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, $\boldsymbol{x_{i\,kj.}}$ All $\boldsymbol{n}$ sample vectors are concatenated into a **training set**, $X = \begin{bmatrix} \boldsymbol{x_{1,k_j}} \\ \vdots \\ \boldsymbol{x_{n,k_j}} \end{bmatrix} = \begin{bmatrix} x_{1,1,k_j} & \cdots & x_{1,b,k_j} \\ \vdots & \vdots & \vdots \\ x_{n,1,k_j} & \cdots & x_{n,b,k_j} \end{bmatrix}$, for *b*, features.

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

$$P(\boldsymbol{x}|k_j) = P(x_1, \cdots, x_i|k_j) \textbf{ [Eq. 1]}$$

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

$$\boldsymbol{\mu_{k_j}} = \boldsymbol{E}\left[\boldsymbol{B_{k_j}}\right] = \left[E\left[B_{1,k_j}\right], \ldots, E\left[B_{m,k_j}\right]\right] \textbf{ [Eq. 2]}$$

Where the **expected value** of a vector, $\boldsymbol{B_{kj}}$, for, *m*, observations is the arithmetic mean,

$$E\left[\boldsymbol{B_{k_j}}\right] = \overline{\boldsymbol{B_{k_j}}} = \frac{1}{m}\sum_{i=1}^{m} B_i \textbf{ [Eq. 3]}$$

The maximum likelihood estimator of the **covariance matrix**, $\boldsymbol{\Sigma_{k_j}}$, for a vector, $\boldsymbol{B_{k_j}}$,with *m* observations is,

$$\boldsymbol{\Sigma_{k_j}} = \frac{1}{m}\sum_{i=1}^{m}\left(B_i - \overline{\boldsymbol{B_{k_j}}}\right)\left(x_i - \overline{\boldsymbol{B_{k_j}}}\right)^T \textbf{ [Eq. 4]}$$

Step 2: Label assignment

After model parameters have been calculated for each class, labels can be assigned to a **test set** with *t* samples, $Y = \begin{bmatrix} \boldsymbol{y_1} \\ \vdots \\ \boldsymbol{y_t} \end{bmatrix} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,b} \\ \vdots & \vdots & \vdots \\ y_{t,1} & \cdots & y_{t,b} \end{bmatrix}$.

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, $\boldsymbol{y}$, for each class, k_j,

$$f_{k_j}(\boldsymbol{y}) = \frac{\exp\left(-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{\mu_{k_j}}\right)^T \boldsymbol{\Sigma_{k_j}}^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu_{k_j}}\right)\right)}{\sqrt{2\pi^i\left|\boldsymbol{\Sigma_{k_j}}\right|}} \textit{ or } \boldsymbol{y} \sim \mathcal{M}(\boldsymbol{\mu_{k_j}}, \boldsymbol{\Sigma_{k_j}}) \textbf{ [Eq. 5]}$$

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

$$C_i \sim \frac{\left[f_{\boldsymbol{k_1}}(\boldsymbol{y}), \ldots, f_{\boldsymbol{k_j}}(\boldsymbol{y})\right]}{\sum_{All\ classes} f_{\boldsymbol{k_j}}(\boldsymbol{y})} \ [\boldsymbol{Eq.6}]$$

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 $f_B(\boldsymbol{B})$. 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,

$$P(\boldsymbol{x}|k_j) = P(x_1|k_j)P(x_2|k_j) \ldots P(x_b|k_j) = \prod_{i=1}^{b} P(x_i|k_j) \ [\boldsymbol{Eq.7}]$$

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,

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

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, $\boldsymbol{\sigma_{k_j}}$,

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

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

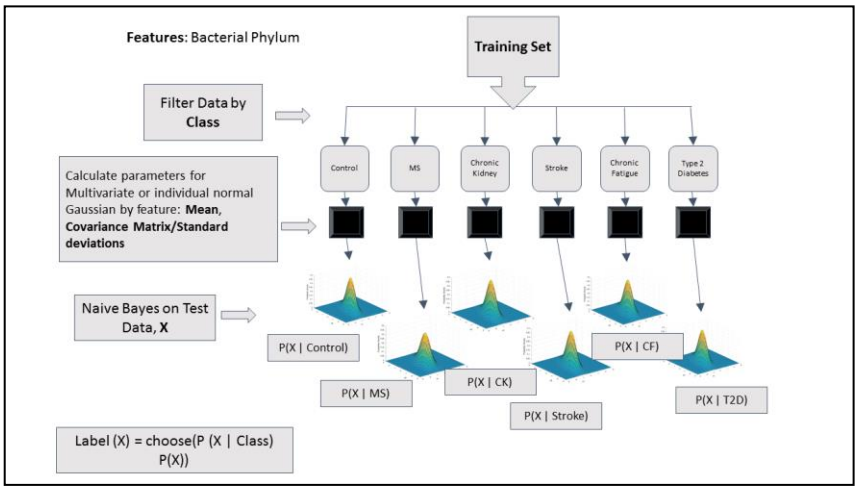

Figure 1 Bayesian Classification (Naive and Non-Naive) Workflow