

Michael Silverstein

Thanks Wikipedia <3

Aim 2: General Disease Predictor – Multiclass Bayesian Classifier

Step 1: Parameter training

Observed bacterial abundance values for all features (bacterial phylums) for a given sample, i , with class label, k_j , is denoted by the vector, $\mathbf{x}_{i k_j}$. All n sample vectors are concatenated into a **training set**, $X =$

$$\begin{bmatrix} \mathbf{x}_{1,k_j} \\ \vdots \\ \mathbf{x}_{n,k_j} \end{bmatrix} = \begin{bmatrix} x_{1,1,k_j} & \cdots & x_{1,b,k_j} \\ \vdots & \vdots & \vdots \\ x_{n,1,k_j} & \cdots & x_{n,b,k_j} \end{bmatrix}, \text{ for } b, \text{ features.}$$

The **joint probability** for a given sample, \mathbf{x} , with class, k_j , is,

$$P(\mathbf{x}|k_j) = P(x_1, \dots, x_i|k_j) \text{ [Eq. 1]}$$

A **mean vector**, $\boldsymbol{\mu}_{b,k_j}$, can be constructed to describe the expected value of feature vector, \mathbf{B}_{k_j} , for feature b , for class, k_j , and for all samples within a given class, m ,

$$\boldsymbol{\mu}_{k_j} = E[\mathbf{B}_{k_j}] = \left[E[B_{1,k_j}], \dots, E[B_{m,k_j}] \right] \text{ [Eq. 2]}$$

Where the **expected value** of a vector, \mathbf{B}_{k_j} , for, m , observations is the arithmetic mean,

$$E[\mathbf{B}_{k_j}] = \overline{\mathbf{B}_{k_j}} = \frac{1}{m} \sum_{i=1}^m B_i \text{ [Eq. 3]}$$

The maximum likelihood estimator of the **covariance matrix**, $\boldsymbol{\Sigma}_{k_j}$, for a vector, \mathbf{B}_{k_j} , with m observations is,

$$\boldsymbol{\Sigma}_{k_j} = \frac{1}{m} \sum_{i=1}^m (B_i - \overline{\mathbf{B}_{k_j}})(x_i - \overline{\mathbf{B}_{k_j}})^T \text{ [Eq. 4]}$$

Step 2: Label assignment

After model parameters have been calculated for each class, labels can be assigned to a **test set** with t

samples, $Y = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_t \end{bmatrix} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,b} \\ \vdots & \vdots & \vdots \\ y_{t,1} & \cdots & y_{t,b} \end{bmatrix}.$

Assuming that each feature is normally distributed, the joint probability from **Eq. 1** can be described by a **multivariate normal probability density function** (PDF) for a vector of observables, \mathbf{y} , for each class, k_j ,

$$f_{k_j}(\mathbf{y}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_{k_j})^T \boldsymbol{\Sigma}_{k_j}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{k_j})\right)}{\sqrt{2\pi^i |\boldsymbol{\Sigma}_{k_j}|}} \text{ or } \mathbf{y} \sim \mathcal{M}(\boldsymbol{\mu}_{k_j}, \boldsymbol{\Sigma}_{k_j}) \text{ [Eq. 5]}$$

A label, C_i , is then stochastically assigned to sample, \mathbf{y} , after the PDF of \mathbf{y} has been calculated from each class. C_i is drawn from the normalized distribution of each PDF, f_{k_j} ,

$$C_i \sim \frac{[f_{k_1}(\mathbf{y}), \dots, f_{k_j}(\mathbf{y})]}{\sum_{\text{All classes}} f_{k_j}(\mathbf{y})} \quad [\text{Eq. 6}]$$

Naïve Bayes Assumption

Notice that the multivariate PDF requires taking the inverse of the covariance matrix, therefore the covariance matrix must be non-singular in order to calculate the probability $f_B(\mathbf{B})$. This requires that no columns of the covariance matrix be linearly dependent, the probability of which increases as the amount of “missing data” (zero-inflated data) increases. Intuitively, this makes sense: the covariance matrix must span *all* b dimensions of the feature space in order for the variance between each dimension to be represented.

The Naïve Bayes assumption addresses this by assuming class conditional independence, which alters **Eq. 1** to,

$$P(\mathbf{x}|k_j) = P(x_1|k_j)P(x_2|k_j) \dots P(x_b|k_j) = \prod_{i=1}^b P(B_i|k_j) \quad [\text{Eq. 7}]$$

With the Naïve Bayes assumption, given a class features are considered to be independent of one another. Instead of using multivariate distribution to describe the PDF over each class, each class, k_j , is described by the product of **normal univariate distributions**, for each feature b , altering **Eq. 5** to,

$$f_{k_j}(\mathbf{y}) = \prod_{i=1}^b \frac{1}{\sqrt{2\pi\sigma_{i,k_j}^2}} \exp\left(-\frac{(y_i - \mu_{i,k_j})^2}{2\sigma_{i,k_j}^2}\right) \text{ or } \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{k_j}, \boldsymbol{\sigma}_{k_j}) \quad [\text{Eq. 8}]$$

Where, μ_{i,k_j} is the **mean** (an element from the mean vector $\boldsymbol{\mu}_{k_j}$ calculated in **[Eq. 2]**) and σ_{i,k_j} is the **standard deviation** of feature, i , for class, k_j , and is contained within a set of all standard deviations for class k_j in the vector, $\boldsymbol{\sigma}_{k_j}$,

$$\boldsymbol{\sigma}_{k_j} = \bigcup_{i=1}^b \sqrt{\frac{1}{m} \sum_{l=1}^m (B_l - \mu_{i,k_j})^2} \quad [\text{Eq. 9}]$$

DISCUSS WHY NAÏVE AND NON-NAÏVE IN RESULTS/DISCUSSION

A summary of the above equations can be seen in Table 1 and a graphical representation of the workflow can be seen in Figure 1.

Method	Expected value of class	Variance of class	PDF
Non-Naïve Bayes	$\boldsymbol{\mu}_{k_j}$ [Eq. 3]	$\boldsymbol{\Sigma}_{k_j}$ [Eq. 4]	$\mathcal{M}(\boldsymbol{\mu}_{k_j}, \boldsymbol{\Sigma}_{k_j})$ [Eq. 5]
Naïve Bayes	$\boldsymbol{\mu}_{k_j}$ [Eq. 3]	$\boldsymbol{\sigma}_{k_j}$ [Eq. 9]	$\mathcal{N}(\boldsymbol{\mu}_{k_j}, \boldsymbol{\sigma}_{k_j})$ [Eq. 8]

Table 1 Description of statistical building blocks of the Naive and Non-Naive Bayesian classifiers

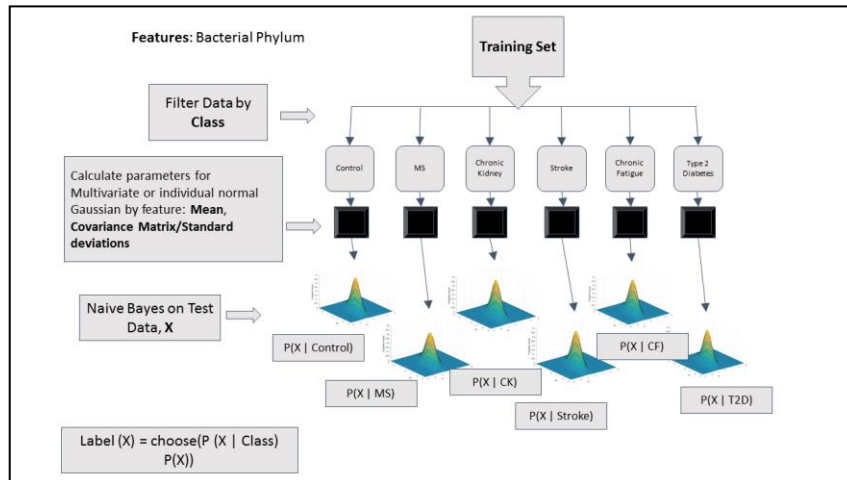


Figure 1 Bayesian Classification (Naive and Non-Naive) Workflow