# Naive Bayes Classifier

Ricco RAKOTOMALALA

# Maximum a posteriori rule

## Calculating the posterior probability

Bayes theorem

$$P(Y = y_k / \aleph) = \frac{P(Y = y_k) \times P(\aleph / Y = y_k)}{P(\aleph)}$$

$$= \frac{P(Y = y_k) \times P(\aleph / Y = y_k)}{\sum_{l=1}^{K} P(Y = y_l) \times P(\aleph / Y = y_l)}$$

## MAP - Maximum a posteriori rule

$$y_{k^*} = \arg\max_k P(Y = y_k / \aleph)$$

$$\Leftrightarrow$$

$$y_{k^*} = \arg\max_k P(Y = y_k) \times P(\aleph / Y = y_k)$$

Prior probability of class k : $P(Y = y_k)$
Estimated by empirical frequency $n_k/n$

How to estimate $P(X/Y=y_k)$ ?

Assumptions are introduced in order to obtain a convenient calculation of this likelihood

# Naive Bayes for Discrete Predictors (Categorical predictors)

# Conditional independence assumption

Conditional independence for the calculation of the likelihood

$$P(\aleph / Y = y_k) = \prod_{j=1}^{J} P(X_j / Y = y_k)$$

The attributes are all conditionally independent of one another given the value of Y !

For a categorical attribute X, the conditional probability for the value $x_l$ is computed as follows…

$$P(X = x_l / Y = y_k) = \frac{P(X = x_l \wedge Y = y_k)}{P(Y = y_k)}$$

The probability is estimated using the conditional relative frequency

$$\hat{P}\left(X = x_l / Y = y_k\right) = \frac{\#\left\{\omega \in \Omega, X(\omega) = x_l \wedge Y(\omega) = y_k\right\}}{\#\left\{\omega \in \Omega, Y(\omega) = y_k\right\}} = \frac{n_{kl}}{n_k}$$

| $Y \setminus X$ | $x_l$ | $\Sigma$ |
|---|---|---|
| | | |
| $y_k$ | $n_{kl}$ | $n_k$ |
| | | |
| $\Sigma$ | | $n$ |

The Laplace rule of succession is often used to estimate the conditional probability

$$\hat{P}\left(X = x_l / Y = y_k\right) = p_{l/k} = \frac{n_{kl} + 1}{n_k + K}$$

This is a kind of smoothing; it enables also to overcome the ($n_{kl}$ = 0) problem.

# An example using a toy dataset

| Maladie | Marié | Etud.Sup |
|---|---|---|
| Présent | Non | Oui |
| Présent | Non | Oui |
| Absent | Non | Non |
| Absent | Oui | Oui |
| Présent | Non | Oui |
| Absent | Non | Non |
| Absent | Oui | Non |
| Présent | Non | Oui |
| Absent | Oui | Non |
| Présent | Oui | Non |

## Direct estimation of the posterior probability

$$\hat{P}(Maladie = Absent\,/\,Marié = oui, Etu = oui) = \frac{1}{1} = 1$$

$$\hat{P}(Maladie = \text{Présent}\,/\,Marié = oui, Etu = oui) = \frac{0}{1} = 0$$

→ If Etu = oui and Marié = oui Then Maladie = Absent !

(+) No assumptions, (-) small number of covered examples

| NB Maladie | |
|---|---|
| **Maladie** | **Total** |
| **Absent** | **5** |
| Présent | 5 |
| Total général | 10 |

| NB Maladie | Marié | | |
|---|---|---|---|
| Maladie | Non | **Oui** | Total général |
| **Absent** | 2 | **3** | 5 |
| Présent | 4 | 1 | 5 |
| Total général | 6 | 4 | 10 |

| NB Maladie | Etud.Sup | | |
|---|---|---|---|
| Maladie | Non | **Oui** | Total général |
| **Absent** | 4 | **1** | 5 |
| Présent | 1 | 4 | 5 |
| Total général | 5 | 5 | 10 |

## Conditional independence assumption

$$\hat{P}(Maladie = Absent\,/\,Marié = oui, Etu = oui)$$
$$= \hat{P}(Maladie = Absent) \times \hat{P}(Marié = oui\,/\,M = Abs.) \times \hat{P}(Etu = oui\,/\,M = Abs.)$$
$$= \frac{5+1}{10+2} \times \frac{3+1}{5+2} \times \frac{1+1}{5+2} = 0.082$$

$$\hat{P}(Maladie = présent\,/\,Marié = oui, Etu = oui)$$
$$= \hat{P}(Maladie = présent) \times \hat{P}(Marié = oui\,/\,M = Abs.) \times \hat{P}(Etu = oui\,/\,M = Abs.)$$
$$= \frac{5+1}{10+2} \times \frac{1+1}{5+2} \times \frac{4+1}{5+2} = 0.102$$

→ If Etu = oui and Marié = oui Then Maladie = Présent !

(-) Questionable assumption, (+) more reliable estimation of probabilities

# Advantage and shortcoming (end of the course?)

**+**

>> Simplicity, quickness, ability to handle very large dataset, no possible crash during the calculations

>> Incrementality (we store only the contingency tables)

>> Statistically robust (even if the assumption is very questionable)

>> This is a linear classifier → similar classification performance
(see the numerous experiments described in scientific papers)

**−**

>> No indication about the relevance of the attributes (really ?)

>> Very high number of rules
(in practice, the logical rules are not computed, the contingency tables for the calculation of the conditional frequency are deployed e.g. PMML format)

>> Not explicit model (really ?) → not used in marketing domain, etc.

We see often these conclusions in the literature…
Is it possible to go beyond that?

# Extracting an explicit model from a Naive Bayes classifier

Logarithmic transformation

$$y_{k^*} = \arg\max_k P(Y = y_k) \times \prod_{j=1}^{J} P(X_j / Y = y_k)$$

$$\Leftrightarrow y_{k^*} = \arg\max_k \left[ \ln P(Y = y_k) + \sum_{j=1}^{J} \ln P(X_j / Y = y_k) \right]$$

# Model using one predictive attribute

A discrete attribute X with L levels

$$d(y_k, X) = \ln P(Y = y_k) + \ln P(X / Y = y_k)$$

From X, we can create L dummy variables

$$\begin{aligned} d(y_k, X) &= \ln P(Y = y_k) + \sum_{l=1}^{L} \ln P(X = x_l / Y = y_k) \times I_l \\ &= \ln P(Y = y_k) + \sum_{l=1}^{L} \ln P(X = x_l / Y = y_k) \times I_l \\ &= a_{0,k} + \sum_{l=1}^{L} a_{l,k} \times I_l \end{aligned}$$

We obtain a linear combination of the dummy variables i.e. an explicit model which is easy to deploy

→ K linear classification functions (such as linear discriminant analysis)

# An example (Y : Maladie; X : Etu.Sup)

| NB Maladie | |
|---|---|
| Maladie | Total |
| **Absent** | **5** |
| Présent | 5 |
| Total général | 10 |

| NB Maladie | Etud.Sup | | |
|---|---|---|---|
| Maladie | Non | **Oui** | Total général |
| **Absent** | 4 | **1** | 5 |
| Présent | 1 | 4 | 5 |
| Total général | 5 | 5 | 10 |

$$d(absent, X) = \ln \frac{5+1}{10+2} + \ln \frac{4+1}{5+2} \times (X = non) + \ln \frac{1+1}{5+2} \times (X = oui)$$
$$= -0.6931 - 0.3365 \times (X = non) - 1.2528 \times (X = oui)$$

$$d(present, X) = -0.6931 - 1.2528 \times (X = non) - 0.3365 \times (X = oui)$$

For an instance (Etu.Sup = NON) →

$$\begin{cases} d(absent, X) = -0.6931 - 0.3365 = -1.0296 \\ d(present, X) = -0.9631 - 1.2528 = -1.9495 \end{cases}$$

Prediction : Maladie = non

## Implemented solution into TANAGRA
(Using [L-1] dummy variables for an attribute X with L levels)

### Prior distribution of class attribute "Maladie"

| Values | Count | Percent | Histogram |
|---|---|---|---|
| Absent | 5 | 50.00 % | |
| Présent | 5 | 50.00 % | |

### Model description

| | Classification functions | |
|---|---|---|
| Descriptors | Absent | Présent |
| Etud.Sup = Oui | -0.916291 | 0.916291 |
| constant | -1.029619 | -1.945910 |

$$d(y_k, X) = \ln P(Y = y_k) + \sum_{l=1}^{L} \ln P(X = x_l / Y = y_k) \times I_l$$

$$= \ln P(Y = y_k) + \ln P(X = x_L / Y = y_k) + \sum_{l=1}^{L-1} \ln \frac{P(X = x_l / Y = y_k)}{P(X = x_L / Y = y_k)} \times I_l$$

$$= b_{0,k} + \sum_{l=1}^{L-1} b_{l,k} \times I_l$$

since

$$I_1 + I_2 + \cdots + I_L = 1$$

One level [$x_L$] becomes the reference level

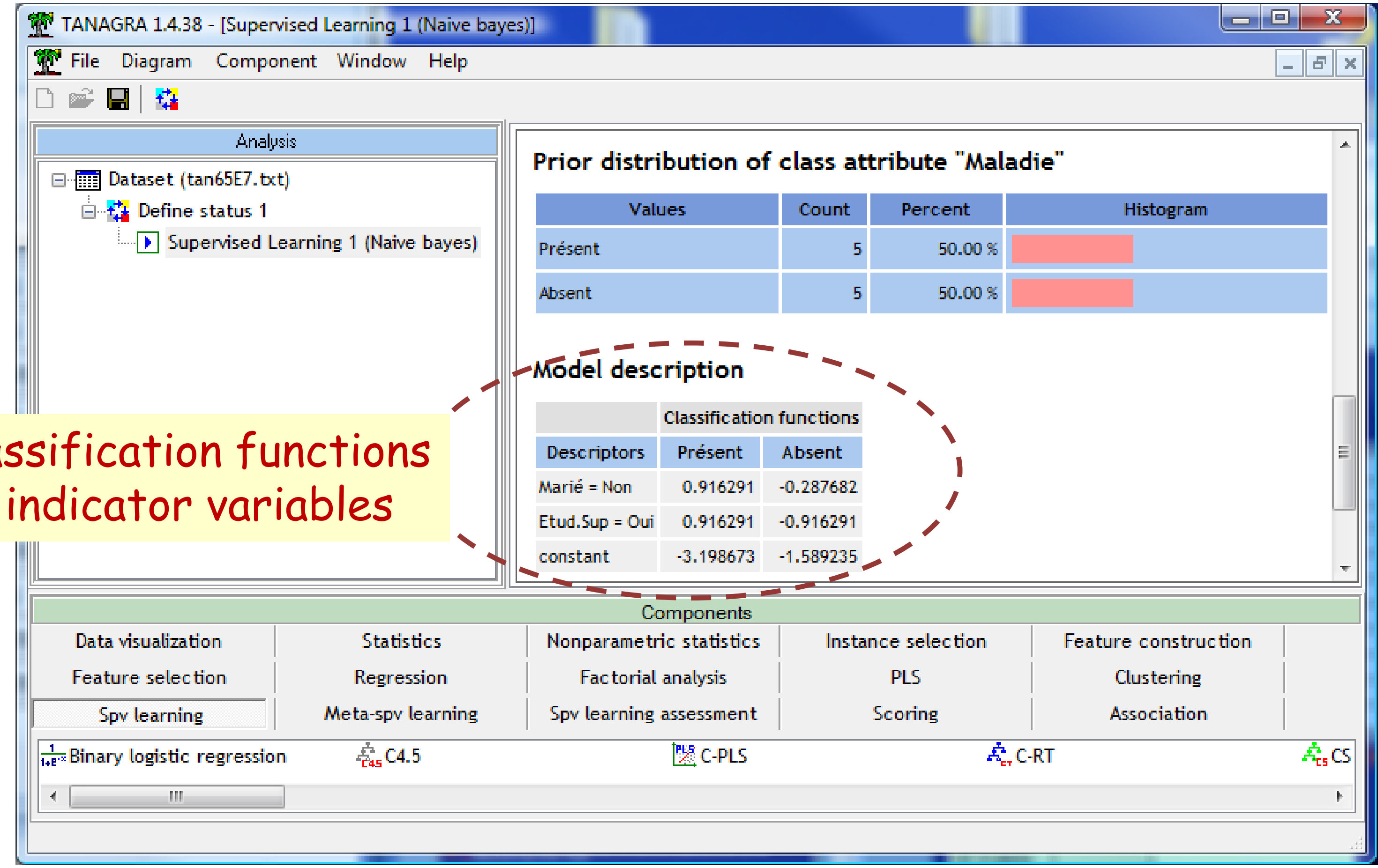
The dummy coding is the most commonly used coding scheme

| Maladie | Marié | Etud.Sup |
|---|---|---|
| Présent | Non | Oui |
| Présent | Non | Oui |
| Absent | Non | Non |
| Absent | Oui | Oui |
| Présent | Non | Oui |
| Absent | Non | Non |
| Absent | Oui | Non |
| Présent | Non | Oui |
| Absent | Oui | Non |
| Présent | Oui | Non |

# Extension to J predictive attributes

Dummy coding scheme

$X_j$ with $L_j$ levels → $(L_j-1)$ dummy variables




Linear classification functions using the indicator variables

## Construction of the SCORE function

The class attribute has 2 levels :: Y={+,-}

$$
-\begin{cases} d(+,X) = a_{+,0} + a_{+,1}X_1 + a_{+,2}X_2 + \cdots + a_{+,J}X_J \\ d(-,X) = a_{-,0} + a_{-,1}X_1 + a_{-,2}X_2 + \cdots + a_{-,J}X_J \end{cases}
$$

$$
d(X) = c + c_1 X_1 + c_2 X_2 + \ldots + c_J X_J
$$

Decision rule

D(X) > 0 → Y = +

Interpretation

>> D(X) is the SCORE function. It assigns a score proportional to positive class probability estimate to the instances

>> The sign of the coefficients allows to interpret the influence of the descriptors

| Notre exemple : | Classification functions | | SCORE |
|---|---|---|---|
| **Descriptors** | **Présent** | **Absent** | **D(X)** |
| Marié = Non | 0.916291 | -0.287682 | 1.203973 |
| Etud.Sup = Oui | 0.916291 | -0.916291 | 1.832582 |
| constant | -3.198673 | -1.589235 | -1.609438 |

Not being married makes sick…

To study makes sick…

# Reading of the coefficients in the classification functions

## Estimation of the conditional probabilities

| Nombre de Maladie | Marié | | |
|---|---|---|---|
| Maladie | Non | Oui | Total général |
| Présent | 0.8 | 0.2 | 1.0 |
| Absent | 0.4 | 0.6 | 1.0 |
| Total général | 0.6 | 0.4 | 1.0 |

## Naives Bayes Classifier (explicit representation)

| | Classification functions | |
|---|---|---|
| **Descriptors** | **Présent** | **Absent** |
| Marié = Non | 1.38629 | -0.4055 |
| constant | -2.3026 | -1.204 |

$$odds(M = N / M = O; Y = present) = \frac{0.8}{0.2} = 4 \Rightarrow \ln(odds) = 1.386294$$

The sick individuals (maladie = présent) have 4 times more chance to be not married than to be married

The coefficient of the classification function corresponds to the logarithm of the odds

$$odds(M = N / M = O; M = absent) = \frac{0.4}{0.6} = 0.667 \Rightarrow \ln(odds) = -0.4055$$

For the non-sick individuals, they have (1/0.667) = 1.5 times more chance to be married than not to be married.

# Reading of the coefficients in the score function (binary problem)

| Nombre de Maladie | Marié | | |
|---|---|---|---|
| Maladie | Non | Oui | Total général |
| Présent | 0.8 | 0.2 | 1.0 |
| Absent | 0.4 | 0.6 | 1.0 |
| Total général | 0.6 | 0.4 | 1.0 |

| | Classification functions | | |
|---|---|---|---|
| Descriptors | Présent | Absent | SCORE |
| Marié = Non | 1.38629 | -0.40547 | 1.79176 |
| constant | -2.30259 | -1.20397 | -1.09861 |

$$odds-ratio(M = N / M = O; Y = P / Y = A) = \frac{odds(M = N / M = O; Y = P)}{odds(M = N / M = O; Y = A)} = \frac{4}{0.66} = 6$$

$$\ln(6) = 1.79176$$

The sick individuals have 6 times more chance to be married than the non-sick individuals.

The coefficient of the score function corresponds to the odds-ratio

**Comments**

- The reading of the odds-ratio is inverted compared with the logistic regression
- This interpretation is relevant if only if the association between X and Y is significant

# Feature selection

Checking the relevance of the variables
Removing the irrelevant variables
Removing the redundancy between the variables

## Amazing consequence of the conditional independence assumption

By nature, the coefficients associated to a variable are estimated independently to the other predictive variables

➔ thus the addition or the removal of one predictive variable does not modify the coefficients related to the other variables.

| | Classification functions | |
|---|---|---|
| Descriptors | Présent | Absent |
| Marié = Non | 0.916291 | -0.287682 |
| constant | -1.94591 | -1.252763 |

Classifier with 1 variable

| | Classification functions | |
|---|---|---|
| Descriptors | Présent | Absent |
| Marié = Non | 0.916291 | -0.287682 |
| Etud.Sup = Oui | 0.916291 | -0.916291 |
| constant | -3.198673 | -1.589235 |

Classifier with 2 variables ("Etu.Sup" is added)

It is not needed to recalculate the other coefficients when we add or we remove a variable.

# Relevance of an attribute (1)

A variable is influent if it enables to increase the differences between the classification functions $d(y_k, X)$ (according to $y_k$)

⇔ If the conditional distributions $P(X/y_k)$ are different according to $y_k$

⇔ If the conditional distributions $P(X/y_k)$ are different to the marginal distribution $P(X)$

| Nombre de Marié | Etud.Sup | | |
|---|---|---|---|
| Maladie | Non | Oui | Total général |
| Absent | 0.8 | 0.2 | 1.0 |
| Présent | 0.2 | 0.8 | 1.0 |
| Total général | 0.5 | 0.5 | 1.0 |

| Nombre de Marié | Marié | | |
|---|---|---|---|
| Maladie | Non | Oui | Total général |
| Absent | 0.4 | 0.6 | 1.0 |
| Présent | 0.8 | 0.2 | 1.0 |
| Total général | 0.6 | 0.4 | 1.0 |

$$H(X) = \sum_{l=1}^{L} p_{.l} \log_2 p_{.l}$$

~ total variance

$$H(X/Y) = \sum_{k=1}^{K} p_{k.} \sum_{l=1}^{L} p_{l/k} \log_2 p_{l/k}$$

~ within variance

$$H(X) - H(X/Y) = I(Y, X)$$

$$= \sum_{l=1}^{L} \sum_{k=1}^{K} p_{kl} \log_2 \frac{p_{kl}}{p_{l.} \times p_{k.}}$$

~ Between Variance
i.e. explained variance

Mutual information

# Relevance of an attribute (2)

We can establish a hierarchy between the predictive variables

$$I(Y,ES)=0.2781$$

| Nombre de Marié | Etud.Sup | | |
|---|---|---|---|
| Maladie | Non | Oui | Total général |
| Absent | 0.4 | 0.1 | 0.5 |
| Présent | 0.1 | 0.4 | 0.5 |
| Total général | 0.5 | 0.5 | 1.0 |

$$I(Y,M)=0.1245$$

| Nombre de Marié | Marié | | |
|---|---|---|---|
| Maladie | Non | Oui | Total général |
| Absent | 0.2 | 0.3 | 0.5 |
| Présent | 0.4 | 0.1 | 0.5 |
| Total général | 0.6 | 0.4 | 1.0 |

We can even determine the statistical significance of the association

Statistical test ($H_0$ : the variables are independent)

$$G=2\times n\times \ln 2\times I(Y,X)$$
$$\sim \chi^2[(K-1)\times(L-1)]$$

$$G(ES)=3.85$$
$$\Rightarrow p.value=0.0496$$

The association between Y and ES is significant

$$G(M)=1.73$$
$$\Rightarrow p.value=0.1889$$

The association between Y and M is not significant

# Ranking using the symmetrical uncertainty measure

Defined between [0 ; 1]

$$s_{Y,X} = 2 \times \frac{I(Y,X)}{H(Y)+H(X)}$$

E.g.« kr-vs-kp » dataset (19 selected pour $\alpha$ = 0.001)

**Calculations details**

| N° | Attribute | Values | Statistic | Statistic (Histogram) | p-value |
|---|---|---|---|---|---|
| 1 | rimmx | 2 | 0.452284 | | 0.000000 |
| 2 | bxqsq | 2 | 0.380101 | | 0.000000 |
| 3 | wknck | 2 | 0.365265 | | 0.000000 |
| 4 | bkxwp | 2 | 0.232908 | | 0.000000 |
| 5 | wkna8 | 2 | 0.196557 | | 0.000000 |
| 6 | r2ar8 | 2 | 0.164526 | | 0.000000 |
| 7 | bkxcr | 2 | 0.163828 | | 0.000000 |
| 8 | mulch | 2 | 0.158337 | | 0.000000 |
| 9 | wkpos | 2 | 0.146543 | | 0.000000 |
| 10 | bkxbq | 2 | 0.139802 | | 0.000000 |
| 11 | skrxp | 2 | 0.130476 | | 0.000000 |
| 12 | stlmt | 2 | 0.127724 | | 0.000000 |
| 13 | wkcti | 2 | 0.126350 | | 0.000000 |
| 14 | rkxwp | 2 | 0.101402 | | 0.000000 |
| 15 | bkon8 | 2 | 0.091163 | | 0.000000 |
| 16 | rxmsq | 2 | 0.087799 | | 0.000001 |
| 17 | bkxwp | 2 | 0.085171 | | 0.000001 |

RANKING:

1. Calculating s for each predictive variable
2. Sort them in a decreasing order
3. Retain only the variables significantly related to Y

Shortcoming

- Choosing the right significance level « alpha » is difficult
- All the associations are significant when the database size « n » increase

→Possible solution : "elbow rule"

Unacceptable shortcoming

This solution does not take into account the redundancy between the variables

# Feature selection which handles the redundancy – CFS approach

The MERIT of a subset of "p" attributes is defined as follows

$$merit = \frac{p \times \bar{s}_{Y,X}}{\sqrt{p + p \times (p+1) \times \bar{s}_{X,X}}}$$

Numerator: association of the predictive attributes with the target variable (relevance)

Denominator : association between the predictive attributes (redundancy)

→ The aim is to obtain a subset of attributes which are strongly related to the target attribute and weakly related to each other.

Results

## INPUT attribute selection

| INPUT selection | |
|---|---|
| Before filtering | 34 |
| After filtering | 3 |

E.g. « kr-vs-kp » - only 3 selected var.

## Keeped into INPUT selection

| | Attributes |
|---|---|
| 1 | bxqsq |
| 2 | rimmx |
| 3 | wknck |

## Calculations details

| Selected attribute | MERIT(S) |
|---|---|
| rimmx | 0.235390 |
| bxqsq | 0.246590 |
| wknck | 0.257278 |

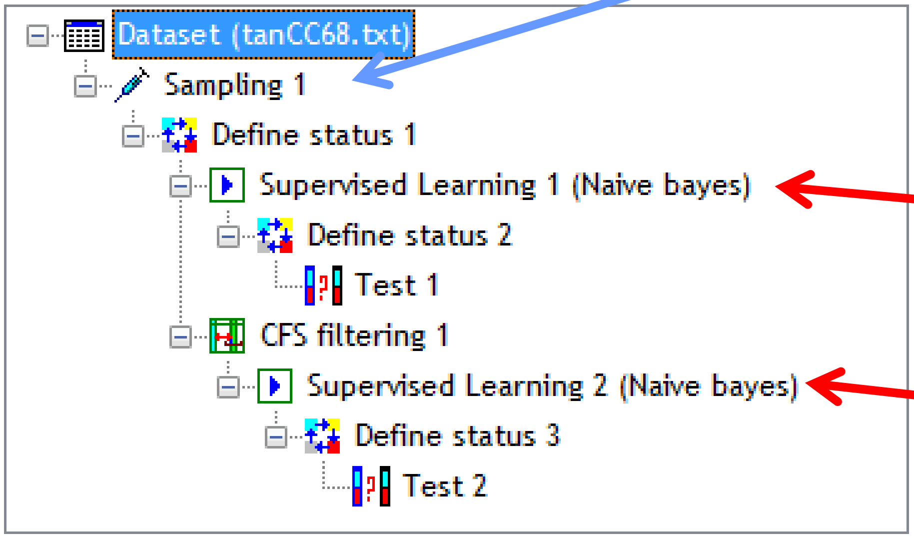
« FORWARD » strategy

- Beginning with 0 variables
- Adding sequentially one variable i.e. c.-à-d. the variable which maximizes the increase of MERIT is added
- Etc.

→Stopping rule: stop when the additional variable does not increase the MERIT

# The selection is justified and appropriate?

1500 training set size
1696 test set size




34 descriptors
Test error rate = 14.80%

3 descriptors
Test error rate = 9.67%

The variable selection enables to reduce the number of variables by maintaining the performance level

Sometimes, it increases the prediction performance (e.g. here, but this is rare)

Sometimes, it is wrong (when we remove too many variables)

# Naive Bayes Classifier for continuous predictors (1)

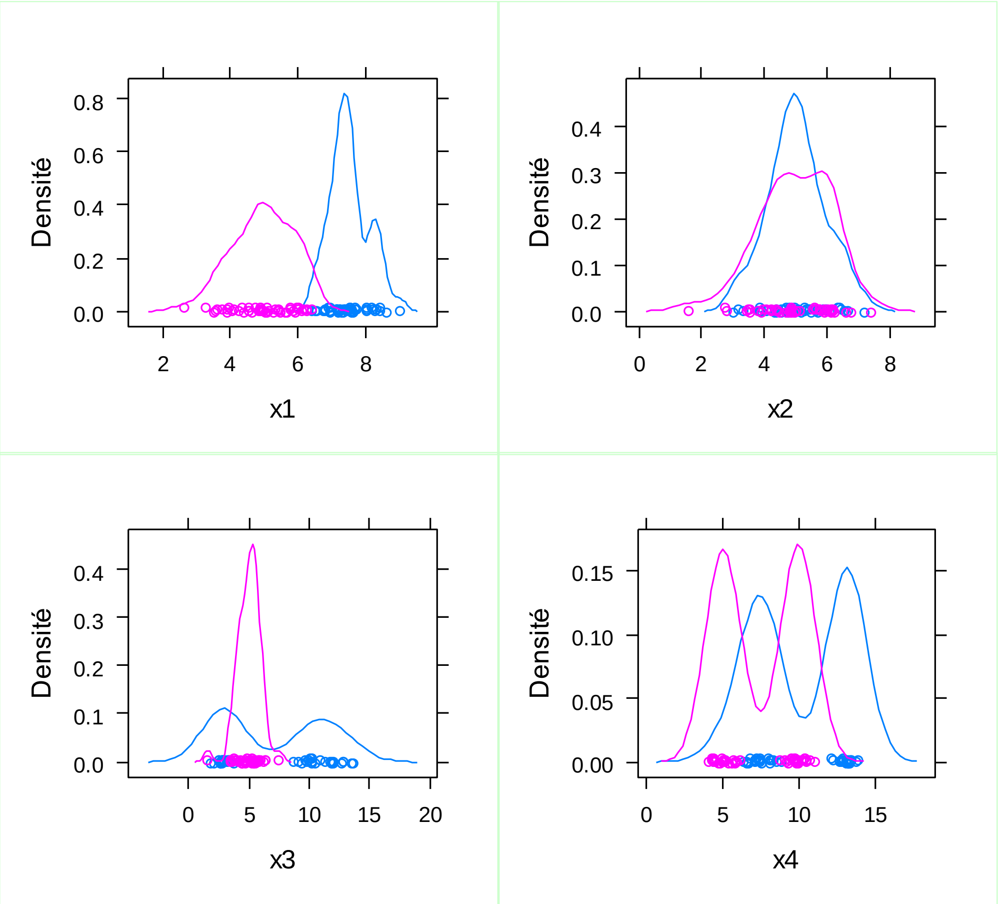
Getting the previous situation (discrete variables) by discretizing the continuous variables e.g. using a supervised approach such as MDLPC (Fayyad & Irani, 1993)

→ Empirical studies shows that this is a good solution
→ This is the best solution when we have a mix of continuous and discrete predictive attributes

# Discretization of continuous attributes
## Using a specific supervised algorithm

The well-known unsupervised approaches (e.g. equal width, equal frequency) do not consider the target attribute. They are not adapted to the supervised learning context.

4 examples of conditional distributions

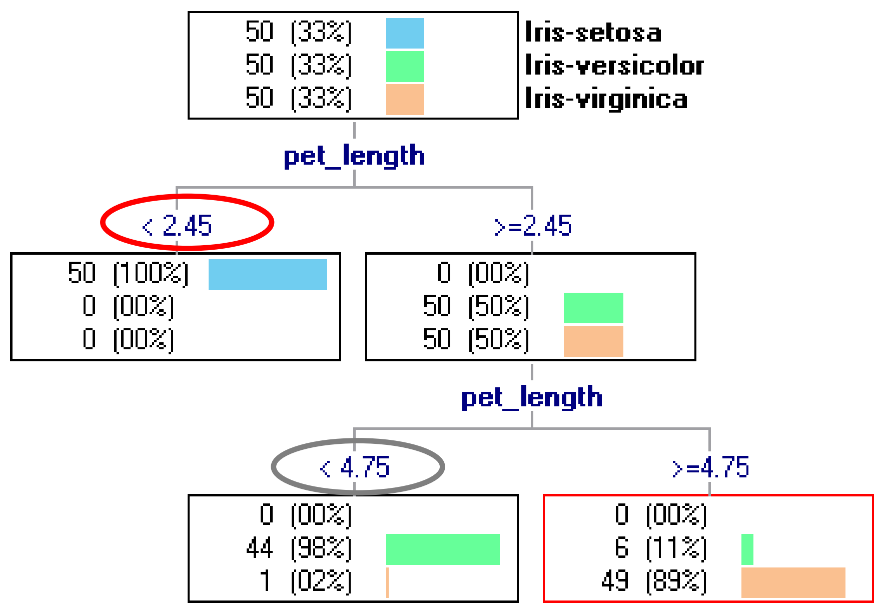



Why supervised algorithms (MDLPC, Fayyad et Irani, 1993 ; Chi-merge, Kerber, 1992) are more convenient?

- Detecting the intervals where one of the classes is overrepresented
- Detecting automatically the right number of intervals

# Discretization of continuous attributes using a decision tree learning algorithm

The variable to discretize is the only one predictive variable used in the decision tree learning. The variable is used - with different cut points - in the various splitting process.

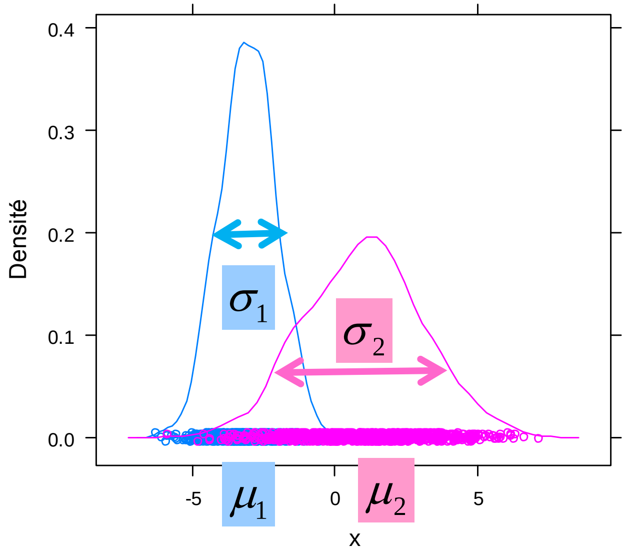
# Naive Bayes Classifier for continuous predictors (2)

Parametric approach
Making assumptions about the conditional distributions

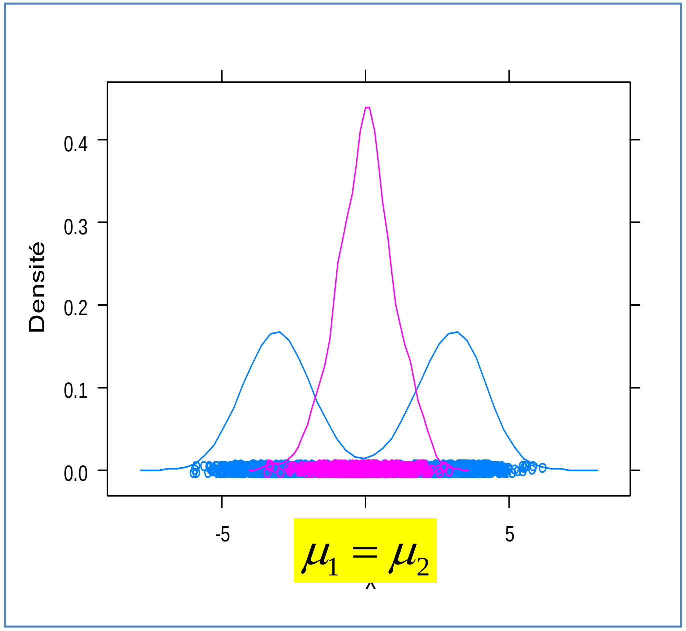
# Assumption.1 – Gaussian conditional distribution

$$P[X_j / Y = y_k] = f_k(X_j) = \frac{1}{\sigma_{k,j}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_j - \mu_{k,j}}{\sigma_{k,j}}\right)^2}$$

Normal distribution for X conditionally to $y_k$




Compatible with the Gaussian assumption




Not compatible with the Gaussian assumption
→ possible solution: discretization

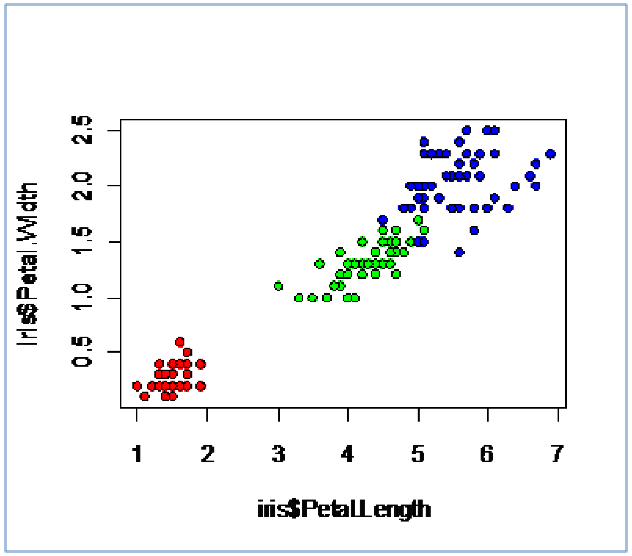
Note: This is a particular case of the discriminant analysis where we consider than the values outside of the main diagonal of the covariance matrix are zero (see Linear Discriminant Analysis).

## Quadratic classifier

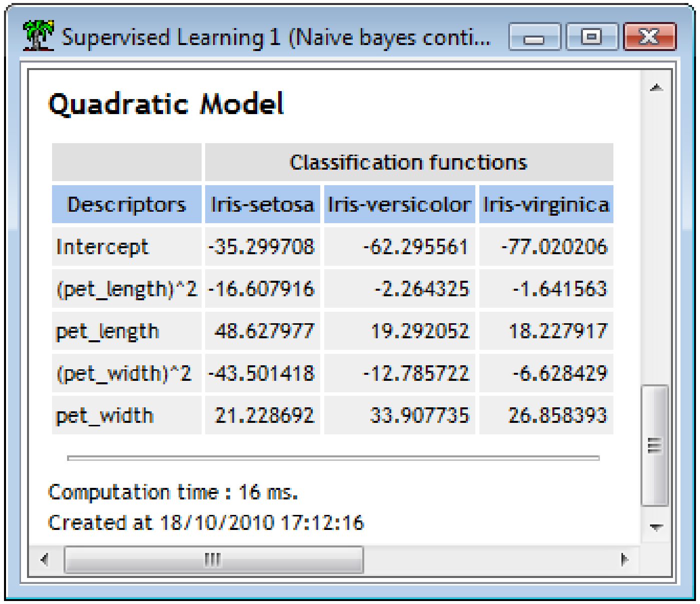
$$d(y_k, \aleph) \propto \ln p_k + \sum_j \left\{ -\frac{1}{2\times\sigma_{k,j}^2} x_j^2 + \frac{\mu_{k,j}}{\sigma_{k,j}^2} x_j - \left( \frac{\mu_{k,j}^2}{2\times\sigma_{k,j}^2} + \ln(\sigma_{k,j}) \right) \right\}$$

$$\propto \ln p_k + \sum_j a_{k,j} x_j^2 + b_{k,j} x_j + c_{k,j}$$

The decision rule is not modified i.e. $\hat{y}(\omega) = y_{k^*} \Leftrightarrow y_{k^*} = \arg\max_k d\left[y_k, \aleph(\omega)\right]$




IRIS dataset (2 predictive variables)




**Quadratic Model**

| | Classification functions | | |
|---|---|---|---|
| Descriptors | Iris-setosa | Iris-versicolor | Iris-virginica |
| Intercept | -35.299708 | -62.295561 | -77.020206 |
| (pet_length)^2 | -16.607916 | -2.264325 | -1.641563 |
| pet_length | 48.627977 | 19.292052 | 18.227917 |
| (pet_width)^2 | -43.501418 | -12.785722 | -6.628429 |
| pet_width | 21.228692 | 33.907735 | 26.858393 |

Computation time : 16 ms.
Created at 18/10/2010 17:12:16

The interpretation is not easy.

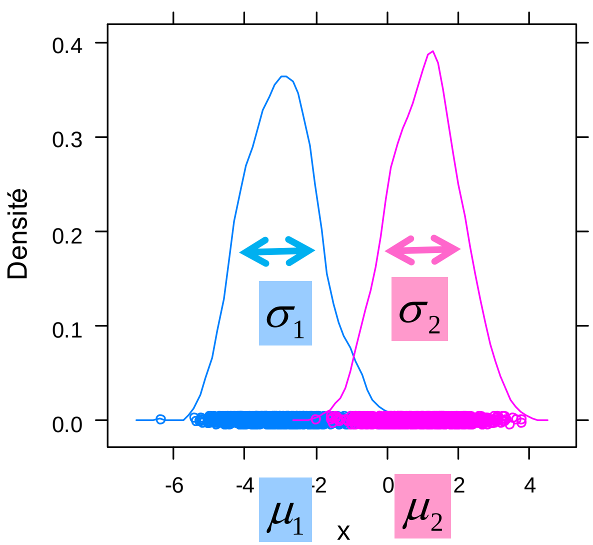
# Assumption.2 - Homoscedasticity

The conditional variances are the same over the classes

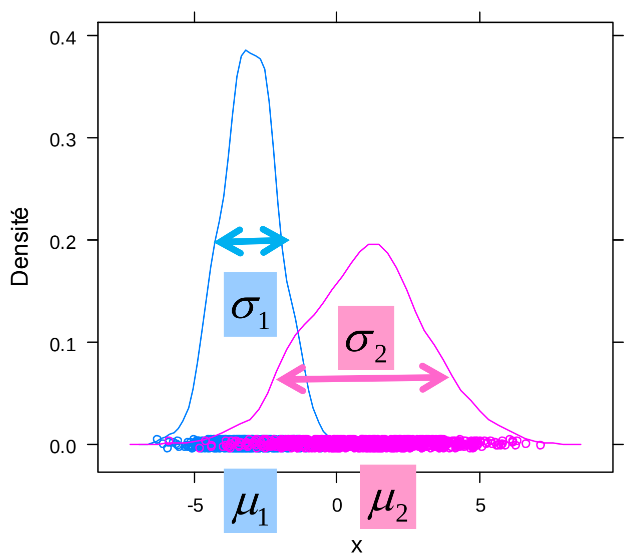
$$\sigma_{k,j} = \sigma_j \ , \ \forall k$$

The common variance is estimated with the within variance

$$P\left[X_j / Y = y_k\right] = f_k(X_j) = \frac{1}{\sigma_j\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_j - \mu_{k,j}}{\sigma_j}\right)^2}$$




Compatible with the homoscedasticity assumption




Not compatible with the assumption

→ But the approach is robust

## Linear classifier

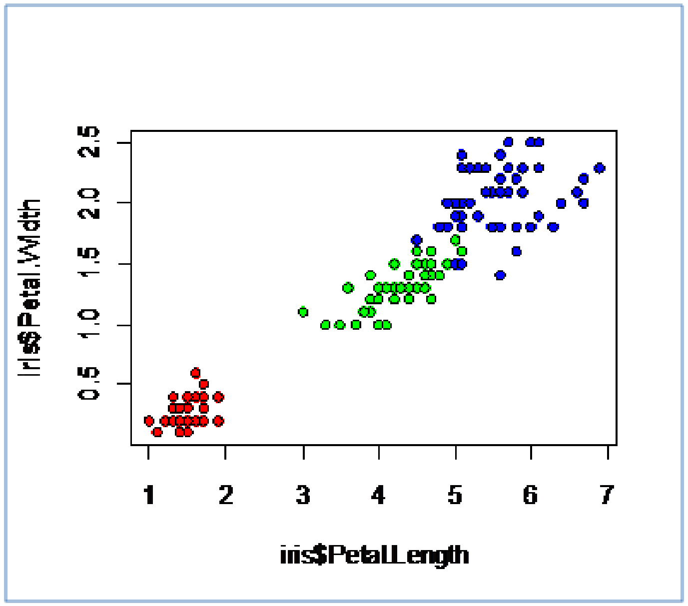
$$d(y_k, \aleph) \propto \ln p_k + \sum_j \left\{ \frac{\mu_{k,j}}{\sigma_j^2} x_j - \frac{\mu_{k,j}^2}{2 \times \sigma_j^2} \right\}$$

$$\propto a_{k,0} + a_{k,1} x_1 + a_{k,2} x_2 + \cdots + a_{k,J} x_J$$

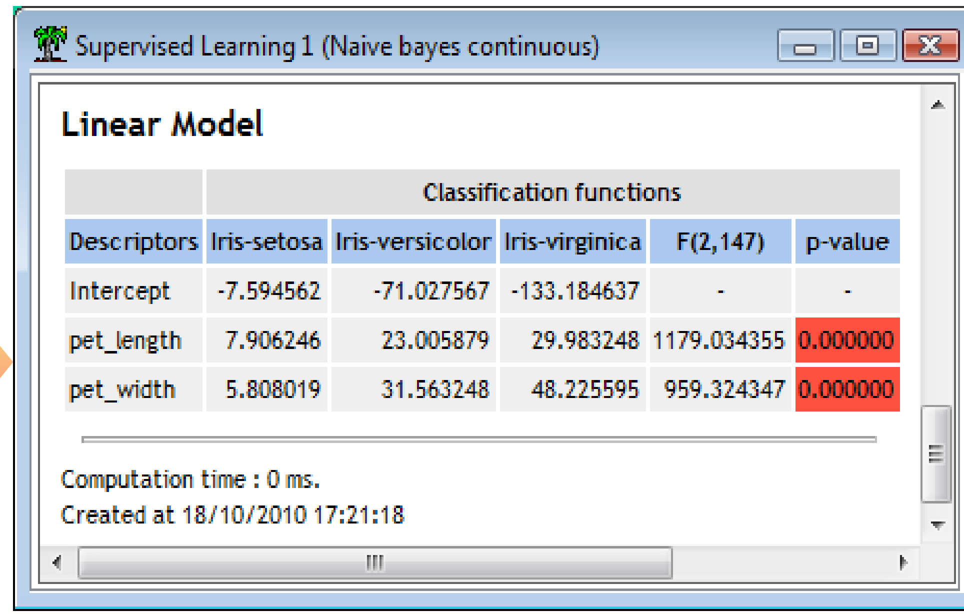
The decision rule is not modified i.e.

$$\hat{y}(\omega) = y_{k^*} \Leftrightarrow y_{k^*} = \arg\max_k d[y_k, \aleph(\omega)]$$

If K=2 (binary problem), we can calculate the SCORE function -- D(X)




Fichier IRIS (2 variables)




### Linear Model

| | Classification functions | | | | |
|---|---|---|---|---|---|
| Descriptors | Iris-setosa | Iris-versicolor | Iris-virginica | F(2,147) | p-value |
| Intercept | -7.594562 | -71.027567 | -133.184637 | - | - |
| pet_length | 7.906246 | 23.005879 | 29.983248 | 1179.034355 | 0.000000 |
| pet_width | 5.808019 | 31.563248 | 48.225595 | 959.324347 | 0.000000 |

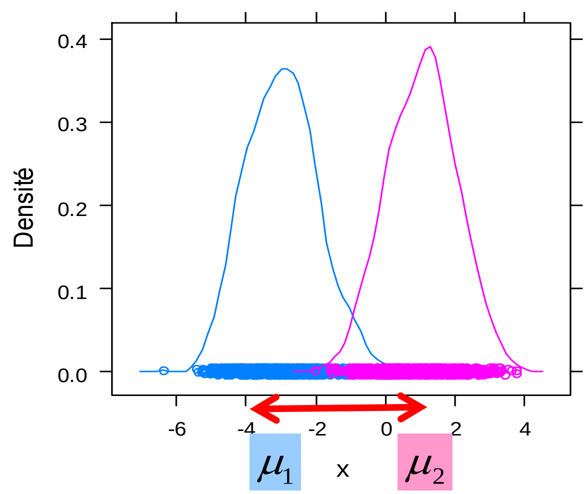
Computation time : 0 ms.
Created at 18/10/2010 17:21:18

**The interpretation is easier**
PET.LENGTH low -> Setosa
PET.LENGTH middle -> Versicolor
PET.LENGTH high -> Virginica

# Variables importance

Evaluate the relevance of the variables
Remove the irrelevant variables
Removing the redundancies

# Variable importance – One way ANOVA scheme




Comparison of conditional means

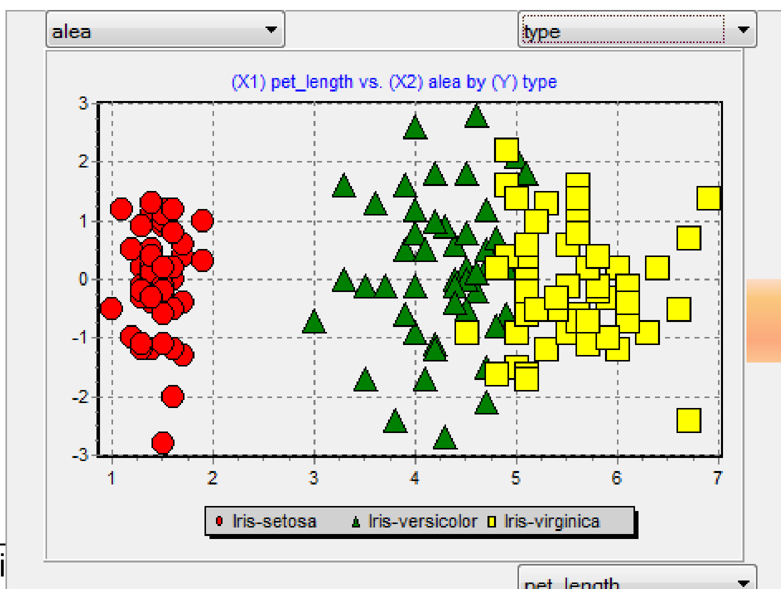
$$H_0 : \mu_{k,j} = \mu, \forall k$$

Test statistic F

$$F = \frac{\dfrac{\sum_k n_k (\hat{\mu}_k - \hat{\mu})^2}{K-1}}{\dfrac{\sum_k (n_k - 1)\hat{\sigma}_k^2}{n-K}}$$

Between Variance
--------------------
Within Variance

Under H0, F ~ Fisher (K-1, n-K) d.f.




**Linear Model**

| | Classification functions | | | | |
|---|---|---|---|---|---|
| Descriptors | Iris-setosa | Iris-versicolor | Iris-virginica | F(2,147) | p-value |
| Intercept | -6.886948 | -50.112224 | -84.338242 | - | - |
| alea | -0.041928 | 0.142192 | -0.105733 | 0.909119 | 0.405132 |
| pet_length | 7.906246 | 23.005879 | 29.983248 | 1179.034355 | 0.000000 |

## RANKING :approach

1. Calculating F for all the variables
2. Sort them according F in a decreasing order
3. Retain only the variables with a significant association

IRIS + 1 ALEA (variable generated randomly)

### Keeped into INPUT selection

| | Attributes |
|---|---|
| 1 | pet_length |
| 2 | pet_width |
| 3 | sep_length |
| 4 | sep_width |

### Calculations details

| N° | Attribute | F | F (max normalized) | p-value (2,147) |
|---|---|---|---|---|
| 1 | pet_length | 1179.03 | | 0.000000 |
| 2 | pet_width | 959.32 | | 0.000000 |
| 3 | sep_length | 119.26 | | 0.000000 |
| 4 | sep_width | 47.36 | | 0.000000 |
| 5 | alea | 0.91 | | 0.405132 |

Same problems than for the discrete attributes

→ Choosing the significance level "alpha"
→ Dealing with redundancy

Extension of the CFS approach to continuous predictors

$$merit = \frac{p \times \bar{s}_{Y,X}}{\sqrt{p + p \times (p+1) \times \bar{s}_{X,X}}}$$

- Measure 1 : Measuring the association between Y (discrete) et X (continuous)
- Measure 2 : Measuring the association between $X_j$ (continuous) et $X_{j'}$ (continuous)

Problem → Measure 1 and Measure 2 must be comparable !

## Other approaches

→ STEPDISC algorithm for linear discriminant analysis (Multivariate analysis of variance – MANOVA)
But the calculations are costly.

→ Using embedded approach of other learning algorithms (e.g. decision tree).
But the relevant variables for a method are not necessarily the same for the naive bayes classifier

→ Discretize the predictive variables and use the selection approaches for discrete attributes

# Conclusion

>> Very often used in the research domain (text mining, etc.)

+

>> Strong advantages (Incrementality, ability to handle very large database)

>> We can extract an explicit model (completely unknown)

−

>> Not used in some domains (e.g. marketing)... because the users do not know that we can extract an explicit model than we can deploy easily

# References


Tanagra – "Naive Bayes Classifier for discrete predictors"

http://data-mining-tutorials.blogspot.fr/2010/07/naive-bayes-classifier-for-discrete.html

Tanagra – "Naive Bayes Classifier for continuous predictors"

http://data-mining-tutorials.blogspot.fr/2010/11/naive-bayes-classifier-for-continuous.html

Wikipedia, « Naive Bayes Classifier »

http://en.wikipedia.org/wiki/Naive_Bayes_classifier

STATSOFT e-books, « Naive Bayes Classifier » (see. other distribution assumptions)

http://www.statsoft.com/textbook/naive-bayes-classifier/