Genetic Algorithm for Support Vector Machines Optimization in Probability of Default Prediction

Wolfgang Härdle
Dedy Dwi Prastyo

Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. – Center for Applied Statistics
and Economics
Humboldt–Universität zu Berlin
http://lvb.wiwi.hu-berlin.de
http://www.case.hu-berlin.de

Classifier

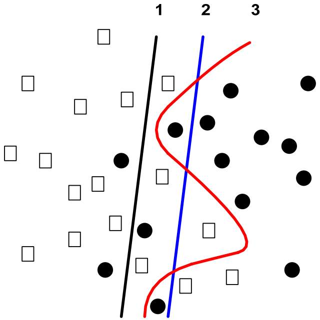


Figure 1: Linear classifier functions (1 and 2) and a non-linear one (3)

Loss

- Nonlinear classifier function f be described by a function class $\mathcal{F}$ fixed a priori, i.e. class of linear classifiers (hyperplanes)
- Loss

$$L(x,y) = \frac{1}{2}\,|f(x) - y| = \begin{cases} 0, & \text{if classification is correct,} \\ 1, & \text{if classification is wrong.} \end{cases}$$

Expected and Empirical Risk

- Expected risk – expected value of loss under the true probability measure

$$R(f) = \int \frac{1}{2} |f(x) - y| \, dF(x, y)$$

- Empirical risk – average value of loss over the training set

$$\widehat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} |f(x_i) - y_i|$$

VC bound

Vapnik-Chervonenkis (VC) bound – there is a function ϕ (monotone increasing in VC dimension h) so that for all $f \in \mathcal{F}$ with probability $1 - \eta$ hold

$$R(f) \leq \widehat{R}(f) + \phi\left(\frac{h}{n}, \frac{\log(\eta)}{n}\right)$$

Outline

1. Introduction ✓
2. Support Vector Machine (SVM)
3. Feature Selection
4. Application
5. Conclusions

SVM

- Classification
  Data $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\} : \Omega \to (\mathcal{X} \times \mathcal{Y})^n$
  $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \in \{-1, 1\}$

- Goal – to predict $\mathcal{Y}$ for new observation, $x \in \mathcal{X}$, based on information in D_n

Linearly (Non-) Separable Case

detail

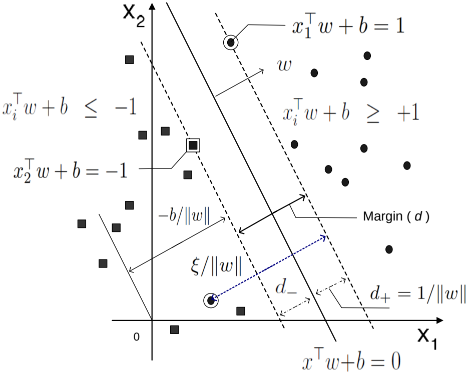


Figure 2: Hyperplane and its margin in linearly (non-) separable case

SVM Dual Problem

$$
\begin{aligned}
\max_{\alpha} L_D(\alpha) = \quad & \max_{\alpha} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \right\}, \\
\text{s.t.} \quad & 0 \le \alpha_i \le C \\
& \sum_{i=1}^{n} \alpha_i y_i = 0
\end{aligned}
$$

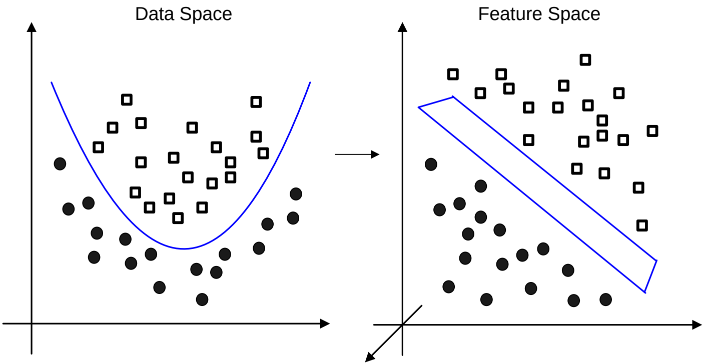


Figure 3: Mapping two dimensional data space into a three dimensional feature space, $\mathbb{R}^2 \mapsto \mathbb{R}^3$. The transformation $\Psi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^\top$ corresponds to $K(x_i, x_j) = (x_i^\top x_j)^2$

Non-Linear SVM

$$\max_{\alpha} L_D(\alpha) = \max_{\alpha} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right\}$$

$$\text{s.t.} \qquad 0 \le \alpha_i \le C, \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

- Gaussian RBF kernel – $K(x_i, x_j) = exp\left(-\frac{1}{\sigma} \|x_i - x_j\|^2\right)$
- Polynomial kernel – $K(x_i, x_j) = \left(x_i^\top x_j + 1\right)^p$

Structural Risk Minimization (SRM)

Search for the model structure $\mathcal{S}_h$,

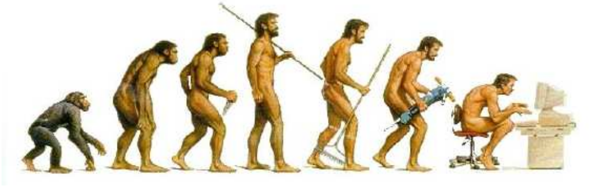
$$\mathcal{S}_{h_1} \subseteq \mathcal{S}_{h_2} \subseteq \ldots \subseteq \mathcal{S}_{h*} \subseteq \ldots \subseteq \mathcal{S}_{h_k} = \mathcal{F}$$

such that $f \in \mathcal{S}_{h*}$ minimises the expected risk bound, with $f \subseteq \mathcal{F}$ is class of linear function and h is VC dimension
i.e.

$$SVM(h_1) \subseteq \ldots \subseteq SVM(h*) \subseteq \ldots \subseteq SVM(h_k) = \mathcal{F}$$

with h correspond to the value of SVM (kernel) parameter

Evolutionary Feature Selection ▸ GA



- Featured selection – SVM parameters optimization
- Evolutionary optimization – Genetic Algorithm (GA)
- GA finds global optimum solution

GA - SVM

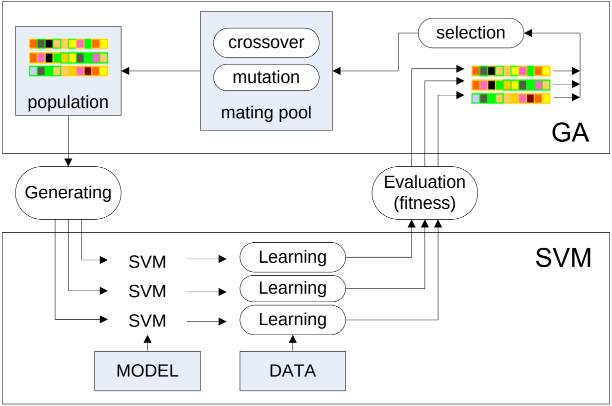


Figure 4: Iteration (generation) in GA-SVM

Credit Scoring & Probability of Default

- Score (Sc) from SVM method

$$Sc(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x)$$

- Probability of Default (PD)

$$f(y = 1|Sc) = \frac{1}{1 + \exp(\beta_0 + \beta_1 Sc)}$$

β_0 and β_1 are estimated by minimizing the negative log-likelihood function (Karatzoglou and Meyer, 2006)

Validation of Scores

Discriminatory power (of the score)

- Cumulative Accuracy Profile (CAP) curve
- Receiver Operating Characteristic (ROC) curve
- Accuracy, Specificity, Sensitivity

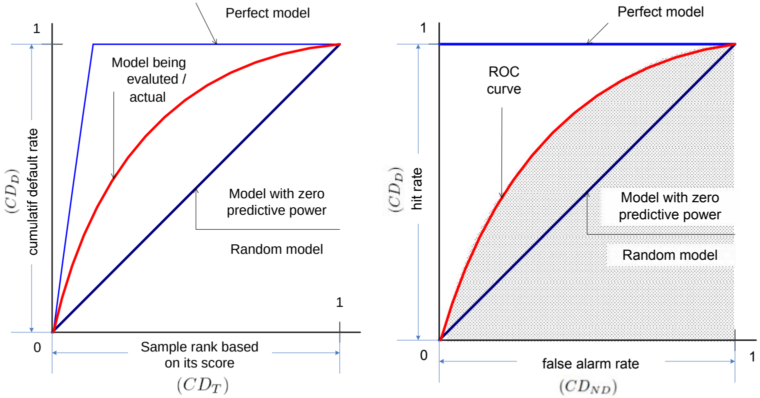


Figure 5: CAP curve (left) and ROC curve (right)

Discriminatory power

- **Cumulative Accuracy Profile** (CAP) curve
  - CAP/Power/Lorenz curve → Accuracy Ratio (AR)
  - Total sample vs. default sample

- **Receiver Operating Characteristic** (ROC) curve
  - ROC curve → Area Under Curve (AUC)
  - Non-default sample vs. default sample

- Relationship: **AR** = 2 **AUC** - 1

Discriminatory power (cont'd)

		sample	
		default (1)	non-default (-1)
predicted	(1)	True Positive (TP)	False Positive (FP)
	(-1)	False Negative (FN)	True Negative (TN)
total		P	N

- Accuracy, $\mathrm{P}(\widehat{Y} = Y) = \frac{TP+TN}{P+N}$
- Specificity, $\mathrm{P}(\widehat{Y} = -1 | Y = -1) = \frac{TN}{N}$
- Sensitivity, $\mathrm{P}(\widehat{Y} = 1 | Y = 1) = \frac{TP}{P}$

Examples – Small Sample

- 100 solvent and insolvent companies
- X3 – Operating Income / Total Asset
- X24 – Account Payable / Total Asset

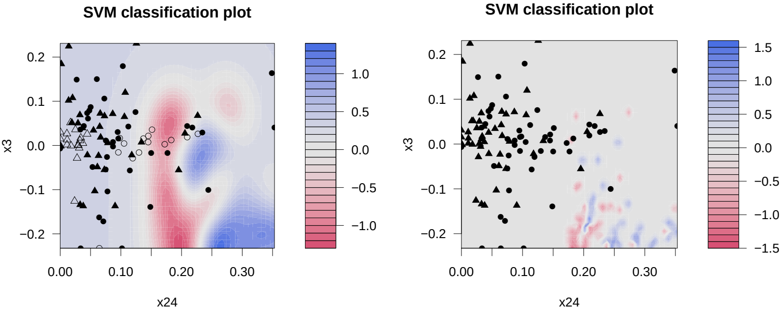


Figure 6: SVM plot, $C = 1$ and $\sigma = 1/2$, training error 0.19 (left) and GA-SVM, $C = 14.86$ and $\sigma = 1/121.61$, training error 0 (right).

Credit reform data

type	solvent (%)	insolvent (%)	total (%)
Manufacturing	27.37 (26.06)	25.70 (1.22)	27.29
Construction	13.88 (13.22)	39.70 (1.89)	15.11
Wholesale and retail	24.78 (23.60)	20.10 (0.96)	24.56
Real estate	17.28 (16.46)	9.40 (0.45)	16.90
total	83.31 (79.34)	94.90 (4.52)	83.86
others	16.69 (15.90)	5.10 (0.24)	16.14
#	20,000	1,000	21,000

Table 1: Credit reform data

Pre-processing

year	solvent # (%)	insolvent # (%)	total # (%)
1997	872 (9.08)	86 (0.90)	958 (9.98)
1998	928 (9.66)	92 (0.96)	1020 (10.62)
1999	1005 (10.47)	112 (1.17)	1117 (11.63)
2000	1379 (14.36)	102 (1.06)	1481 (15.42)
2001	1989 (20.71)	111 (1.16)	2100 (21.87)
2002	2791 (29.07)	135 (1.41)	2926 (30.47)
total	8964 (93.36)	638 (6.64)	9602 (100)

Table 2: Pre-processed credit reform data

Scenario

scenario	training set	testing set
Scenario-1	1997	1998
Scenario-2	1997-1998	1999
Scenario-3	1997-1999	2000
Scenario-4	1997-2000	2001
Scenario-5	1997-2001	2002

Table 3: Training and testing data set

Full model, $X_1, \ldots, X_{28}$

- Predictors – 28 financial ratio variables
- Population (# solutions) – 20
- Evolutionary iteration (generation) – 100
- Elitism – 0.2 of population
- Crossover rate – 0.5, mutation rate – 0.1
- Optimal SVM parameters – $\sigma = 1/178.75$ and $C = 63.44$

Quality of classification (1/2)

		sample	
		training	testing
Disc. power	AR	1	1
	AUC	1	1
	Accuracy	1	1
	Specificity	1	1
	Sensitivity	1	1

Table 4: Discriminatry power of Scenario-1, 2, 3, 4, 5

Quality of classification (2/2)

training	TE (CV)	testing	TE (CV)
1997	0 (8.98)	1998	0 (9.02)
1997-1998	0 (8.99)	1999	0 (10.03)
1997-1999	0 (9.37)	2000	0 (6.89)
1997-2000	0 (8.57)	2001	0 (5.29)
1997-2001	0 (4.55)	2002	0 (4.61)

Table 5: Percentage of Training Error (TE) and Cross-Validation (CV, with group=5)

Conclusion

- Optimal feature selection (via Genetic Algorithm) leads to perfect classification
- Cross validation – overcome the overfiting in training & testing error

Genetic Algorithm for Support Vector Machines Optimization in Probability of Default Prediction

Wolfgang Härdle
Dedy Dwi Prastyo

Ladislaus von Bortkiewicz Chair of Statistics
C.A.S.E. – Center for Applied Statistics and Economics
Humboldt–Universität zu Berlin
http://lvb.wiwi.hu-berlin.de
http://www.case.hu-berlin.de

References

Chen, S., Härdle, W. and Moro, R.
Estimation of Default Probabilities with Support Vector Machines
Quantitative Finance, 2011, 11, 135 - 154

Holland, J.H.
Adaptation in Natural and Artificial Systems
University of Michigan Press, 1975

References

Karatzoglou, A. and Meyer, D.
Support Vector Machines in R
Journal of Statistical Software, 2006, 15:9, 1-28

Zhang, J. L. and Härdle, W.
The Bayesian Additive Classification Tree Applied to Credit Risk Modelling
Computational Statistics and Data Analysis, 2010, 54, 1197-1205

Linearly Separable Case

back

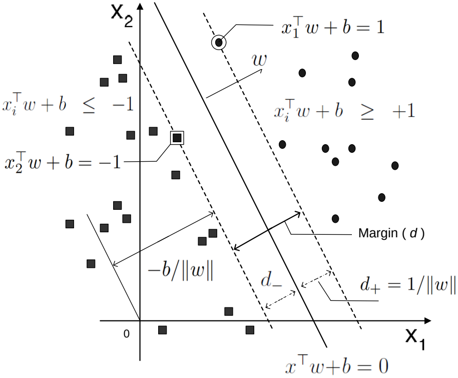


Figure 7: Separating hyperplane and its margin in linearly separable case

- Choose $f \in \mathcal{F}$ such that margin $(d_- + d_+)$ is maximal
- No error separation, if all $i = 1, 2, ..., n$ satisfy

$$\begin{aligned} x_i^\top w + b &\geq +1 \qquad \text{for} \quad y_i = +1 \\ x_i^\top w + b &\leq -1 \qquad \text{for} \quad y_i = -1 \end{aligned}$$

- Both constraints are combined into

$$y_i(x_i^\top w + b) - 1 \geq 0 \qquad i = 1, 2, ..., n$$

- Distance between margins and the separating hyperplane is $d_+ = d_- = 1/\|w\|$
- Maximize the margin, $d_+ + d_- = 2/\|w\|$, could be attained by minimizing $\|w\|$ or $\|w\|^2$
- Lagrangian for the primal problem

$$L_P(w, b) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \{y_i(x_i^\top w + b) - 1\}$$

Karush-Kuhn-Tucker (KKT) first order optimality conditions

$$\frac{\partial L_P}{\partial w_k} = 0: \qquad w_k - \sum_{i=1}^{n} \alpha_i y_i x_{ik} = 0 \qquad k = 1, \dots, d$$

$$\frac{\partial L_P}{\partial b} = 0: \qquad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$y_i(x_i^\top w + b) - 1 \geq 0 \qquad i = 1, \dots, n$$

$$\alpha_i \geq 0$$

$$\alpha_i \{ y_i(x_i^\top w + b) - 1 \} = 0$$

- Solution $w = \sum_{i=1}^{n} \alpha_i y_i x_i$, therefore

$$
\begin{aligned}
\frac{1}{2}\|w\|^2 &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \\
-\sum_{i=1}^{n} \alpha_i \{ y_i (x_i^\top w + b) - 1 \} &= -\sum_{i=1}^{n} \alpha_i y_i x_i^\top \sum_{j=1}^{n} \alpha_j y_j x_j + \sum_{i=1}^{n} \alpha_i \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j + \sum_{i=1}^{n} \alpha_i
\end{aligned}
$$

- Lagrangian for the dual problem

$$
L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j
$$

- Primal and dual problems

$$\min_{w,b} \quad L_P\,(w,b)$$

$$\max_{\alpha} \quad L_D\,(\alpha) \qquad\qquad \text{s.t.} \quad \alpha_i \geq 0, \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

- Optimization problem is convex, therefore the dual and primal formulations give the same solution
- Support vector, a point i for which $y_i(x_i^\top w + b) = 1$ holds

Linearly Non-separable Case

back

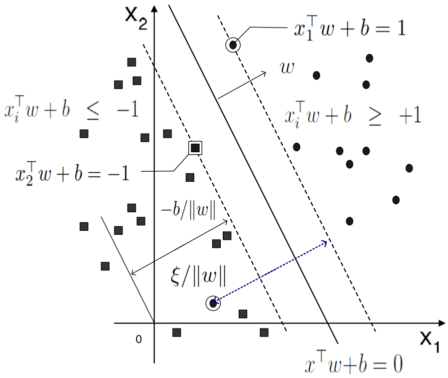


Figure 8: Hyperplane and its margin in linearly non-separable case

- Slack variables ξ_i represent the violation from strict separation

$$
\begin{aligned}
x_i^\top w + b &\geq 1 - \xi_i \qquad &\text{for} \qquad & y_i = 1, \\
x_i^\top w + b &\leq -1 + \xi_i \qquad &\text{for} \qquad & y_i = -1, \\
\xi_i &\geq 0
\end{aligned}
$$

- constraints are combined into

$$
y_i(x_i^\top w + b) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0
$$

- If $\xi_i > 0$, the objective function is

$$
\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i
$$

- Lagrange function for the primal problem

$$L_P(w,b,\xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\{y_i\left(x_i^\top w + b\right) - 1 + \xi_i\} - \sum_{i=1}^{n}\mu_i\xi_i,$$

where $\alpha_i \geq 0$ and $\mu_i \geq 0$ are Lagrange multipliers

- Primal problem

$$\min_{w,b,\xi} L_P(w,b,\xi)$$

First order conditions

$$\frac{\partial L_P}{\partial w_k} = 0 : \qquad w_k - \sum_{i=1}^{n} \alpha_i y_i x_{ik} = 0$$

$$\frac{\partial L_P}{\partial b} = 0 : \qquad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial L_P}{\partial \xi_i} = 0 : \qquad C - \alpha_i - \mu_i = 0$$

$$\begin{aligned} \text{s.t.} \quad & \alpha_i \geq 0, \quad \mu_i \geq 0, \quad \mu_i \xi_i = 0 \\ & \alpha_i \{ y_i (x_i^\top w + b) - 1 + \xi_i \} = 0 \end{aligned}$$

- Note that $\sum_{i=1}^{n} \alpha_i y_i b = 0$. Translate primal problem into

$$L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j + \sum_{i=1}^{n} \xi_i (C - \alpha_i - \mu_i)$$

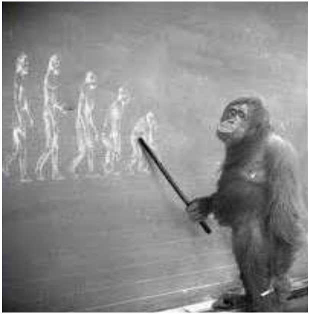
- Last term is 0, therefore the dual problem is

$$\max_{\alpha} L_D(\alpha) = \max_{\alpha} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \right\},$$

$$\text{s.t.} \qquad 0 \le \alpha_i \le C, \qquad \sum_{i=1}^{n} \alpha_i y_i = 0$$

back

What is a Genetic Algorithm ?



Genetics algorithm is search and optimization technique based on Darwin's principle on natural selection (Holland, 1975)

GA – Initialization

Back

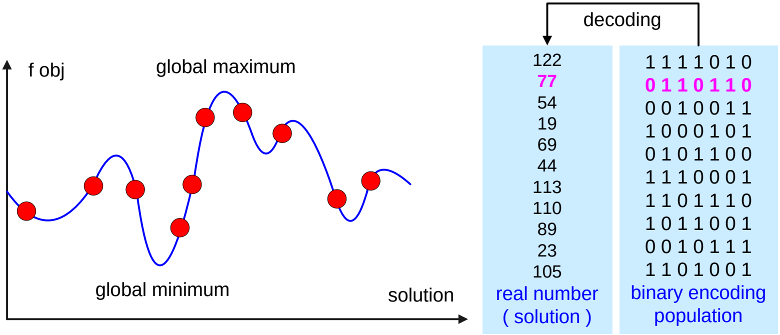


Figure 9: GA at first generation

GA – Convergency

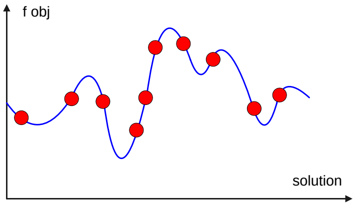


Figure 10: Solutions at 1^{st} generation (left) and r^{th} generation (right)

GA – Decoding

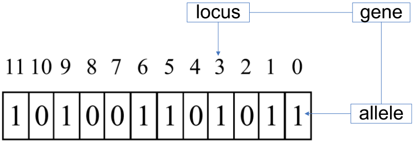


Figure 11: Decoding

$$\theta = \theta_{lower} + (\theta_{upper} - \theta_{lower}) \frac{\sum_{i=0}^{l-1} a_i 2^i}{2^l}$$

where θ is solution (i.e. parameter C or σ), a is allele

GA – Fitness evaluation

- Calculate $f(\theta_i)$, $\quad i = 1, \ldots, popsize$
- Evaluate fitness, $f_{dp}(\theta_i)$
  $f_{dp}(\theta_i)$ – AR, AUC, accuracy, specificity, sensitivity
- Relative fitness, $p_i = \frac{f_{dp}(\theta^i)}{\sum_{k=i}^{popsize} f_{dp}(\theta^i)}$

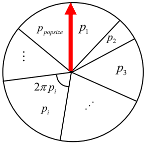
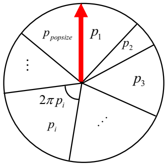


Figure 12: Proportion to be choosen in the next iteration (generation)

GA – Roulette wheel



- $rand \sim \mathrm{U}(0,1)$
- Select i^{th} chromosome if $\sum_{i=1}^{k} p_i < rand < \sum_{i=1}^{k+1} p_i$
- Repeat *popsize* times to get *popsize* new chromosomes

GA – Crossover

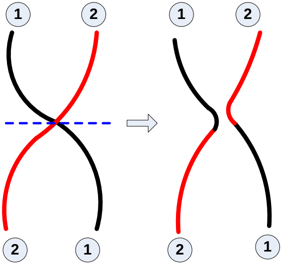


Figure 13: Crossover in nature

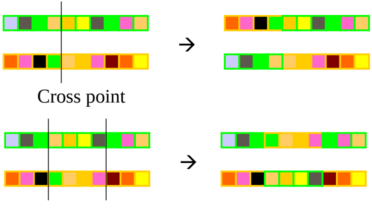


Figure 14: Randomly choosen one-point crossover (top) and two-points crossover (bottom)

GA – Reproductive operator

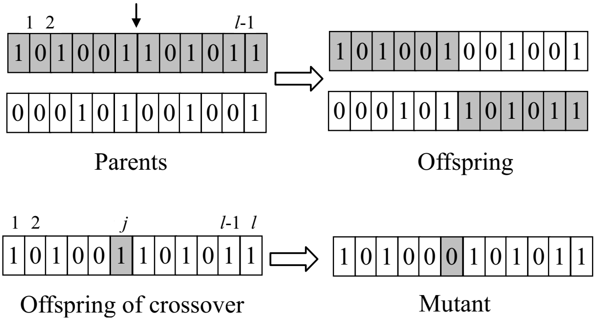


Figure 15: One-point crossover (top) and bit-flip mutation (bottom)

GA – Elitism

- Best solution in each iteration is maintained in another memory place
- New population replaces the old one, check whether best solution is in the population

  If not, replace any one in the population with best solution

Nature to Computer Mapping

Back

Nature	GA-SVM
Population	Set of parameter
Individual (phenotype)	Parameters
Fitness	Discriminatory power
Chromosome (genotype)	Encoding of parameter
Gene	Binary encoding
Reproduction	Crossover
Generation	Iteration

Table 6: Nature to GA-SVM mapping