Bayesian inference for LS-SVMs

Johan Suykens

KU Leuven, ESAT-STADIUS
Kasteelpark Arenberg 10
B-3001 Leuven (Heverlee), Belgium
Email: johan.suykens@esat.kuleuven.be
http://www.esat.kuleuven.be/stadius

Lecture 7

Contents

- LS-SVM: parameters and hyperparameters; 3 Levels of inference
- Level 1: maximum posterior, decision making, moderated outputs, probabilistic interpretations; unbalanced data sets and bias term correction
- Level 2: inference of hyperparameters
- Level 3: inference of kernel parameters and model comparison
- Algorithms
- Input selection by Automatic Relevance Determination (ARD)
- Case study: Preoperative prediction of malignancy of ovarian tumors
- Case study: Financial time series prediction

LS-SVM model, parameters and hyperparameters (1)

- LS-SVM classifier model

$$\begin{aligned}\min_{w,b,e_c} J_{\mathrm{P}}(w,e_c) = & \quad \mu \frac{1}{2} w^T w + \zeta \frac{1}{2} \sum_{k=1}^{N} e_{c,k}^2 \\ \text{subject to} & \quad y_k \left[w^T \varphi(x_k) + b \right] = 1 - e_{c,k}, \;\; k = 1, ..., N\end{aligned}$$

  Two tuning parameters μ, ζ (called **hyperparameters**) corresponding to $\gamma = \zeta / \mu$, error variables denoted as $e_{c,k}$.

- Classifier in the primal weight space $y(x) = \mathrm{sign}[w^T \varphi(x) + b]$

- This is essentially a **regression problem** with targets + 1 and -1:

$$\sum_{k=1}^{N} e_{c,k}^2 = \sum_{k=1}^{N} (y_k e_{c,k})^2 = \sum_{k=1}^{N} e_k^2 \;\; (\text{where} \;\; e_k = y_k e_{c,k} \;\; \text{and} \;\; y_k^2 = 1)$$

LS-SVM model, parameters and hyperparameters (2)

- Write now

$$\begin{array}{rl} \min\limits_{w,b,e} J_{\mathrm{P}}(w,e) = & \mu\, E_W + \zeta\, E_D \\ \text{subject to} & e_k = y_k - [w^T \varphi(x_k) + b]\,,\ \ k = 1, ..., N \end{array}$$

with $E_W = \frac{1}{2} w^T w$, $E_D = \frac{1}{2} \sum_{k=1}^{N} e_k^2 = \frac{1}{2} \sum_{k=1}^{N} \left(y_k - [w^T \varphi(x_k) + b] \right)^2$.

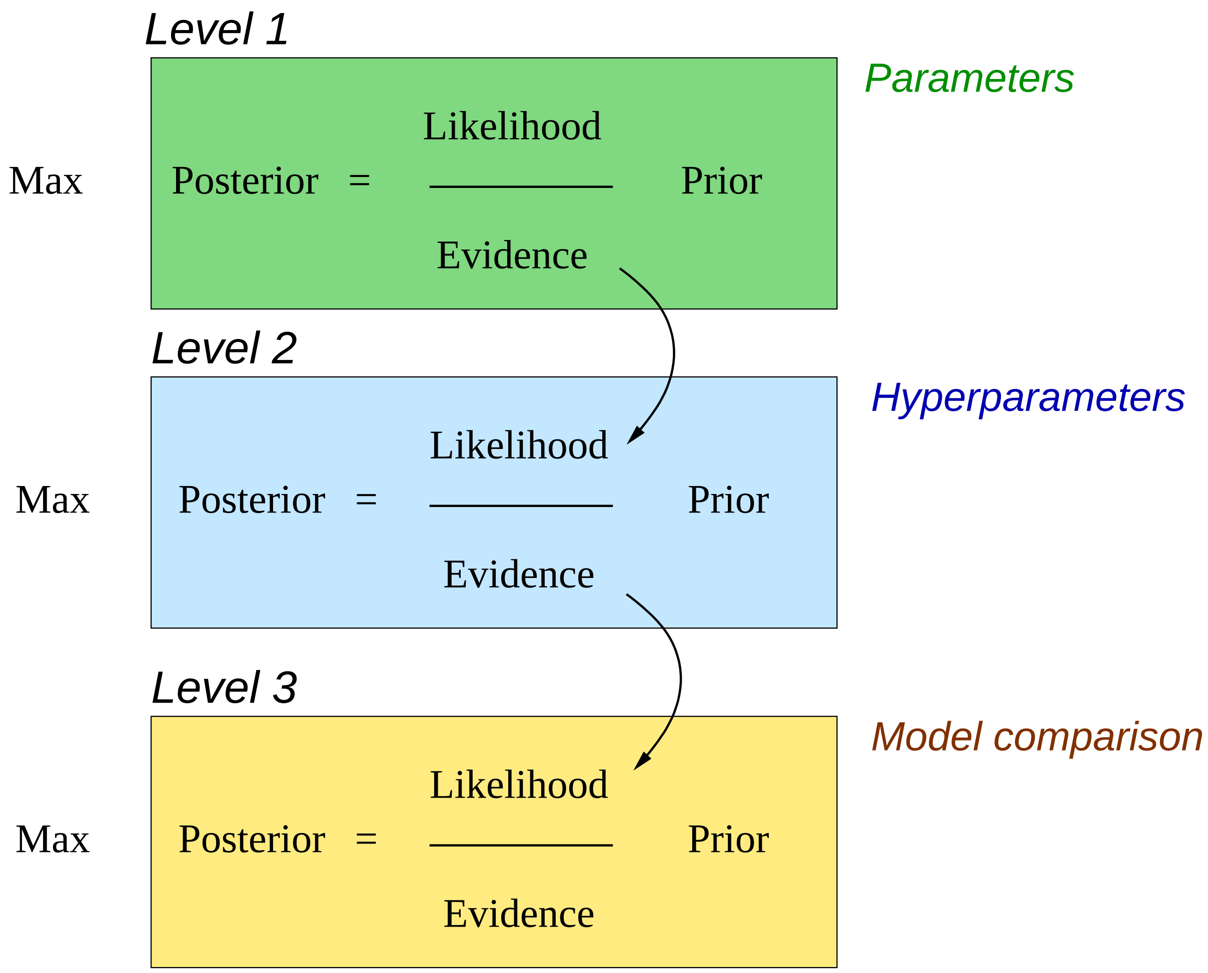
- Dual problem: solve in α, b:

$$\left[\begin{array}{c|c} 0 & 1_v^T \\ \hline 1_v^T & \frac{1}{\mu}\Omega + \frac{1}{\zeta} I \end{array} \right] \left[\begin{array}{c} b \\ \hline \alpha \end{array} \right] = \left[\begin{array}{c} 0 \\ \hline y \end{array} \right]$$

with $y = [y_1; ...; y_N]$, $1_v = [1; ...; 1]$ and $\Omega_{kl} = K(x_k, x_l) = \varphi(x_k)^T \varphi(x_l)$

- Dual representation: $y(x) = \mathrm{sign}[\frac{1}{\mu} \sum_{k=1}^{N} \alpha_k K(x, x_k) + b]$

Bayesian inference



Level 1

Parameters

$$\text{Max} \quad \text{Posterior} = \frac{\text{Likelihood}}{\text{Evidence}} \quad \text{Prior}$$

Level 2

Hyperparameters

$$\text{Max} \quad \text{Posterior} = \frac{\text{Likelihood}}{\text{Evidence}} \quad \text{Prior}$$

Level 3

Model comparison

$$\text{Max} \quad \text{Posterior} = \frac{\text{Likelihood}}{\text{Evidence}} \quad \text{Prior}$$

Bayesian inference (3 levels) (1)

- Notation: i-th model $\mathcal{H}_i$, number of considered models $n_{\mathcal{H}}$, models $\mathcal{H}_\sigma$ depending on σ of RBF kernel, training data $\mathcal{D} = \{x_k, y_k\}_{k=1}^N$.
- Levels of inference:
  - *Level 1* [inference of parameters w, b]

$$p(w, b|\mathcal{D}, \mu, \zeta, \mathcal{H}_\sigma) = \frac{p(\mathcal{D}|w, b, \mu, \zeta, \mathcal{H}_\sigma)}{p(\mathcal{D}|\mu, \zeta, \mathcal{H}_\sigma)} p(w, b|\mu, \zeta, \mathcal{H}_\sigma)$$

  - *Level 2* [inference of hyperparameters μ, ζ]

$$p(\mu, \zeta|\mathcal{D}, \mathcal{H}_\sigma) = \frac{p(\mathcal{D}|\mu, \zeta, \mathcal{H}_\sigma)}{p(\mathcal{D}|\mathcal{H}_\sigma)} p(\mu, \zeta|\mathcal{H}_\sigma)$$

  - *Level 3* [inference of kernel parameter σ and model comparison]

$$p(\mathcal{H}_\sigma|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{H}_\sigma)}{p(\mathcal{D})} p(\mathcal{H}_\sigma)$$

Bayesian inference (3 levels) (2)

- At each of these levels one has

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

- The likelihood at a certain level equals the evidence at the previous level. In this way, by gradually integrating out the parameters at different levels, the subsequent levels are linked to each other.

- An important aspect of Bayesian methods is that the assumptions are made very explicit in the prior.

Bayesian LS-SVM framework

- *Level 1 inference:*
  Inference of w, b
  Probabilistic interpretation of outputs
  Moderated outputs
  Additional correction for prior probabilities and bias term

- *Level 2 inference:*
  Inference of hyperparameter γ
  Effective number of parameters $(< N)$
  Eigenvalues of centered kernel matrix are important

- *Level 3:*
  Model comparison - Occam's razor
  Moderated output with uncertainty on hyperparameters
  Selection of σ width of kernel
  Input selection (ARD at level 3 instead of level 2)

Level 1 - Maximum Posterior (1)

- One has

$$p(w,b|\mathcal{D},\mu,\zeta,\mathcal{H}_\sigma) = \frac{p(\mathcal{D}|w,b,\mu,\zeta,\mathcal{H}_\sigma)}{p(\mathcal{D}|\mu,\zeta,\mathcal{H}_\sigma)} p(w,b|\mu,\zeta,\mathcal{H}_\sigma)$$

  where the evidence $p(\mathcal{D}|\mu,\zeta,\mathcal{H}_\sigma)$ is a normalizing constant, which can be determined by taking the integral over all possible values of w, b of the posterior and setting this integral to one.

- At this level the **prior** corresponds to the **regularization term** and the **sum of the squared error** variables to the **likelihood**.

Level 1 - Maximum Posterior (2)

- For the **prior** we assume that it is separable between w and b (assuming w, b to be statistically independent)

  $$p(w, b|\mu, \zeta, \mathcal{H}_\sigma) = p(w|\mu, \zeta, \mathcal{H}_\sigma)p(b|\mu, \zeta, \mathcal{H}_\sigma)$$

  and that we may simplify this to

  $$p(w, b|\mu, \zeta, \mathcal{H}_\sigma) = p(w|\mu, \mathcal{H}_\sigma)p(b|\sigma_b, \mathcal{H}_\sigma)$$

  where we let $\sigma_b \to \infty$ to approximate a uniform distribution.

- When $\sigma_b \to \infty$ one assumes the prior

  $$\begin{array}{rcl} p(w, b|\mu, \mathcal{H}_\sigma) & = & \left(\frac{\mu}{2\pi}\right)^{n_h/2} \exp\left(-\mu\,\frac{1}{2}w^T w\right) \frac{1}{\sqrt{2\pi}\sigma_b} \exp\left(-\frac{1}{2}\frac{b^2}{\sigma_b^2}\right) \\ & \propto & \left(\frac{\mu}{2\pi}\right)^{n_h/2} \exp\left(-\mu\frac{1}{2}w^T w\right) \end{array}$$

  with n_h the dimension of the feature space for $\varphi(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n_h}$.

Level 1 - Maximum Posterior (3)

- Assuming **independent data** one has the **likelihood**

$$p(\mathcal{D}|w,b,\zeta,\mathcal{H}_\sigma) \;=\; \prod_{k=1}^{N} p(x_k,y_k|w,b,\zeta,\mathcal{H}_\sigma) \propto \prod_{k=1}^{N} p(e_k|w,b,\zeta,\mathcal{H}_\sigma) = \prod_{k=1}^{N} \sqrt{\frac{\zeta}{2\pi}} \exp\left(-\frac{1}{2}\zeta e_k^2\right)$$

The variance around the targets $+1$ and -1 is assumed to be $1/\zeta$.

- Application of Bayes' rule for the **posterior** gives

$$p(w,b|\mathcal{D},\mu,\zeta,\mathcal{H}_\sigma) \propto \exp\left(-\mu\frac{1}{2}w^T w - \zeta\frac{1}{2}\sum_{k=1}^{N} e_k^2\right).$$

One then aims at finding w, b that maximize the posterior which means that one minimizes $J_{\rm P}(w, e_c)$. One rather maximizes the logarithm of the posterior with **maximum posterior solution** denoted as $w_{\rm MP}, b_{\rm MP}$.

Level 1 - Maximum Posterior (4)

- One obtains

$$p(w - w_{\mathrm{MP}}, b - b_{\mathrm{MP}} | \mathcal{D}, \mu, \zeta, \mathcal{H}_\sigma) = \frac{1}{\sqrt{(2\pi)^{(n_h+1)} \mathrm{det} Q}} \exp\left(-\frac{1}{2} g^T Q^{-1} g\right)$$

  with $g = [w - w_{\mathrm{MP}}; b - b_{\mathrm{MP}}]$, $Q = \mathrm{Cov}([w;b],[w;b])$, $H = Q^{-1}$.

- The covariance matrix Q is related to the Hessian of the (quadratic) cost function:

$$Q = H^{-1} = \left[\begin{array}{cc} (\mu I + \zeta G)^{-1} & -(\mu I + \zeta G)^{-1} H_{12} H_{22}^{-1} \\ -H_{22}^{-1} H_{12}^T (\mu I + \zeta G)^{-1} & H_{22}^{-1} + H_{22}^{-1} H_{12}^T (\mu I + \zeta G)^{-1} H_{12} H_{22}^{-1} \end{array} \right]$$

  where

$$H_{11} = \mu I + \zeta \Upsilon \Upsilon^T, H_{12} = \zeta \Upsilon 1_v, H_{22} = \zeta N$$

  and $\Upsilon = [\varphi(x_1), ..., \varphi(x_N)]$, $G = \Upsilon M_c \Upsilon^T$,
  and centering matrix $M_c = I - (1/N) 1_v 1_v^T$.

Decision making (1)

- The following **assumption** is meaningful because of the link with **kernel FDA**:

$$
\begin{aligned}
p(x|y=+1,w,b,\zeta,\mathcal{H}_\sigma) &= \sqrt{\frac{1}{2\pi\zeta^{-1}}}\exp\left(-\frac{1}{2}\frac{e^2}{\zeta^{-1}}\right) \\
p(x|y=-1,w,b,\zeta,\mathcal{H}_\sigma) &= \sqrt{\frac{1}{2\pi\zeta^{-1}}}\exp\left(-\frac{1}{2}\frac{e^2}{\zeta^{-1}}\right)
\end{aligned}
$$

with

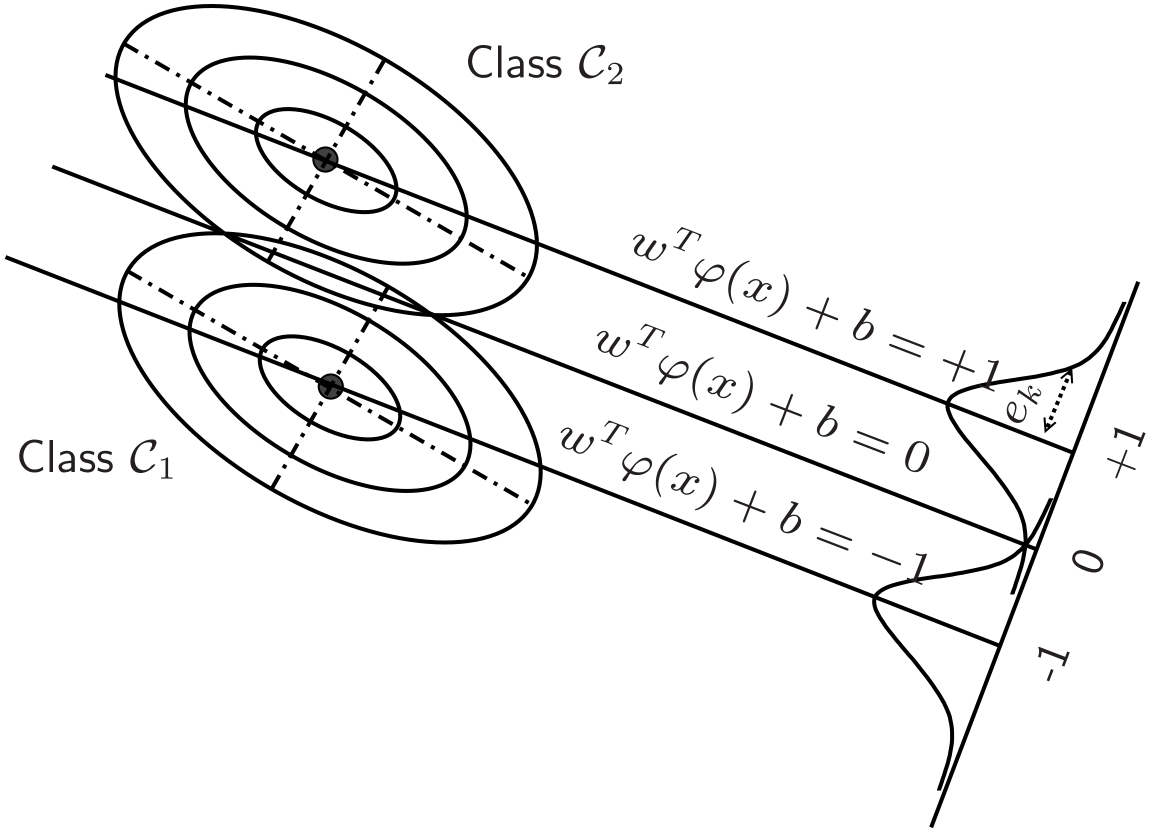
$$
\begin{aligned}
e_k &= w^T\left(\varphi(x_k)-\hat{\mu}^{(1)}\right), \quad k=1,...,N_1 \\
e_l &= w^T\left(\varphi(x_l)-\hat{\mu}^{(2)}\right), \quad l=1,...,N_2
\end{aligned}
$$

and

$$
\begin{aligned}
\hat{\mu}^{(1)} &= \tfrac{1}{N_1}\textstyle\sum_{k=1}^{N_1}\varphi(x_k) \\
\hat{\mu}^{(2)} &= \tfrac{1}{N_2}\textstyle\sum_{l=1}^{N_2}\varphi(x_l)
\end{aligned}
$$

with N_1, N_2 the number of data points of class 1 and 2.

Decision making (2)



As explained in the kernel Fisher discriminant interpretation of LS-SVM classifiers, one aims at minimizing the within scatter for Class $\mathcal{C}_1$ and $\mathcal{C}_2$.

Decision making (3)

- By marginalization over w, b (in fact it is more accurate to marginalize also over the hyperparameters and the kernel parameters, but this has less influence on the result), one obtains:

$$
\begin{aligned}
p(x|y=+1,\mathcal{D},\mu,\zeta,\mathcal{H}_\sigma) &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\zeta^{-1}+\sigma^2_{(1)}(x)}}\exp\left(-\frac{1}{2}\frac{m^2_{(1)}(x)}{\zeta^{-1}+\sigma^2_{(1)}(x)}\right)\\
p(x|y=-1,\mathcal{D},\mu,\zeta,\mathcal{H}_\sigma) &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\zeta^{-1}+\sigma^2_{(2)}(x)}}\exp\left(-\frac{1}{2}\frac{m^2_{(2)}(x)}{\zeta^{-1}+\sigma^2_{(2)}(x)}\right)
\end{aligned}
$$

with $\sigma^2_{(1)}(x) = [\varphi(x)-\mu^{(1)}]H_{11}^{-1}[\varphi(x)-\mu^{(1)}]$, $\sigma^2_{(2)}(x) = [\varphi(x)-\mu^{(2)}]H_{11}^{-1}[\varphi(x)-\mu^{(2)}]$

and

$$
\begin{aligned}
m_{(1)}(x) &= w_{\mathrm{MP}}^T(\varphi(x)-\mu^{(1)}) \simeq \frac{1}{\mu}\sum_{k=1}^{N}\alpha_k K(x,x_k) - \hat{\mu}_{d1}\\
m_{(2)}(x) &= w_{\mathrm{MP}}^T(\varphi(x)-\mu^{(2)}) \simeq \frac{1}{\mu}\sum_{k=1}^{N}\alpha_k K(x,x_k) - \hat{\mu}_{d2}
\end{aligned}
$$

Decision making (4)

- Application of **Bayesian decision theory** gives

$$P(y|x,\mathcal{D},\mu,\zeta,\mathcal{H}_\sigma) = \frac{p(x|y,\mathcal{D},\mu,\zeta,\mathcal{H}_\sigma)}{p(x|\mathcal{D},\mu,\zeta,\mathcal{H}_\sigma)} P(y|\mathcal{D},\mu,\zeta,\mathcal{H}_\sigma)$$

where $P(y)$ denotes the prior class probability.

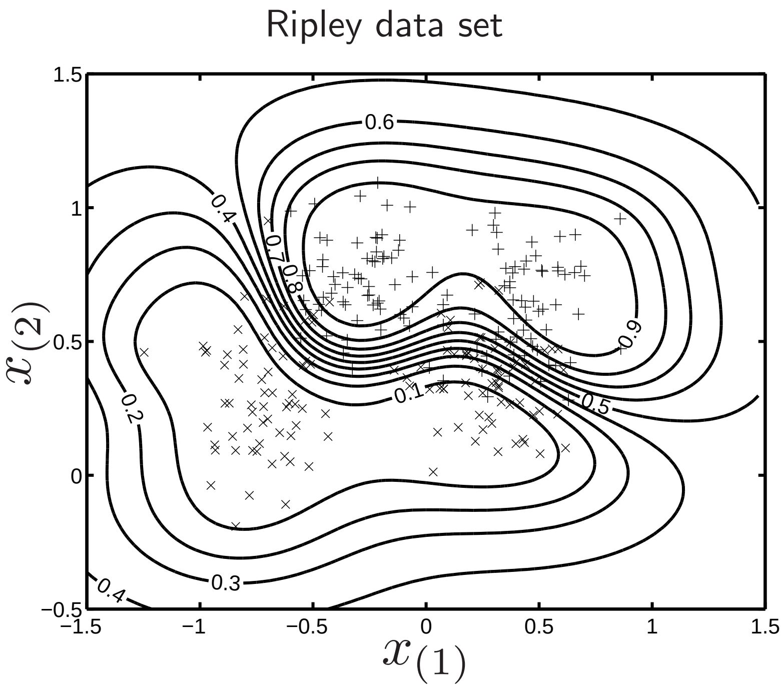
- One selects the class with **maximal posterior class probability**. Assuming $\sigma^2_{(1)} = \sigma^2_{(2)}$ and defining $\sigma_e^2 = \sqrt{\sigma^2_{(1)}\sigma^2_{(2)}}$ one has

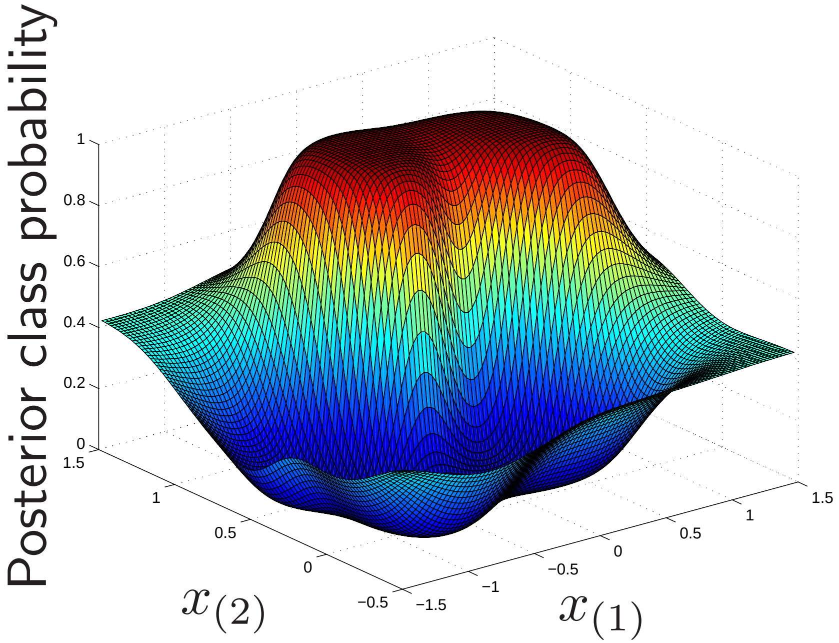
$$y(x) \quad = \quad \mathrm{sign}[\frac{1}{\mu}\sum_{k=1}^{N}\alpha_k K(x,x_k) - \frac{\mu_{d1}+\mu_{d2}}{2} + \frac{\zeta^{-1}+\sigma_e^2(x)}{\mu_{d1}-\mu_{d2}} \log\frac{P(y=+1)}{P(y=-1)}]$$

This results in a classifier of the form

$$y(x) = \mathrm{sign}[\frac{1}{\mu}\sum_{k=1}^{N}\alpha_k K(x,x_k) + \Delta b(x)]$$

Example: Ripley data set





- Probabilistic interpretation with moderated output
- Bias term correction for unbalanced and/or small data sets

Moderated outputs

- A simple way to obtain a **probabilistic interpretation** of the LS-SVM classifier is to apply the **softmax function** to the latent variable $z = w^T \varphi(x) + b$. It results in

$$\begin{aligned} P(y=+1|x,w,b,\mathcal{D},\mu,\zeta,\mathcal{H}_\sigma) &= \frac{1}{1+\exp\left(-(w^T\varphi(x)+b)\right)} \\ P(y=-1|x,w,b,\mathcal{D},\mu,\zeta,\mathcal{H}_\sigma) &= 1-P(y=+1|x,w,b,\mathcal{D},\mu,\zeta,\mathcal{H}_\sigma) \\ &= \frac{1}{1+\exp\left(w^T\varphi(x)+b\right)}. \end{aligned}$$

- The functional form of this expression is consistent with a Gaussian assumption of the error variables around the targets $+1$ and -1.

Unbalanced data sets (1)

- It may happen in real-life situations that the given training data set is unbalanced in the sense that $N_1 \gg N_2$ or $N_2 \gg N_1$. Bayesian decision theory provides a method for taking into account this unbalancing by having **different prior class probabilities** by

$$P(y=+1) = \frac{N_1}{N_1+N_2} \text{ and } P(y=-1) = \frac{N_2}{N_1+N_2}$$

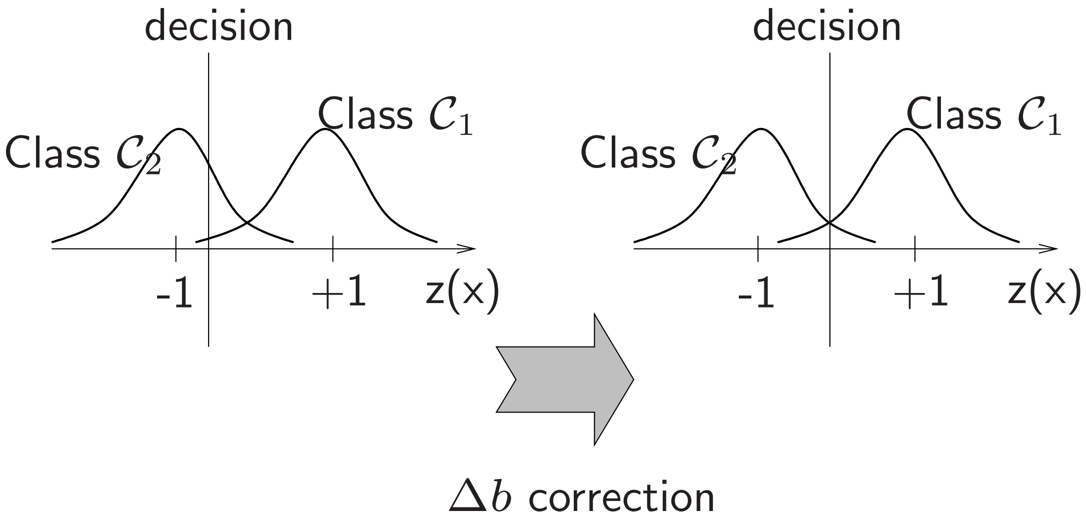
- A minimal requirement of any classifier is that it should be able to perform **better than the majority rule** which decides $y=+1$ if $N_1 > N_2$ and else $y=-1$.

Unbalanced data sets (2)

The Bayesian LS-SVM classifier can take into account a **bias term correction**:

$$y(x) = \text{sign}[\frac{1}{\mu}\sum_{k=1}^{N}\alpha_k K(x, x_k) + \Delta b(x)]$$

with bias term correction $\Delta b(x)$. (one can also do a bias term correction in a non-Bayesian context after computing the LS-SVM classifier and changing the bias term afterwards).



Level 2 - inference of hyperparameters (1)

- Calculation of **posterior**:

$$p(\mu, \zeta | \mathcal{D}, \mathcal{H}_\sigma) = \frac{p(\mathcal{D}|\mu, \zeta, \mathcal{H}_\sigma)}{p(\mathcal{D}|\mathcal{H}_\sigma)} p(\mu, \zeta | \mathcal{H}_\sigma)$$

One obtains (assuming uniform prior on μ, ζ)

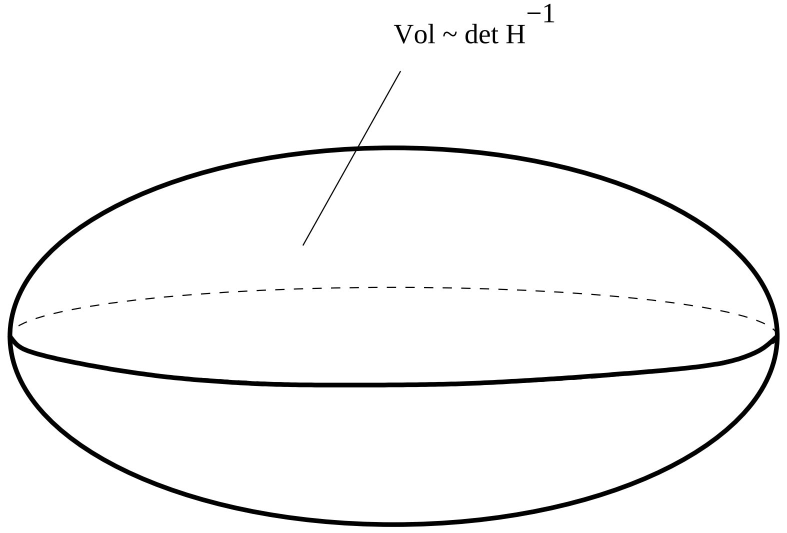
$$p(\mu, \zeta | \mathcal{D}, \mathcal{H}_\sigma) \propto \exp\left(-J_{\mathrm{P}}(w_{\mathrm{MP}}, b_{\mathrm{MP}})\right) \times \sqrt{\mu^{n_h} \zeta^N} \sqrt{\det H^{-1}}$$

with

$$\begin{aligned} J_{\mathrm{P}}(w, b) &= J_{\mathrm{P}}(w_{\mathrm{MP}}, b_{\mathrm{MP}}) + \tfrac{1}{2}([w; b] - [w_{\mathrm{MP}}; b_{\mathrm{MP}}])^T H ([w; b] - [w_{\mathrm{MP}}; b_{\mathrm{MP}}]) \\ J_{\mathrm{P}}(w_{\mathrm{MP}}, b_{\mathrm{MP}}) &= \mu E_W(w_{\mathrm{MP}}) + \zeta E_D(w_{\mathrm{MP}}, b_{\mathrm{MP}}) \end{aligned}$$

Level 2 - inference of hyperparameters (2)

- $\det H^{-1}$ characterizes the volume of an ellipsoid centered at $[w_{\rm MP}; b_{\rm MP}]$



Note: An ellipsoid $\mathcal{E} = \{x | \, \|Ax - b\| \leq 1\}$ has a volume proportional to $\det A^{-1}$

Level 2 - inference of hyperparameters (3)

- Finding **optimal** μ, ζ by **maximizing log posterior**:

$$\begin{aligned}\min_{\mu,\zeta} J(\mu,\zeta) &= \mu E_W(w_{\mathrm{MP}}) + \zeta E_D(w_{\mathrm{MP}}, b_{\mathrm{MP}}) \\ &\quad + \frac{1}{2}\sum_{i=1}^{N_{\mathrm{eff}}} \log(\mu + \zeta\lambda_{G,i}) - \frac{N_{\mathrm{eff}}}{2}\log\mu - \frac{N-1}{2}\log\zeta,\end{aligned}$$

  where $\det H^{-1} = 1/\det H$ with $\det H = N\mu^{n_h - N_{\mathrm{eff}}}\zeta\prod_{i=1}^{N_{\mathrm{eff}}}(\mu+\zeta\lambda_{G,i})$ and $\lambda_{G,i}$ are the eigenvalues of matrix G and N_{eff} is the number of non-zero eigenvalues of the matrix $M_c\Omega M_c$.

- Optimal solution: from $\partial J(\mu,\zeta)/\partial\mu = 0$, $\partial J(\mu,\zeta)/\partial\zeta = 0$ one obtains set of equations in μ, ζ:

$$\left\{\begin{array}{c} 2\mu\, E_W(w_{\mathrm{MP}};\mu,\zeta) = d_{\mathrm{eff}}(\mu,\zeta) - 1 \\ 2\zeta\, E_D(w_{\mathrm{MP}}, b_{\mathrm{MP}};\mu,\zeta) = N - d_{\mathrm{eff}}(\mu,\zeta) \end{array}\right.$$

Level 2 - inference of hyperparameters (4)

- **Effective number of parameters:**

$$d_{\mathrm{eff}} = \sum_i \lambda_{i,u} / \lambda_{i,r} = 1 + \sum_{i=1}^{N_{\mathrm{eff}}} \frac{\zeta \lambda_{G,i}}{\mu + \zeta \lambda_{G,i}}$$

with
$\lambda_{i,u}$: eigenvalues Hessian of unregularized cost fu. $J_{\mathrm{P},u} = \zeta E_D$
$\lambda_{i,r}$: eigenvalues Hessian of regularized cost fu. $J_{\mathrm{P},r} = \mu E_W + \zeta E_D$

- Note that d_{eff} is smaller than N.

Level 3 - inference of kernel parameters (1)

- Level 3 posterior:

$$p(\mathcal{H}_\sigma|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{H}_\sigma)}{p(\mathcal{D})} p(\mathcal{H}_\sigma)$$

  If $p(\mathcal{H}_\sigma)$ prior uniform then $p(\mathcal{H}_\sigma|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{H}_\sigma)$.

- Consider two model (RBF kernel with σ_1 and σ_2):

$$\frac{p(\mathcal{H}_{\sigma_1}|\mathcal{D})}{p(\mathcal{H}_{\sigma_2}|\mathcal{D})} = \frac{p(\mathcal{D}|\mathcal{H}_{\sigma_1})}{p(\mathcal{D}|\mathcal{H}_{\sigma_2})} \cdot \frac{p(\mathcal{H}_{\sigma_1})}{p(\mathcal{H}_{\sigma_2})}$$

  When $p(\mathcal{H}_{\sigma_1}) = p(\mathcal{H}_{\sigma_2})$ then

$$\frac{p(\mathcal{H}_{\sigma_1}|\mathcal{D})}{p(\mathcal{H}_{\sigma_2}|\mathcal{D})} = \frac{p(\mathcal{D}|\mathcal{H}_{\sigma_1})}{p(\mathcal{D}|\mathcal{H}_{\sigma_2})} = \text{Bayes factor}$$

  This **Bayes factor** characterizes the improvement of model $\mathcal{H}_{\sigma_1}$ with respect to $\mathcal{H}_{\sigma_2}$.

Level 3 - inference of kernel parameters (2)

- From the normalization $\int p(\theta|\mathcal{D},\mathcal{H}_\sigma)d\theta = 1$ one finds that the evidence equals

$$p(\mathcal{D}|\mathcal{H}_\sigma) = \int p(\mathcal{D}|\theta,\mathcal{H}_\sigma)p(\theta|\mathcal{H}_\sigma)d\theta$$

  meaning that one marginalizes over the likelihood and the prior.

- **Laplace approximation** method:

$$\begin{aligned} p(\mathcal{D}|\mathcal{H}_\sigma) &= \int \exp\left(-h(\theta)\right) d\theta \\ &\simeq \exp(-h(\theta_{\mathrm{MP}})) \int \exp\left(-\frac{1}{2}(\theta-\theta_{\mathrm{MP}})^T H(\theta-\theta_{\mathrm{MP}})\right) d\theta \\ &= \exp(-h(\theta_{\mathrm{MP}}))\,\sqrt{(2\pi)^{n_p}}\,\sqrt{\det H^{-1}} \end{aligned}$$

  with Hessian H evaluated at the maximal posterior and $\theta \in \mathbb{R}^{n_p}$ and $\exp(-h(\theta)) = p(\mathcal{D}|\theta,\mathcal{H}_\sigma)p(\theta|\mathcal{H}_\sigma)$.

Level 3 - inference of kernel parameters (3)

- When assuming that posterior is sharply peaked at the maximum posterior solution θ_{MP}:

$$\begin{aligned} p(\mathcal{D}|\mathcal{H}_\sigma) &\simeq p(\mathcal{D}|\theta_{\mathrm{MP}}, \mathcal{H}_\sigma) \times p(\theta_{\mathrm{MP}}|\mathcal{H}_\sigma) \sqrt{(2\pi)^{n_p}} \sqrt{\det H^{-1}} \\ \text{Evidence} &\simeq \text{Likelihood}|_{\theta=\theta_{\mathrm{MP}}} \times \text{Occam factor} \end{aligned}$$

with H the Hessian of a quadratic function (or approximation) around θ_{MP}.

Bayesian inference algorithm for LS-SVM classifiers

1. **Normalize and standardize** inputs to zero mean and unit variance.

2. Select a model $\mathcal{H}_i$ by **choosing a kernel** K_i (e.g. kernel parameter σ_i in the RBF kernel case), with $i = 1, ..., n_{\mathcal{H}}$ number of models to be compared.

3. Compute the **effective number of parameters** d_{eff}, from the eigenvalue decomposition of the centered Gram matrix.

4. Find the **optimal hyperparameters** $\mu_{\text{MP}}, \zeta_{\text{MP}}$ by solving the optimization problem in $\gamma = \zeta/\mu$, related to maximizing the Level 2 posterior.

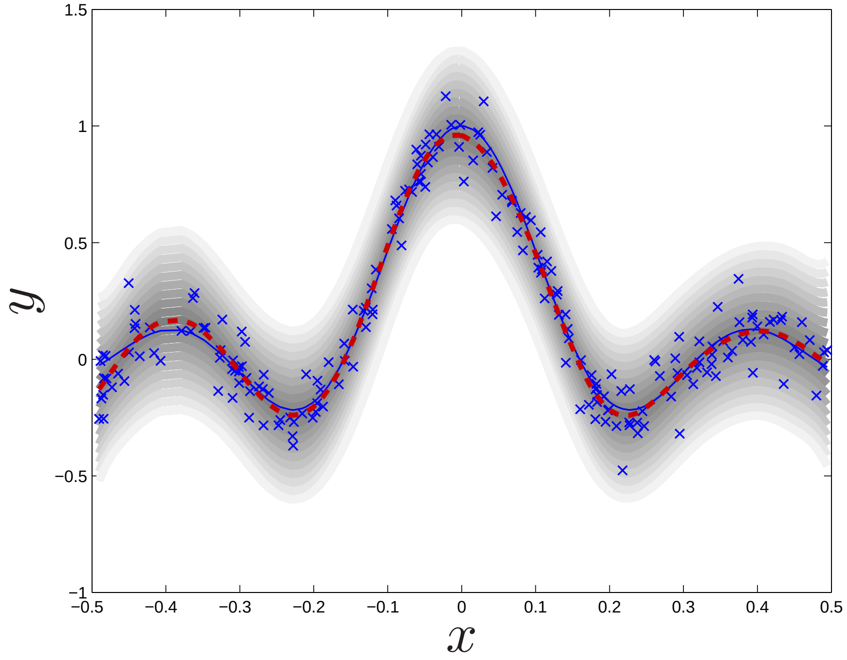
5. Use the expression for Level 3 **model comparison** based on the evidence.

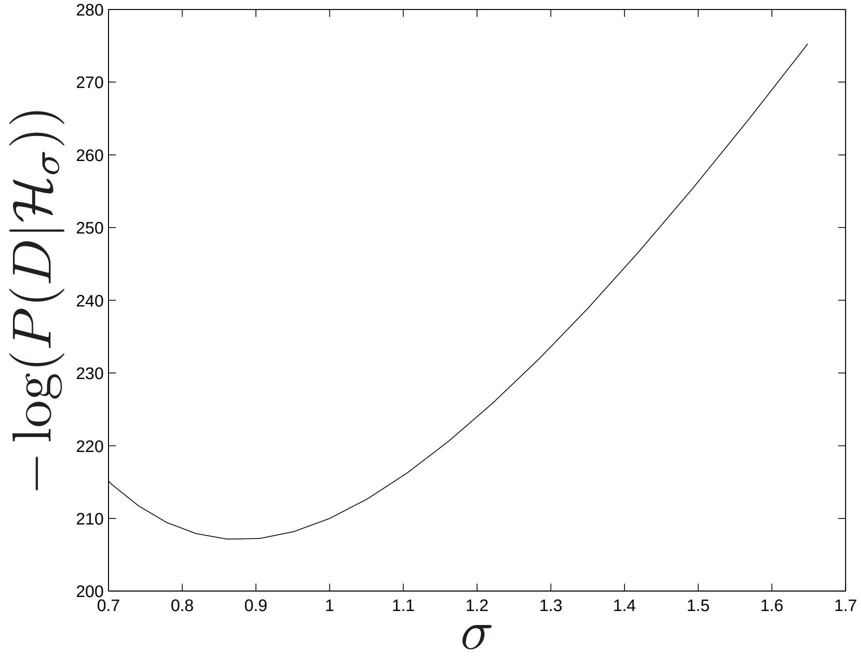
6. **Refine the kernel tuning parameters** (σ_i in the RBF kernel case) and go back to step 2 as long as one can improve.

UCI Benchmarking results

	n	N	N_{test}	N_{tot}	LS-SVM	LS-SVM	LS-SVM	SVM	GP	GP_{b}	GP
					(BayM)	(Bay)	(CV10)	(CV10)	(Bay)	(Bay)	(CV10)
bld	6	230	115	345	**69.4**(2.9)	**69.4**(3.1)	**69.4**(3.4)	**69.2**(3.5)	**69.2**(2.7)	**68.9**(3.3)	**69.7**(4.0)
cra	6	133	67	200	**96.7**(1.5)	**96.7**(1.5)	**96.9**(1.6)	**95.1**(3.2)	**96.4**(2.5)	94.8(3.2)	**96.9**(2.4)
gcr	20	666	334	1000	*73.1*(3.8)	73.5(3.9)	**75.6**(1.8)	*74.9*(1.7)	**76.2**(1.4)	**75.9**(1.7)	**75.4**(2.0)
hea	13	180	90	270	**83.6**(5.1)	**83.2**(5.2)	**84.3**(5.3)	**83.4**(4.4)	**83.1**(5.5)	**83.7**(4.9)	**84.1**(5.2)
ion	33	234	117	351	95.6(0.9)	**96.2**(1.0)	**95.6**(2.0)	95.4(1.7)	*91.0*(2.3)	*94.4*(1.9)	*92.4*(2.4)
pid	8	512	256	768	**77.3**(3.1)	**77.5**(2.8)	**77.3**(3.0)	**76.9**(2.9)	**77.6**(2.9)	**77.5**(2.7)	**77.2**(3.0)
rsy	2	250	1000	1250	**90.2**(0.7)	**90.2**(0.6)	89.6(1.1)	**89.7**(0.8)	**90.2**(0.7)	**90.1**(0.8)	**89.9**(0.8)
snr	60	138	70	208	76.7(5.6)	**78.0**(5.2)	**77.9**(4.2)	76.3(5.3)	**78.6**(4.9)	**75.7**(6.1)	**76.6**(7.2)
tit	3	1467	734	2201	**78.8**(1.1)	**78.7**(1.1)	**78.7**(1.1)	**78.7**(1.1)	**78.5**(1.0)	77.2(1.9)	**78.7**(1.2)
wbc	9	455	228	683	95.9(0.6)	*95.7*(0.5)	**96.2**(0.7)	**96.2**(0.8)	95.8(0.7)	*93.7*(2.0)	**96.5**(0.7)
AP					**83.7**	**83.9**	**84.1**	83.6	**83.7**	83.2	**83.7**
AR					**2.3**	**2.5**	**2.5**	**3.8**	**3.2**	**4.2**	**2.6**
P_{ST}					**1.000**	**0.754**	**1.000**	**0.344**	**0.754**	**0.344**	**0.508**

Bayesian LS-SVM regression





Input selection by automatic relevance determination

- RBF kernel can be used to determine relevance of inputs:

$$K(x,z) = \exp\left(-(x-z)^T S^{-1}(x-z)\right)$$

  where $S = \mathrm{diag}([s_1^2; ...; s_n^2])$. Small values of $1/s_i^2$ indicate that the corresponding input is not very relevant in the chosen model.

- *LS-SVM input selection by ARD:*
  - Normalize the inputs to zero mean and unit variance
  - Start with equal elements $s_1 = s_2 = ... = s_n$ or $S = s^2 I$.
    This kernel parameter is optimized by Level 3 inference based on the evidence formula. The result of this optimization is taken as starting point for initializing a nonlinear optimization problem for the diagonal matrix S by optimizing the same evidence formula at Level 3.
  - Inspect the results from Level 3 inference and remove the least relevant input. Reduce the dimensionality of the problem and set $n := n - 1$.
  - Go to step 1 and remove inputs as long as one improves according to the Level 3 evidence formula.

Case study: prediction of malignancy of ovarian tumors (1)

- Ovarian masses: a common problem in **gynecology**
  - Types : benign and malignant
    (borderline, primary, metastatic invasive)
  - Ovarian cancer : the highest mortality rate among gynecologic cancers
  - Early detection of ovarian cancer is difficult
  - Treatment and management of different types is different

- **Preoperative distinction** between benign and malignant tumors.

- **Medical techniques** for preoperative evaluation
  - Serum tumor marker: CA125 blood test
  - Transvaginal ultrasonography
  - Color Doppler imaging and blood flow indexing ...

- **Goal:** automate the classification process from experience to AI tools. In this work the use of Bayesian LS-SVM classifiers is focused.

Case study: prediction of malignancy of ovarian tumors (2)

Patient data collected at Univ. Hospitals Leuven, Belgium, 1994 - 1999
425 records, 25 features
291 benign tumors, 134 (32%) malignant tumors

Table 1: 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	Normal blood flow (Colsc3)	15.8 %	35.8 %
	Strong blood flow (Colsc4)	4.5 %	20.3 %
Morphologic	Abdominal fluid (Asc)	32.7 %	67.3 %
	Bilateral mass (Bilat)	13.3 %	39.1 %
	Solid tumor (Sol)	8.3 %	37.6 %
	Irregular wall (Irreg)	33.8 %	73.2 %
	Papillations (Pap)	13.0 %	53.2 %
	Acoustic shadows (Shadows)	12.2 %	5.7 %

Note: for continuous variables, mean±SD in case of a benign and malignant tumor respectively are reported; for binary variables, the occurrences (%) of the corresponding features are reported.

Case study: prediction of malignancy of ovarian tumors (3)

- Select the input variables according to the model evidence $p(D|\mathcal{H}_j)$.
- The heuristic search strategy: e.g. forward, backward, stepwise ...
- Given a certain type of kernel, a forward selection (greedy search) was performed:
  - Starting from no variables
  - Choose each time the variable which gives the greatest increase in the current model evidence
  - Stop the selection when the addition of any remaining variables can no longer increase the model evidence
- 10 variables were selected based on the training set (first treated 265 patient data), using an RBF kernel: *L_CA125, Pap, Sol, Colsc3, Bilat, Meno, Asc, Shadows, Colsc4, Irreg*

Case study: prediction of malignancy of ovarian tumors (4)

Sparse approximation:

- Due to the choice of the L_2 loss function, LS-SVM is loosing sparseness compared with standard SVMs.

- Sparseness can be imposed to LS-SVM by a pruning procedure based upon the support values $\alpha_i = \gamma e_i$.

- Propose to prune the data points which have negative support values:
  - Intuitively, pruning of easy examples will focus the model on the harder cases which lie around the decision boundary.
  - Iteratively prune the data with negative α_i, the hyper-parameters are retuned several times based on the reduced data set using the Bayesian evidence framework.
  - Stop when no more support values are negative.

Case study: prediction of malignancy of ovarian tumors (5)

Model evaluation:

- Preprocessing: normalize the training data into zero mean and variance one, also normalized the test set using the mean and variance estimated from the training set

- The model performance is assessed by ROC analysis:
  - Area under the ROC curve (AUC)
  - *Sensitivity* and *specificity*: the correct classification rate for the malignant and benign class, respectively.
  - Goal is to find a model with high sensitivity for malignant (high cancer detection rate) while maintaining a low false positive rate.

Case study: prediction of malignancy of ovarian tumors (6)

- **Compared models:**
  - Risk of malignancy Index (RMI) (Jacobs *et al.*), a widely used score: RMI $=$ Score$_{\mathrm{morph}}$ $\times$ Score$_{\mathrm{meno}}$ $\times$ CA125
  - Logistic regression model (LR)

$$\mathrm{logit}(\mathcal{P}) = \log(\frac{\mathcal{P}}{1-\mathcal{P}}) = w^T x + b,$$

where $\mathcal{P} = \mathcal{P}(x)$ denote the probability that the tumor is malignant, given the input data x. Prior class probabilities can be incorporated by adjusting the bias term b.

Case study: prediction of malignancy of ovarian tumors (7)

Results:

Model Type (AUC)	Cutoff value	Accuracy (%)	Sensitivity (%)	Specificity (%)
RMI (0.8733)	100	78.13	74.07	80.19
	75	**76.88**	**81.48**	**74.53**
LR1 (0.9111)	0.5	81.25	74.07	84.91
	0.4	80.63	75.96	83.02
	0.3	**80.63**	**77.78**	**82.08**
LS-SVM$_{\text{Lin}}$ (0.9141)	0.5	82.50	77.78	84.91
	0.4	81.25	77.78	83.02
	0.3	**81.88**	**83.33**	**81.13**
LS-SVM$_{\text{RBF}}$ (0.9184)	0.5	84.38	77.78	87.74
	0.4	83.13	81.48	83.96
	0.3	**84.38**	**85.19**	**83.96**

Case study: prediction of malignancy of ovarian tumors (8)

Randomized cross-validation:
Randomly separating training set and test set; Stratified ($\#malignant : \#benign \simeq 2:1$) for each training and test set; Repeat the hold-out cross-validation 30 times.

Table 2: Averaged performance on the test set from 30 runs of randomized cross validation($N_{\text{train}} = 265, N_{\text{test}} = 160$)

Model Type	$\overline{\mathrm{AUC}}$	Cutoff	Accuracy	Sensitivity	Specificity
	$\pm SD$	value	(%)	(%)	(%)
RMI	0.8882	100	82.65	81.73	83.06
	±0.0284	**80**	**81.10**	**83.87**	**79.85**
LR1	0.9397	0.5	83.29	89.33	80.55
	±0.0209	**0.4**	**81.94**	**91.60**	**77.55**
LS-SVM1$_{\text{Lin}}$	0.9405	0.5	84.31	87.40	82.91
	±0.0199	**0.4**	**82.77**	**90.47**	**79.27**
LS-SVM1$_{\text{RBF}}$	0.9424	0.5	84.85	86.53	84.09
	±0.0207	**0.4**	**83.52**	**90.00**	**80.58**

Case study: prediction of malignancy of ovarian tumors (9)

References:

- C. Lu, T. Van Gestel, J.A.K. Suykens, S. Van Huffel, I. Vergote, D. Timmerman, "Preoperative prediction of malignancy of ovarium tumor using least squares support vector machines", in *Artificial Intelligence in Medicine*, 28(3):281-306, 2003.
- J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002.
- T. Van Gestel *et al.*, "A Bayesian framework for Least Squares Support Vector Machine classifiers, Gaussian processes and kernel Fisher discriminant analysis," *Neural Computation*, vol. 15, no.5, pp. 1115-1148, 2002.
- D. Timmerman *et al.*, "Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses," *Ultrasound Obstet Gynecol*, vol. 13, pp. 17-25, 1999.
- D. Timmerman *et al.*, "Terms, Definitions and Measurements to describe the ultrasonographic features of adnexal tumors: a consensus opinion from the international ovarian tumor analysis (IOTA) group," *Ultrasound Obstet Gynecol*, vol. 16, pp. 500-505, 2000.

Case study: financial time series prediction (1)

- **Introduction:** Financial Time Series prediction has been a traditional research area for statistics, econometrics and lately, machine learning techniques. The advantages of having a good forecasting capability within the financial business is straightforward, although there is an important theoretical discussion about 'forecastability' of the financial markets (the so-called **Efficient Market Hypothesis**).

- In the financial community, the variance of a financial time series is known as **volatility**, and it has been shown that assuming a constant volatility may not be realistic enough. Here, a volatility estimation is obtained as a by-product of the LS-SVM model for the level of the series, and then it is used to improve the prediction in a final stage.

- **Reference:** Van Gestel, T., Suykens, J., Baestaens, D., Lambrechts, A., Lanckriet, G., Vandaele, B., De Moor, B., Vandewalle, J. "Financial Time Series Prediction Using Least Squares Support Vector Machines Within the Evidence Framework," *IEEE Transactions on Neural Networks*, 12(4):809-821, 2001.

Case study: financial time series prediction (2)

- **Problem and Data:** Prediction of the German DAX30 index, corrected by volatility considerations, in a one-step-ahead basis. The explanatory variables to be included in the model:
  - Lagged values of DAX30
  - US-30 years bond
  - S&P 500, FTSE, CAC40 (stocks indices)

- **Goal:** To predict the one-step-ahead DAX30 value, with volatility correction.

- **Methodology and Implementation:** Using the variables described above, the LS-SVM model was trained using Bayesian selection of hyperparameters. The procedure gives a probabilistic interpretation to the parameters, therefore it is possible to compute the volatility of the series. This volatitily is modelled by means of a second LS-SVM model.

Case study: financial time series prediction (3)

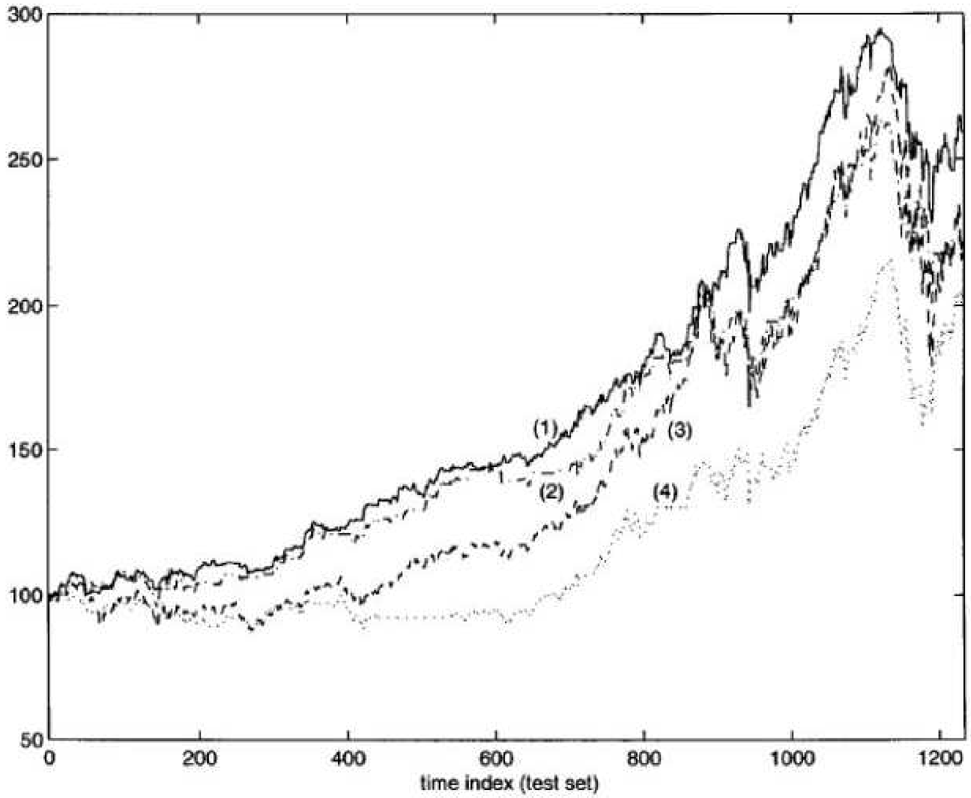
Results: The LS-SVM with Bayesian selection of hyperparameters allows to compute a prediction for the DAX30 index. The figure shows the cumulative profit for different trading strategies based on different models:

(1) LS-SVM, RBF kernel with volatility correction

(2) ARX model

(3) Buy-Hold strategy

(4) AR model


time index (test set)
(1)
(2)
(3)
(4)