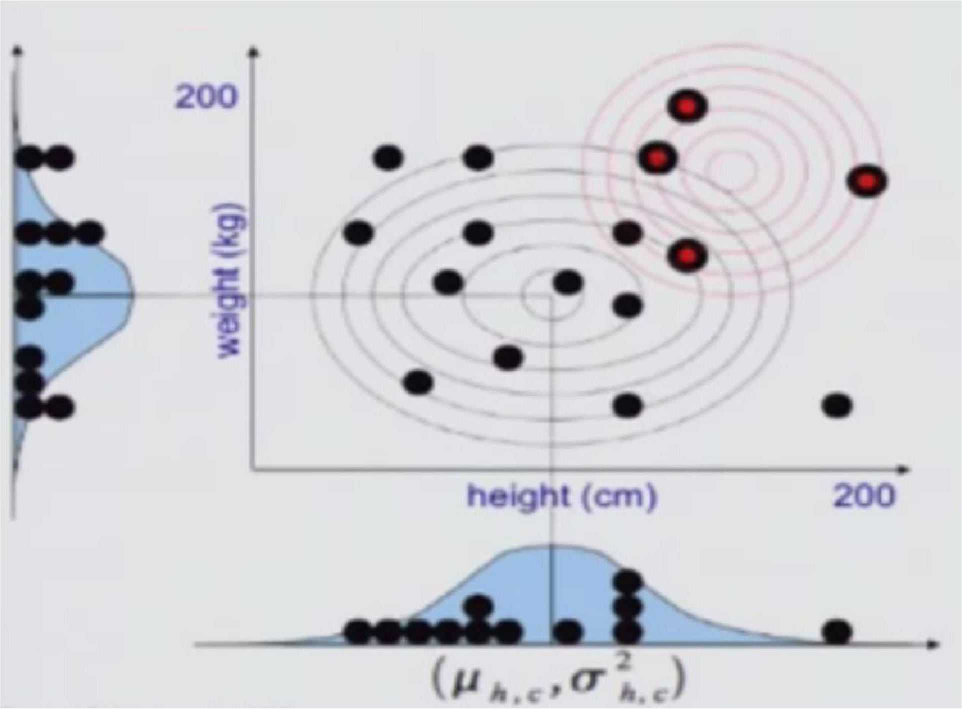


Figure © https://www.ma.utexas.edu/

Independence Assumption

- Multiplying two Gaussian distributions



- The same thing for the adult's features

Figure © Victor Lavrenko, Lecture on GNB, University of Edinburgh

Test Phase

- $X =< h_X, w_X >$ (test data)
  - $p(h_X|Y = a) = \frac{1}{\sqrt{2\pi\sigma_{h,a}^2}} \exp(-\frac{(h_X-\mu_{h,a})^2}{2\sigma_{h,a}^2})$
  - $p(w_X|Y = a) = \frac{1}{\sqrt{2\pi\sigma_{w,a}^2}} \exp(-\frac{(w_X-\mu_{w,a})^2}{2\sigma_{w,a}^2})$
- Similarly compute $p(h_X|Y = c)$ an d$p(w_X|Y = c)$

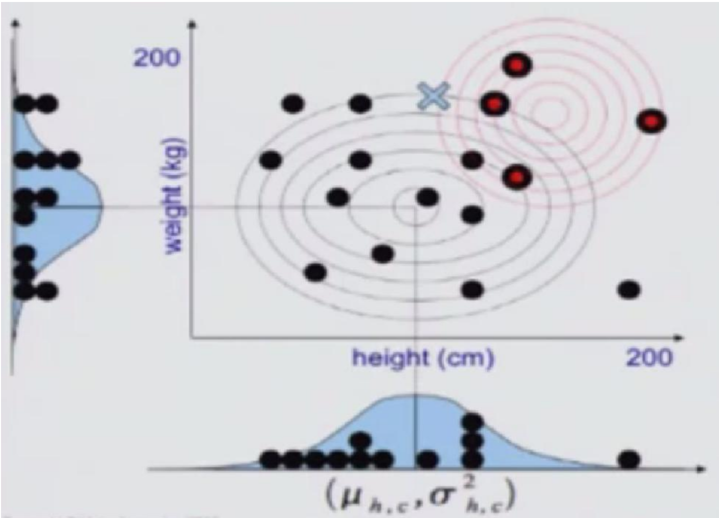


Figure © Victor Lavrenko, Lecture on GNB, University of Edinburgh

Classification

- Classification Rule:

$$y_{MLE} = \underset{y \in \{a,c\}}{\operatorname{argmax}}\ p(Y = y|X) = \underset{y \in \{a,c\}}{\operatorname{argmax}}\ p(X|Y = y)\, p(Y = y)$$

- Test data: $X =< h_X, w_X >$

If $p(< h_X, w_X > |Y = a)\, p(Y = a) > p(< h_X, w_X > |Y = c)\, p(Y = c)$

$$\xrightarrow{\text{In depen cassumption}}$$

If $p(h_X|Y = a)\, p(w_X|Y = a)\, p(Y = a) > p(h_X|Y = c)\, p(w_X|Y = c)\, p(Y = c)$

Then Class is <u>a</u>dult

Else Class is <u>c</u>hildren

A Gentle Correction

- Population mean vs. Sample mean

$$\mu_{h,c} = \frac{1}{12} \sum_{i:y_i=c} h^i \rightarrow \bar{X}_{h,c} = \frac{1}{12} \sum_{i:y_i=c} h^i$$

- Population variance vs. Sample variance

$$\sigma_{h,c}^2 = \frac{1}{12-1} \sum_{i:y_i=c} \left(h^i - \mu_{h,c}\right)^2 \rightarrow S_{h,c}^2 = \frac{1}{12-1} \sum_{i:y_i=c} \left(h^i - \bar{X}_{h,c}\right)^2$$

Parametric Form of P(Y|X) for GNB Classifier

- Assumptions
  - Y is boolean, governed by a Bernoulli distribution
  - $\boldsymbol{X} =< X_1, X_2, \dots, X_n >$, where X_i is a continuous random variable
  - For each X_i, $P(X_i|Y = y_k)$ is a Gaussian distribution of the form $N(\mu_{ik}, \sigma_i)$
  - For all i and $j \neq i$, X_i and X_j are conditionally independent given Y

GNB Parametric Form of P(Y|X)

- We derive the parametric form of $p(Y|X)$ that follows from the set of GNB assumptions:

$$p(Y=1|\boldsymbol{X}) = \frac{p(X|Y=1)p(Y=1)}{p(X)} = \frac{p(\boldsymbol{X}|Y=1)p(Y=1)}{p(\boldsymbol{X}|Y=1)p(Y=1)+p(\boldsymbol{X}|Y=0)p(Y=0)}$$

$$= \frac{1}{1+\frac{p(\boldsymbol{X}|Y=0)p(Y=0)}{p(\boldsymbol{X}|Y=1)p(Y=1)}}$$

GNB Parametric Form of P(Y|X)

- We derive the parametric form of $p(Y|X)$ that follows from the set of GNB assumptions:

$$p(Y=1|\boldsymbol{X}) = \frac{p(X|Y=1)p(Y=1)}{p(X)} = \frac{p(\boldsymbol{X}|Y=1)p(Y=1)}{p(\boldsymbol{X}|Y=1)p(Y=1)+p(\boldsymbol{X}|Y=0)p(Y=0)}$$

$$= \frac{1}{1+\frac{p(\boldsymbol{X}|Y=0)p(Y=0)}{p(\boldsymbol{X}|Y=1)p(Y=1)}} = \frac{1}{1+\exp[\ln\frac{p(\boldsymbol{X}|Y=0)p(Y=0)}{p(\boldsymbol{X}|Y=1)p(Y=1)}]}$$

$$= \frac{1}{1+\exp[\ln\frac{p(\boldsymbol{X}|Y=0)}{p(\boldsymbol{X}|Y=1)}+\ln\frac{p(Y=0)}{p(Y=1)}]} = \frac{1}{1+\exp[\sum_{i=1}^{n}\ln\frac{P(x_i|Y=0)}{P(x_i|Y=1)}+\ln\frac{p(Y=0)}{p(Y=1)}]} \quad \text{(Eq. I)}$$

GNB Parametric Form of P(Y|X)

- The sigma term in the denominator:

$$\sum_{i=1}^{n} \ln \frac{p(x_i|Y=0)}{p(x_i|Y=1)} = \sum_i \ln \frac{\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-(X_i-\mu_{i1})^2}{2\sigma_i^2}\right)}$$

$$= \sum_i \ln\left(\exp\left(\frac{(X_i-\mu_{i1})^2-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right)\right) = \sum_i \frac{\left(X_i^2-2X_i\mu_{i1}+\mu_{i1}^2\right)-\left(X_i^2-2X_i\mu_{i0}+\mu_{i0}^2\right)}{2\sigma_i^2}$$

$$= \sum_i \frac{2X_i(\mu_{i0}-\mu_{i1})+\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}$$

GNB Parametric Form of P(Y|X)

- The sigma term in the denominator:

$$\sum_{i=1}^{n} \ln \frac{p(x_i|Y=0)}{p(x_i|Y=1)} = \sum_i \ln \frac{\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(\frac{-(X_i-\mu_{i0})^2}{2\sigma_i^2})}{\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(\frac{-(X_i-\mu_{i1})^2}{2\sigma_i^2})}$$

$$= \sum_i \ln\left(\exp\left(\frac{(X_i-\mu_{i1})^2-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right)\right) = \sum_i \frac{(X_i^2-2X_i\mu_{i1}+\mu_{i1}^2)-(X_i^2-2X_i\mu_{i0}+\mu_{i0}^2)}{2\sigma_i^2}$$

$$= \sum_i \frac{2X_i(\mu_{i0}-\mu_{i1})+\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2} = \sum_i \left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2} X_i + \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right) \quad \text{(Eq. II)}$$

GNB Parametric Form of P(Y|X)

$$\mathrm{I,II} \Rightarrow p(\mathrm{Y}=1|\mathbf{X}) = \frac{1}{1+\exp[\sum_{i=1}^{n}\left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}X_i + \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right) + \ln\frac{p(Y=0)}{p(Y=1)}]}$$

GNB Parametric Form of P(Y|X)

$$\mathrm{I,II} \Rightarrow p(\mathrm{Y}=1|\mathbf{X}) = \frac{1}{1+\exp[\sum_{i=1}^{n}\left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}\mathrm{X_i}+\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right)+\ln\frac{p(Y=0)}{p(Y=1)}]}$$

Or equivalently:

$$p(\mathrm{Y}=1|\mathbf{X}) = \frac{1}{1+\exp(\sum_{i=1}^{n} w_i\mathrm{X_i}+w_0)}$$

where $w_i = \frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}$, $w_0 = ln\frac{1-P(Y=1)}{P(Y=1)}+\sum_{i=1}^{n}\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}$

GNB Parametric Form of P(Y|X)

$$\mathrm{I,II} \Rightarrow p(\mathrm{Y}=1|\mathbf{X}) = \frac{1}{1+\exp[\sum_{i=1}^{n}\left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}\mathrm{X_i}+\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right)+\ln\frac{p(Y=0)}{p(Y=1)}]}$$

Or equivalently:

$$p(\mathrm{Y}=1|\mathbf{X}) = \frac{1}{1+\exp(\sum_{i=1}^{n} w_i\mathrm{X_i}+w_0)}$$

where $w_i = \frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}$, $w_0 = ln\frac{1-p(Y=1)}{p(Y=1)} + \sum_{i=1}^{n}\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}$

$$p(\mathrm{Y}=0|\mathbf{X}) = 1-p(\mathrm{Y}=1|\mathbf{X}) = \frac{\exp(\sum_{i=1}^{n} w_i\mathrm{X_i}+w_0)}{1+\exp(\sum_{i=1}^{n} w_i\mathrm{X_i}+w_0)}$$

GNB Parametric Form of P(Y|X)

If $p(Y = 1|\boldsymbol{X}) \geq p(Y = 0\ |\boldsymbol{X})$ Then Class is +

If $\frac{p(Y=1\ |\boldsymbol{X})}{p(Y = 0\ |\boldsymbol{X})} = \frac{1}{\exp(\sum_{i=1}^{n} w_i \mathrm{X_i} + w_0)} \geq 1$ Then Class is +

GNB Parametric Form of P(Y|X)

If $p(Y = 1|\boldsymbol{X}) \geq p(Y = 0\,|\boldsymbol{X})$ Then Class is +

If $\frac{p(Y=1\,|\boldsymbol{X})}{p(Y = 0\,|\boldsymbol{X})} = \frac{1}{\exp(\sum_{i=1}^{n} w_i \mathrm{X_i} + w_0)} \geq 1$ Then Class is +

If $\exp(\sum_{i=1}^{n} w_i \mathrm{X_i} + w_0) \leq 1$ Then Class is +

If $\boldsymbol{\sum_{i=1}^{n} w_i \mathrm{X_i} + w_0 \leq 0}$ Then Class is +; Else Class is –

GNB Parametric Form of P(Y|X)

If $p(Y = 1|\boldsymbol{X}) \geq p(Y = 0\,|\boldsymbol{X})$ Then Class is +

If $\frac{p(Y=1\,|\boldsymbol{X})}{p(Y = 0\,|\boldsymbol{X})} = \frac{1}{\exp(\sum_{i=1}^{n} w_i \mathrm{X_i} + w_0)} \geq 1$ Then Class is +

If $\exp(\sum_{i=1}^{n} w_i \mathrm{X_i} + w_0) \leq 1$ Then Class is +

If $\sum_{i=1}^{n} \boldsymbol{w_i} \mathbf{X_i} + \boldsymbol{w_0} \leq \mathbf{0}$ Then Class is +; Else Class is –

Therefore, naïve Bayes learning algorithm can be viewed as a linear classifier, under the assumption that variance of each feature is independent of the class (i.e if $\sigma_{ik} = \sigma_i$).

References

- Victor Lavrenko, Lecture on Gaussian Naïve Bayes, University of Edinburgh

- Tom Mitchell, Generative and Discriminative Classifiers: Naïve Bayes and Logistic Regression, Machine Learning (2nd ed.), Chapter 3, McGraw Hill, 2015