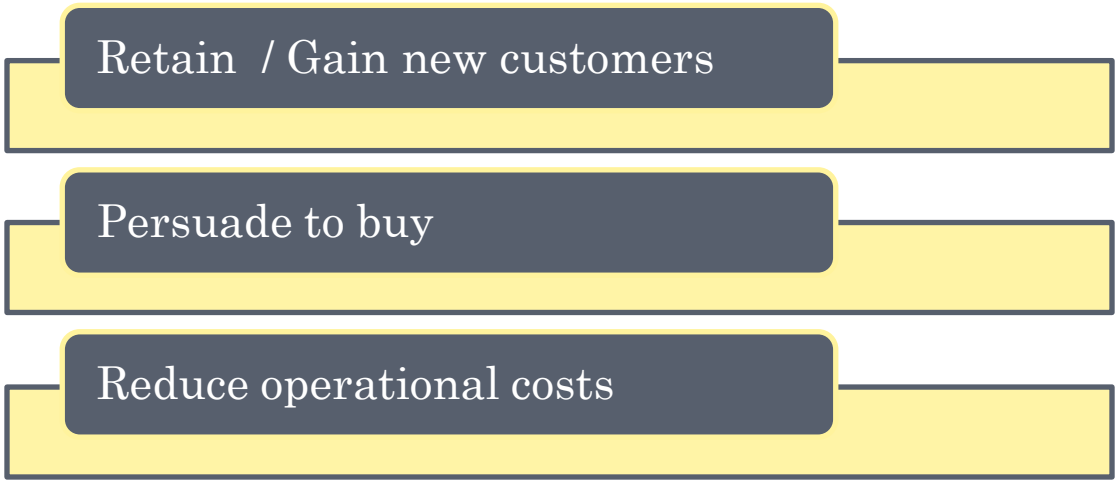
Customer Relationship Management in marketing programs: A machine learning approach for decision

Fernanda Alcantara
F.Alcantara@cs.ucl.ac.uk

CRM Goal

Support the decision taking

Prediction, pattern recognition

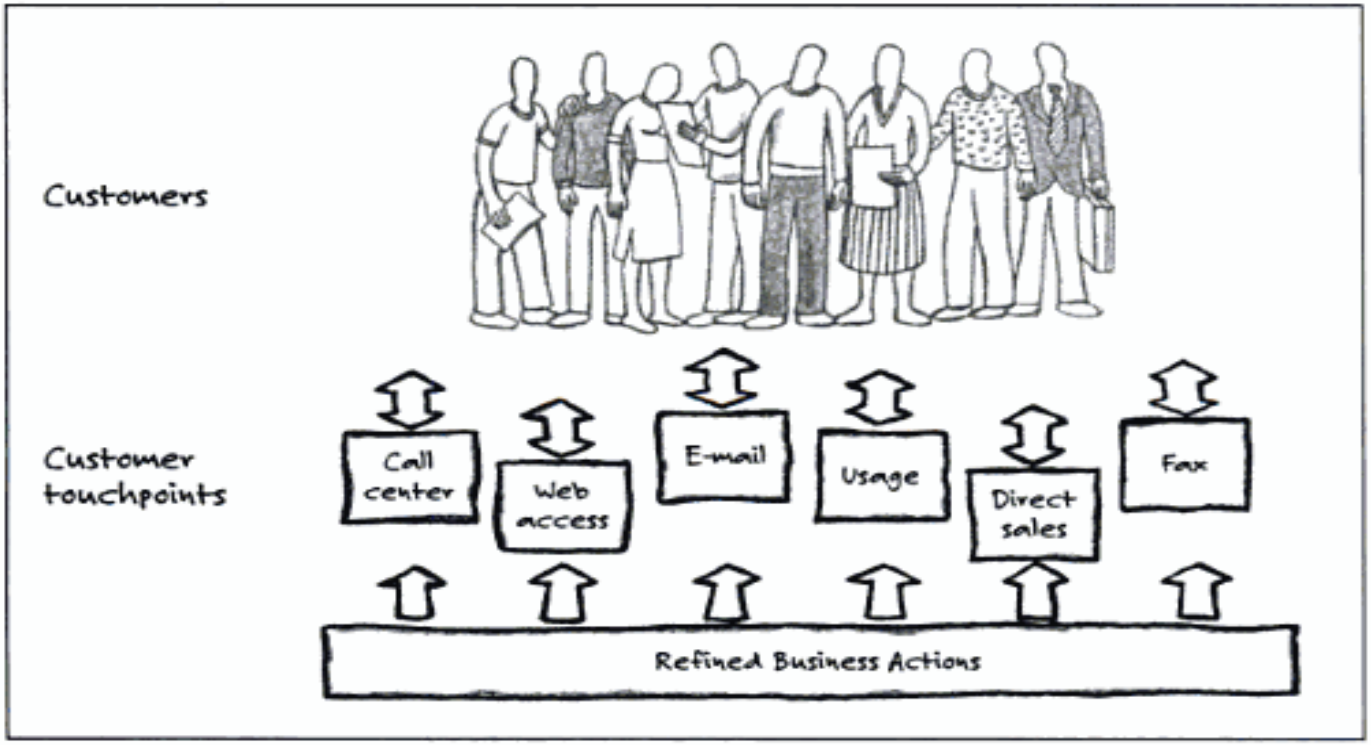


Channel / period of the year / localization / value of the product

CRM: Initial Steps

Rule Engines
- Nomail

Response Tracking
- Response Rate (RR)



(Dyché, 2002)

Example contract

25. Personal Data

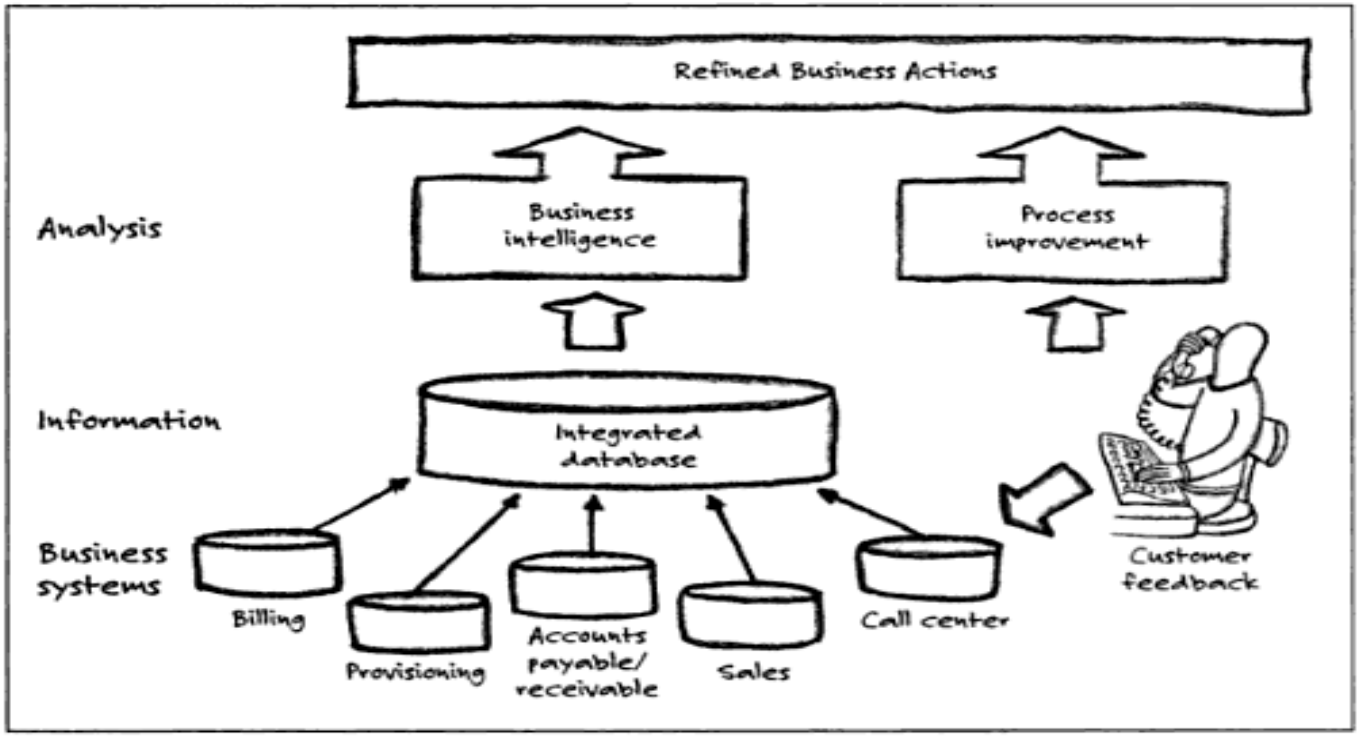
25.1 The personal information given for the on-line sales is obligatory, necessary for recording and processing the delivery of the orders as well as for the billing team. This information is strictly confidential. Giving false information will result in the automatic cancellation of the order.

25.2 You have the right to access, to modify, to change or correct the information given. To exert this right, you can do this online or if you are unable to do this for any technical reasons then you may contact our Customer Service team.

25.3 **This can be used with certain partners**, in particular in order to measure and improve the effectiveness of certain adverts. The information obtained is strictly anonymous and simply makes it possible to gather statistics on the frequency of use of certain pages of the Website.

Data warehouses

Central directory for data from various sources
Data storage
Processed by the BI team to analyse business performance
Db: Oracle/Sybase Financial tool: Business Object (BO)



(Dyché, 2002)

Data Mining

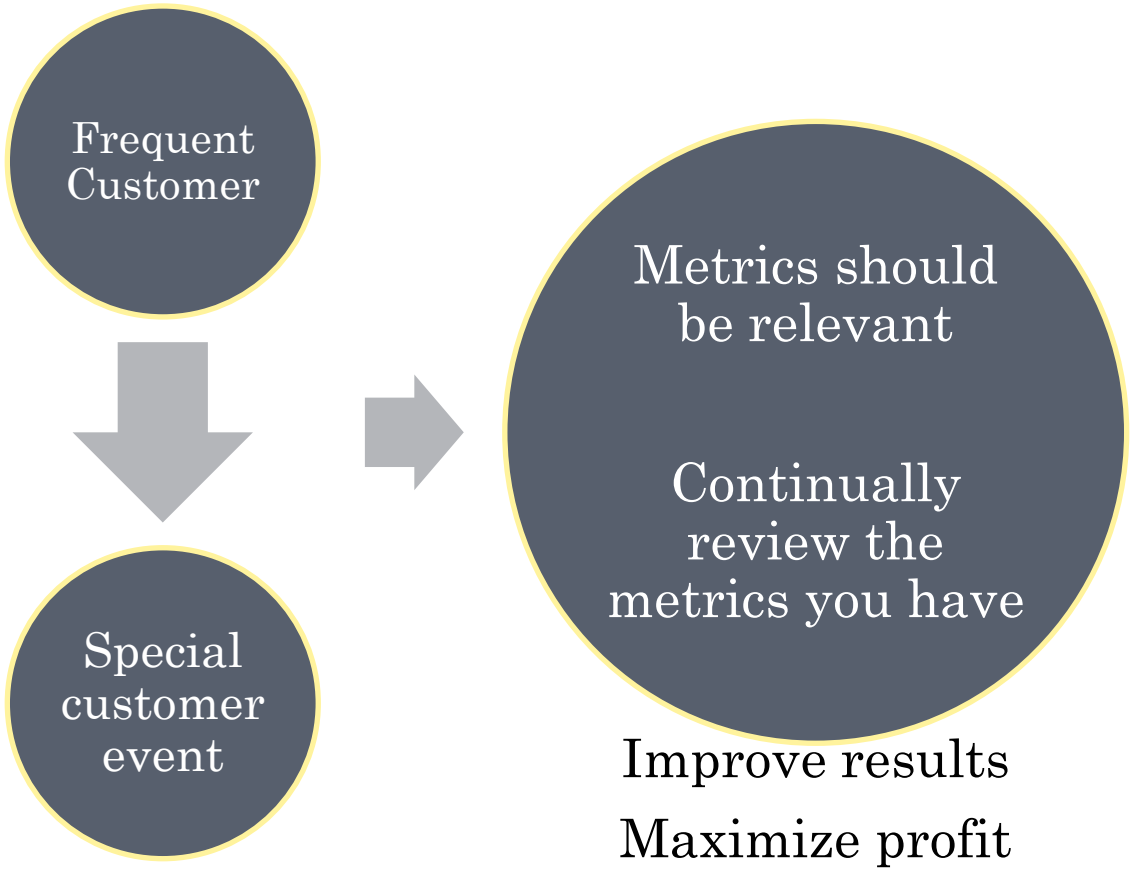
Use information to deliver the right content, in the right time, for increase engagement/ sales.

Mining through customer profile and historical behaviour (offline data)

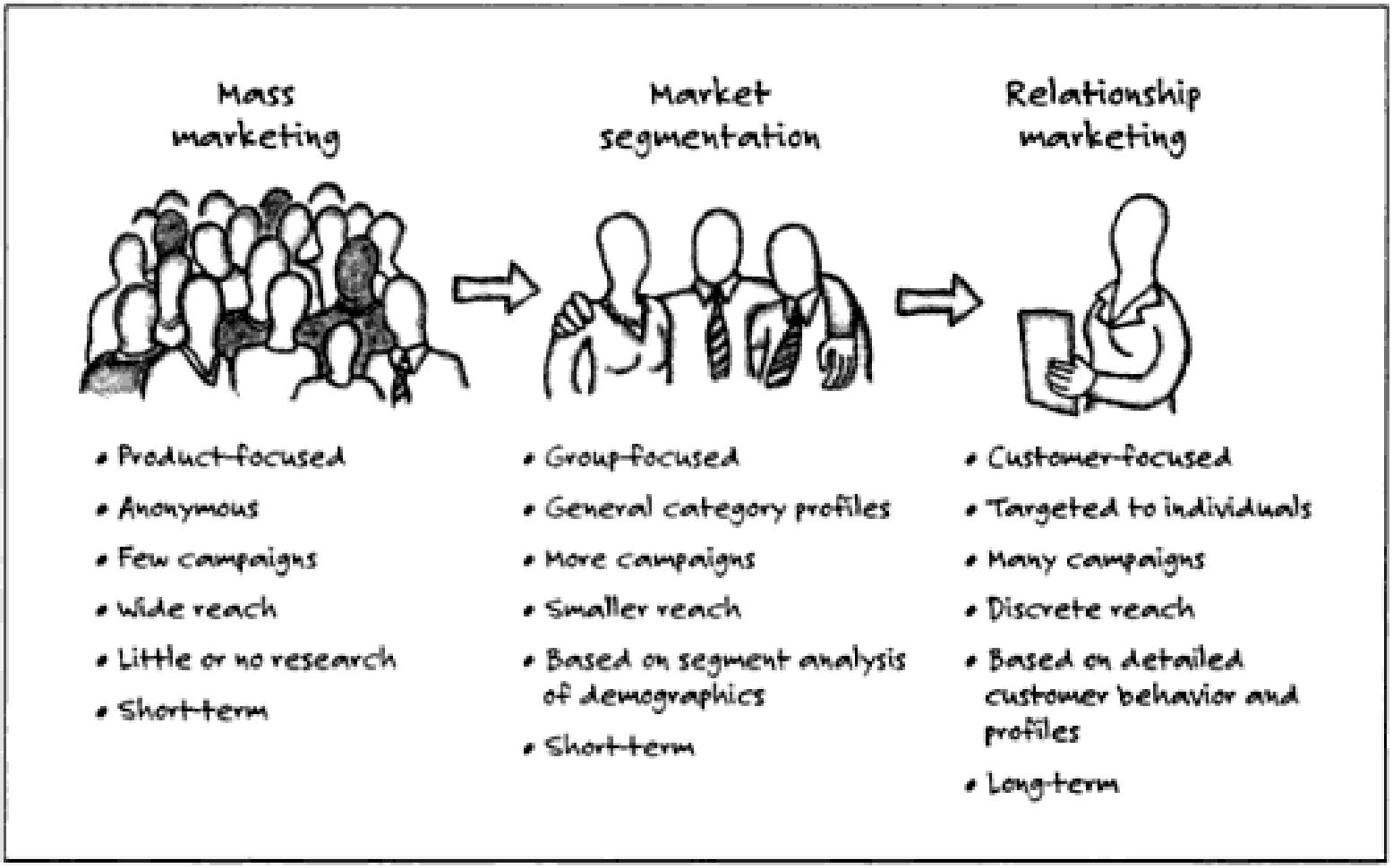
- Classification
- Pattern
- Identify target
- Clustering

Marketing Campaign

Example : "who visit and spend frequently on a site might be sent invites to a "Special Customers Only" event, while frequent visitors who rarely buy may be prompted into purchase by a discount email on products they often look at."

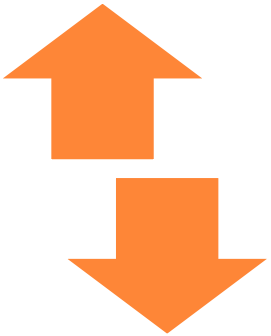


Marketing Segmentation



(Dyché, 2002)

Machine Learning



Reward

Cost

Evaluates solutions for business needs using prediction, recommendation analyses and tracking result control

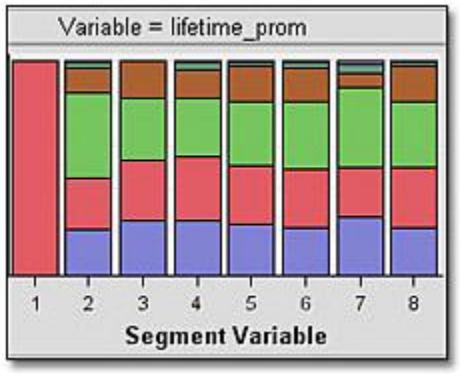
Identify individuals with high response rate and target-profit potential.

Tools: SAS

Data mining with SAS® Enterprise Miner

Market basket analysis
Sequence and Web path analysis
Dimension reduction techniques:
- Variable selection
- LARS (Least Angle Regression)
- Principal components
- Variable clustering
- Time series mining
- Manage time metrics with descriptive data

Linear and logistic regression.
Decision trees
Gradient boosting
Neural Networks
Partial least squares regression
Two-stage modeling
Memory-based reasoning
Model ensembles, including bagging and boosting
Statistical diagnostics and ROI metrics

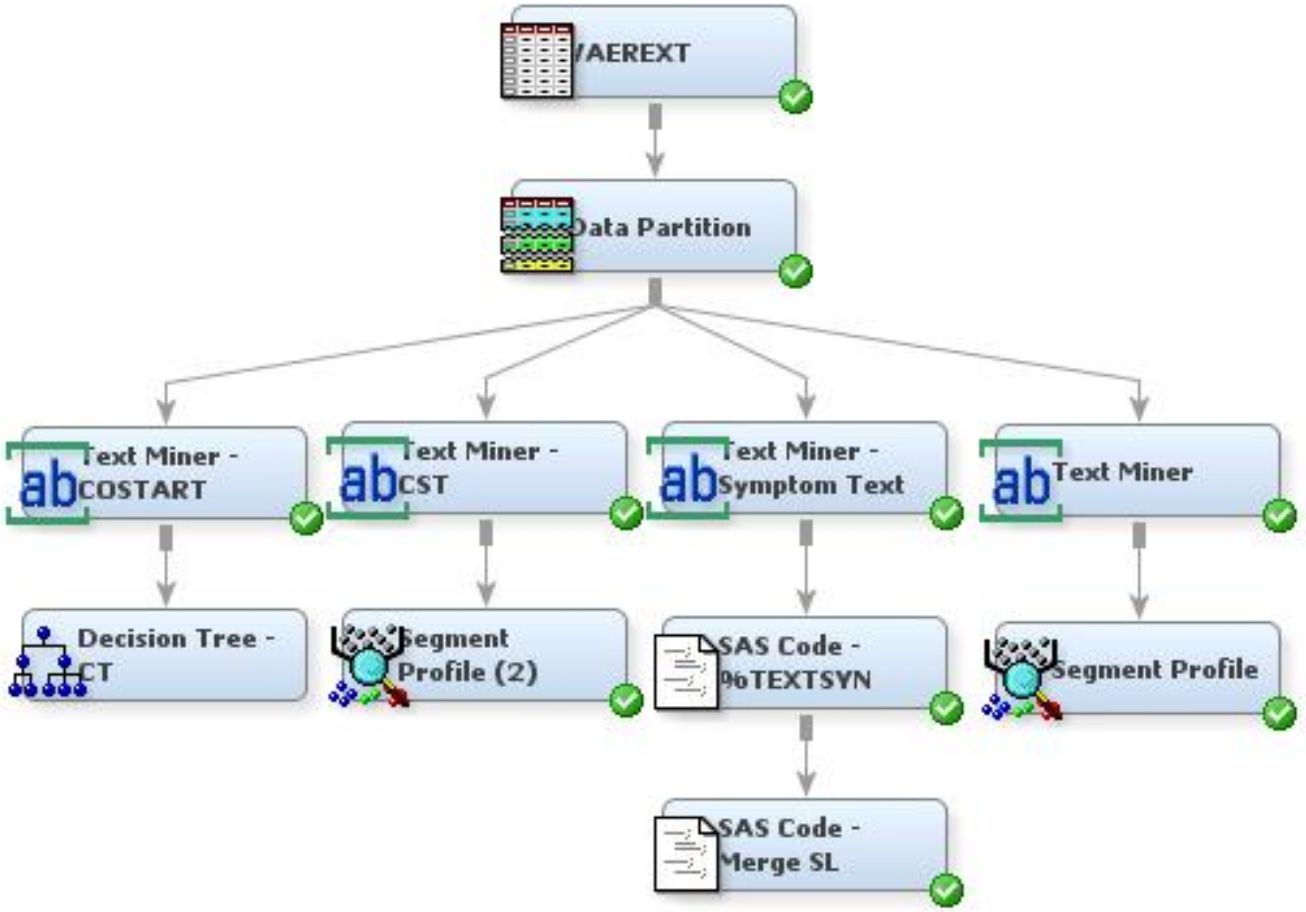


Tools: SAS

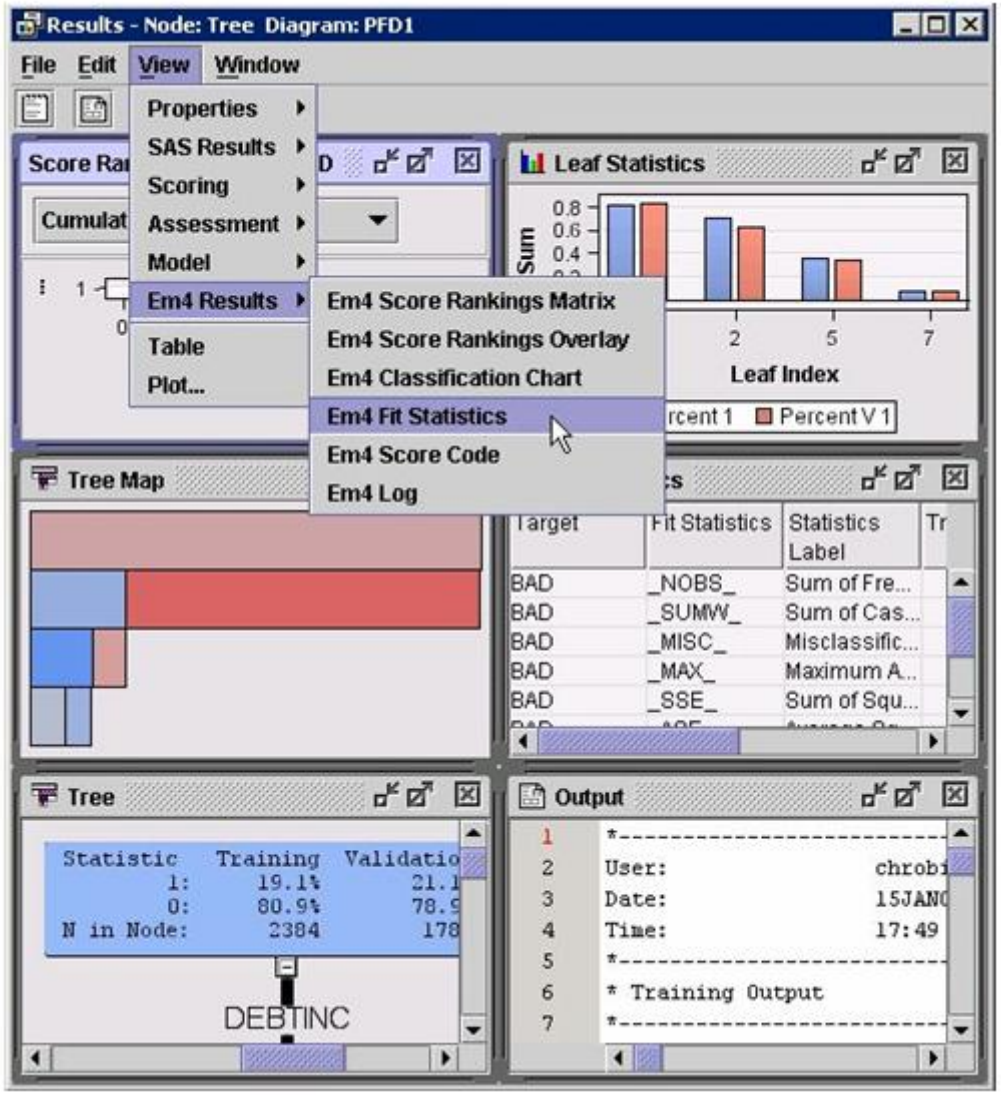
Data mining with SAS® Enterprise Miner

Example of Output: Distribution, decision tree and clustering.

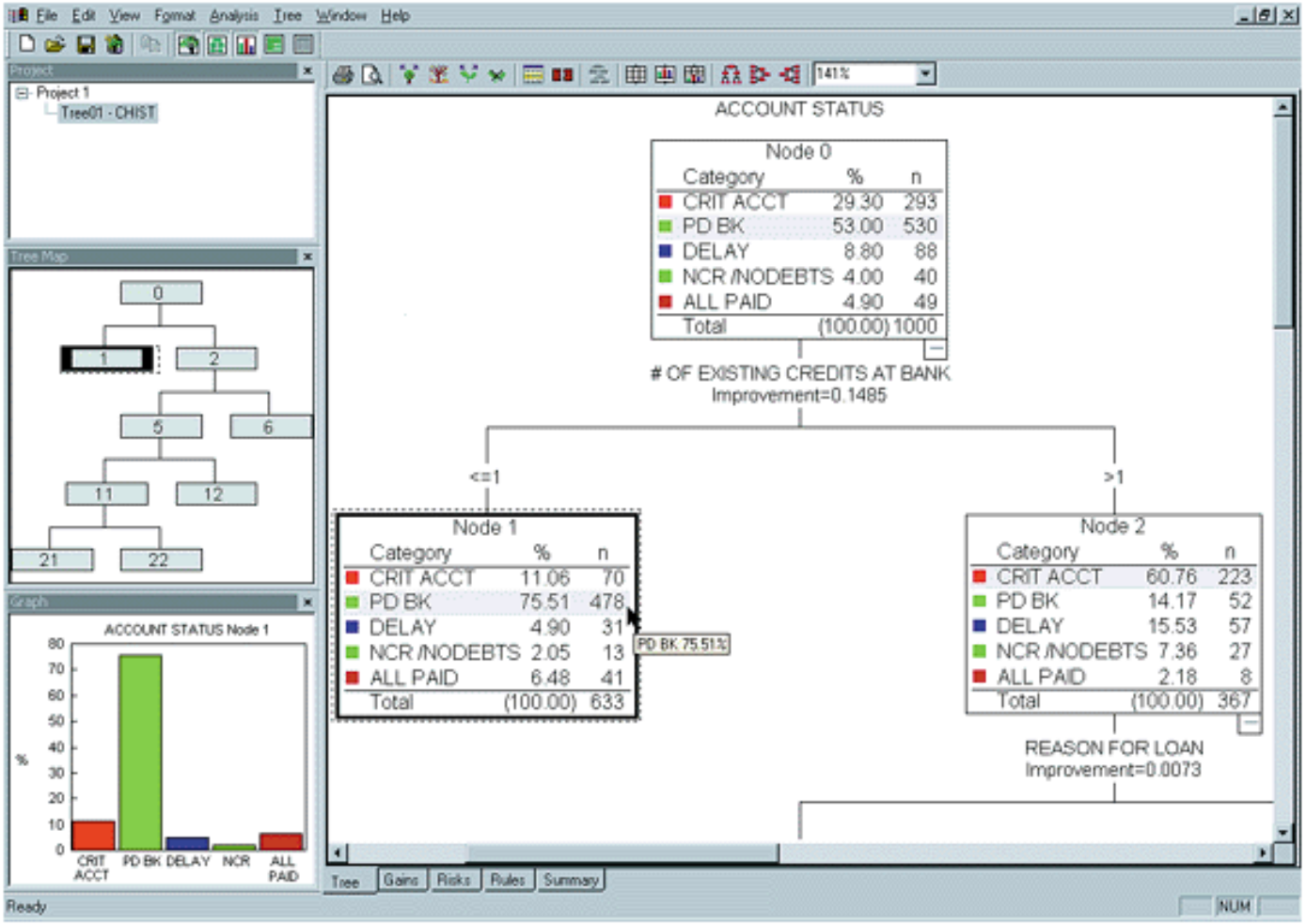
Tools: SAS Miner Structure



Decision Tree: SAS Output



Decision Tree: SPSS Output



Decision Tree

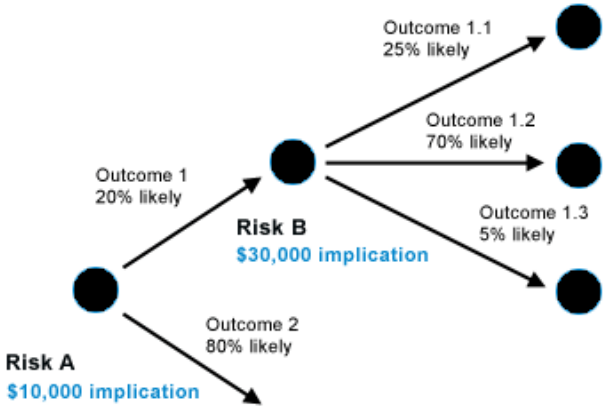
p (portfolio = converge) = 0.4
p (portfolio = not converge) = 0.6

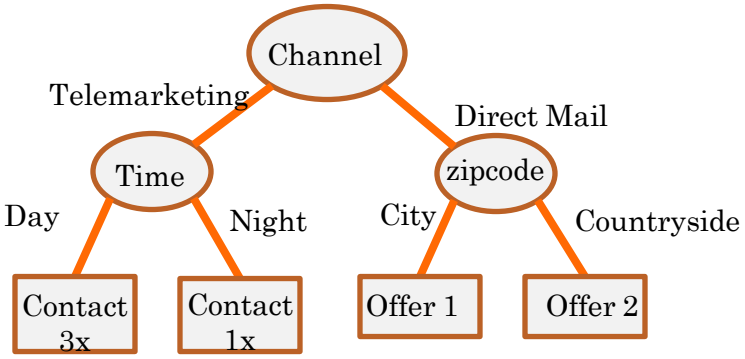
U (TMK, conv) = £ +20
U (TMK, not conv) = £ -20

U (DM, conv) = £ +10
U (DM, not conv) = £ -10

U (TMK) = 20*0.4 + -20*0.6 = -4
U (DM) = 10*0.4 + -10*0.6 = -2

Go for Direct mail





p (portfolio = converge) = 0.6
p (portfolio = not converge) = 0.4

U (TMK, conv) = £ +20
U (TMK, not conv) = £ -20

U (DM, conv) = £ +10
U (DM, not conv) = £ -10

U (TMK) = 20*0.6 + -20*0.4 = 4
U (DM) = 10*0.6 + -10*0.4 = 2

Go for Telemarketing

Naive Bayes

It assumes that if you know the correct label (expected class) for each component from your dataset you can predict the label for a new customer.

f1	f2	f3	y
0	1	0	1
0	0	1	1
1	0	1	1
0	0	1	1
0	1	0	1
1	0	1	0
1	1	1	0
1	1	0	0
1	1	1	0
1	0	1	0

Prediction for all features when the label is positive:

F1 (1,1) = 1/5 F2(1,1) = 2/5 F3(1,1) = 3/5
F1 (0,1) = 4/5 F2(0,1) = 3/2 F3(0,1) = 2/5

and negative:

F1 (1,0) = 5/5 F2(1,0) = 3/5 F3(1,0) = 4/5
F1 (0,0) = 0/5 F2(0,0) = 2/5 F3(0,0) = 1/5

New x : <0,1,1>

S(1) = F1(0,1)*F2(1,1)*F3(1,1) = 9/5 or 1.8
S(0) = F1(0,0)*F2(1,0)*F3(1,0) = 0

Prediction: If S(1)>S(0) the predict class is 1else class=0

Features are independent variables (y/n):

"central localization"
"contacted in the last six months"
"target A"

Naive Bayes

The probability from the portfolio of clients target A is given by:

Sex:	(70%) Men	and	(30%) women
Acquisition:	(95%) 1-6 months	and	(13%) 6+ months
Localization:	(90%) Central	and	(10%) countryside

The probability from the portfolio of clients target B is given by:

Sex:	(80%) Men	and	(20%) women
Acquisition:	(75%) 1-6 months	and	(25%) 6+ months
Localization:	(45%) Central	and	(55%) countryside

For the composition of a marketing campaign 70% from the customers are target B. The probability the new customer being target A given:

$p(y \mid x_*) = \{target=A \mid Sex=men, Aq= yes, Local=central\}$

$$p(y \mid x_*) = \frac{p(x_* \mid y)p(y)}{p(x_*)} = \frac{p(x_* \mid y)p(y)}{\sum_y p(x_* \mid y)\, p(y)}$$

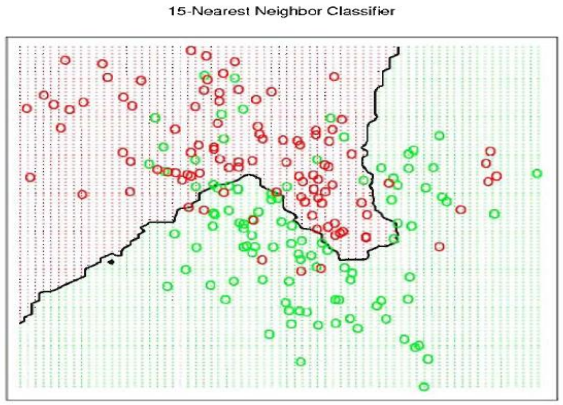
Clustering: *K-Nearest Neighbour*

To calculate the distance of the new point x compared to all training dataset **S** the Euclidean distance is computed by the formula:

$$dist\ (p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

After that, the algorithm will select a group of the closest neighbours (with size defined by K) and classify the new test point according to the classification of the majority of the neighbours

$$f(x) = \begin{cases} 0 & \text{if } \frac{1}{k}\sum_{i \epsilon Ix} y_i < \frac{1}{2} \\ 1 & \text{if } \frac{1}{k}\sum_{i \epsilon Ix} y_i \geq \frac{1}{2} \end{cases}$$



(Hastie, Tibshirani, & Friedman, 2001)

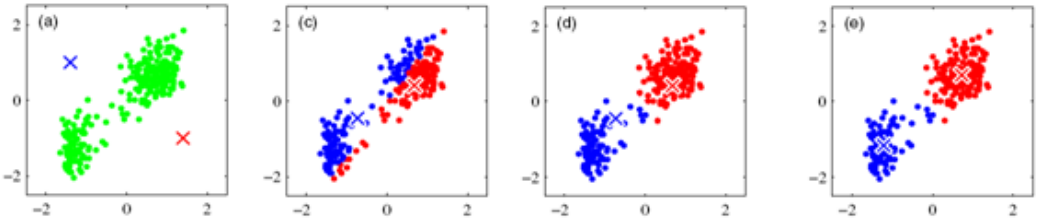
Clustering: *K-means*

• unknown labels

Given a set of unlabeled points x_i initialize k clusters centroids c randomically and start to calculate the Euclidean distances distance = $\text{argmin}_n \, |x_i - c_n|^2$.

After that, the centroids must to be updated to the mean position of the points assign to it : $c_n^{new} = \frac{1}{i} \sum x^i$

Because the number of clusters is restricting, this algorithm will eventually arrive at minima in finite steps.
The measure of how well centroids are place is given by residual sum of square or RSS. Thus for centroid k of a particular cluster w we have: $RSS_k = \sum |x - c_n(w_k)|^2$



(Bishop, 2006)

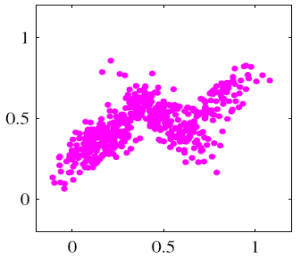
Clustering: *Gaussian Mixture Models (GMM's)*

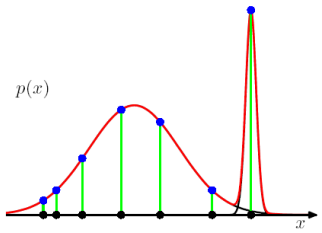
• Unknown labels

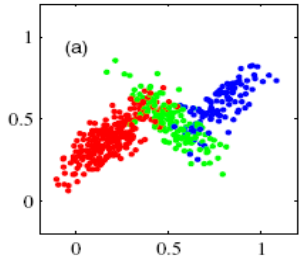
• When the data is distributed very close to each other is difficult to identify an ideal separation between the data.

The goal is to group the data in coherent clusters according the density concentration.

Assumption: each cluster is defined by a different Gaussian
Training set is the sum of all Gaussians distributions (distribution μ covariance σ)







(Bishop, 2006)

To fit this mixture of Gaussians, the model uses a latent (or hidden) variable z where x, z have the joint distribution: $P(x, z) = P(x \mid z)P(z)$

$z \sim \text{Multinomial}(\emptyset)$ where $\emptyset \geq 0$ and thus, $p(z=j)$
$x \mid z = j \sim N(\mu_j, \sum_j)$

Clustering: *Gaussian Mixture Models (GMM's)*

• Expectation-Maximization algorithm (EM) for statistical estimation of maximum likelihood

• The parameters Ø will learn from data such that $P(X|\theta) = P(X|z, \theta)P(z|\theta)$ and so we can maximize the log likelihood function $L(\theta)$

E-step: the missing labels z will be guessed, using the conditional expectation, by the observed data x and the current estimation of the model parameters.

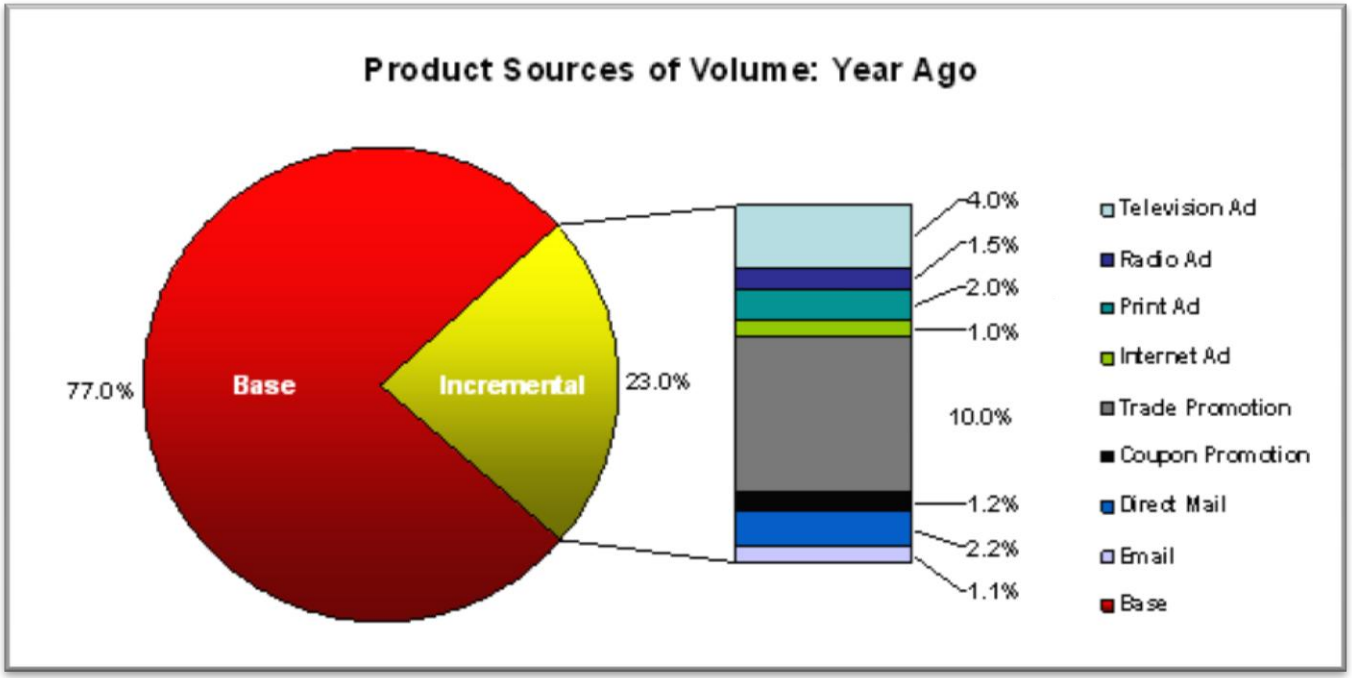
The probability that the label z come from a Gaussian distribution j

$$\frac{P(X|z)P(z)}{\sum_z P(X|z)P(z)} \quad => \quad p(z = j \mid x; \emptyset, \mu, \Sigma) \quad => \quad \frac{P(x|z = j; \mu, \Sigma)\, P(z = j; \emptyset)}{\sum_{l=1}^{k} P(x|z = l; \mu, \Sigma)\, P(z = l; \emptyset)}$$

M-step: It updates the model parameters based on the previous guesses. The likelihood function is maximized under the assumption that the missing data are known. The algorithm will increase the likelihood every each iteration.

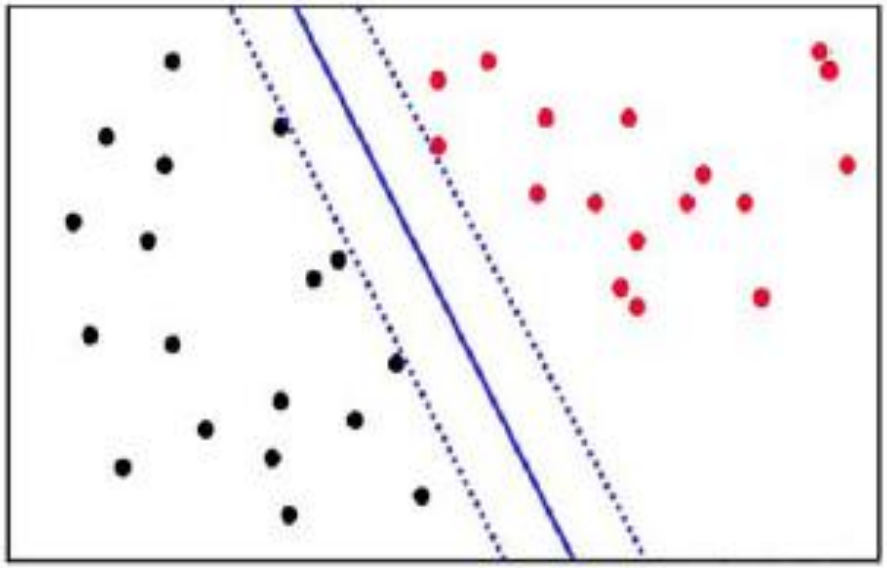
$$\emptyset_j = \frac{1}{m}\sum w_j \qquad \mu_j = \frac{\sum w_j x}{\sum w_j} \qquad \Sigma_j = \frac{\sum w_j (x - \mu_j)\,(x - \mu_j)^T}{\sum w_j}$$

Clustering: *Gaussian Mixture Models (GMM's)*



(Hollensen, 2003)

Linear Classifier



Linear Classifier: *Perceptron*

It works with a bias b and weight vector w which will be adapted during the learning process every time any point x_i is misclassified.

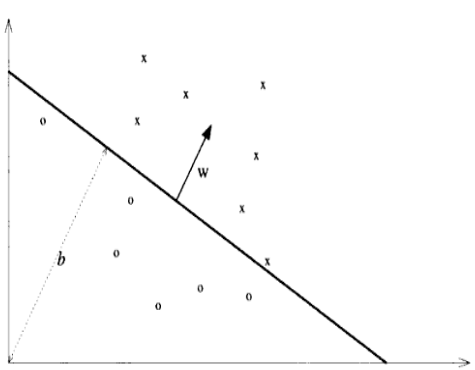
The linear function that interpolates the training set S is defined by f(x) = <w*x> +b

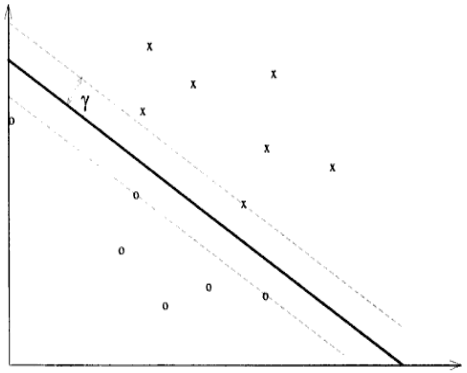
The function error ξ is calculated by the distance between the point misclassified and the line.

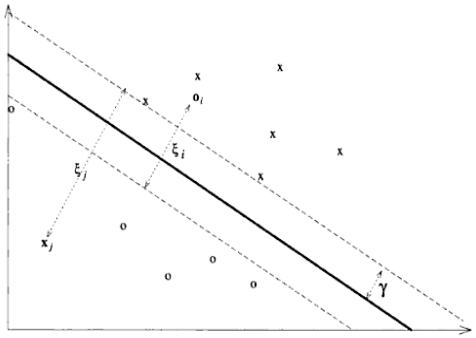
The primal Lagrangian $L(\mathbf{w},b) = \sum(y_i - < w \cdot x_i > - b)^2$

Dual form: $h(x) = \text{sign}\left(\sum \propto y < x \cdot x > + b\right)$ $\qquad w = \sum \propto y\, x$

Funcional margin: $\gamma i = \left(\frac{1}{\|w\|} w, \frac{1}{\|w\|} b\right)$







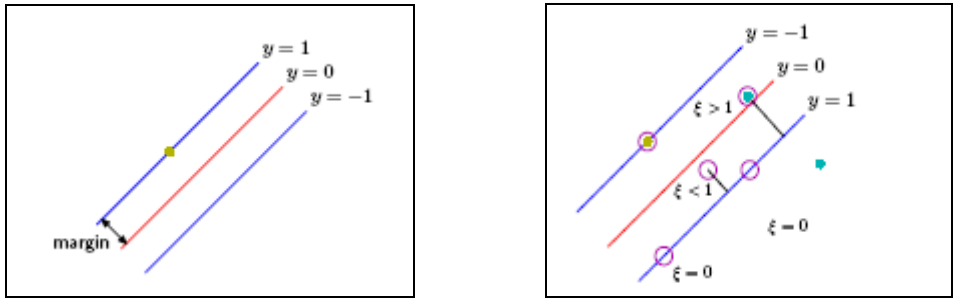
(Cristianini & Shawe-Taylor, 2000)

Linear Classifier: *SVM*

Maximal margin classifier

The weight vector w will be normalised and the computation of the functional margin will separate the points in positive and negative by the decision boundary:

$(\langle w \cdot x \rangle + b) \geq 1 \quad \Rightarrow$ class + $(\langle w \cdot x \rangle + b) < 1 \quad \Rightarrow$ class –



(Bishop, 2006)

Funcional margin: $\gamma i = \frac{1}{\|w\|^2} = (\sum \propto)^{-1/2}$

Assumption: The classification will work only with the value of the dot products from data vectors and not the high dimensionality.
It means that the binary classifier does not need the whole training data but only with the points that indicate how to specify the boundary between the two classes (support vector).

Linear Classifier: *SVM*

Function: $h(x) = \text{sign} < w \cdot x > + b$

Primal Lagragian: $L(\mathbf{w}, b, \propto) = \frac{1}{2} < w \cdot w > - \sum \propto [y(< w \cdot x > + b) - 1]$

Differentiating with respect to w and b will return: $w = \sum y \propto x = 0 \qquad \sum y \propto = 0$

Into the primal: $L(\mathbf{w}, b, \propto) = \frac{1}{2} - \sum_{i,j}^{l} y_i y_j \propto_i \propto_j < x_i \cdot x_j > - \sum_{i,j}^{l} y_i y_j \propto_i \propto_j < x_i \cdot x_j > - \sum_{i}^{l} \propto_i$

$$L(\mathbf{w}, b, \propto) = \sum_{i}^{l} \propto_i - \frac{1}{2} \sum_{i,j}^{l} y_i y_j \propto_i \propto_j < x_i \cdot x_j >$$

Dual representation: $f(x, \propto, b) = \sum_{i}^{l} y_i \propto_i^* < x_i \cdot x > + b^*$

Regularisation parameter *C*

Minimise $_{\xi, w, b}$ $<w \cdot w> + C \sum \xi$ subject to $y\ (<w \cdot x> + b) \geq 1 - \xi$

Conclusion

Why Data Mining is the future of CRM?

"... over 65% of database marketers in a recent Forrester survey say they use response rates as a key metric" (Forrester, 2008)

Simple x Sophisticated techniques
From tracking methods (RR) to learning algorithms

References

Berry, M. J., & Linoff, G. S. (2000). *Mastering data mining.* Wiley.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.

Cristianini, N., & Shawe-Taylor, J. (2000). *Support Vector Machines and other kernel-based learning models.* Cambridge: Cambridge University Press.

Dyché, J. (2002). *The CRM handbook : a business guide to customer relationship management.* Addison-Wesley.

Forrester. (2008). *http://www.omniture.com.* Fonte: Omniture.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and Prediction.* Springer.

Hollensen, S. (2003). *Marketing management: A relationship approach.* Pearson.

MIT, E. E. (2006). *MIT.* Acesso em 01 de 02 de 2010, disponível em MIT: http://ocw.mit.edu/OcwWeb/web/courses/courses/index.htm?r=iTunes#ElectricalEngineeringandComputerScience