KATHOLIEKE UNIVERSITEIT LEUVEN

Probabilistic Machine Learning Approaches to Medical Classification Problems

Chuan LU

Jury:

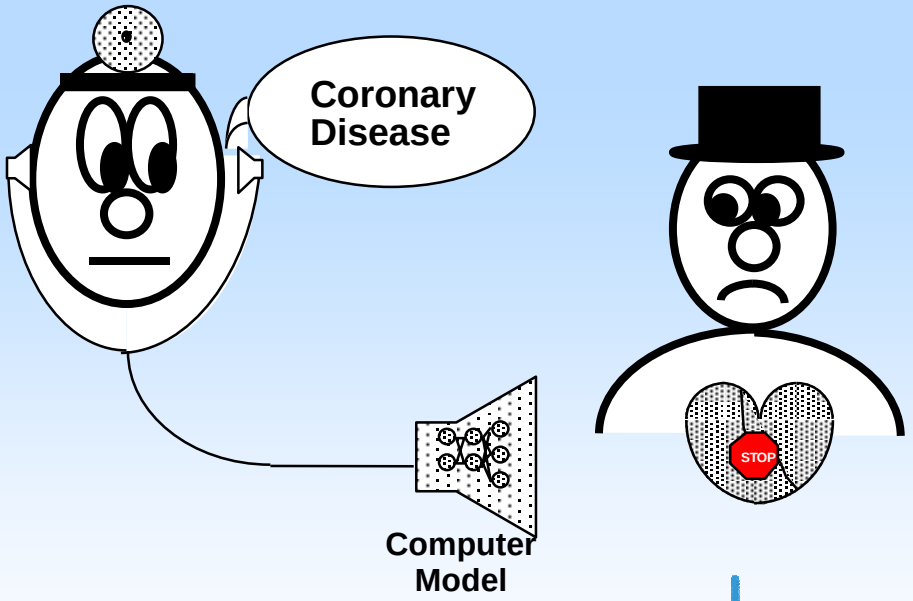
Prof. L. Froyen, chairman
Prof. S. Van Huffel, promotor
Prof. J.A.K. Suykens, promotor
Prof. D. Timmerman
Prof. J. Vandewalle
Prof. J. Beirlant
Prof. P.J.G. Lisboa
Prof. Y. Moreau

ESAT-SCD/SISTA
Katholieke Universiteit Leuven

SISTA

Clinical decision support systems

- Advances in technologies facilitate data collection
- computer based decision support systems
- Human beings: subjective, experience dependent.
- Artificial intelligence (AI) in medicine
- Expert system
- Machine learning
  - Diagnostic modelling
  - Knowledge discovery



Medical classification problems

- Essential for clinical decision making
- Constrained diagnosis problem
  - e.g. benign -, malignant + (for tumors).
- Classification
  - Find a rule to assign an obs. into one of the existing classes
  - supervised learning, pattern recognition

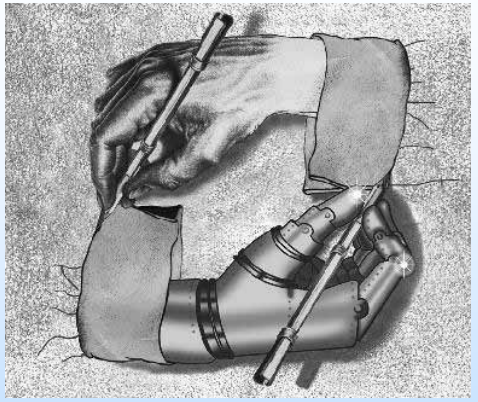
Medical classification problems

- Essential for clinical decision making
- Constrained diagnosis problem
  - e.g. benign -, malignant + (for tumors).
- Classification
  - Find a rule to assign an obs. into one of the existing classes
  - supervised learning, pattern recognition
- Our applications:
  - Ovarian tumor classification with patient data
  - Brain tumor classification based on MRS spectra
  - Benchmarking cancer diagnosis based on microarray data
- Challenge: uncertainty, validation, curse of dimensionality



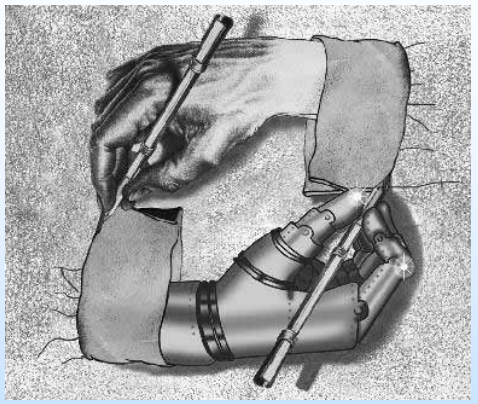
SISTA

Machine learning

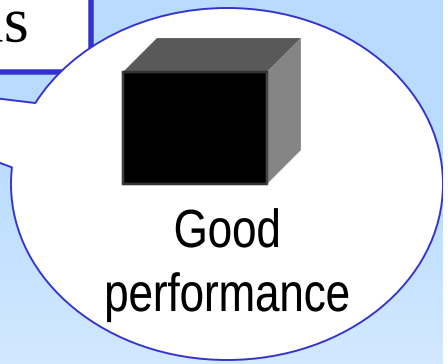


- Apply learning algorithms, autonomous acquisition and integration of knowledge
- Approaches
  - Conventional statistical learning algorithms
  - Artificial neural networks, Kernel-based models
  - Decision trees
  - Learning sets of rules
  - Bayesian networks

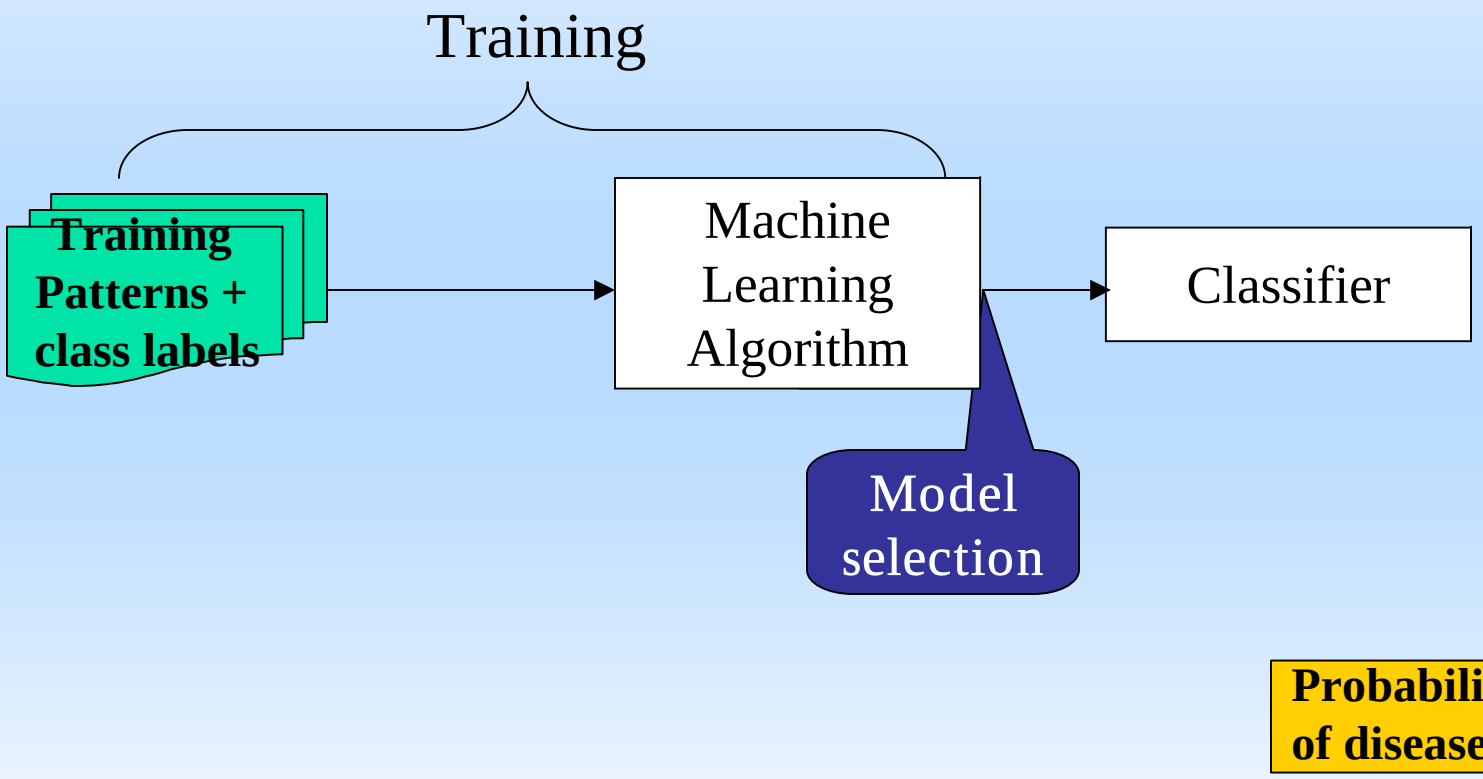
Machine learning



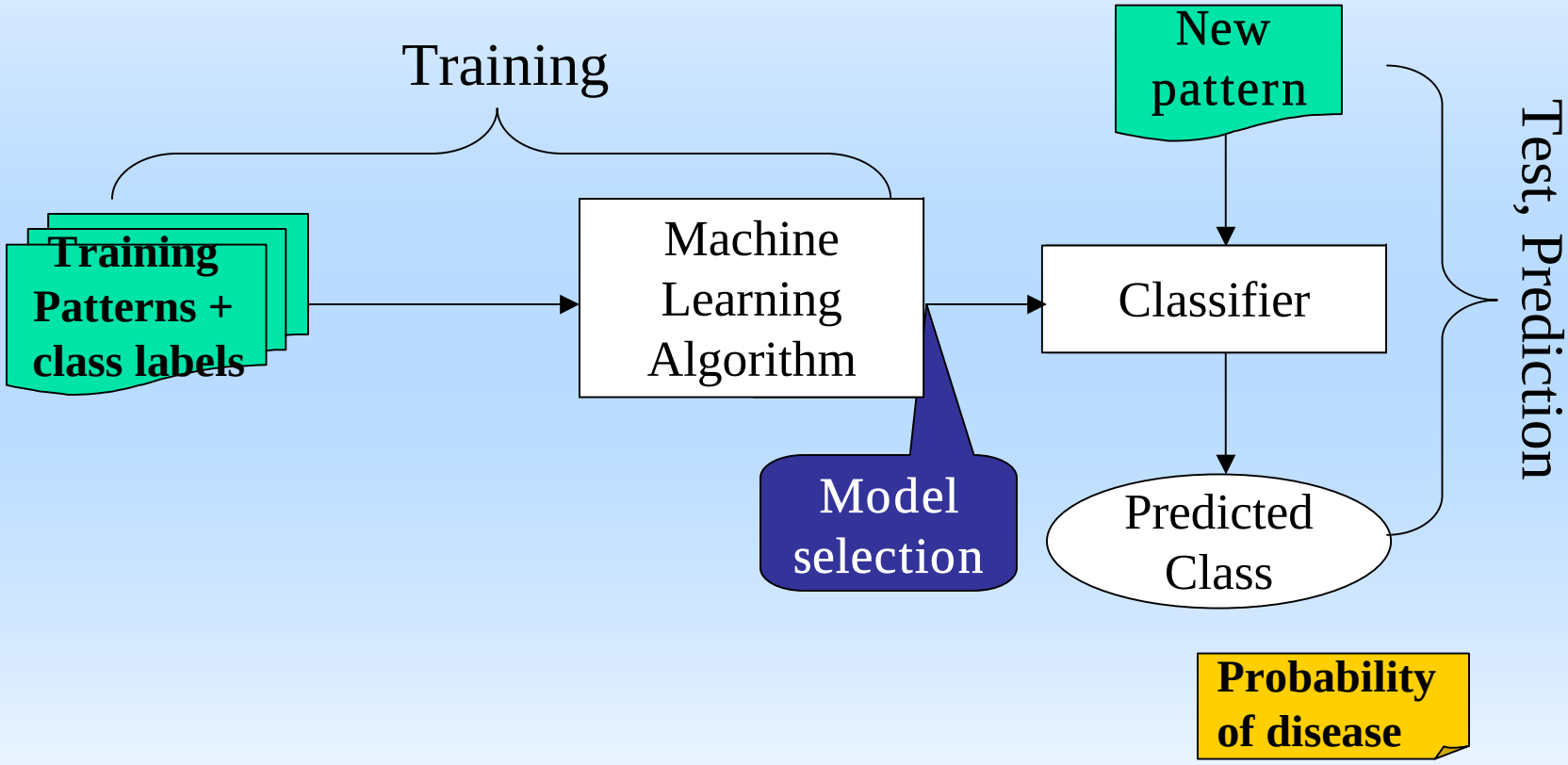
- Apply learning algorithms, autonomous acquisition and integration of knowledge
- Approaches
  - Conventional statistical learning algorithms
  - Artificial neural networks, Kernel-based models
  - Decision trees
  - Learning sets of rules
  - Bayesian networks



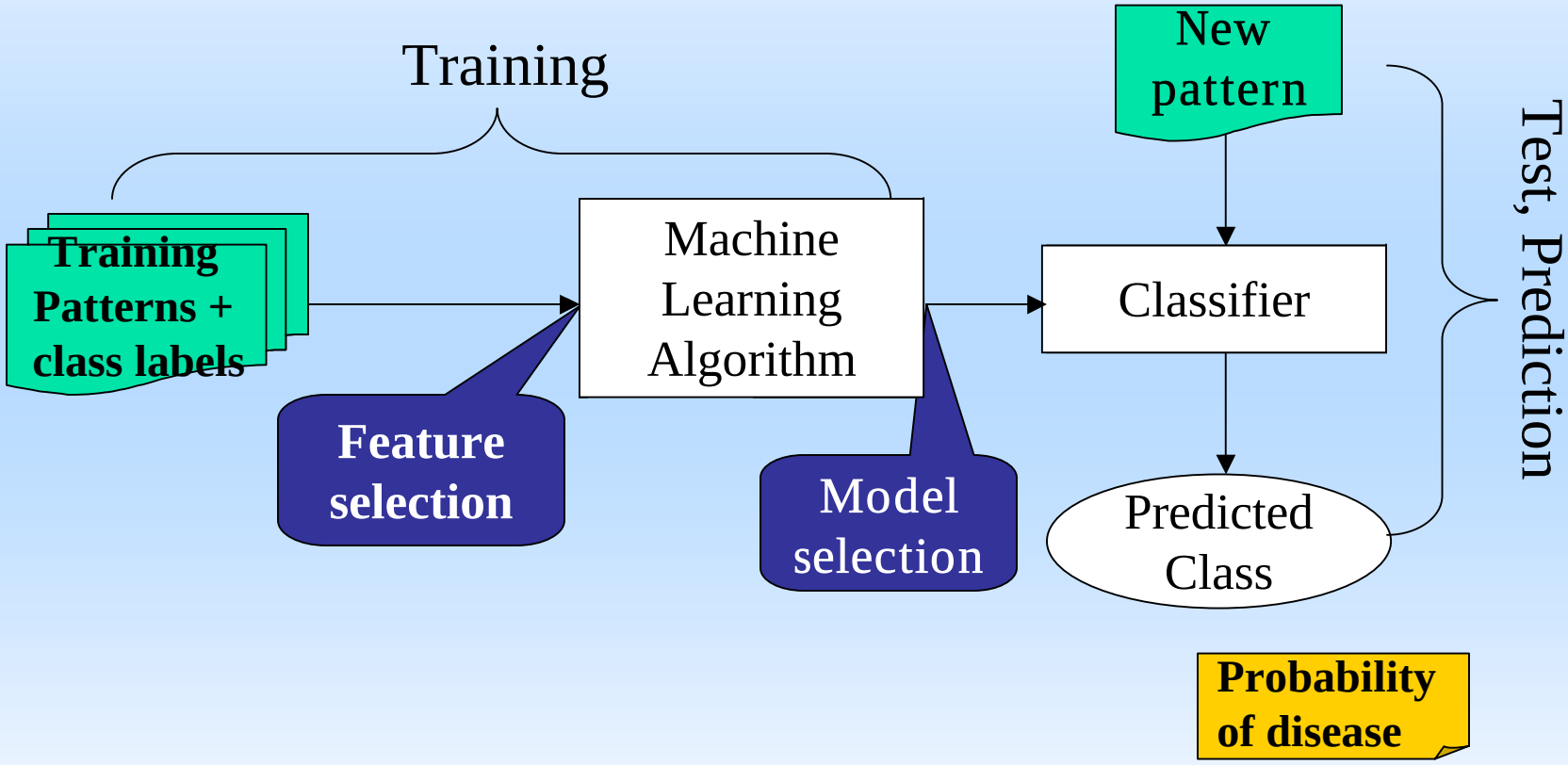
Building classifiers – a flowchart



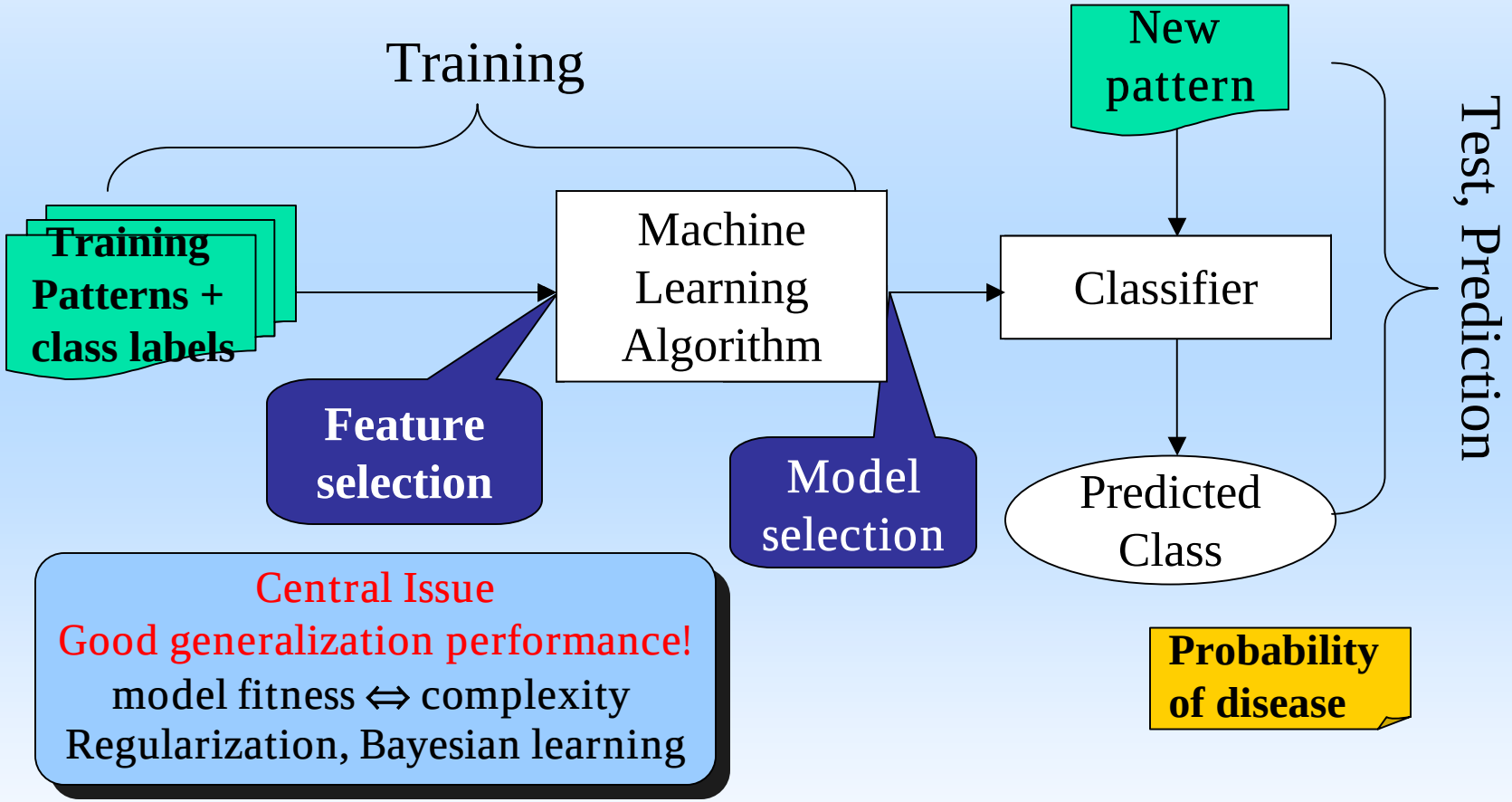
Building classifiers – a flowchart



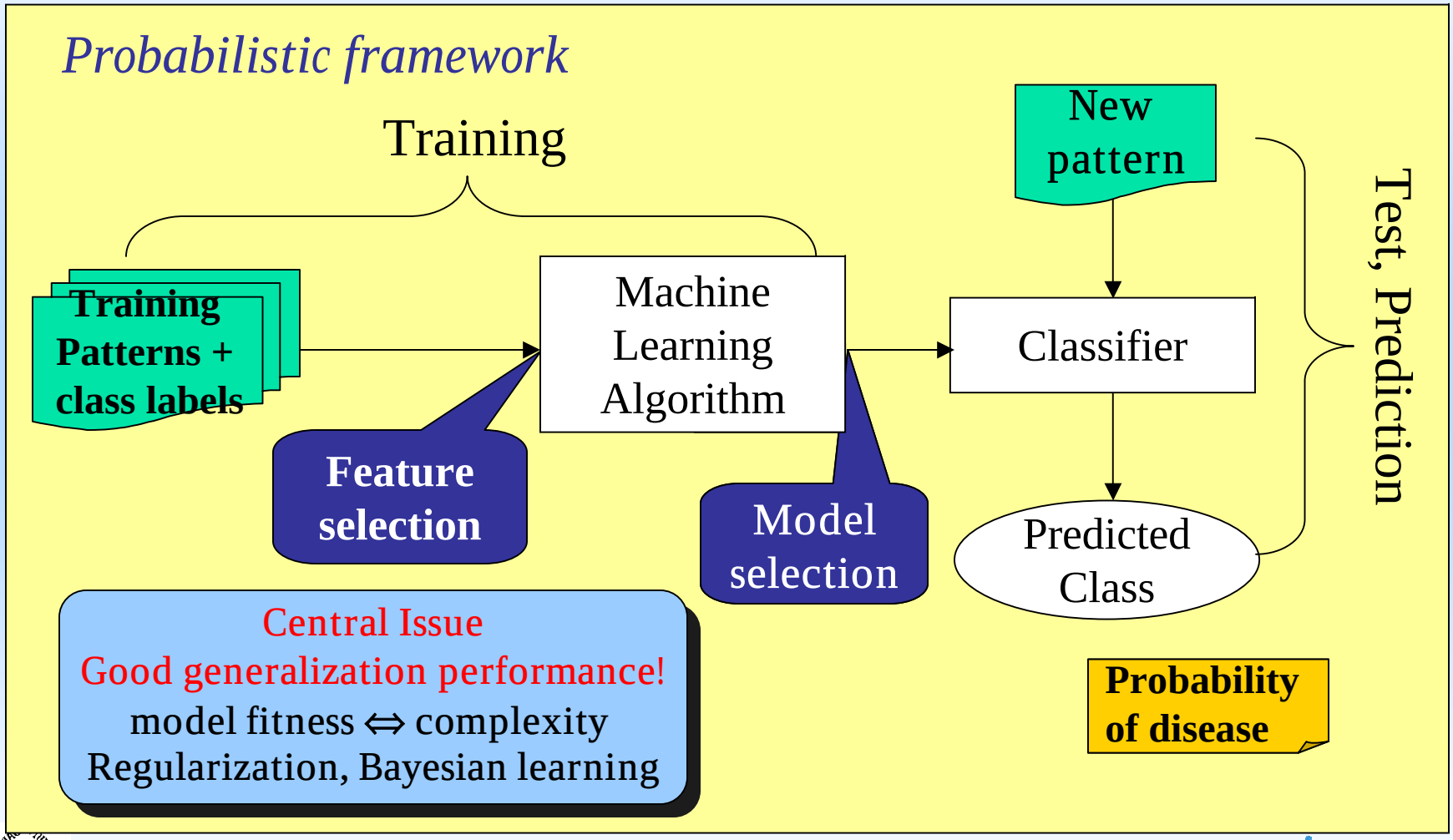
Building classifiers – a flowchart



Building classifiers – a flowchart



Building classifiers – a flowchart



Outline

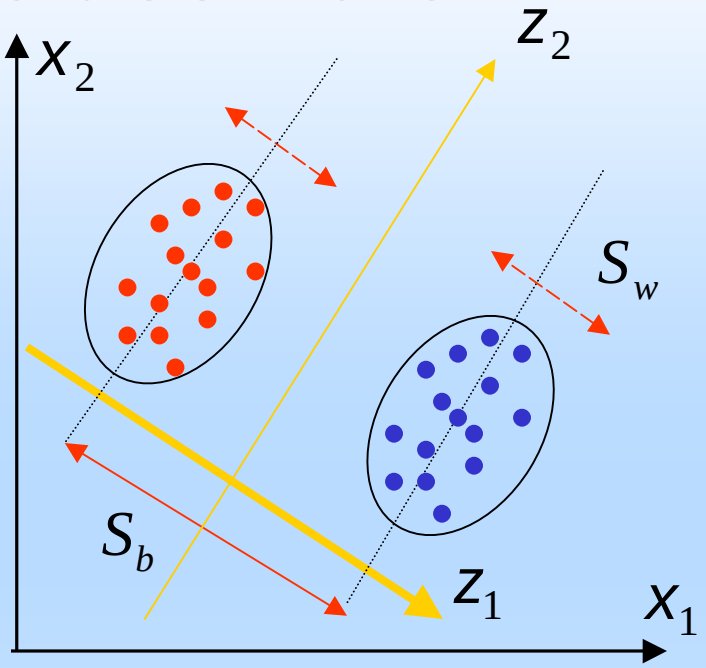
- Supervised learning
- Bayesian frameworks for blackbox models
- Preoperative classification of ovarian tumors
- Bagging for variable selection and prediction in cancer diagnosis problems
- Conclusions

Outline

- Supervised learning
- Bayesian frameworks for blackbox models
- Preoperative classification of ovarian tumors
- Bagging for variable selection and prediction in cancer diagnosis problems
- Conclusions

Conventional linear classifiers

- Linear discriminant analysis (LDA)
  - Discriminating using $z=\boldsymbol{w}^T\boldsymbol{x} \in \mathrm{R}$
  - Maximizing between-class variance while minimizing within-class variance

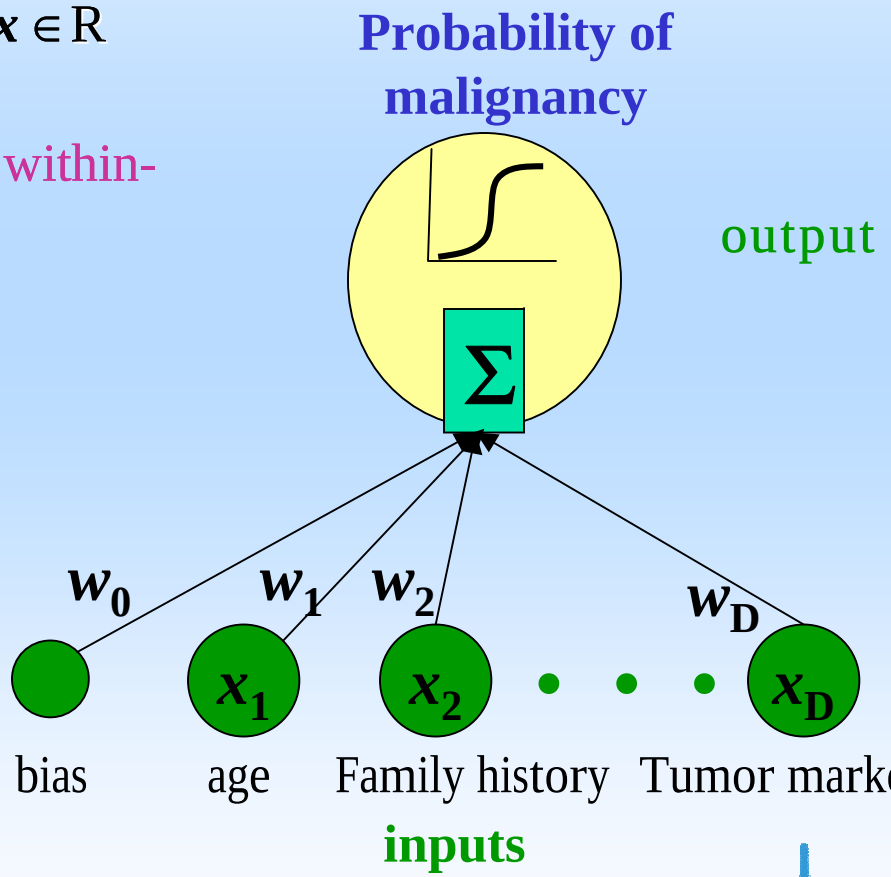


Conventional linear classifiers

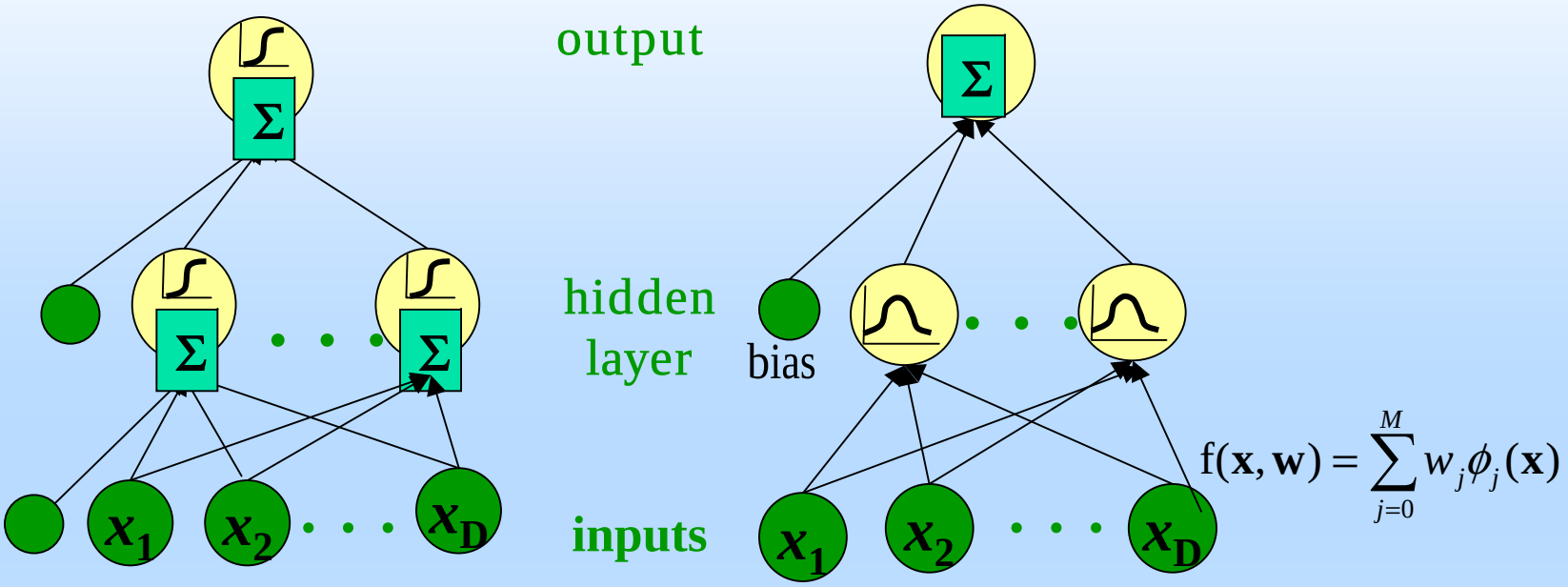
- Linear discriminant analysis (LDA)
  - Discriminating using $z=\boldsymbol{w}^T\boldsymbol{x} \in \mathrm{R}$
  - Maximizing between-class variance while minimizing within-class variance
- Logistic regression (LR)
  - Logit: log (odds)

$$= \log \frac{p}{1-p} = \mathbf{w}^T \mathbf{x} + b$$

  - Parameter estimation: maximum likelihood

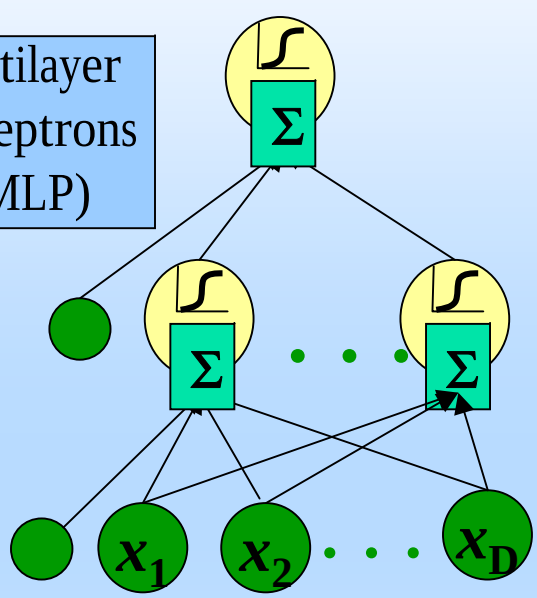


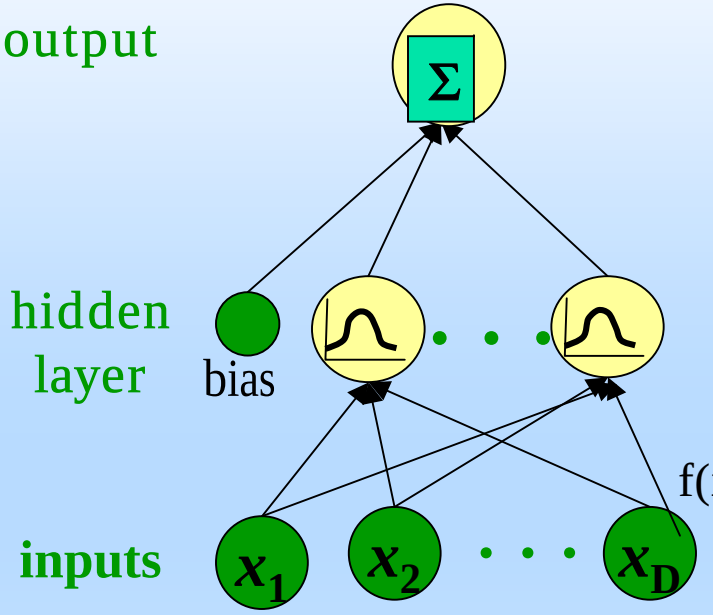
Feedforward neural networks



Feedforward neural networks

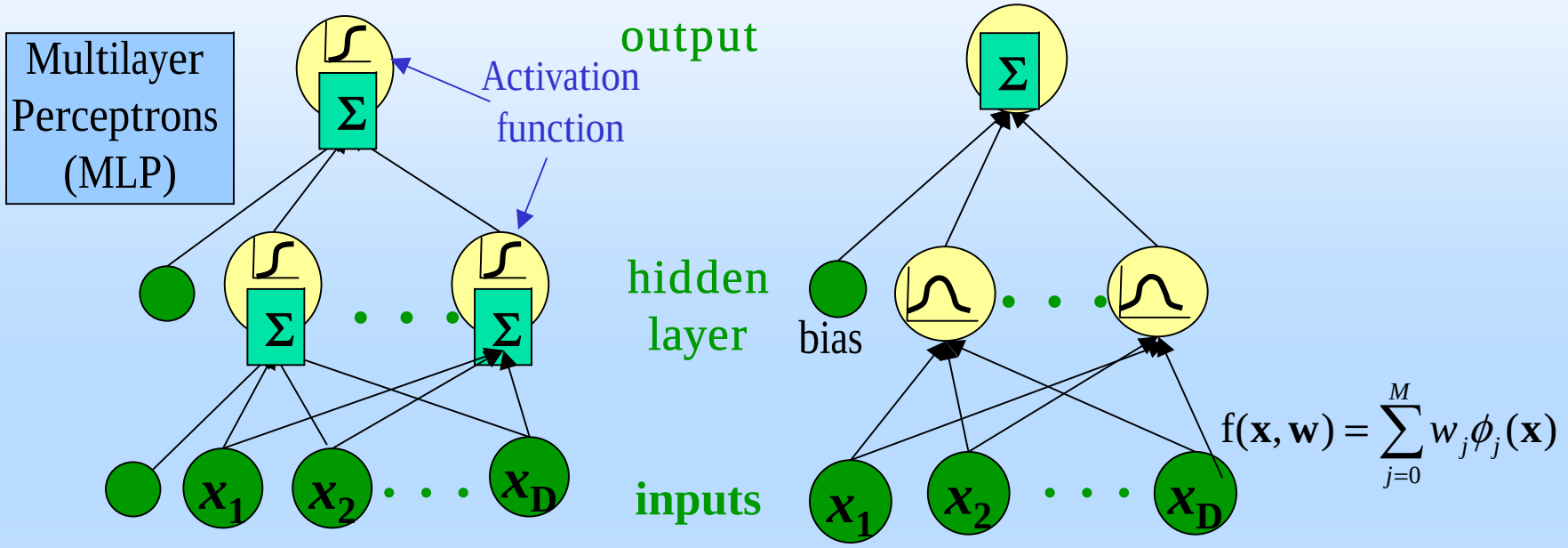
Multilayer Perceptrons (MLP)



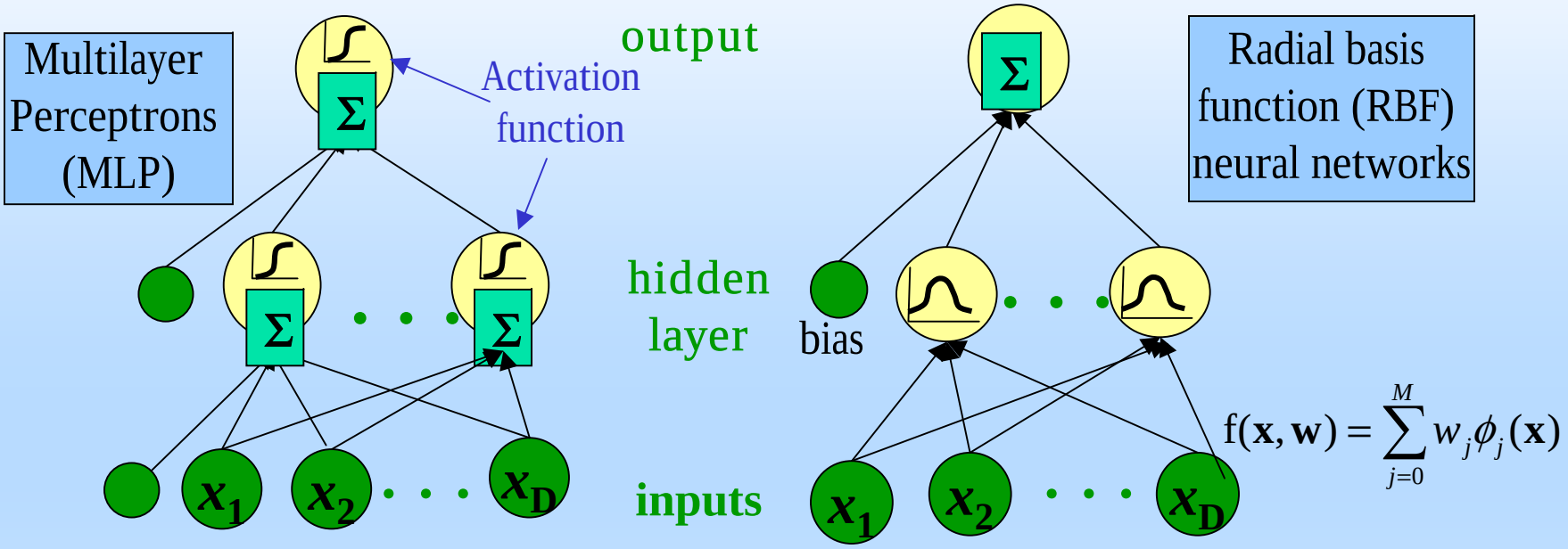


$$f(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M} w_j \phi_j(\mathbf{x})$$

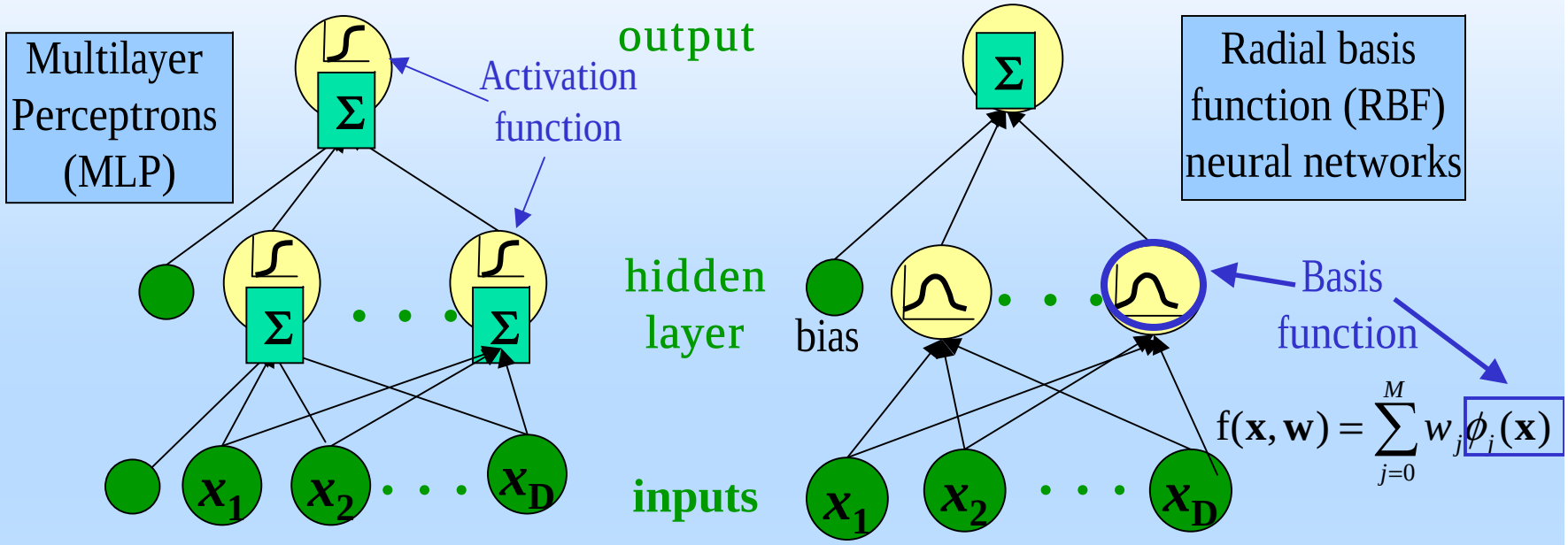
Feedforward neural networks



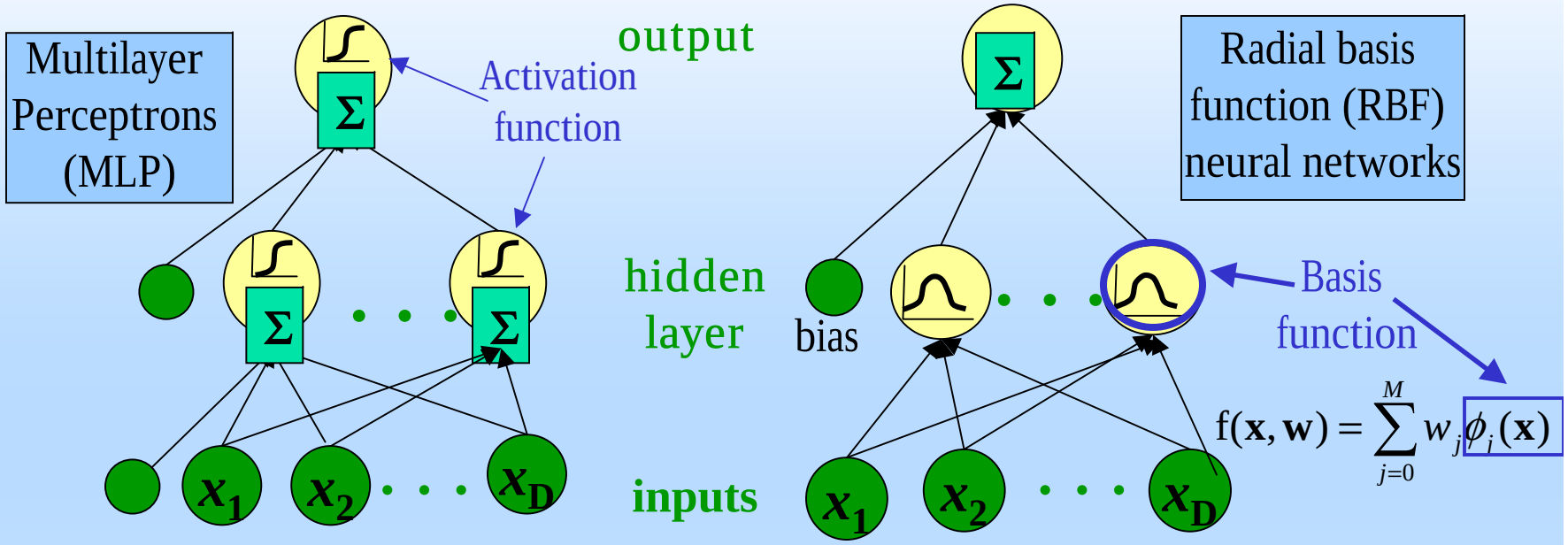
Feedforward neural networks



Feedforward neural networks



Feedforward neural networks



- Training (Back-propagation, L-M, CG,…), validation, test
- Regularization, Bayesian methods
- Automatic relevance determination (ARD)
  - Applied to MLP ⇒ variable selection
  - Applied to RBF-NN⇒ relevance vector machines (RVM)
- Local minima problem

Support vector machines (SVM)

- For classification: functional form

$$y(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} y_i \alpha_i \, k(\mathbf{x}, \mathbf{x}_i) + b \right)$$

kernel function

- Statistical learning theory [Vapnik95]

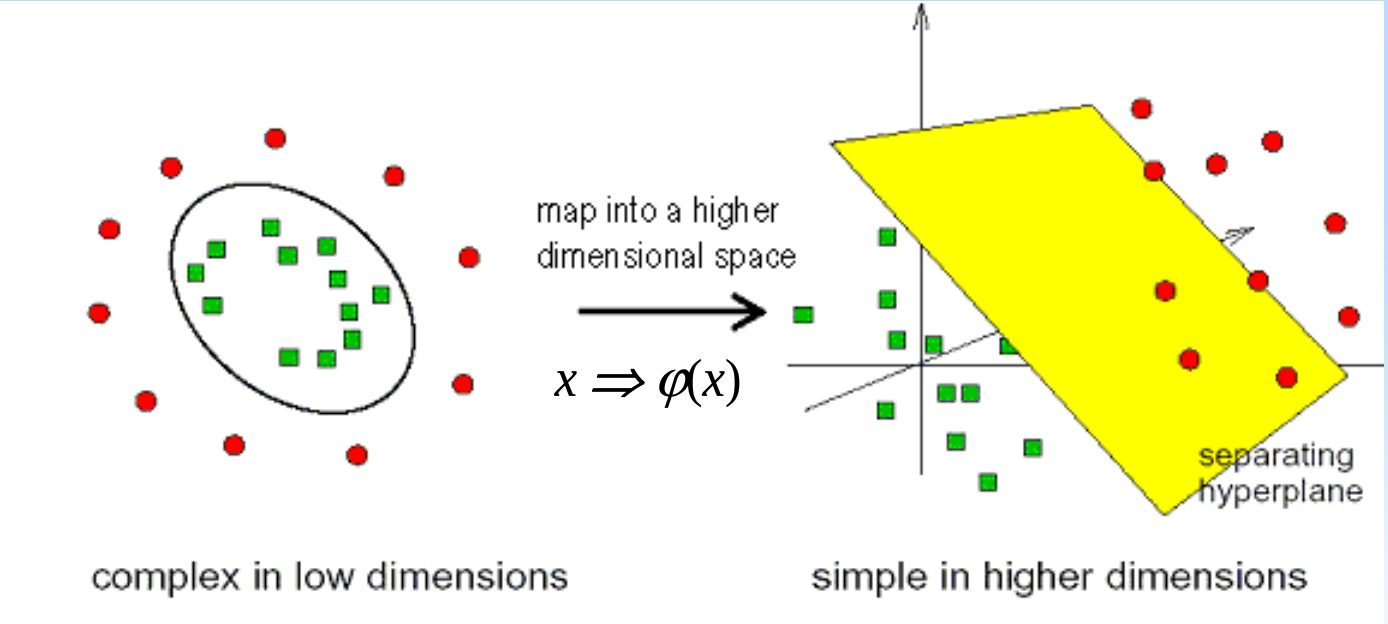
Support vector machines (SVM)

- For classification: functional form

$$y(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} y_i \alpha_i \, k(\mathbf{x}, \mathbf{x}_i) + b\right)$$

kernel function

- Statistical learning theory [Vapnik95]



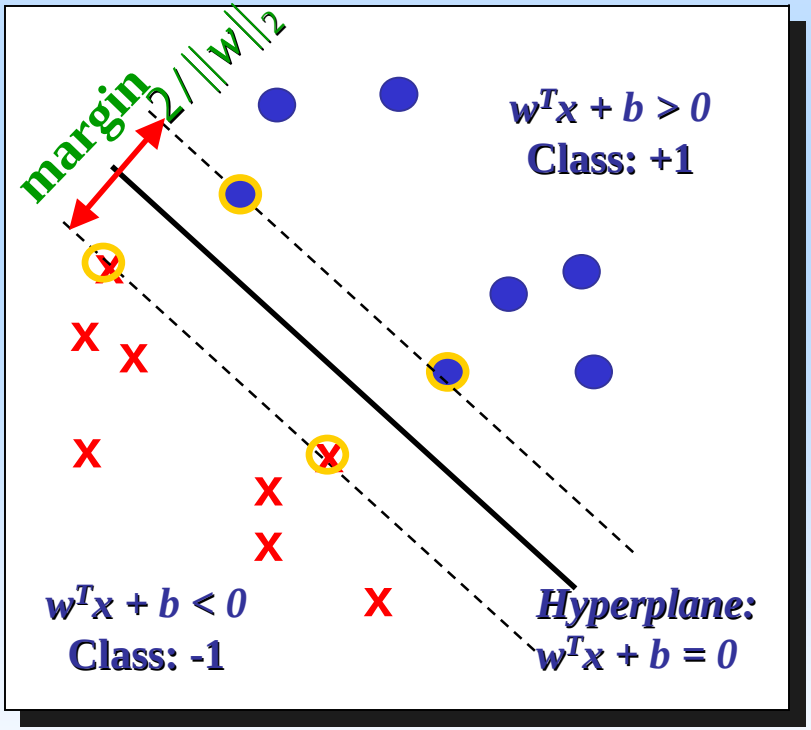
Support vector machines (SVM)

- For classification: functional form

$$y(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} y_i \alpha_i \, k(\mathbf{x}, \mathbf{x}_i) + b \right)$$

kernel function

- Statistical learning theory [Vapnik95]
- Margin maximization



Support vector machines (SVM)

- For classification, functional form

$$y(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} y_i \alpha_i \, k(\mathbf{x}, \mathbf{x}_i) + b \right)$$

kernel function

- Statistical learning theory [Vapnik95]
- Margin maximization
- Kernel trick

Support vector machines (SVM)

- For classification, functional form

$$y(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} y_i \alpha_i \, k(\mathbf{x}, \mathbf{x}_i) + b\right)$$

kernel function

- Statistical learning theory [Vapnik95]
- Margin maximization
- Kernel trick

Feature space

$$f(\mathbf{x}) = \mathbf{w}^T \varphi(\mathbf{x}) + b$$

Mercer's theorem

$k(\mathbf{x}, \mathbf{z}) = \langle \varphi(\mathbf{x}), \varphi(\mathbf{z}) \rangle$

Dual space

$$f(\mathbf{x}) = \sum_{i=1}^{N} y_i \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b$$

Support vector machines (SVM)

- For classification, functional form

$$y(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} y_i \alpha_i \, k(\mathbf{x}, \mathbf{x}_i) + b \right)$$

kernel function

- Statistical learning theory
- Margin maximization
- Kernel trick

Positive definite kernel k(.,.)

RBF kernel: $k(\mathbf{x}, \mathbf{z}) = \exp\{-\|\mathbf{x} - \mathbf{z}\|^2 / r^2\}$

Linear kernel: $k(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T \mathbf{z}$

Feature space

$$f(\mathbf{x}) = \mathbf{w}^T \varphi(\mathbf{x}) + b$$

Mercer's theorem

$k(\mathbf{x}, \mathbf{z}) = \langle \varphi(\mathbf{x}), \varphi(\mathbf{z}) \rangle$

Dual space

$$f(\mathbf{x}) = \sum_{i=1}^{N} y_i \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b$$

Support vector machines (SVM)

- For classification, functional form

$$y(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} y_i \alpha_i \, k(\mathbf{x}, \mathbf{x}_i) + b \right)$$

kernel function

Positive definite kernel k(.,.)

RBF kernel: $k(\mathbf{x}, \mathbf{z}) = \exp\{-\|\mathbf{x} - \mathbf{z}\|^2 / r^2\}$

Linear kernel: $k(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T \mathbf{z}$

- Statistical learning theory
- Margin maximization
- Kernel trick

Feature space

$$f(\mathbf{x}) = \mathbf{w}^T \varphi(\mathbf{x}) + b$$

Mercer's theorem

$k(\mathbf{x}, \mathbf{z}) = \langle \varphi(\mathbf{x}), \varphi(\mathbf{z}) \rangle$

Dual space

$$f(\mathbf{x}) = \sum_{i=1}^{N} y_i \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b$$

- Quadratic programming
- Sparseness, unique solution
- Additive kernels $k(\mathbf{x}, \mathbf{z}) = \sum_{j=1}^{D} k_j(x^{(j)}, z^{(j)})$

Support vector machines (SVM)

- For classification, functional form

$$y(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} y_i \alpha_i \, k(\mathbf{x}, \mathbf{x}_i) + b \right)$$

kernel function

Positive definite kernel k(.,.)

RBF kernel: $k(\mathbf{x}, \mathbf{z}) = \exp\{-\|\mathbf{x} - \mathbf{z}\|^2 / r^2\}$

Linear kernel: $k(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T \mathbf{z}$

- Statistical learning theory
- Margin maximization
- Kernel trick

Feature space

$$f(\mathbf{x}) = \mathbf{w}^T \varphi(\mathbf{x}) + b$$

Mercer's theorem

$k(\mathbf{x}, \mathbf{z}) = <\varphi(\mathbf{x}), \varphi(\mathbf{z})>$

Dual space

$$f(\mathbf{x}) = \sum_{i=1}^{N} y_i \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b$$

- Quadratic programming
- Sparseness, unique solution
- Additive kernels $k(\mathbf{x}, \mathbf{z}) = \sum_{j=1}^{D} k_j(x^{(j)}, z^{(j)})$

Additive kernel-based models
Enhanced interpretability

Variable selection!

Least squares SVMs

- LS-SVM classifier [Suykens99]
  - SVM variant
  - Inequality constraint $\Rightarrow$ equality constraint
  - Quadratic programming $\Rightarrow$ solving linear equations



SISTA

Least squares SVMs

- LS-SVM classifier [Suykens99]
  - SVM variant
  - Inequality constraint $\Rightarrow$ equality constraint
  - Quadratic programming $\Rightarrow$ solving linear equations

Primal problem

The following model is taken:

$$\mathbf{f}(\mathbf{x}) = \mathbf{w}^T \varphi(\mathbf{x}) + b$$

$$\min_{w,b} J(\mathbf{w}, b) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} e_i^2,$$

$$\text{s.t. } y_i[\mathbf{w}^T \varphi(\mathbf{x}_i) + b] = 1 - e_i$$

$$i = 1, \ldots, N$$

with regularization const. C.

Least squares SVMs

- LS-SVM classifier [Suykens99]
  - SVM variant
  - Inequality constraint ⇒ equality constraint
  - Quadratic programming ⇒ solving linear equations

Primal problem

The following model is taken:

$$\mathbf{f}(\mathbf{x}) = \mathbf{w}^T \varphi(\mathbf{x}) + b$$

$$\min_{w,b} J(\mathbf{w}, b) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} e_i^2,$$

$$\text{s.t. } y_i[\mathbf{w}^T\varphi(\mathbf{x}_i) + b] = 1 - e_i$$

$$i = 1, \ldots, N$$

with regularization const. C.

solved in dual space

Dual problem

$$\begin{bmatrix} 0 & \mathbf{1}_v^T \\ \mathbf{1}_v & \boldsymbol{\Omega} + C^{-1}\mathbf{I}_N \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{y} \end{bmatrix}$$

$$\mathbf{y} = [y_1, \ldots, y_N]^T, \mathbf{1}_v = [1, \ldots, 1]^T, \mathbf{e} = [e_1, \ldots, e_N]^T,$$

$$\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]^T, \Omega_{ij} = \varphi(\mathbf{x}_i)^T\varphi(\mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j)$$

Resulting classifier:

$$\mathrm{y}(\mathbf{x}) = \mathrm{sign}[\sum_{i=1}^{N} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b]$$

Model evaluation

- Performance measure
  - Accuracy: correct classification rate
  - Receiver operating characteristic (ROC) analysis
    - Confusion table

		True result	
		–	+
Test result	–	TN	FN
	+	FP	TP

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{specficity} = \frac{TN}{TN + FP}$$

    - ROC curve
    - Area under the ROC curve
      AUC=$P[y(\boldsymbol{x}_-)<y(\boldsymbol{x}_+)]$

Model evaluation

- Performance measure
  - Accuracy: correct classification rate
  - Receiver operating characteristic (ROC) analysis
    - Confusion table

		True result	
		–	+
Test result	–	TN	FN
	+	FP	TP

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{specficity} = \frac{TN}{TN + FP}$$

    - ROC curve
    - Area under the ROC curve
      AUC=$P[y(\boldsymbol{x}_-)<y(\boldsymbol{x}_+)]$

Assumption:
equal misclass. cost and constant class distribution in the target environment

Model evaluation

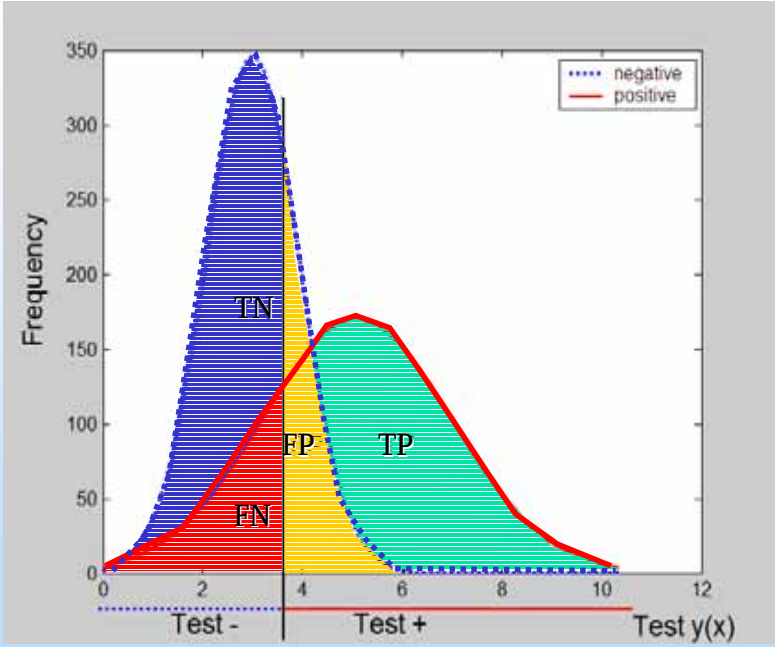
- Performance measure
  - Accuracy: correct classification rate
  - Receiver operating characteristic (ROC) analysis
    - Confusion table

		True result	
		–	+
Test result	–	TN	FN
	+	FP	TP

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{specficity} = \frac{TN}{TN + FP}$$

    - ROC curve
    - Area under the ROC curve
      AUC=$P[y(\boldsymbol{x}_-)<y(\boldsymbol{x}_+)]$



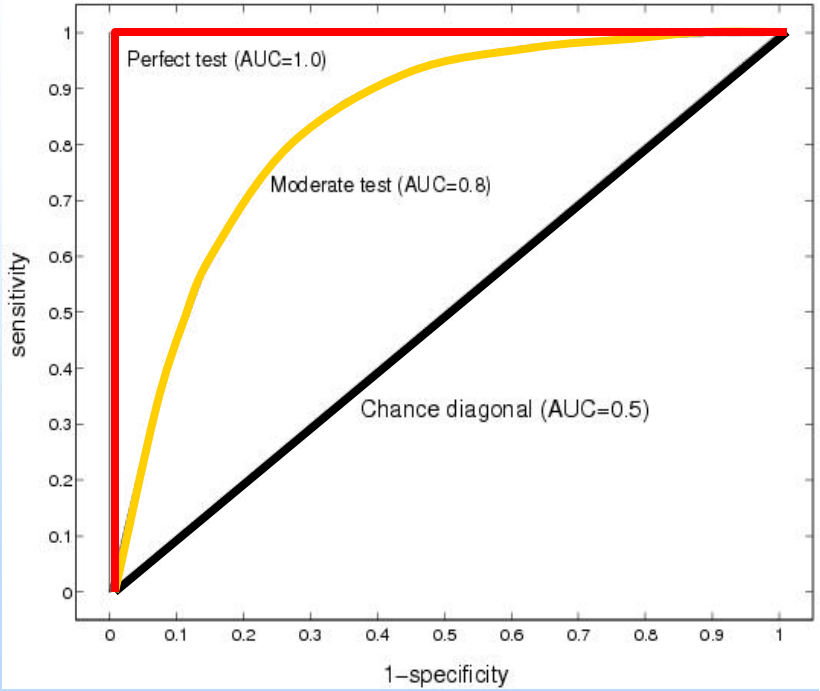
Model evaluation

- Performance measure
  - Accuracy: correct classification rate
  - Receiver operating characteristic (ROC) analysis
    - Confusion table

		True result	
		–	+
Test result	–	TN	FN
	+	FP	TP

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{specficity} = \frac{TN}{TN + FP}$$



    - ROC curve
    - Area under the ROC curve
      AUC=$P[y(\boldsymbol{x}_-)<y(\boldsymbol{x}_+)]$

Model evaluation

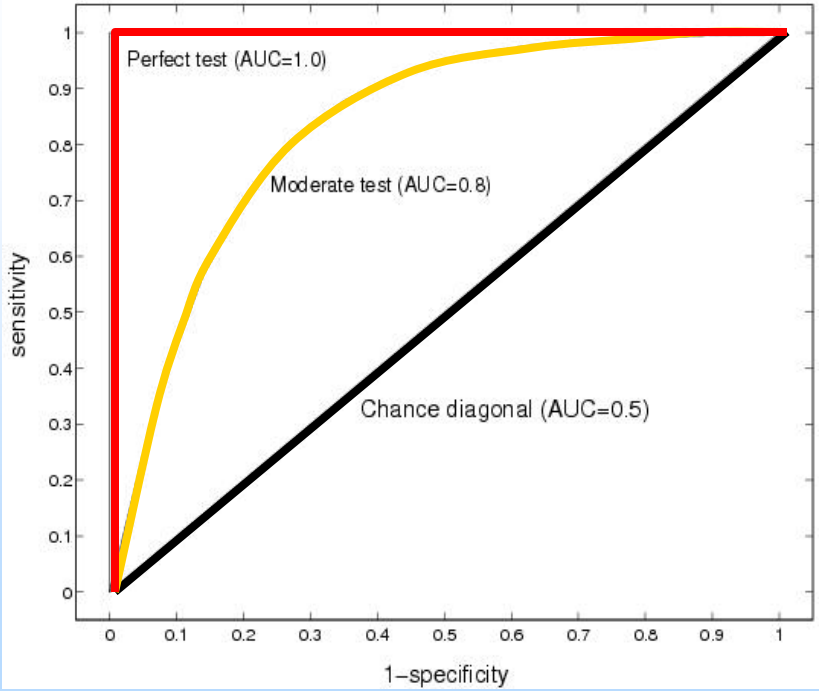
- Performance measure
  - Accuracy: correct classification rate
  - Receiver operating characteristic (ROC) analysis
    - Confusion table

		True result	
		–	+
Test result	–	TN	FN
	+	FP	TP

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{specficity} = \frac{TN}{TN + FP}$$

    - ROC curve
    - Area under the ROC curve
      AUC=$P[y(\boldsymbol{x}_-)<y(\boldsymbol{x}_+)]$



Training
Validation
Test

Outline

- Supervised learning
- Bayesian frameworks for blackbox models
- Preoperative classification of ovarian tumors
- Bagging for variable selection and prediction in cancer diagnosis problems
- Conclusions

Bayesian frameworks for blackbox models

- Advantages
  - Automatic control of model complexity, without CV
  - Possibility to use prior info and hierarchical models for hyperparameters
  - Predictive distribution for output

Principle of Bayesian learning [MacKay95]

- Define the probability **distribution** over all quantities within the model
- Update the distribution given data using **Bayes' rule**
- Construct **posterior probability distributions** for the (hyper)parameters.
- Prediction based on the posterior distributions over all the parameters.



SISTA

Bayesian inference

[MacKay95, Suykens02, Tipping01]

Bayes' rule $$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

$$p\left(\mathbf{w}, b \middle| D, \boldsymbol{\theta}, H\right) = \frac{p(D|\mathbf{w}, b, \boldsymbol{\theta}, H)\, p(\mathbf{w}, b|\boldsymbol{\theta}, H)}{P(D|\boldsymbol{\theta}, H)}$$

$$p(\boldsymbol{\theta}|D, H) = \frac{p(D|\boldsymbol{\theta}, H)\, p(\boldsymbol{\theta}|H)}{p(D|H)} \propto p(D|\boldsymbol{\theta}, H)$$

$$p(H_j|D) = \frac{p(D|H_j)\, p(H_j)}{p(D)} \propto p(D|H_j)$$

Model evidence

(model H : kernel parameter, e.g. RBF kernel width,

$\boldsymbol{\theta}$: hyperparameters, e.g. regularization parameters)



SISTA

Bayesian inference

[MacKay95, Suykens02, Tipping01]

Bayes' rule $\quad \text{Posterior} = \dfrac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$

Level 1:
infer $\mathbf{w}, b$ for given H, $\boldsymbol{\theta}$

$$p(\mathbf{w}, b \mid D, \boldsymbol{\theta}, H) = \frac{p(D \mid \mathbf{w}, b, \boldsymbol{\theta}, H)\, p(\mathbf{w}, b \mid \boldsymbol{\theta}, H)}{P(D \mid \boldsymbol{\theta}, H)}$$

$$p(\boldsymbol{\theta} \mid D, H) = \frac{p(D \mid \boldsymbol{\theta}, H)\, p(\boldsymbol{\theta} \mid H)}{p(D \mid H)} \propto p(D \mid \boldsymbol{\theta}, H)$$

$$p(H_j \mid D) = \frac{p(D \mid H_j)\, p(H_j)}{p(D)} \propto p(D \mid H_j)$$

$p(D \mid H_j)$ → Model evidence

(model H : kernel parameter, e.g. RBF kernel width,

$\boldsymbol{\theta}$: hyperparameters, e.g. regularization parameters)

Bayesian inference

[MacKay95, Suykens02, Tipping01]

Bayes' rule $\quad \text{Posterior} = \dfrac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$

Level 1: infer $\mathbf{w}, b$ for given H, $\boldsymbol{\theta}$

$$p(\mathbf{w}, b \mid D, \boldsymbol{\theta}, H) = \frac{p(D \mid \mathbf{w}, b, \boldsymbol{\theta}, H)\, p(\mathbf{w}, b \mid \boldsymbol{\theta}, H)}{P(D \mid \boldsymbol{\theta}, H)}$$

Level 2: Infer hyperparameter $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta} \mid D, H) = \frac{p(D \mid \boldsymbol{\theta}, H)\, p(\boldsymbol{\theta} \mid H)}{p(D \mid H)} \propto p(D \mid \boldsymbol{\theta}, H)$$

$$p(H_j \mid D) = \frac{p(D \mid H_j)\, p(H_j)}{p(D)} \propto p(D \mid H_j)$$

→ Model evidence

(model H : kernel parameter, e.g. RBF kernel width,

$\boldsymbol{\theta}$: hyperparameters, e.g. regularization parameters)

Bayesian inference

[MacKay95, Suykens02, Tipping01]

Bayes' rule $\quad \text{Posterior} = \dfrac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$

Level 1: infer $\mathbf{w}, b$ for given H, $\boldsymbol{\theta}$

$$p(\mathbf{w}, b \mid D, \boldsymbol{\theta}, H) = \frac{p(D \mid \mathbf{w}, b, \boldsymbol{\theta}, H)\, p(\mathbf{w}, b \mid \boldsymbol{\theta}, H)}{P(D \mid \boldsymbol{\theta}, H)}$$

Level 2: Infer hyperparameter $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta} \mid D, H) = \frac{p(D \mid \boldsymbol{\theta}, H)\, p(\boldsymbol{\theta} \mid H)}{p(D \mid H)} \propto p(D \mid \boldsymbol{\theta}, H)$$

Level 3: Compare models

$$p(H_j \mid D) = \frac{p(D \mid H_j)\, p(H_j)}{p(D)} \propto p(D \mid H_j)$$

$p(D \mid H_j)$ → Model evidence

(model H : kernel parameter, e.g. RBF kernel width,

$\boldsymbol{\theta}$: hyperparameters, e.g. regularization parameters)

Bayesian inference

[MacKay95, Suykens02, Tipping01]

Bayes' rule $\quad \text{Posterior} = \dfrac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$

Level 1: infer $\mathbf{w}, b$ for given H, $\boldsymbol{\theta}$

$$p(\mathbf{w}, b \mid D, \boldsymbol{\theta}, H) = \frac{p(D \mid \mathbf{w}, b, \boldsymbol{\theta}, H)\, p(\mathbf{w}, b \mid \boldsymbol{\theta}, H)}{P(D \mid \boldsymbol{\theta}, H)}$$

Level 2: Infer hyperparameter $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta} \mid D, H) = \frac{p(D \mid \boldsymbol{\theta}, H)\, p(\boldsymbol{\theta} \mid H)}{p(D \mid H)} \propto p(D \mid \boldsymbol{\theta}, H)$$

Level 3: Compare models

$$p(H_j \mid D) = \frac{p(D \mid H_j)\, p(H_j)}{p(D)} \propto p(D \mid H_j)$$

Marginalization

(Gaussian appr.)

Model evidence

(model H : kernel parameter, e.g. RBF kernel width,

$\boldsymbol{\theta}$: hyperparameters, e.g. regularization parameters)

Sparse Bayesian learning (SBL)

- Automatic relevance determination (ARD) applied to $f(\boldsymbol{x})=\boldsymbol{w}^{T}\boldsymbol{\phi}(\boldsymbol{x})$
- Prior for w_m varies
  hierarchical priors $\Rightarrow$ sparseness

- Basis function $\boldsymbol{\phi}(\boldsymbol{x})$
  - Original variable
  - $\Rightarrow$ linear SBL model
  - $\Rightarrow$ **variable selection!**
  - Kernel $\Rightarrow$ relevance vector machines
  - Relevance vectors: prototypical
- Sequential SBL algorithm [Tipping03]

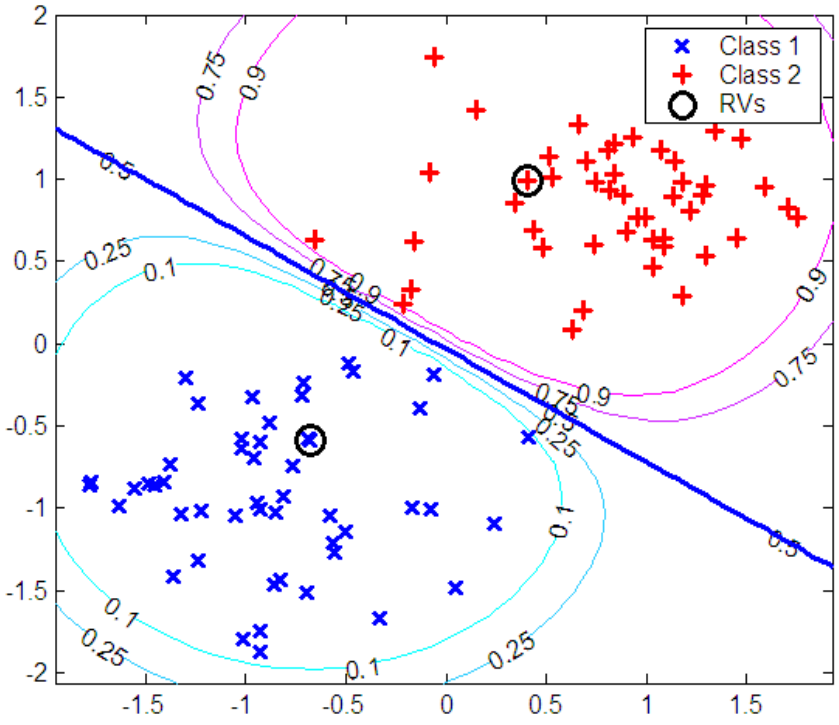


SISTA

Sparse Bayesian learning (SBL)

- Automatic relevance determination (ARD) applied to $f(\boldsymbol{x})=\boldsymbol{w}^{T}\boldsymbol{\phi}(\boldsymbol{x})$
- Prior for w_m varies
  hierarchical priors ⇒ sparseness
- Basis function $\boldsymbol{\phi}(\boldsymbol{x})$
  - Original variable
  ⇒ linear SBL model
  ⇒ **variable selection!**
  - Kernel ⇒
  relevance vector machines
  - Relevance vectors: prototypical
- Sequential SBL algorithm [Tipping03]

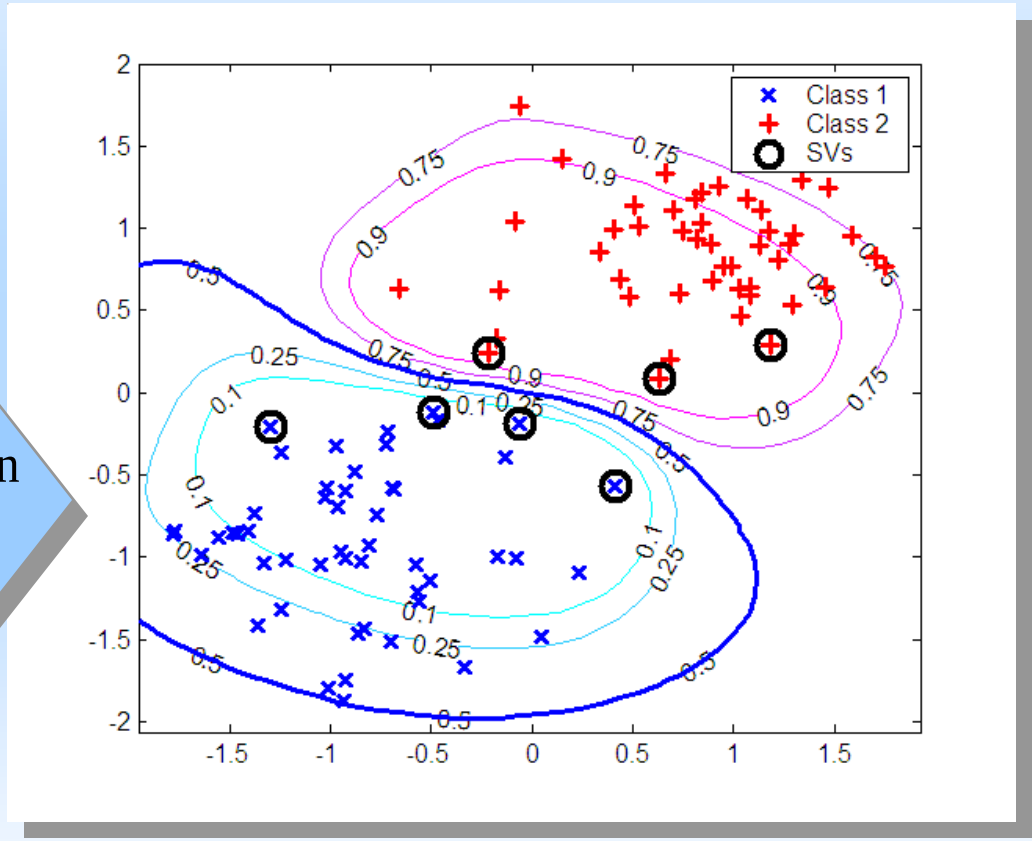
RVM



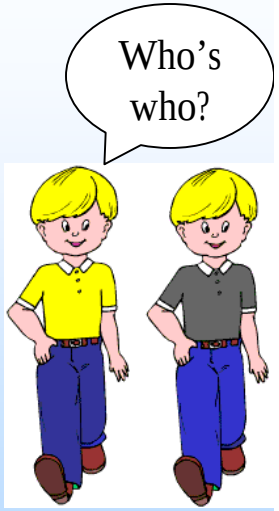
Sparse Bayesian LS-SVMs

- Iteratively pruning of easy cases (support value $\alpha<0$) [Lu02]
- Mimicking margin maximization as in SVM
- Support vectors close to decision boundary

Sparse Bayesian LSSVM



Variable (feature) selection



- Importance in medical classification problems
  - Economics of data acquisition
  - Accuracy and complexity of the classifiers
  - Gain insights into the underlying medical problem
- Filter, wrapper, embedded
- We focus on model evidence based methods within the Bayesian framework [Lu02, Lu04]
  - Forward / stepwise selection
  - Bayesian LS-SVM
  - Sparse Bayesian learning models
  - Accounting for uncertainty in variable selection via sampling methods

Outline

- Supervised learning
- Bayesian frameworks for blackbox models
- Preoperative classification of ovarian tumors
- Bagging for variable selection and prediction in cancer diagnosis problems
- Conclusions

Ovarian cancer diagnosis



- Problem
  - Ovarian masses
    - Ovarian cancer : high mortality rate, difficult early detection
    - Treatment of different types of ovarian tumors differ
  - Develop a reliable diagnostic tool to preoperatively discriminate between malignant and benign tumors.
  - Assist clinicians in choosing the treatment.
- Medical techniques for preoperative evaluation
  - Serum tumor maker: CA125 blood test
  - Ultrasonography
  - Color Doppler imaging and blood flow indexing
- Two-stage study
  - Preliminary investigation: KULeuven pilot project, single-center
  - Extensive study: IOTA project, international multi-center study

Ovarian cancer diagnosis

- Attempts to automate the diagnosis
  - Risk of malignancy Index (RMI) [Jacobs90]
    $RMI = score_{morph} \times score_{meno} \times CA125$
  - Mathematical models

Ovarian cancer diagnosis

- Attempts to automate the diagnosis
  - Risk of malignancy Index (RMI) [Jacobs90]
    $RMI = score_{morph} \times score_{meno} \times CA125$
  - Mathematical models

Logistic Regression
Multilayer perceptrons
Kernel-based models



SISTA

Ovarian cancer diagnosis

- Attempts to automate the diagnosis
  - Risk of malignancy Index (RMI) [Jacobs90]
    $RMI = score_{morph} \times score_{meno} \times CA125$
  - Mathematical models

Logistic Regression
Multilayer perceptrons
Kernel-based models

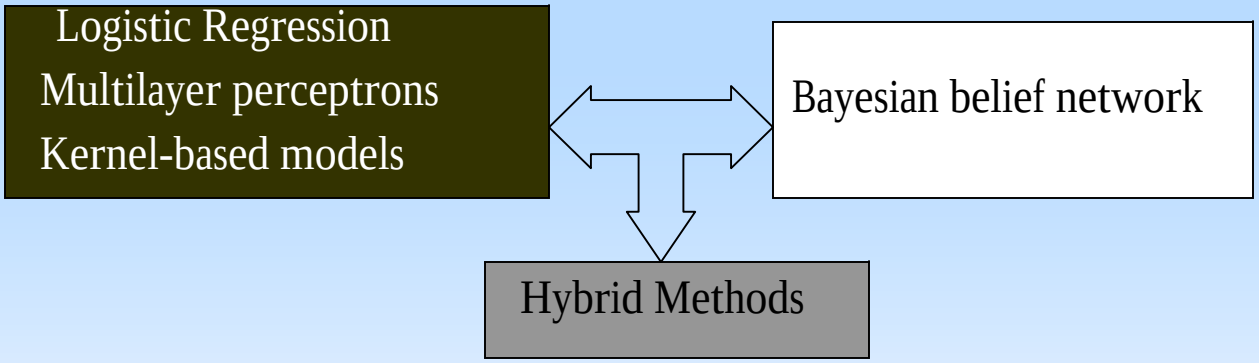
Bayesian belief network



SISTA

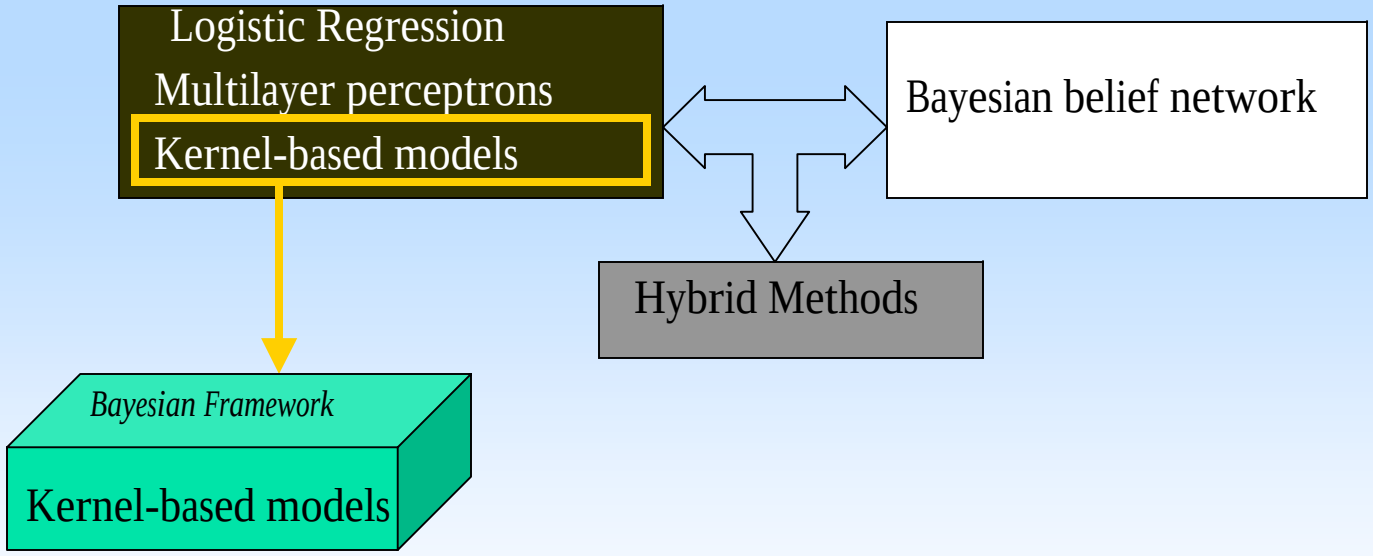
Ovarian cancer diagnosis

- Attempts to automate the diagnosis
  - Risk of malignancy Index (RMI) [Jacobs90]
    $RMI = score_{morph} \times score_{meno} \times CA125$
  - Mathematical models



Ovarian cancer diagnosis

- Attempts to automate the diagnosis
  - Risk of malignancy Index (RMI) [Jacobs90]
    $RMI = score_{morph} \times score_{meno} \times CA125$
  - Mathematical models



Preliminary investigation – pilot project

- Patient data collected at Univ. Hospitals Leuven, Belgium, 1994~1999
- 425 records (data with missing values were excluded), 25 features.
- 291 benign tumors, 134 (32%) malignant tumors
- Preprocessing: e.g.
  - CA_125->log,
  - Color_score {1,2,3,4} -> 3 design variables {0,1}..
- Descriptive statistics



SISTA

Preliminary investigation – pilot project

Demographic, serum marker, color Doppler imaging and morphologic variables

	Variable (symbol)	Benign	Malignant
Demographic	Age (*age*)	45.6 ± 15.2	56.9 ± 14.6
	Postmenopausal (*meno*)	31.0 %	66.0 %
Serum marker	CA 125 (log) (*l_ca125*)	3.0 ± 1.2	5.2 ± 1.5
CDI	High blood flow (*colsc3,4*)	19.0%	77.3 %
Morphologic	Abdominal fluid (*asc*)	32.7 %	67.3 %
	Bilateral mass (*bilat*)	13.3 %	39.0 %
	Unilocular cyst (*un*)	45.8 %	5.0 %
	Multiloc/solid cyst (*mulsol*)	10.7 %	36.2 %
	Solid (*sol*)	8.3 %	37.6 %
	Smooth wall (*smooth*)	56.8 %	5.7 %
	Irregular wall (*irreg*)	33.8 %	73.2 %
	Papillations (*pap*)	12.5 %	53.2 %

Experiment – pilot project

- Desired property for models:
  - **Probability of malignancy**
  - High sensitivity for malign.
    $\leftrightarrow$ low false positive rate.
- Compared models
  - Bayesian LS-SVM classifiers
  - RVM classifiers
  - Bayesian MLPs
  - Logistic regression
  - RMI (reference)
- 'Temporal' cross-validation
  - Training set: 265 data (1994~1997)
  - Test set: 160 data (1997~1999)
- Multiple runs of stratified randomized CV
  - Improved test performance
  - Conclusions for model comparison similar to temporal CV



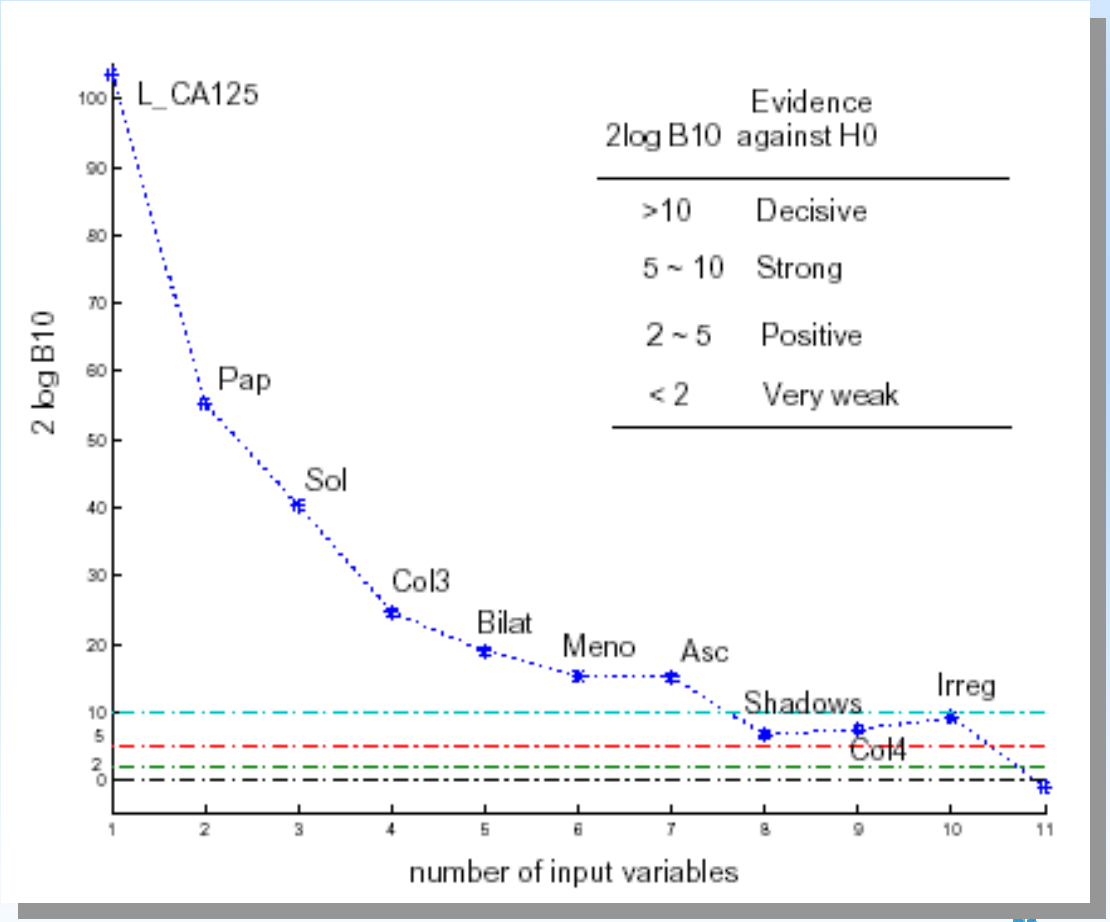
SISTA

Variable selection – pilot project

- Forward variable selection based on Bayesian LS-SVM

10 variables were selected based on the training set (first treated 265 patient data) using RBF kernels.

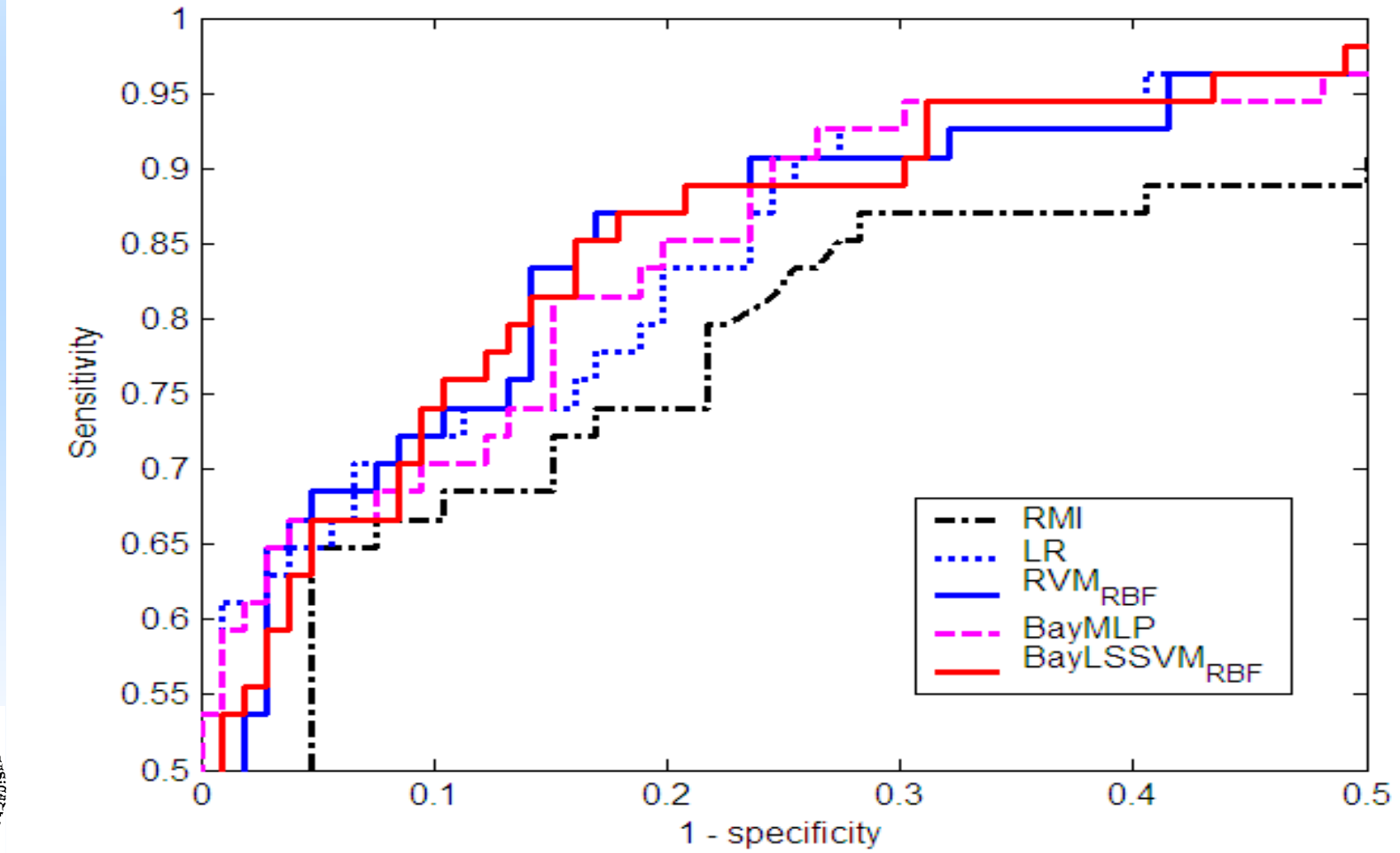
Evolution of the model evidence



Model evaluation – *pilot project*

- Compare the predictive power of the models given the selected variables

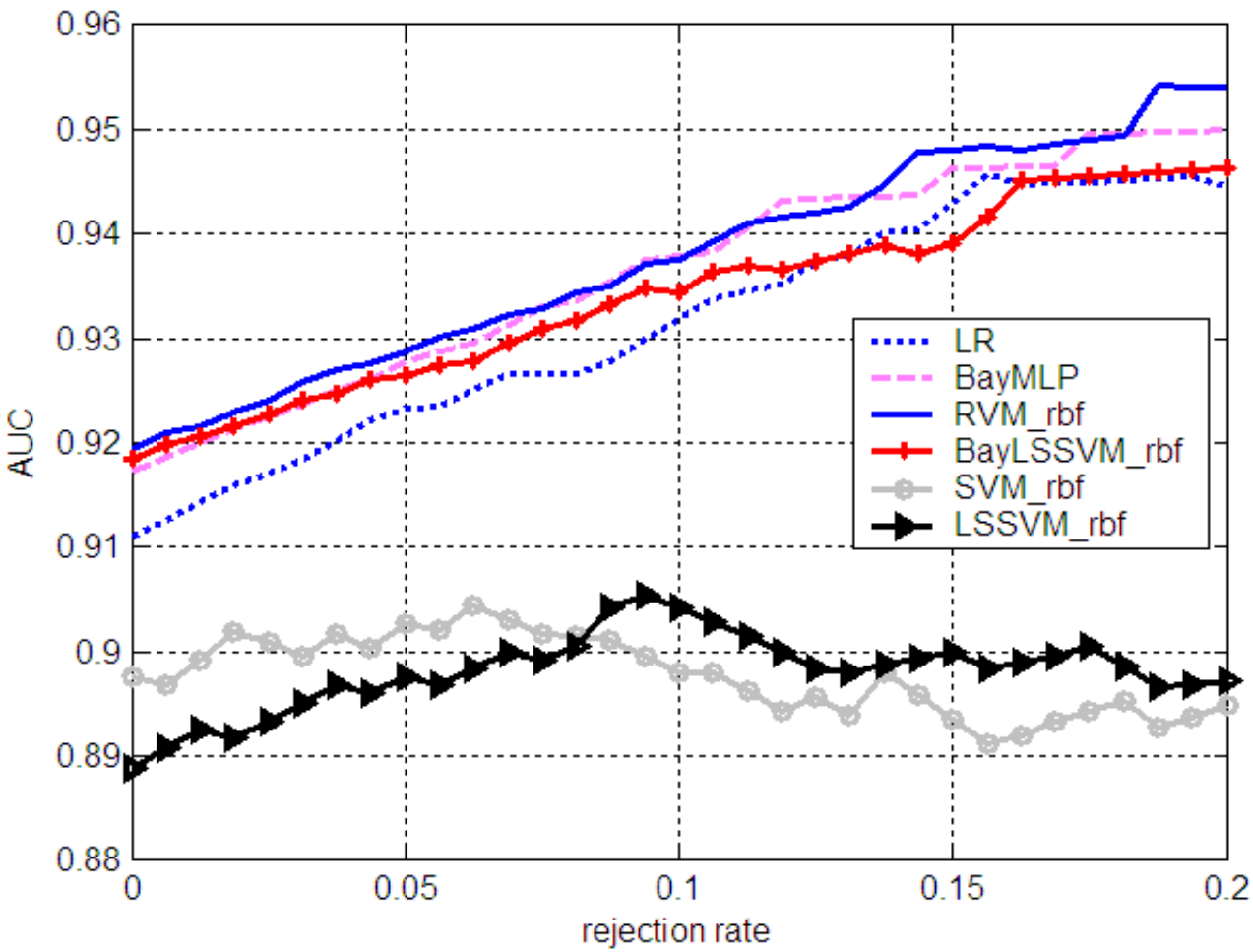
ROC curves on test Set (data from 160 newest treated patients)



Model evaluation – pilot project

- Comparison of model performance on test set with rejection based on

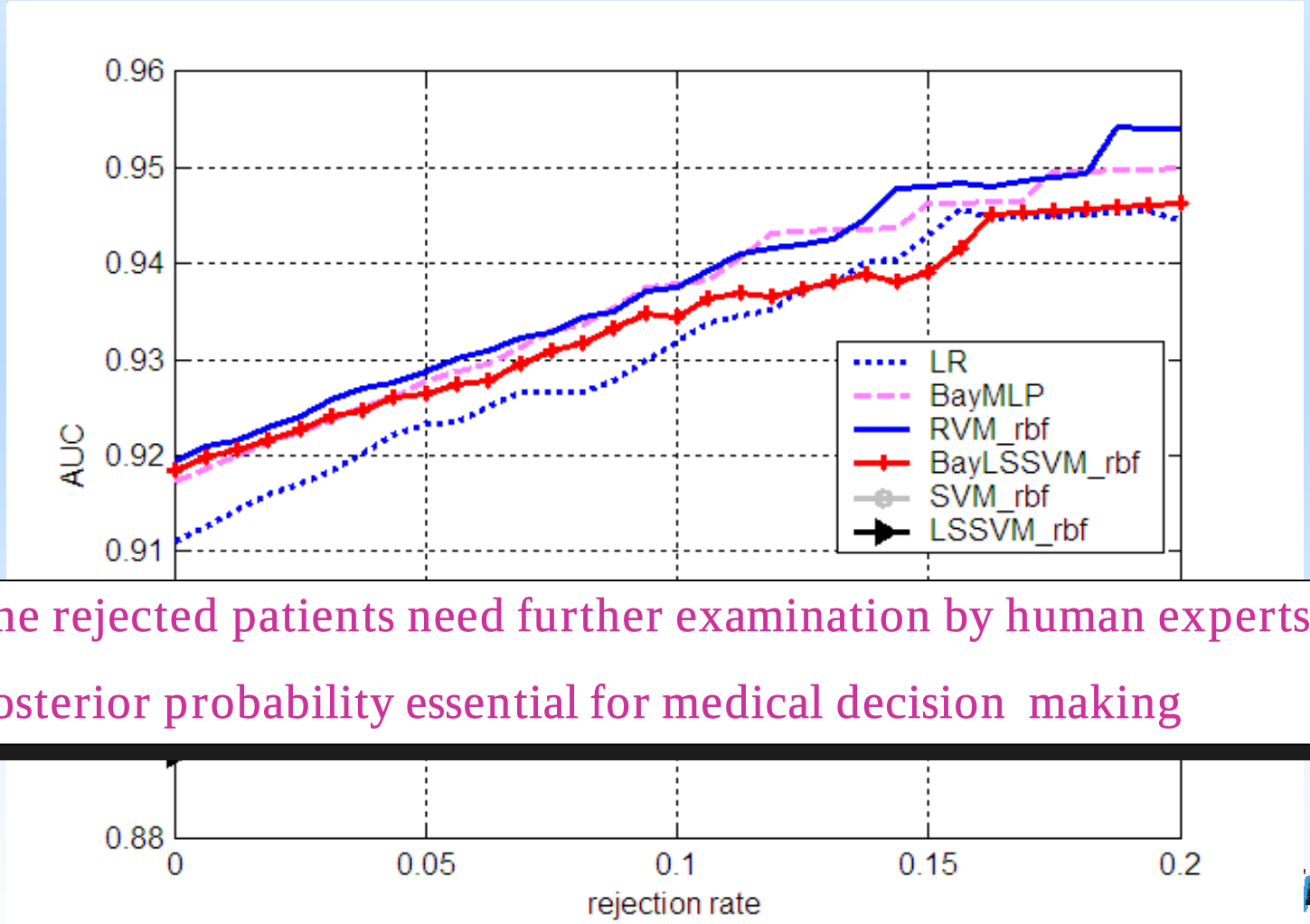
$$|P(y=+1|\mathbf{x}) - 0.5| \propto \text{uncertainty}$$



Model evaluation – pilot project

- Comparison of model performance on test set with rejection based on

$$|P(y=+1|\mathbf{x}) - 0.5| \propto \text{uncertainty}$$



- The rejected patients need further examination by human experts
- Posterior probability essential for medical decision making

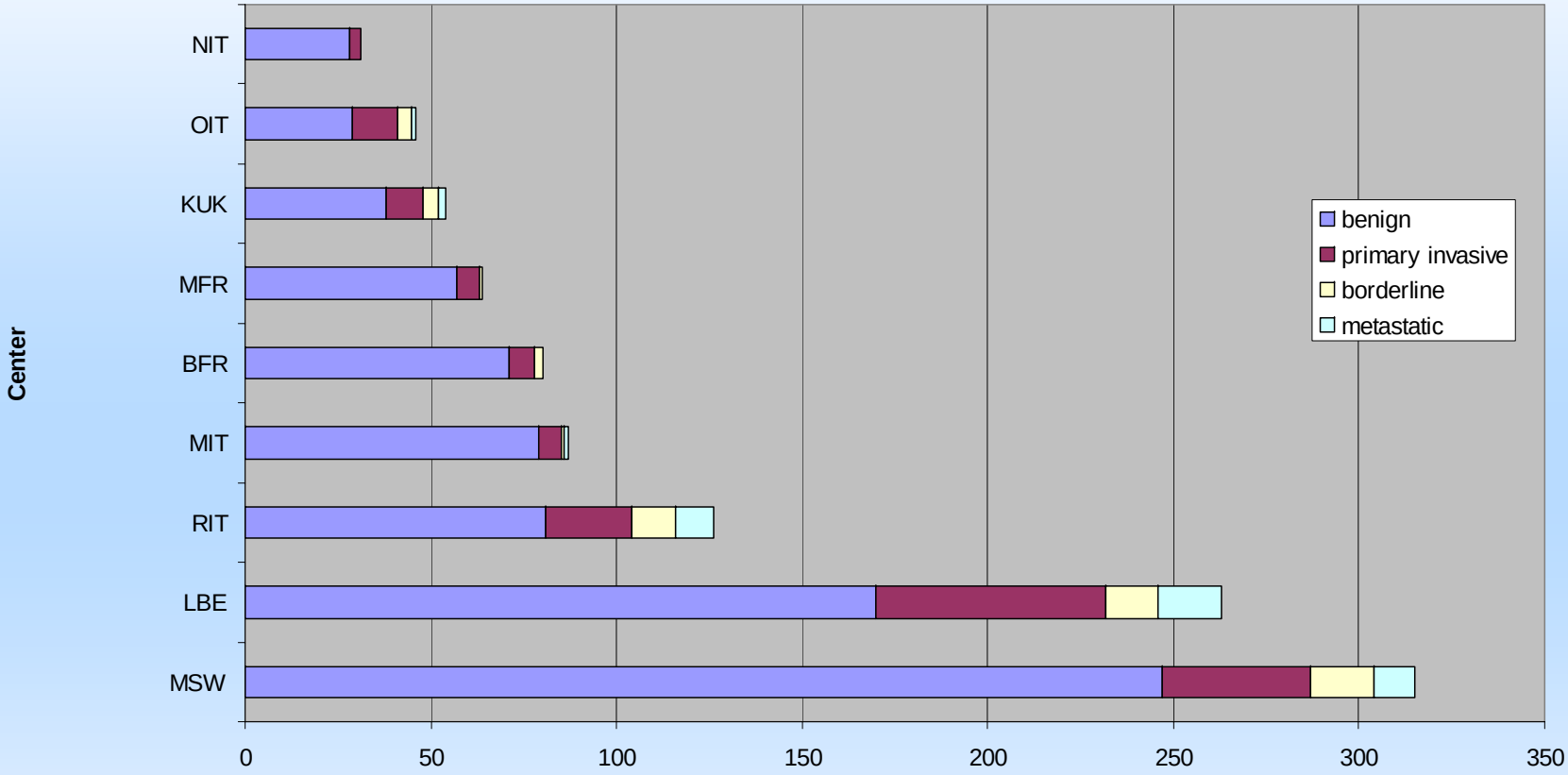


ISTA

Extensive study – IOTA project

- International Ovarian Tumor Analysis
- Protocol for data collection
- A multi-center study
  - 9 centers
  - 5 countries: Sweden, Belgium, Italy, France, UK
- 1066 data of the dominant tumors
  - 800 (75%) benign
  - 266 (25%) malignant
  - About 60 variables after preprocessing

Data – IOTA project



	MSW	LBE	RIT	MIT	BFR	MFR	KUK	OIT	NIT
metastatic	11	17	10	1	0	0	2	1	0
borderline	17	14	12	1	2	1	4	4	0
primary invasive	40	62	23	6	7	6	10	12	3
benign	247	170	81	79	71	57	38	29	28

Number of data

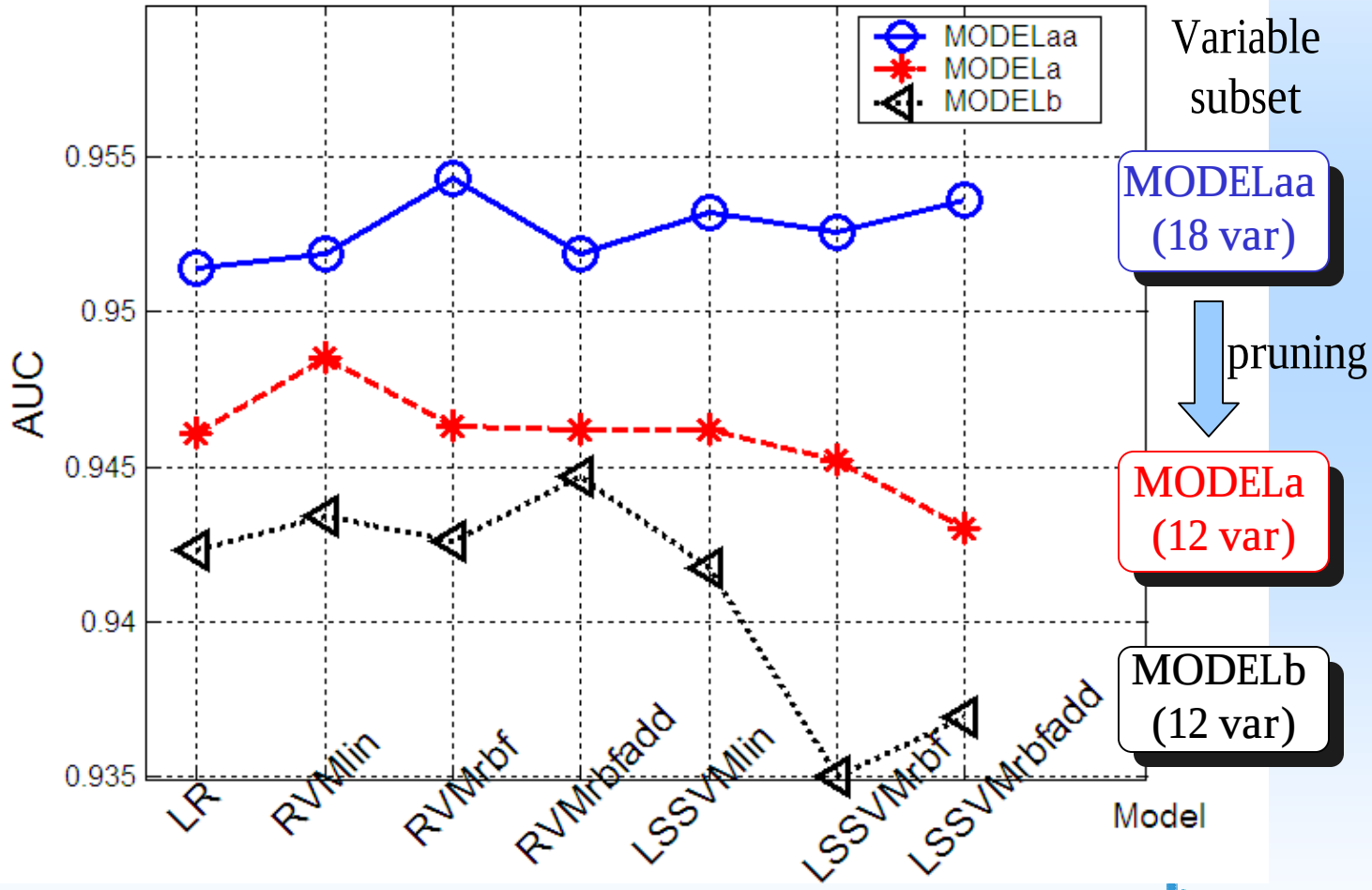
Model development – IOTA project

- Randomly divide data into
  - Training set: N_{train}=754
    Test set: N_{test}=312
  - Stratified for tumor types and centers
- Model building based on the training data
  - Variable selection:
    - with / without CA125
    - Bayesian LS-SVM with linear/RBF kernels
  - Compared models:
    - LRs
    - Bay LS-SVMs, RVMs,
    - Kernels: linear/RB, additive RBF
- Model evaluation
  - ROC analysis
  - Performance of all centers as a whole / of individual centers
  - Model interpretation?

Model evaluation – IOTA project

Comparison of model performance using different variable subsets

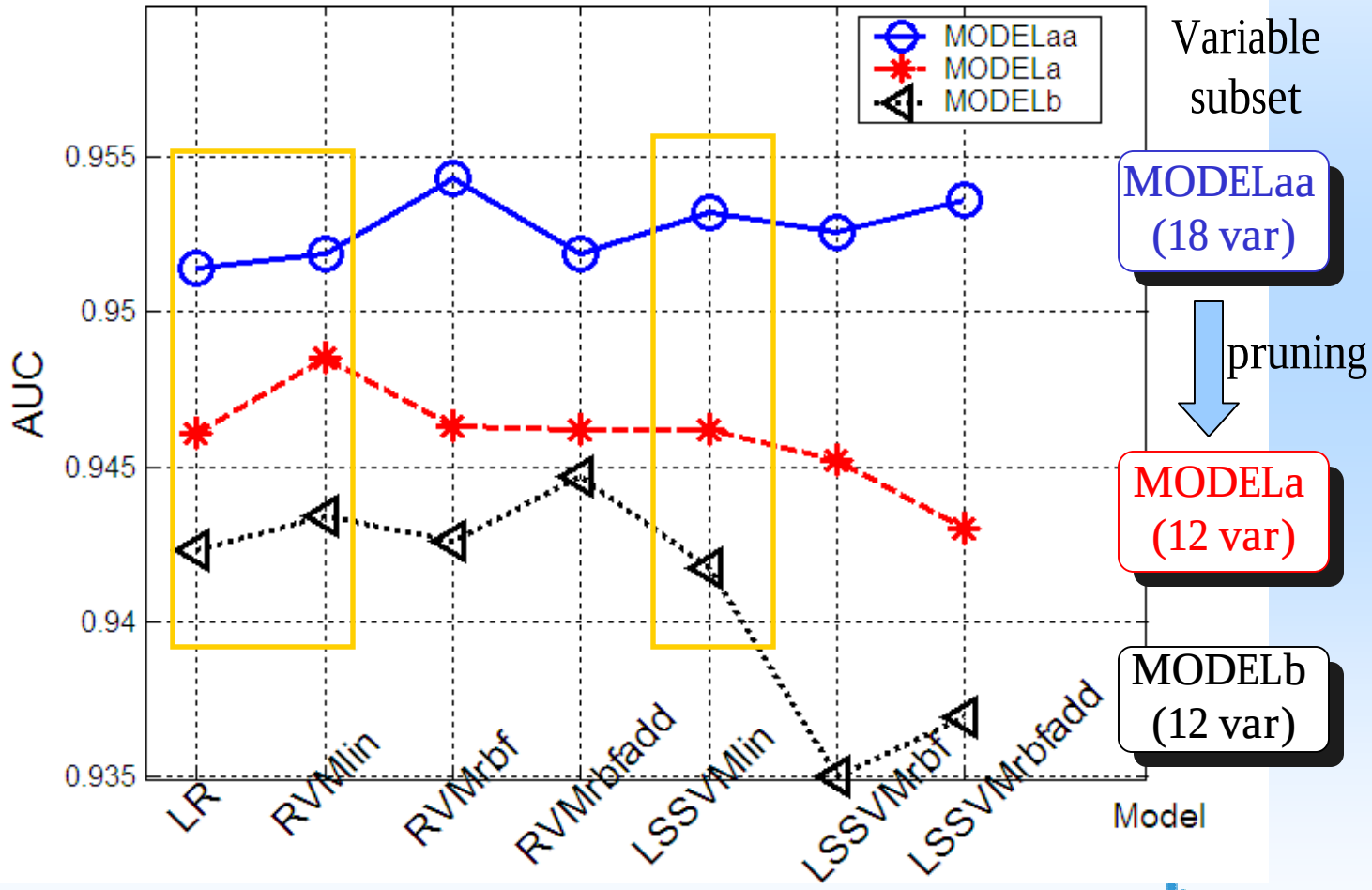
- Variable subset matters more than model type
- Linear models suffice



Model evaluation – IOTA project

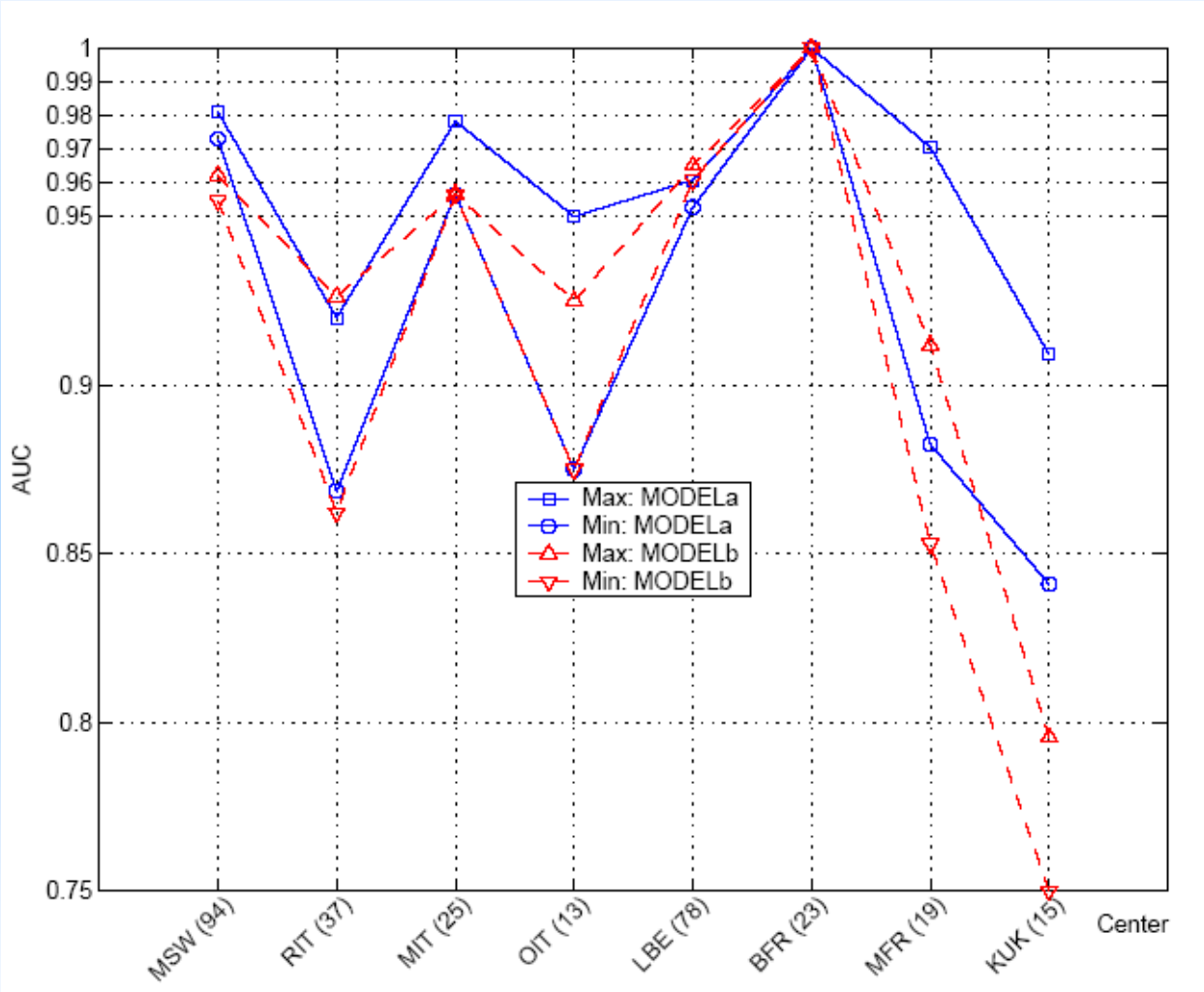
Comparison of model performance using different variable subsets

- Variable subset matters more than model type
- Linear models suffice



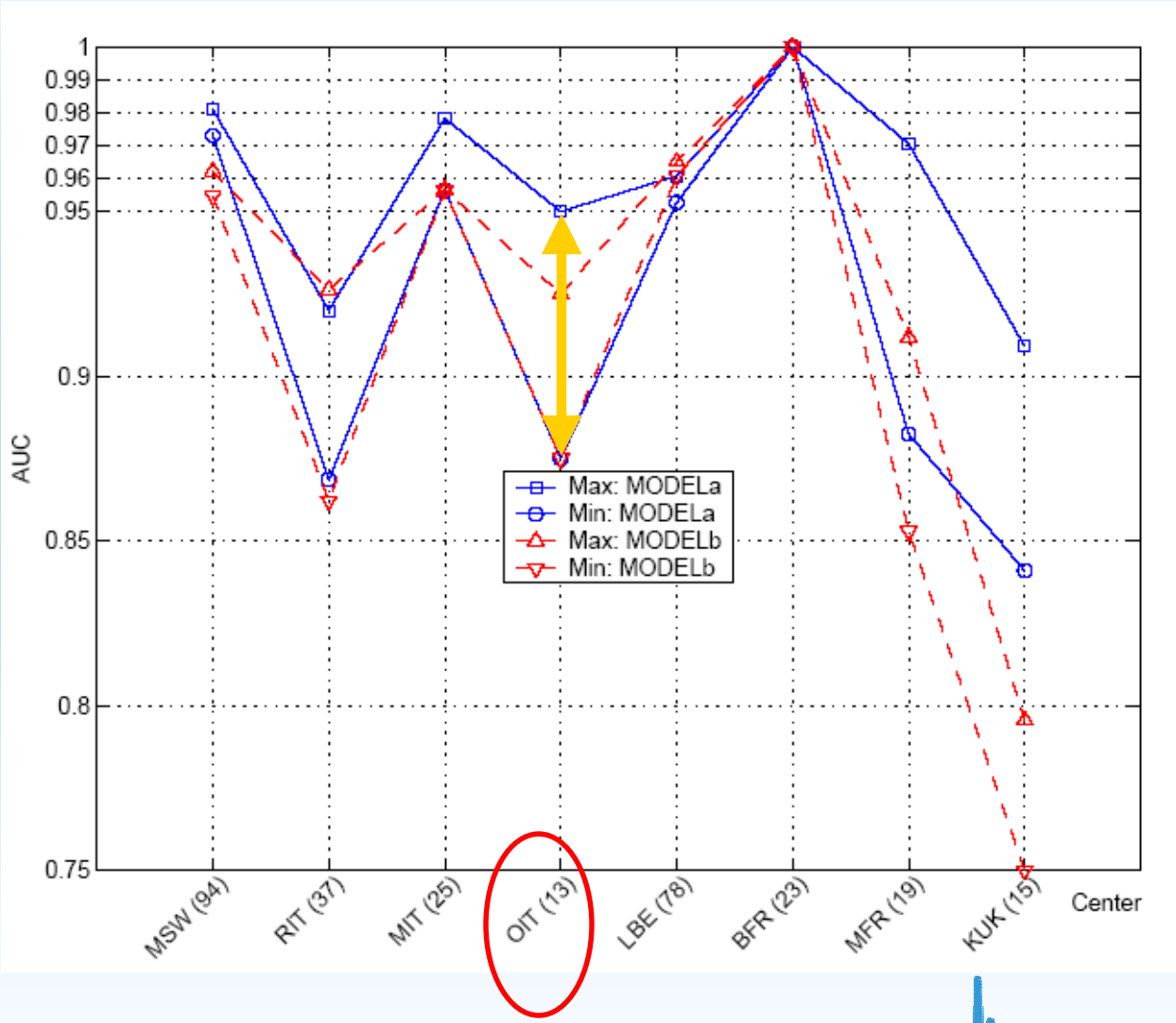
Test in different centers – IOTA project

- Comparison of model performance in different centers using MODELa and MODELb
- AUC range among the various models ~ related to the test set size of the center.
- MODELa performs slightly better than MODELb, but not significant



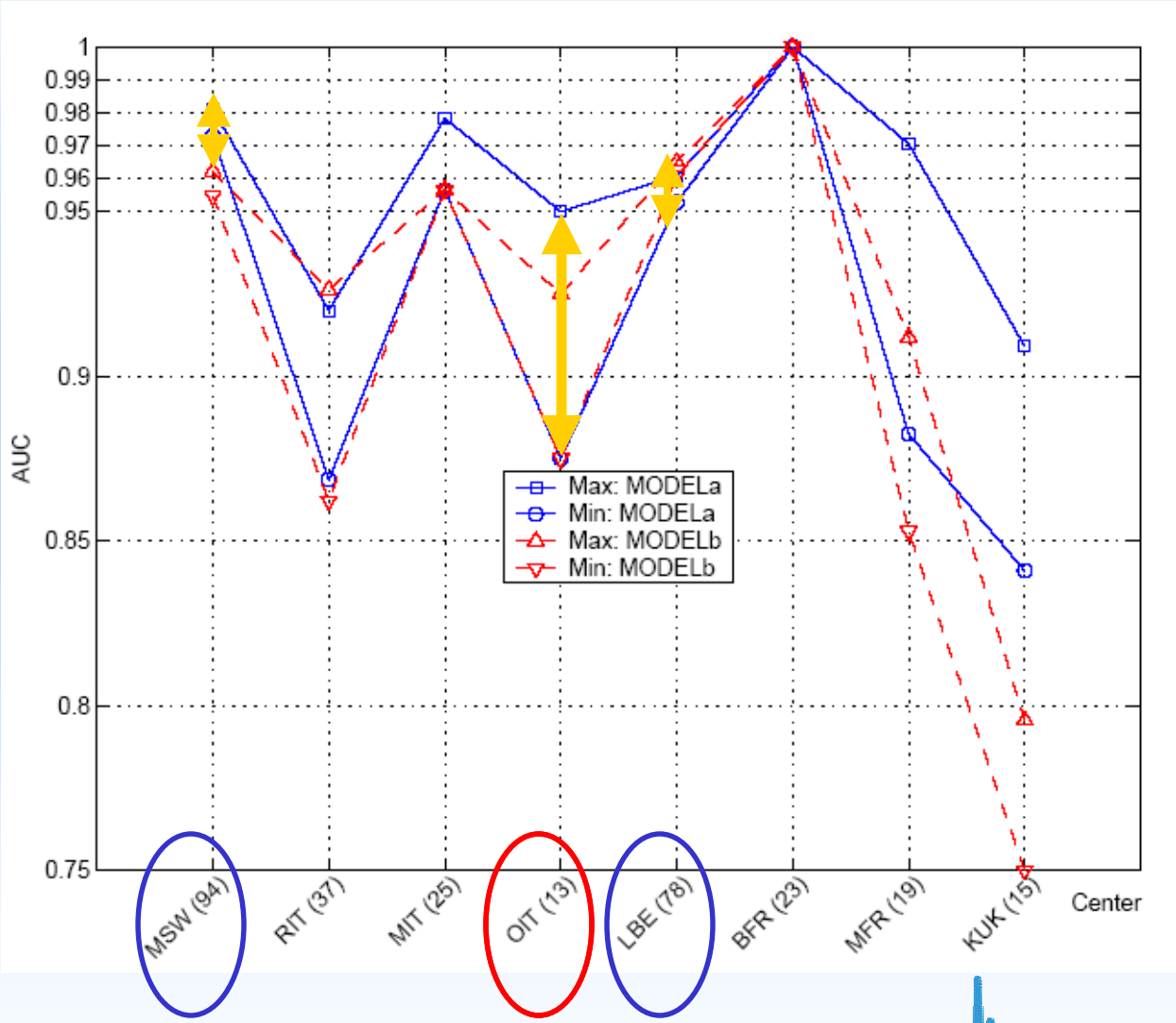
Test in different centers – IOTA project

- Comparison of model performance in different centers using MODELa and MODELb
- AUC range among the various models ~ related to the test set size of the center.
- MODELa performs slightly better than MODELb, but not significant



Test in different centers – IOTA project

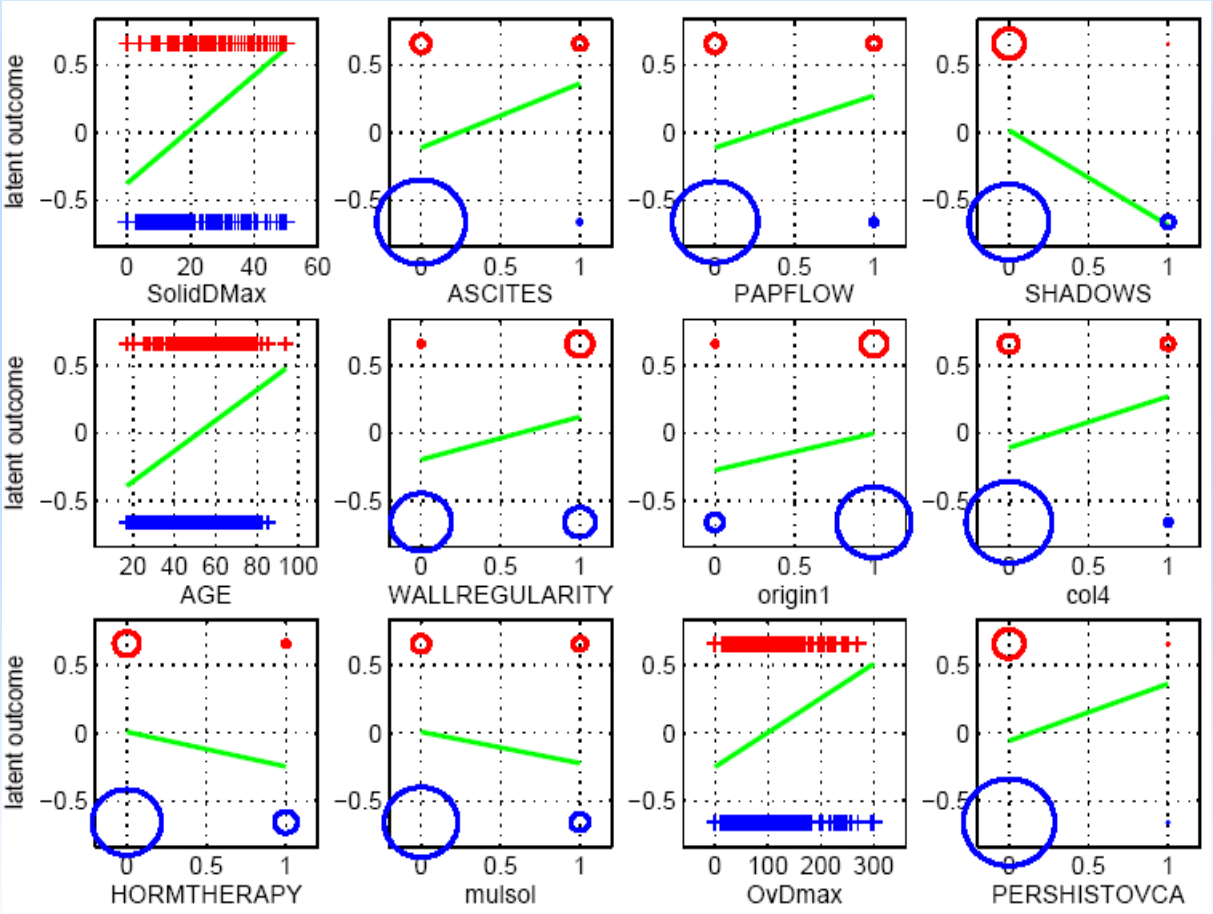
- Comparison of model performance in different centers using MODELa and MODELb
- AUC range among the various models ~ related to the test set size of the center.
- MODELa performs slightly better than MODELb, but not significant

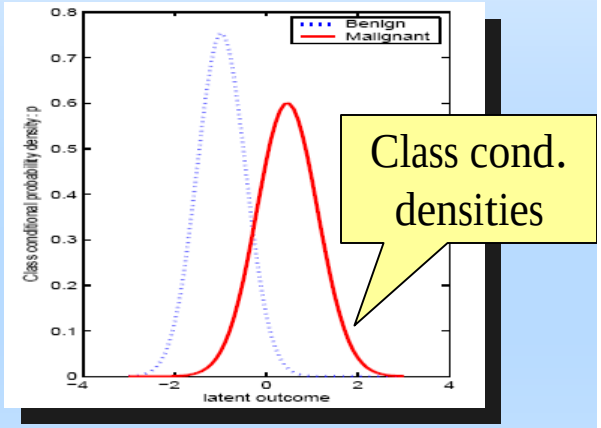


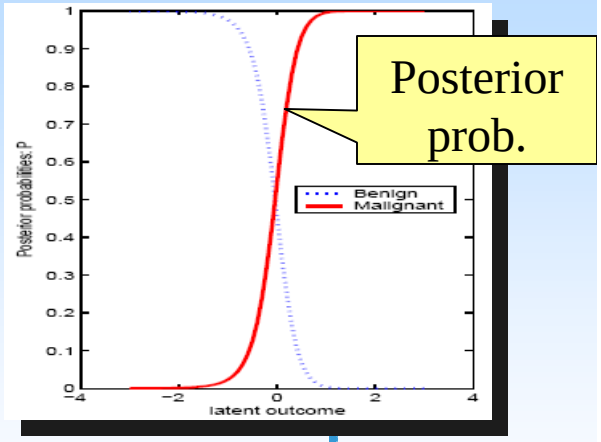
Model visualization – IOTA project

Model fitted using 754 training data. 12 Var from MODELa.

Bayesian LS-SVM with linear kernels



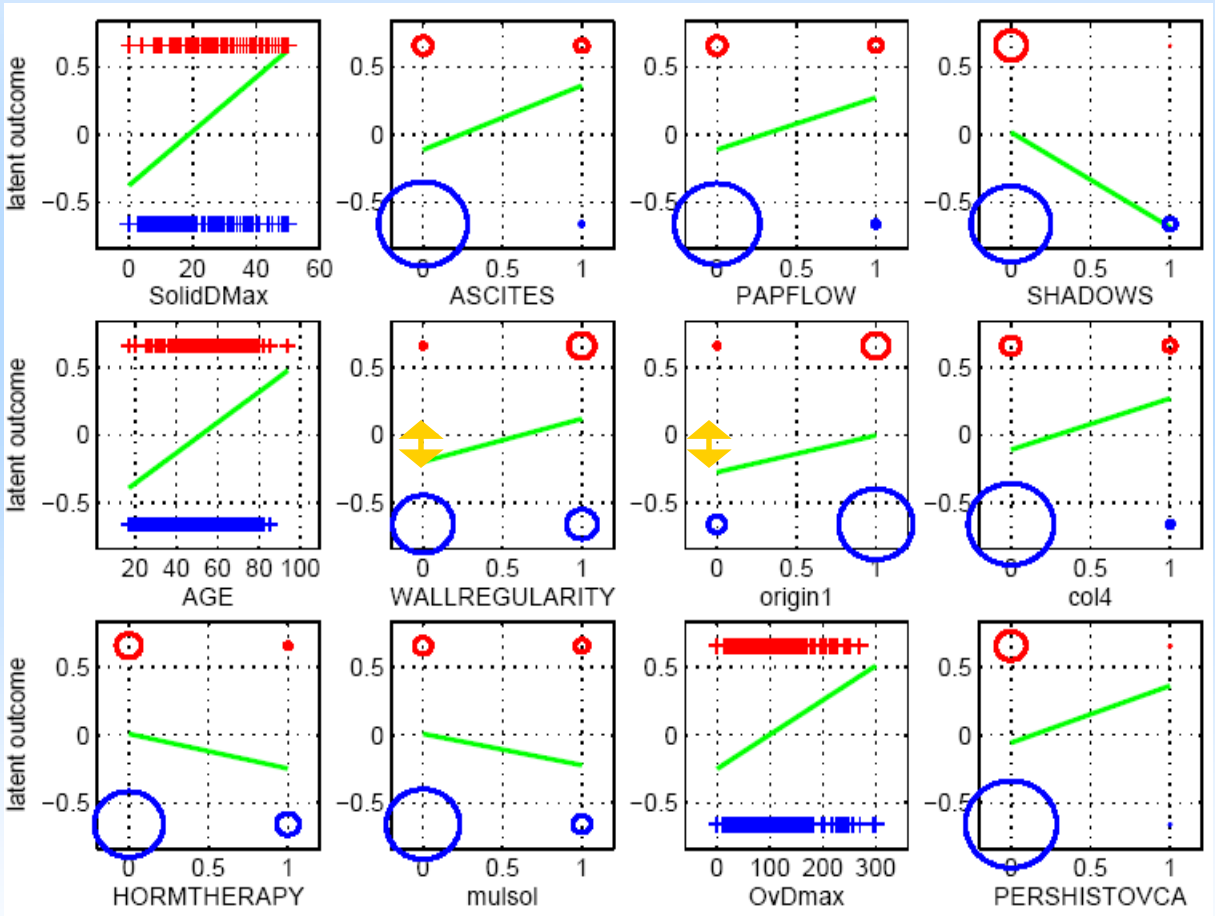


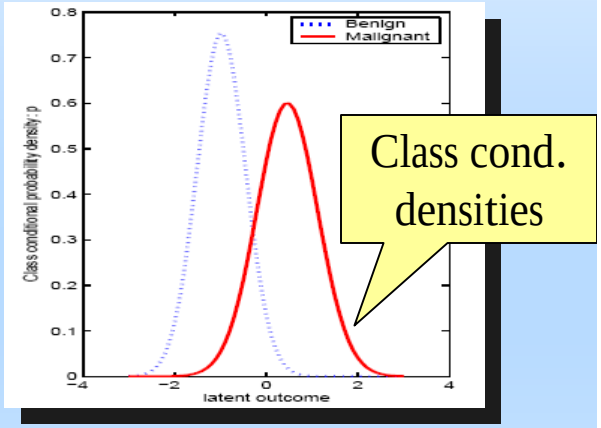


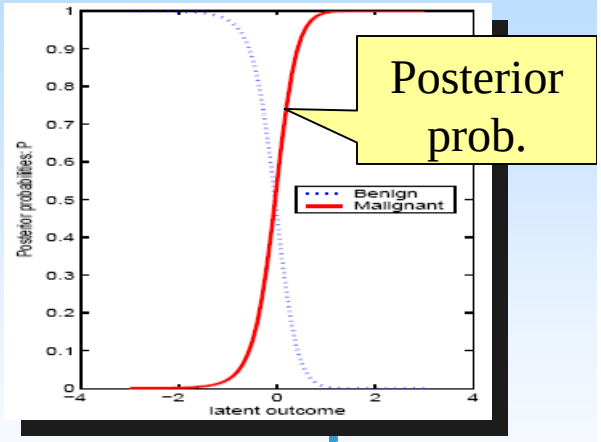
Model visualization – IOTA project

Model fitted using 754 training data. 12 Var from MODELa.

Bayesian LS-SVM with linear kernels



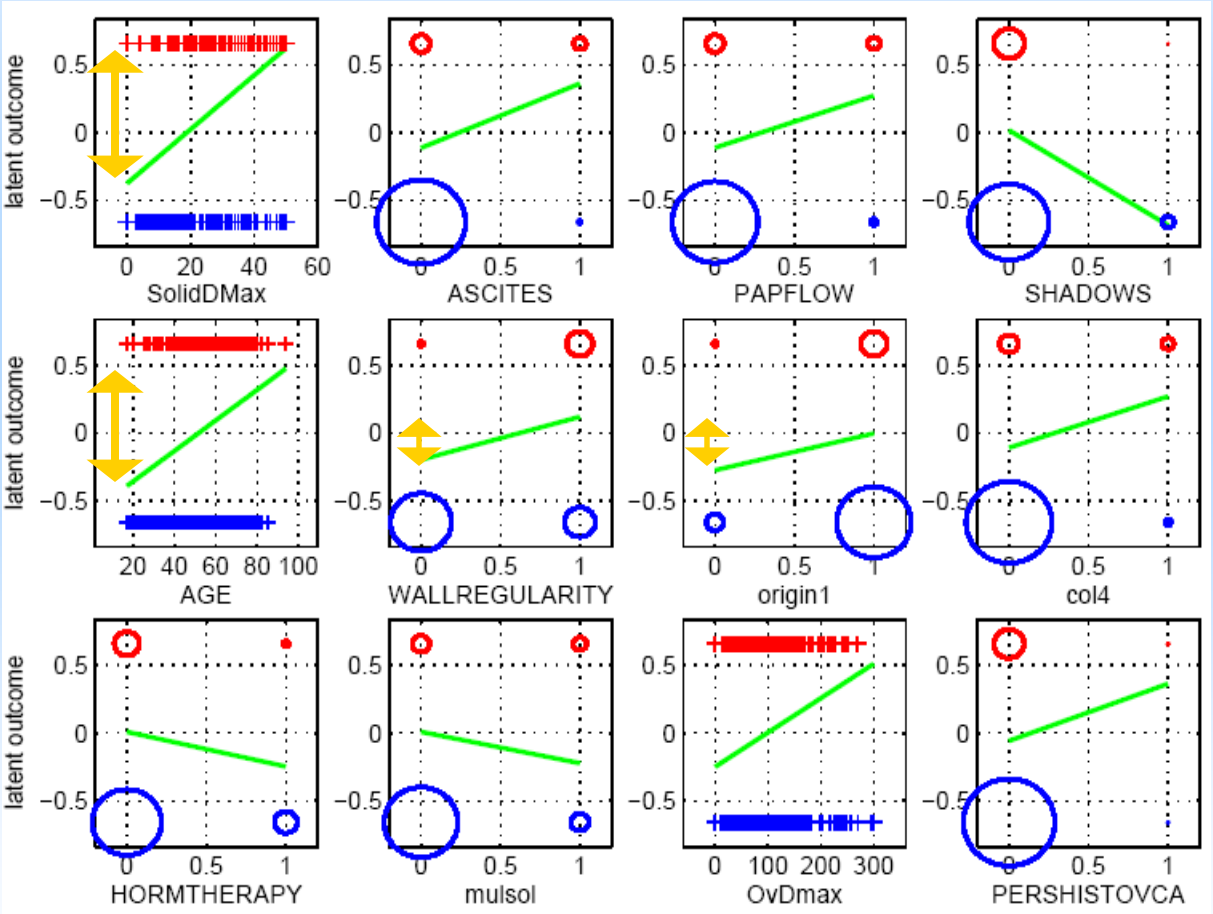


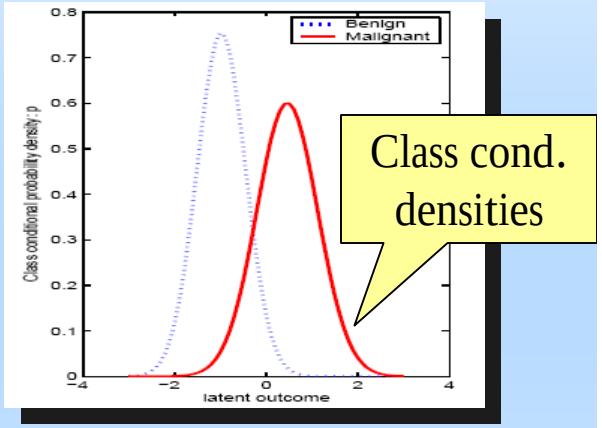


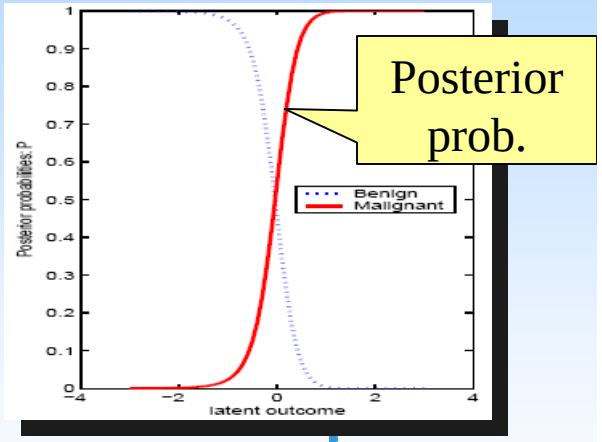
Model visualization – IOTA project

Model fitted using 754 training data. 12 Var from MODELa.

Bayesian LS-SVM with linear kernels





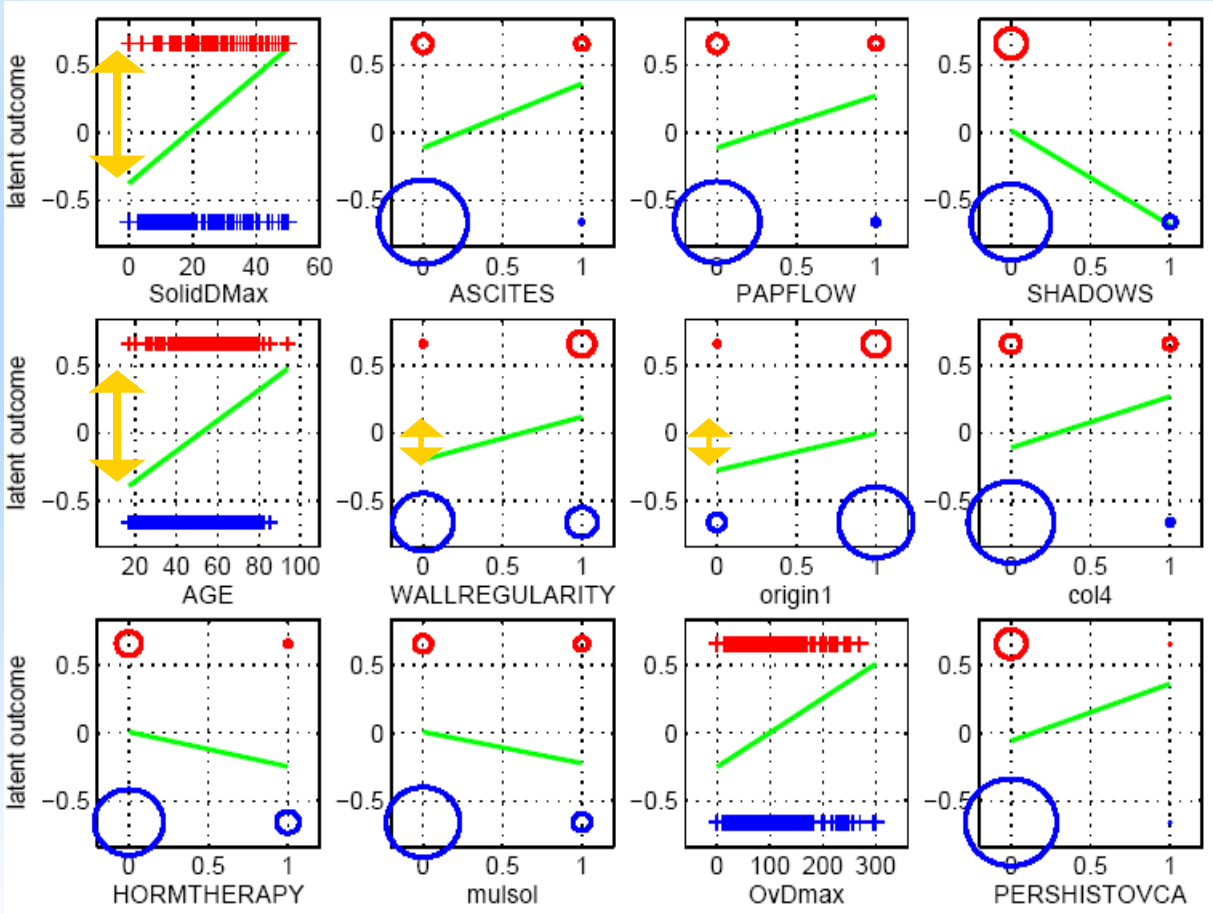


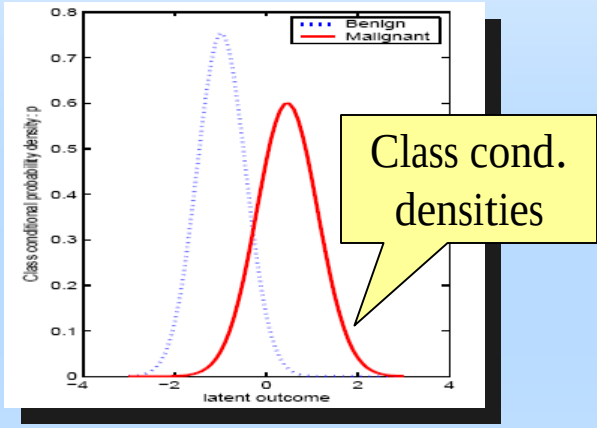
Model visualization – IOTA pro

Test AUC: 0.946
Sensitivity: 85.3%
Specificity: 89.5%

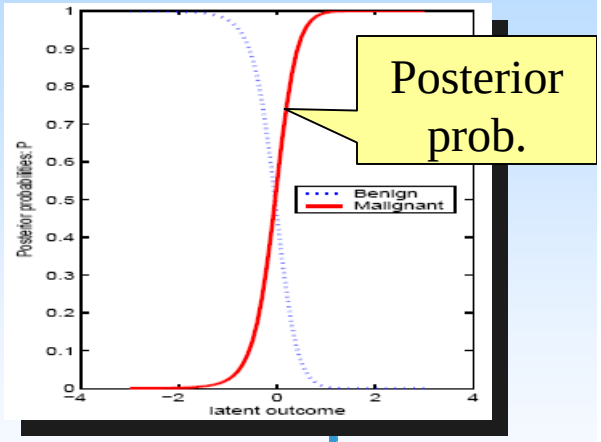
Model fitted using 754 training data. 12 Var from MODELa.

Bayesian LS-SVM with linear kernels





Class cond. densities



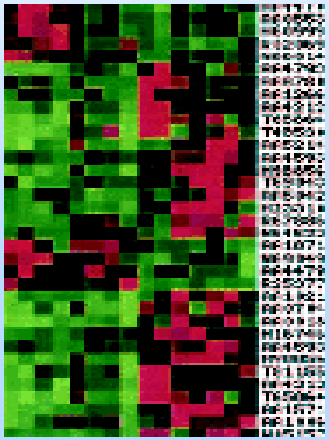
Posterior prob.

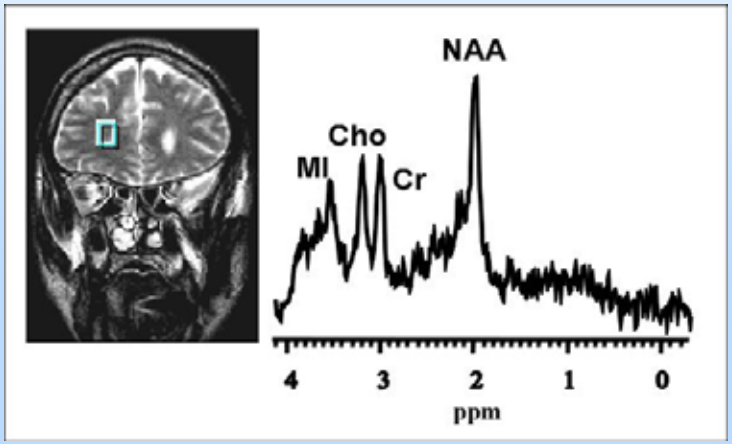
Outline

- Supervised learning
- Bayesian frameworks for blackbox models
- Preoperative classification of ovarian tumors
- Bagging for variable selection and prediction in cancer diagnosis problems
- Conclusions

Bagging linear SBL models for variable selection in cancer diagnosis

- Microarrays and magnetic resonance spectroscopy (MRS)



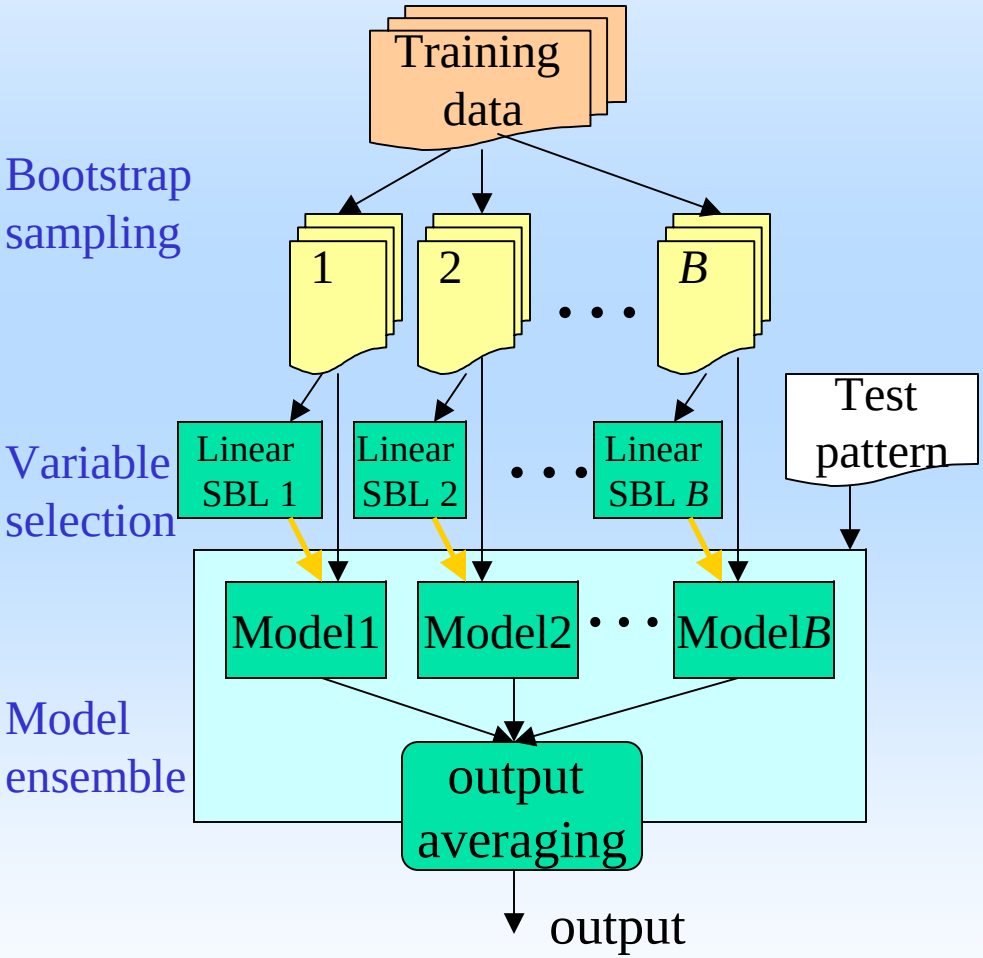


- High dimensionality vs. small sample size
- Data are noisy
- Sequential sparse Bayesian learning algorithm based on logit models (no kernel) as basic variable selection method:

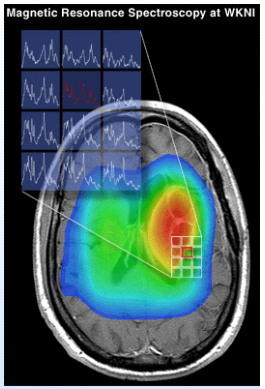
  unstable, multiple solutions => How to **stabilize the procedure**?

Bagging strategy

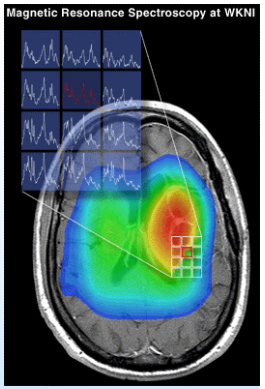
- Bagging: bootstrap + aggregate



Brain tumor classification

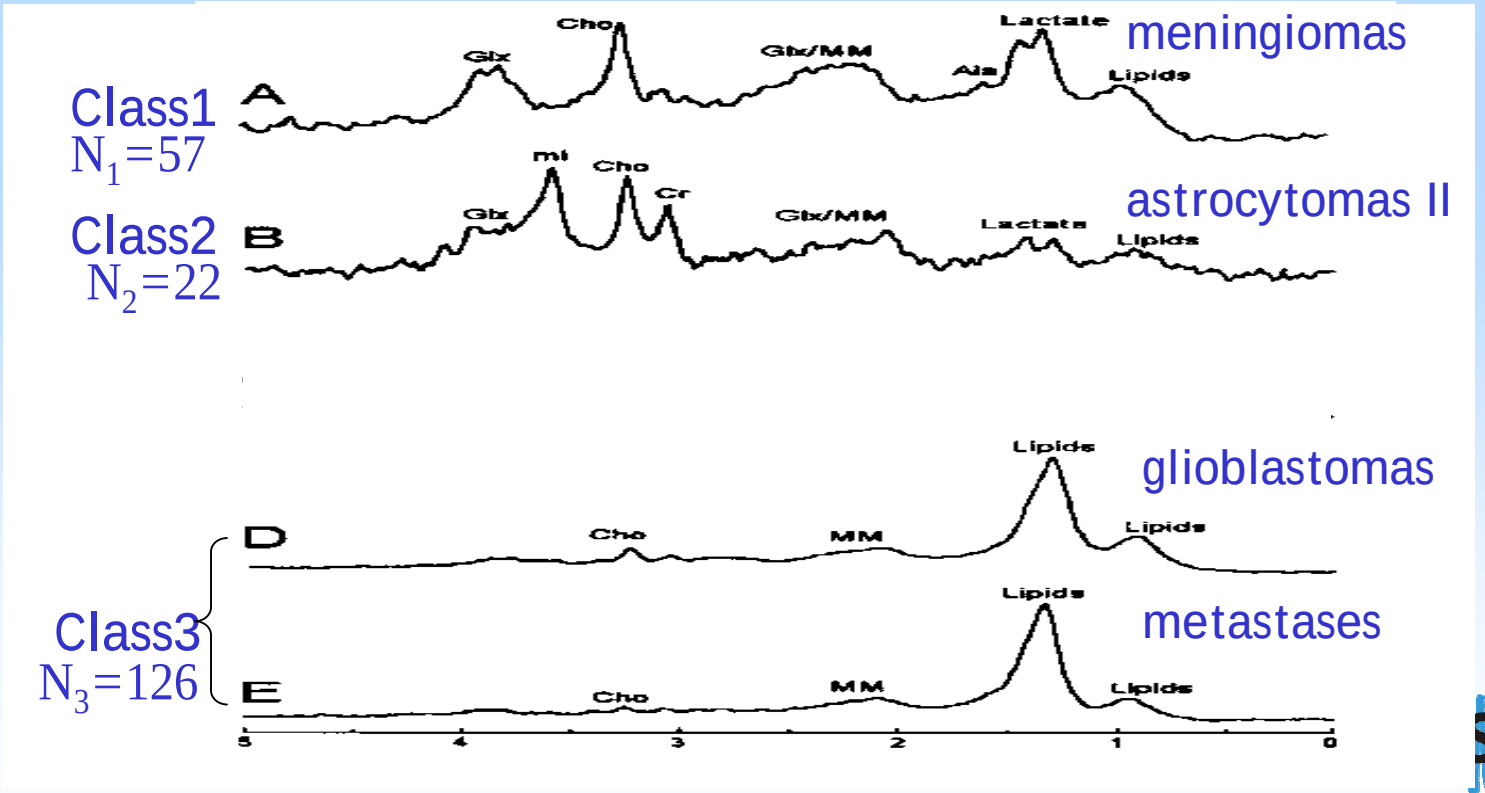


- Based on the ^{1}H short echo magnetic resonance spectroscopy (MRS) spectra data
- 205×138 L2 normalized magnitude values in frequency domain
- 3 classes of brain tumors

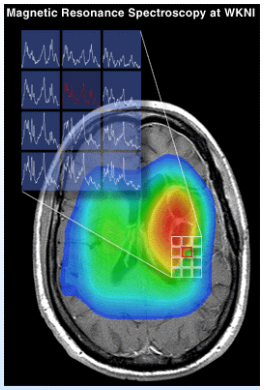


Brain tumor classification

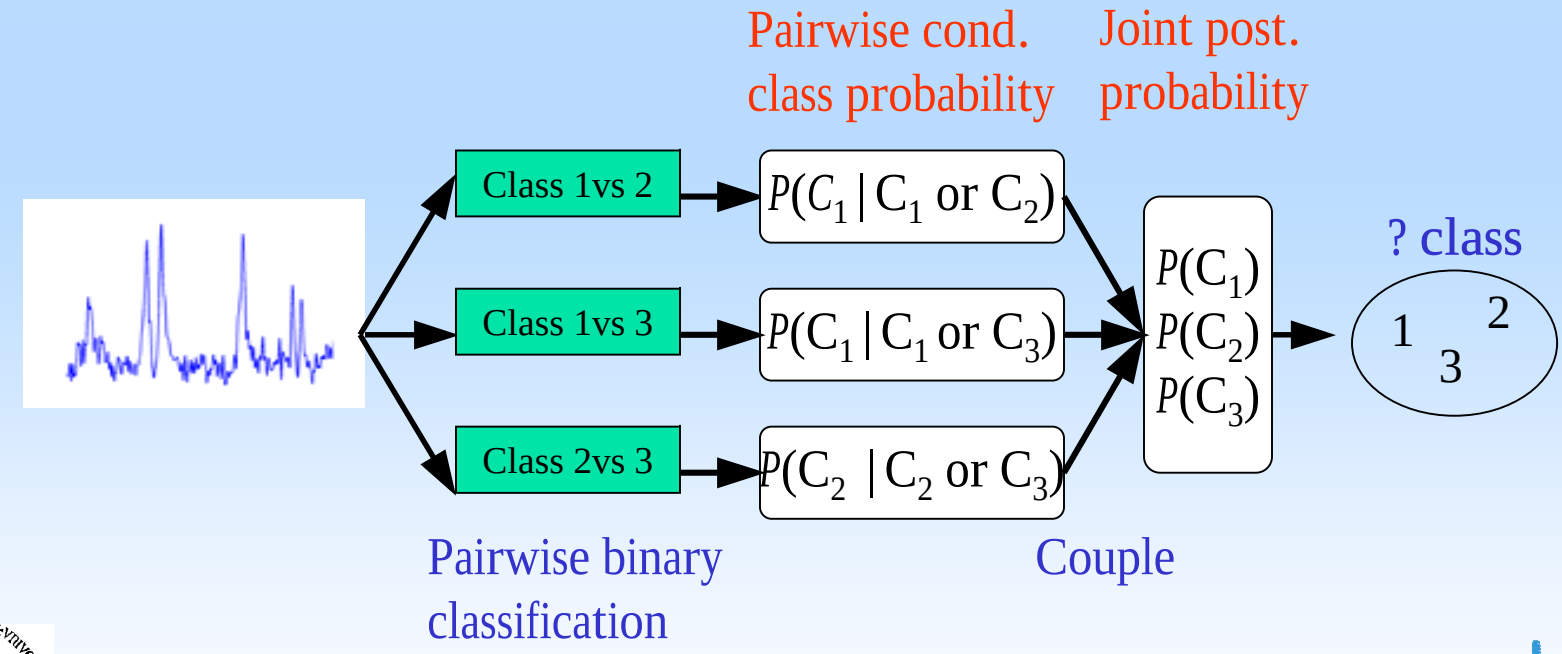
- Based on the ^{1}H short echo magnetic resonance spectroscopy (MRS) spectra data
- 205×138 L2 normalized magnitude values in frequency domain
- 3 classes of brain tumors



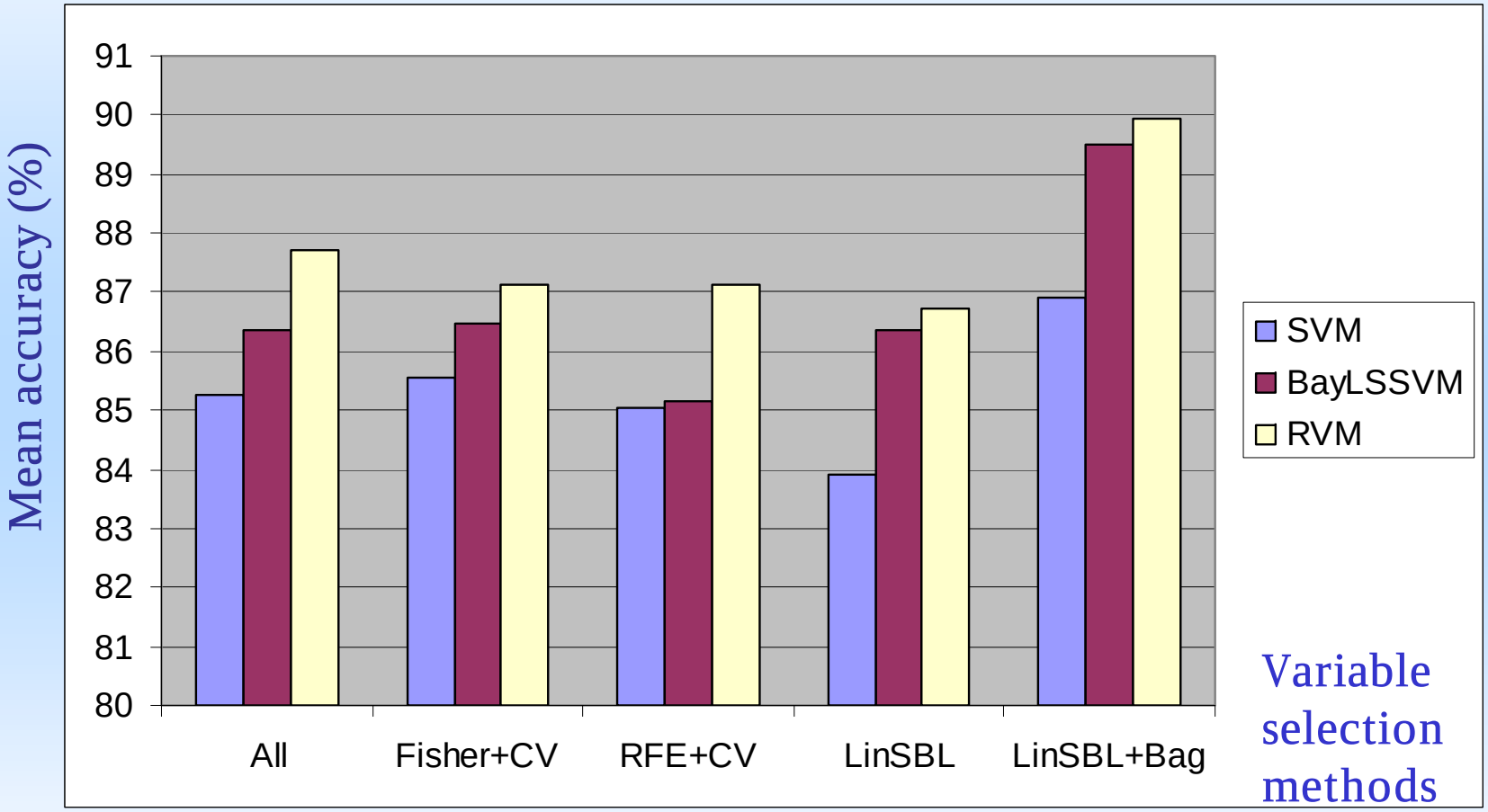
Brain tumor classification



- Based on the ^{1}H short echo magnetic resonance spectroscopy (MRS) spectra data
- 205×138 L2 normalized magnitude values in frequency domain
- 3 classes of brain tumors

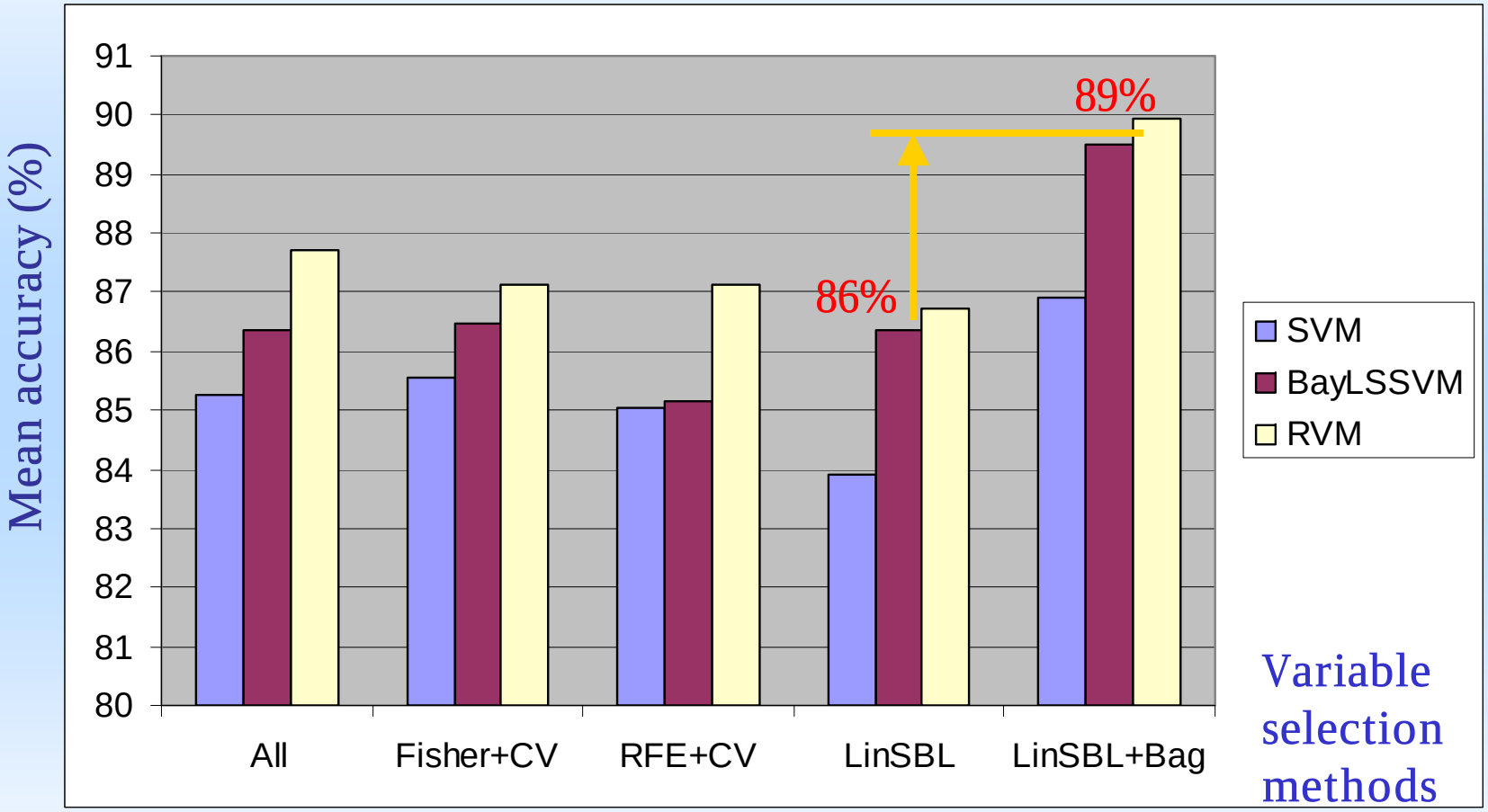


Brain tumor multiclass classification based on MRS spectra data



Mean accuracy from 30 runs of CV

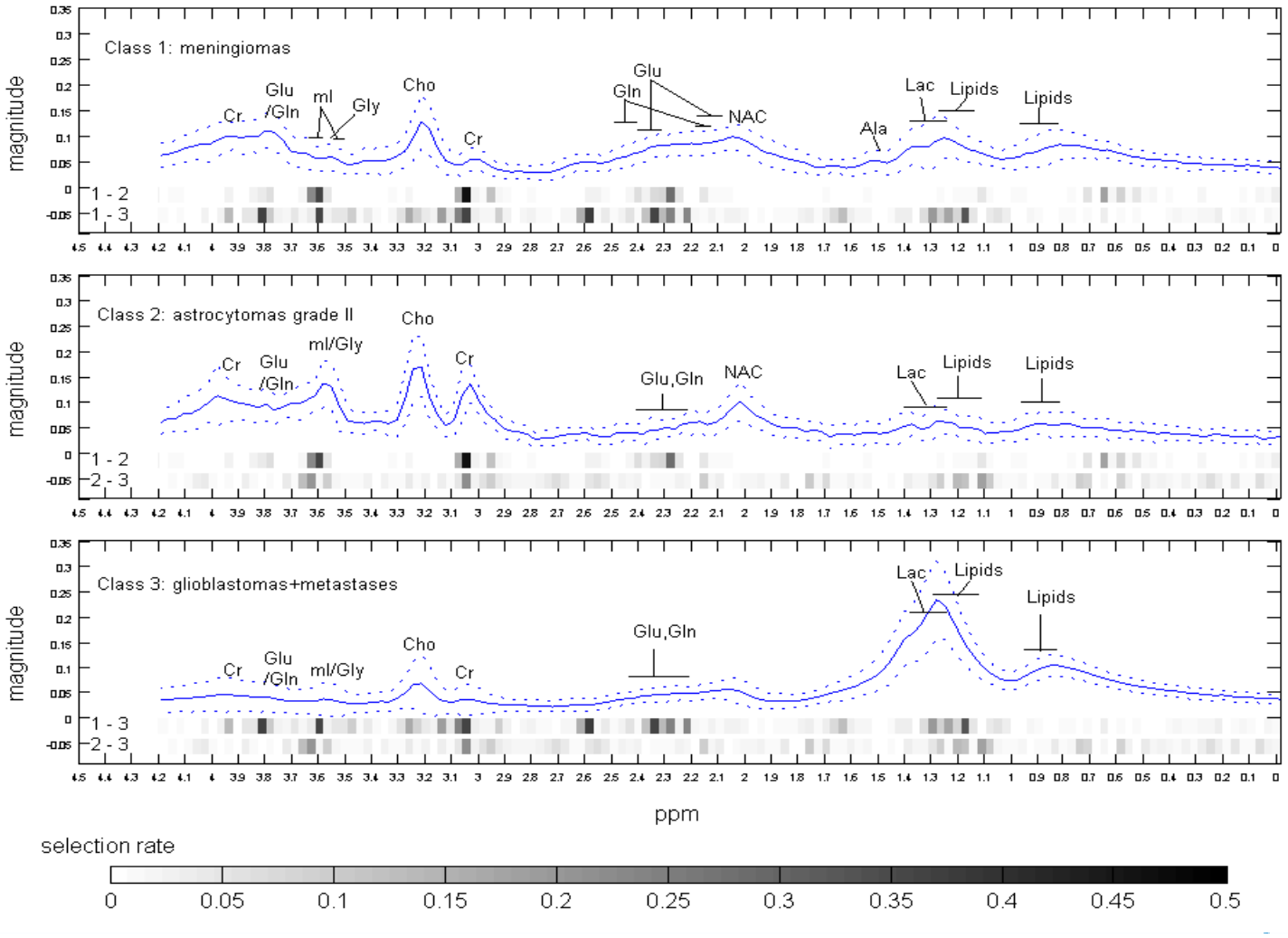
Brain tumor multiclass classification based on MRS spectra data



Mean accuracy from 30 runs of CV

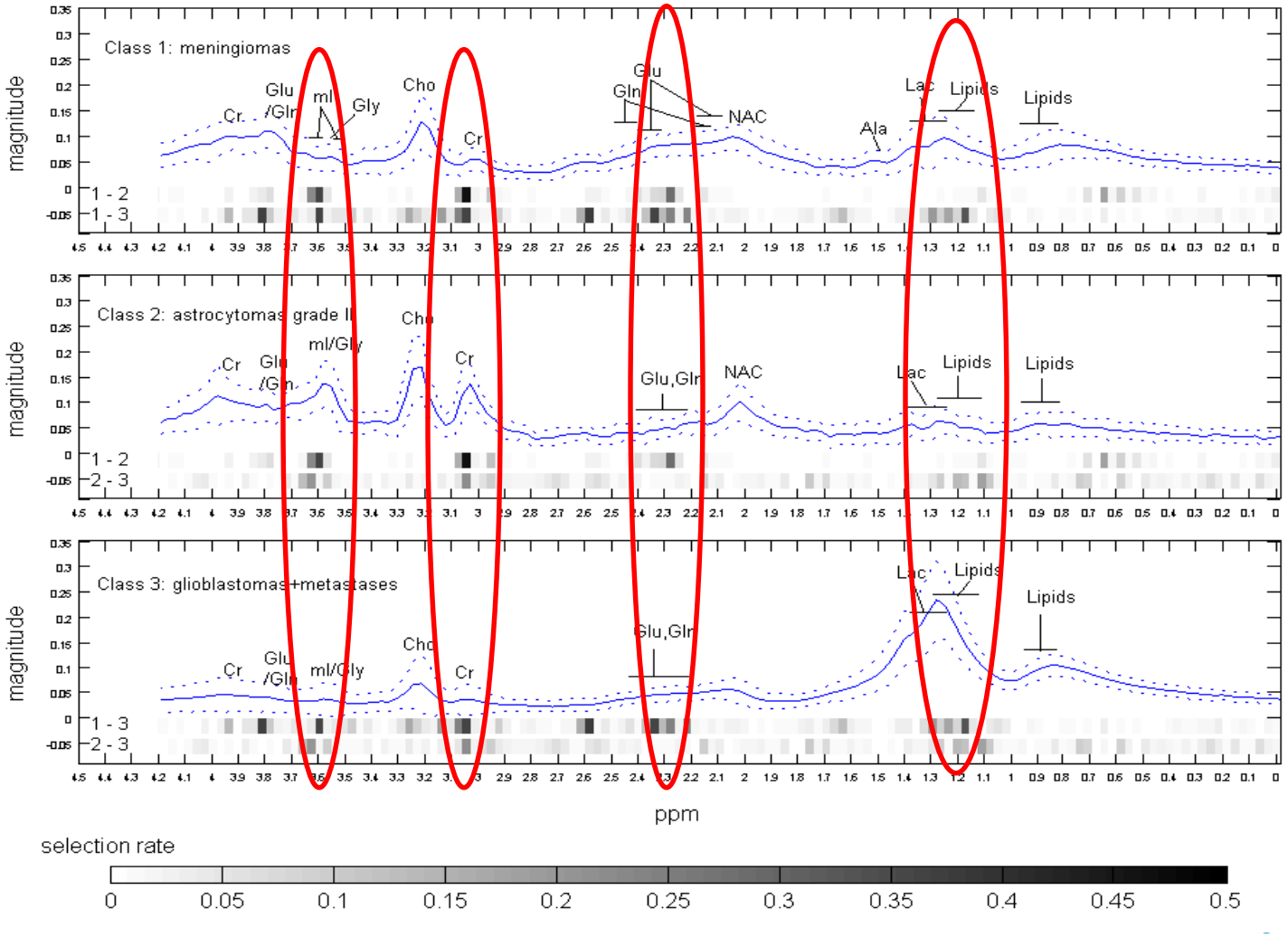
Biological relevance of the selected variables – on MRS spectra

Mean spectrum and selection rate for variables using linSBL+Bag for pairwise binary classification



Biological relevance of the selected variables – on MRS spectra

Mean spectrum and selection rate for variables using linSBL+Bag for pairwise binary classification



Outline

- Supervised learning
- Bayesian frameworks for blackbox models
- Preoperative classification of ovarian tumors
- Bagging for variable selection and prediction in cancer diagnosis problems
- Conclusions

Conclusions

- Bayesian methods: a unifying way for model selection, variable selection, outcome prediction
- Kernel-based models
  - Less hyperparameter to tune compared with MLPs
  - Good performance in our applications.
- Sparseness: good for kernel-based models
  - RVM ← ARD on parametric model
  - LS-SVM ← iterative data point pruning
- Variable selection
  - Evidence based, valuable in applications. Domain knowledge helpful.
  - Variable seection matters more than the model type in our applications.
- Sampling and ensemble: stabilize variable selection and prediction.



SISTA

Conclusions

- Compromise between model interpretability and complexity possible for kernel-based models via additive kernels.
- Linear models suffice in our application.
  Nonlinear kernel-based models worth of trying.

Contributions

- Automatic tuning of kernel parameter for Bayesian LS-SVM
- Sparse approximation for Bayesian LS-SVM
- Proposed two variable selection schemes within Bayesian framework
- Used additive kernels, kPCR and nonlinear biplot to enhance the interpretability of the kernel-based models
- Model development and evaluation of predictive models for ovarian tumor classification, and other cancer diagnosis problems.

Future work

- Bayesian methods: integration for posterior probability, sampling methods or variational methods
- Robust modelling.
- Joint optimization of model fitting and variable selection.
- Incorporate uncertainty, cost in measurement into inference.
- Enhance model interpretability by rule extraction?
- For IOTA data analysis, multi-center analysis, prospective test.
- Combine kernel-based models with belief network (expert knowledge), dealing with missing value problem.



SISTA

Acknowledgments

- Prof. S. Van Huffel and Prof. J.A.K. Suykens
- Prof. D. Timmerman
- Dr. T. Van Gestel, L. Ameye, A. Devos, Dr. J. De Brabanter.
- IOTA project
- EU-funded research project INTERPRET coordinated by Prof. C. Arus
- EU integrated project eTUMOUR coordinated by B. Celda
- EU Network of excellence BIOPATTERN
- Doctoral scholarship of the KUL research council

Thank you!