Introduction to Statistical Machine Learning

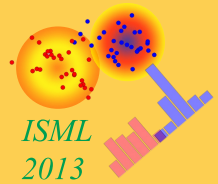
Christfried Webers

Statistical Machine Learning Group
NICTA
and
College of Engineering and Computer Science
The Australian National University

Canberra
February – June 2013

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

Christfried Webers
NICTA
The Australian National University

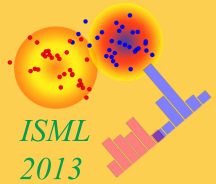


Outlines

Part XII

Sparse Kernel Machines

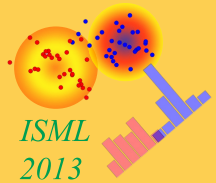
Christfried Webers
NICTA
The Australian National University



Sparse Kernel Machines - Motivation

- Nonlinear kernels extended our toolbox of methods considerably. Wherever an inner product was used in an algorithms, we can replace it with a kernel.
- Kernels act as a kind of 'similarity' measure and can be defined over graphs, sets, strings, and documents.
- But the kernel matrix is a square matrix with dimensions equal to the number of data points N. In order to calculate it, the kernel function must be evaluated for all pairs of training inputs.
- Sparse Kernel Machines implement learning algorithms which are based on sparse kernels. For prediction, these kernels are only evaluated at a subset of the training data.

©2013 Christfried Webers
NICTA
The Australian National University



Maximum Margin Classifiers

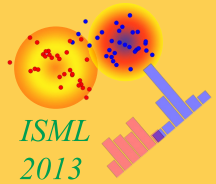
- Return to the two-class classification with a linear model

$$y(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + b$$

  where $\phi(\mathbf{x})$ is some fixed feature space mapping, and b is the bias.
- Training data are N input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ with corresponding targets $t_1, \ldots, t_N$ where $t_n \in \{-1, +1\}$.
- The class of a new point is predicted as $\text{sign}\,(y(\mathbf{x}))$.
- Assume there exists a linear decision boundary. That means, there exist $\mathbf{w}$ and b such that

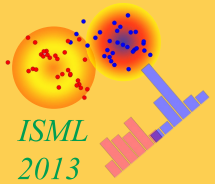
$$t_n \ \text{sign}\,(y(\mathbf{x}_n)) > 0 \qquad\qquad n = 1, \ldots, N.$$

Christfried Webers
NICTA
The Australian National University



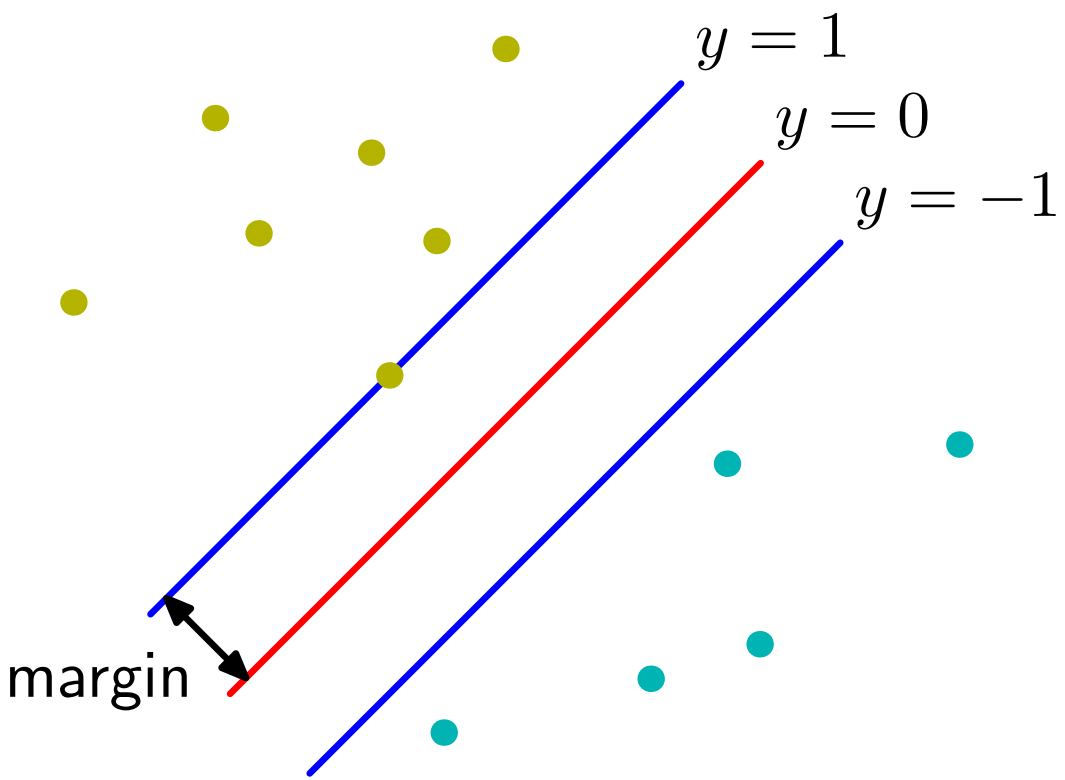
Maximum Margin Classifiers

- There may exist many solutions $\mathbf{w}$ and b for which the linear classifier perfectly separates the two classes.
- The perceptron algorithm can find a solution, but this depends on the initial choice of parameters.
- What is the decision boundary which results in the best generalisation (smallest generalisation error)?



Maximum Margin Classifiers

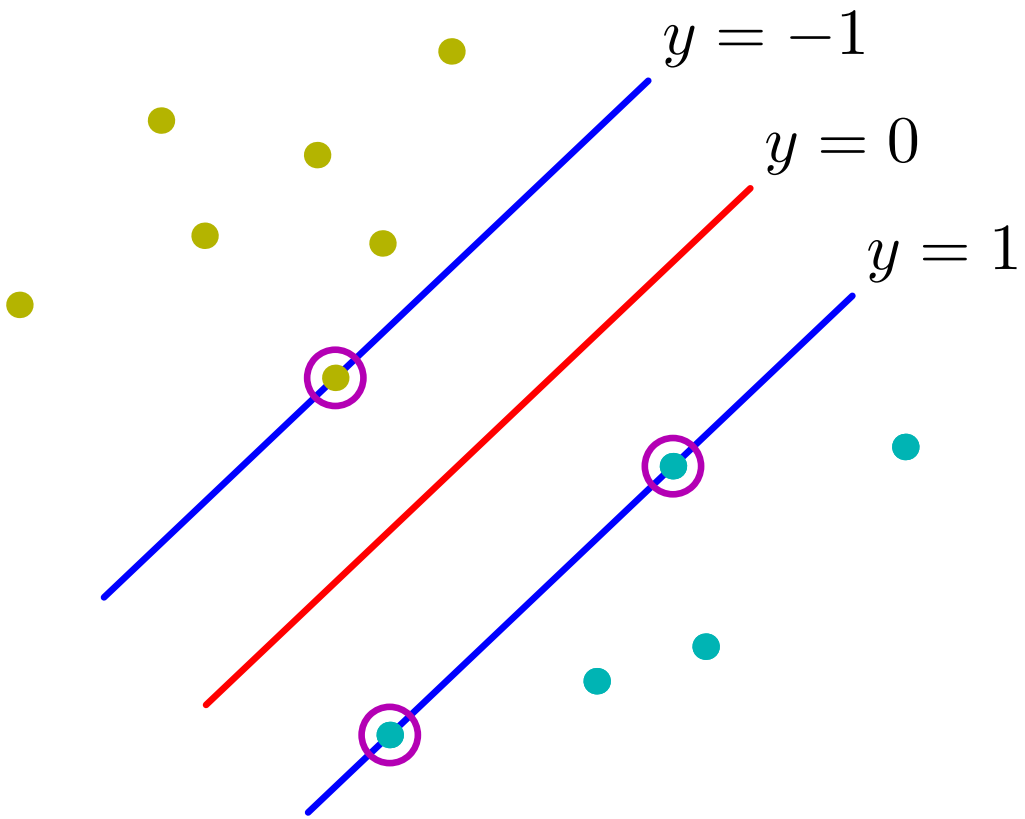
- The margin is the smallest distance between the decision boundary and any of the samples. (We will see later why $y = \pm 1$ on the margin.)

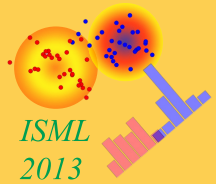


©2013 Christfried Webers
NICTA
The Australian National University

Maximum Margin Classifiers

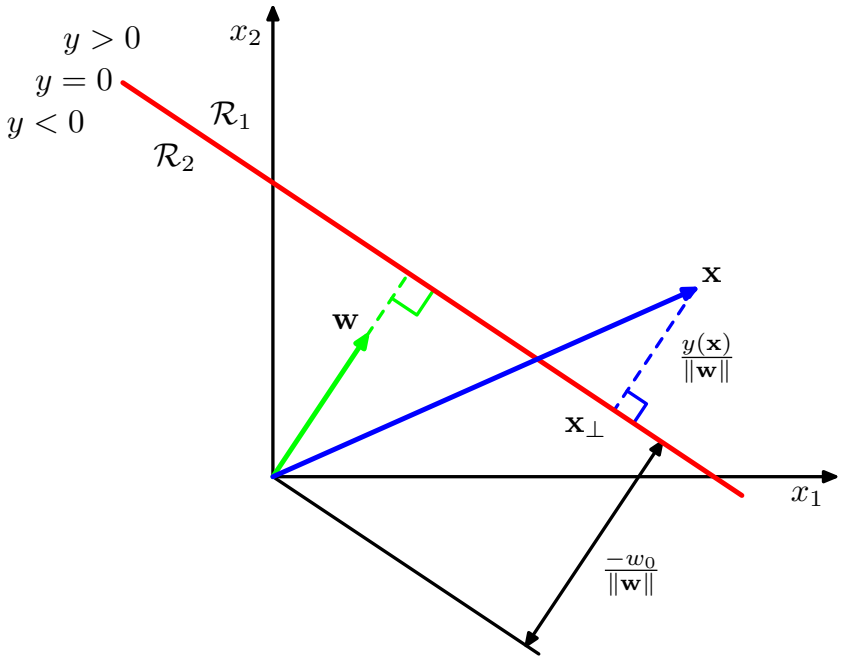
- Support Vector Machines choose the decision boundary which maximises the margin.





Reminder - Linear Classification

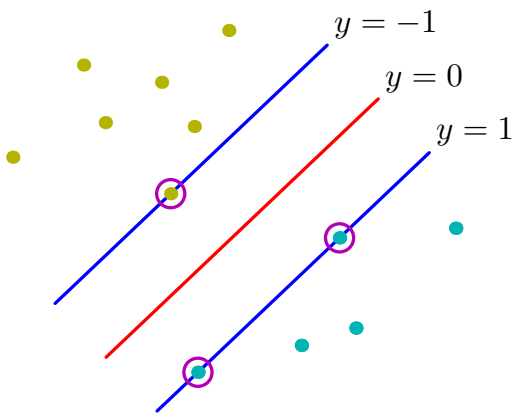
- $\mathbf{w}$ is orthogonal to the decision boundary ($0 = \mathbf{w}^T\mathbf{x_1} + w_0 = \mathbf{w}^T\mathbf{x_2} + w_0$ and $\mathbf{w}^T(\mathbf{x_1} - \mathbf{x_2}) = 0$.)
- Length of the projection of $\mathbf{x}$ onto $\mathbf{w}$ is $\frac{\mathbf{w}^T\mathbf{x}}{\|\mathbf{w}\|}$.
- If $\mathbf{x}$ sits on the decision boundary then $0 = y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ and therefore the distance from the origin orthogonal to the decision boundary (and parallel to $\mathbf{w}$ is $\frac{\mathbf{w}^T\mathbf{x}}{\|\mathbf{w}\|} = -\frac{w_0}{\|\mathbf{w}\|}$.
- The distance of $\mathbf{x}$ from the decision boundary is therefore $\frac{\mathbf{w}^T\mathbf{x}}{\|\mathbf{w}\|} - (-\frac{w_0}{\|\mathbf{w}\|}) = \frac{y(\mathbf{x})}{\|\mathbf{w}\|}$.



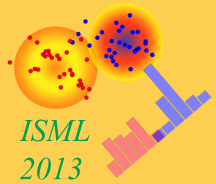
Maximum Margin Classifiers

- Support Vector Machines choose the decision boundary which maximises the smallest distance to samples in both classes.
- We calculated the distance of a point $\mathbf{x}$ from the hyperplane $y(\mathbf{x}) = 0$ as $|y(\mathbf{x})|/\|\mathbf{w}\|$.
- Perfect classification means $t_n y(\mathbf{x}_n) > 0$ for all n.
- Thus, the distance of $\mathbf{x}_n$ from the decision boundary is

$$\frac{t_n\, y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n\, (\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n) + b)}{\|\mathbf{w}\|}$$



©2013
Christfried Webers
NICTA
The Australian National University



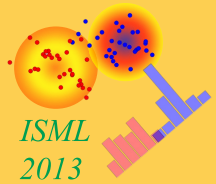
Maximum Margin Classifiers

- Support Vector Machines choose the decision boundary which maximises the smallest distance to samples in both classes.
- For the maximum margin solution, solve

$$\arg\max_{\mathbf{w},b} \left\{ \frac{1}{\|\mathbf{w}\|} \min_n \left[t_n \left(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n) + b \right) \right] \right\}.$$

- How can we solve this?

©2013 Christfried Webers
NICTA
The Australian National University



Maximum Margin Classifiers

- Maximum margin solution

$$\arg\max_{\mathbf{w},b}\left\{\frac{1}{\|\mathbf{w}\|}\min_{n}\left[t_n\left(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)+b\right)\right]\right\}$$

- Observation: Rescaling $\mathbf{w}\to\kappa\mathbf{w}$ and $b\to\kappa b$ does not change the distance from the hypersurface.
- Scale in such a way that for the closest point

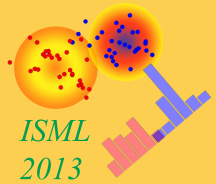
$$t_n\left(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)+b\right)=1.$$

- The canonical representation for the decision hyperplane is therefore

$$t_n\left(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)+b\right)\geq 1 \qquad n=1,\ldots,N.$$

- The constraints are active, if the equality holds, otherwise they are inactive.

Christfried Webers
NICTA
The Australian National University



Maximum Margin Classifiers

- Maximum margin solution

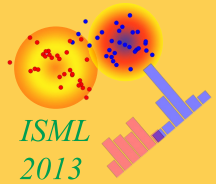
$$\arg\max_{\mathbf{w},b}\left\{\frac{1}{\|\mathbf{w}\|}\min_{n}\left[t_n\left(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)+b\right)\right]\right\}$$

- Transformed once

$$\arg\max_{\mathbf{w},b}\left\{\frac{1}{\|\mathbf{w}\|}\right\} \quad \text{s.t. } t_n\left(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)+b\right)\geq 1.$$

- Transformed again

$$\arg\min_{\mathbf{w},b}\left\{\frac{1}{2}\|\mathbf{w}\|^2\right\} \quad \text{s.t. } t_n\left(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)+b\right)\geq 1.$$



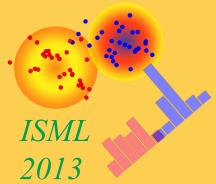
Maximum Margin Classifiers

- Quadratic Programming (QP) problem: minimise a quadratic function subject to linear constraints.

$$\arg\min_{\mathbf{w},b}\left\{\frac{1}{2}\|\mathbf{w}\|^2\right\} \quad \text{s.t. } t_n\left(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)+b\right)\geq 1.$$

- Introduce Lagrange multipliers $\{a_n \geq 0\}$ to get the Lagrangian

$$L(\mathbf{w},b,\mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n\left\{t_n\left(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)+b\right)-1\right\}.$$



Maximum Margin Classifiers

- Lagrangian

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n \left\{ t_n \left(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n) + b \right) - 1 \right\}.$$

- Derivatives with respect to $\mathbf{w}$ and b are

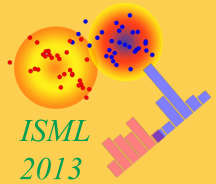
$$\mathbf{w} = \sum_{n=1}^{N} a_n \, t_n \, \boldsymbol{\phi}(\mathbf{x}_n) \qquad\qquad 0 = \sum_{n=1}^{N} a_n \, t_n$$

- Dual representation (again QP problem): Maximise

$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n \, a_m \, t_n \, t_m \, k(\mathbf{x}_n, \mathbf{x}_m)$$

$$\sum_{n=1}^{N} a_n \, t_n = 0$$

$$a_n \geq 0 \qquad\qquad n = 1, \ldots, N.$$



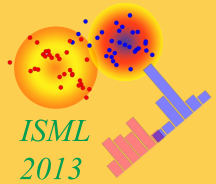
Maximum Margin Classifiers

- Kernel function

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\phi}(\mathbf{x}')$$

  must imply a positive definite kernel (Gram matrix), otherwise the optimisation problem is not bounded from above.
- Solving the optimisation problem is of the order $O(N^3)$ in the number of data points N. The original optimisation problem is of order $O(M^3)$ where M was the number of basis functions.
- Again, kernel allows us to use large or infinite number of basis functions (feature maps), but disadvantegeous for $M < N$.

Christfried Webers
NICTA
The Australian National University



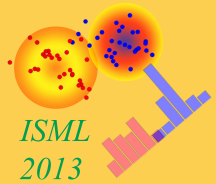
Maximum Margin Classifiers - Prediction

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n \, t_n \, k(\mathbf{x}, \mathbf{x}_n) + b$$

- The Karush-Kuhn-Tucker (KKT) conditions are

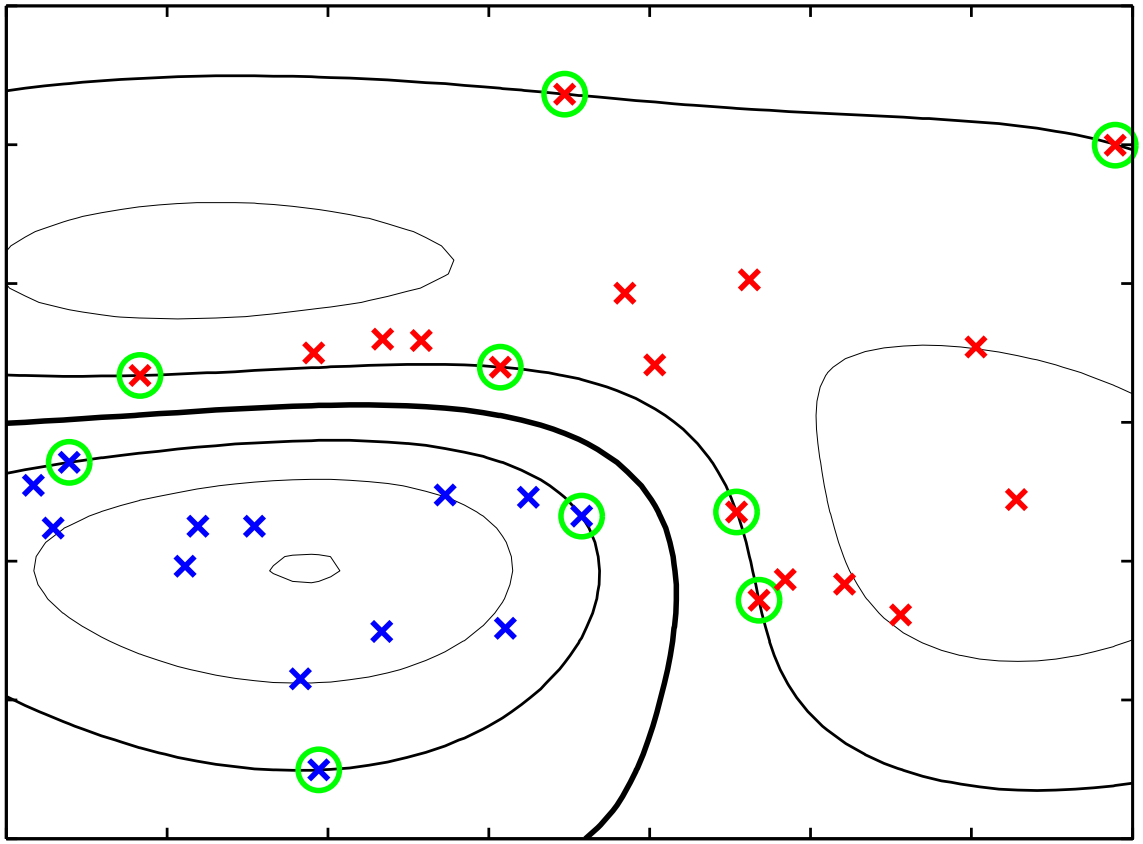
$$\begin{aligned} a_n &\geq 0 \\ t_n \, y(\mathbf{x}_n) - 1 &\geq 0 \\ a_n \left\{ t_n \, y(\mathbf{x}_n) - 1 \right\} &= 0. \end{aligned}$$

- Therefore, either $a_n = 0$ or $t_n \, y(\mathbf{x}_n) - 1 = 0$.
- If $a_n = 0$, no contribution to the prediction!
- After training, use only the set $\mathcal{S}$ of points for which $a_n > 0$ (and therefore $t_n \, y(\mathbf{x}_n) = 1$) holds.
- $\mathcal{S}$ contains only the support vectors.



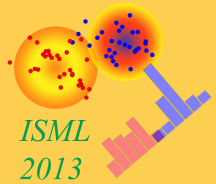
Maximum Margin Classifiers - Support Vectors

- $\mathcal{S}$ contains only the support vectors.



Decision and margin boundaries for a two-class problem using Gaussian kernel functions.

©2013 Christfried Webers
NICTA
The Australian National University



Maximum Margin Classifiers - Support Vectors

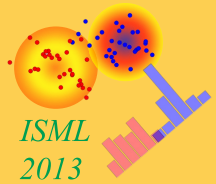
- Now for the bias b.
- Observe that for the support vectors, $t_n\, y(\mathbf{x}_n) = 1$.
- Therefore, we find

$$t_n \left(\sum_{m \in \mathcal{S}} a_m\, t_m\, k(\mathbf{x}_n, \mathbf{x}_m) + b \right) = 1$$

  and can use any support vector to calculate b.
- Numerically more stable: Multiply by t_n, observe $t_n^2 = 1$, and average over all support vectors.

$$b = \frac{1}{N_{\mathcal{S}}} \sum_{n \in \mathcal{S}} \left(t_n - \sum_{m \in \mathcal{S}} a_m\, t_m\, k(\mathbf{x}_n, \mathbf{x}_m) \right)$$

©2013 Christfried Webers
NICTA
The Australian National University



Maximum Margin Classifiers - Error

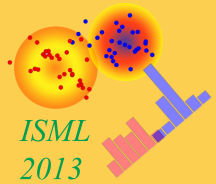
- Express the error for maximum margin classifiers with the help of a function

$$E_\infty(z) = \begin{cases} 0, & z \geq 0 \\ \infty, & z < 0 \end{cases}$$

enforcing the constraints as

$$\sum_{n=1}^{N} E_\infty(y(\mathbf{x}_n)t_n - 1)) + \lambda \|\mathbf{w}\|^2.$$

©2013 Christfried Webers
NICTA
The Australian National University

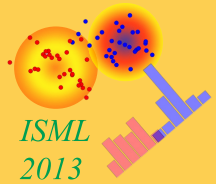


Overlapping Class distributions

- If the training data in the feature space are not linearly separable?
- Allow some data points to be on the 'wrong side' of the decision boundary.
- Increase a penalty with distance from the decision boundary.
- Assume, the penalty increases linearly with distance.
- Introduce slack variable $\xi_n \geq 0$ for each data point n .

$$\xi_n \begin{cases} = 0, & \text{data point is correctly classified and} \\ & \quad \text{on margin boundary or beyond} \\ < 1, & \text{data point is in correct margin} \\ = 1, & \text{data point is on the decision boundary} \\ > 1, & \text{data point is misclassified} \end{cases}$$

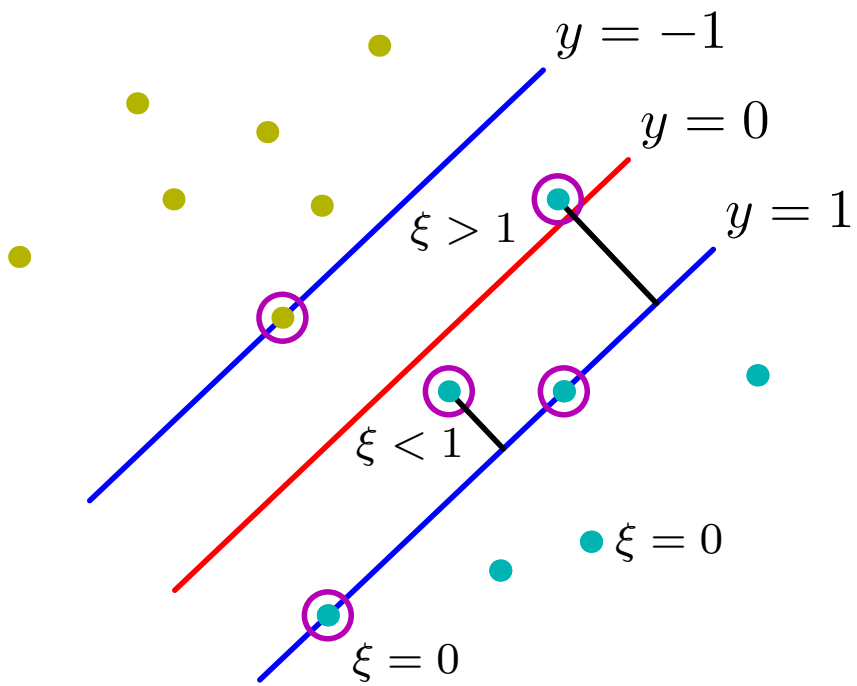
©2013 Christfried Webers
NICTA
The Australian National University



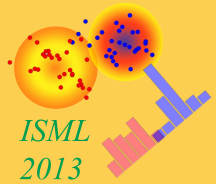
Overlapping Class distributions

- Introduce slack variable $\xi_n \geq 0$ for each data point n .

$$\xi_n = \begin{cases} 0, & \text{data point is correctly classified and} \\ & \text{on margin boundary or beyond} \\ |t_n - y(\mathbf{x})|, & \text{otherwise} \end{cases}$$



©2013
Christfried Webers
NICTA
The Australian National University



Overlapping Class distributions

- Constraints of the separable classification

$$t_n\, y(\mathbf{x}_n) \geq 1, \qquad n = 1, \ldots, N$$

change now to

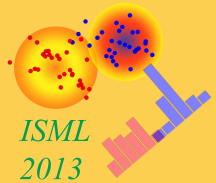
$$t_n\, y(\mathbf{x}_n) \geq 1 - \xi_n, \qquad n = 1, \ldots, N$$

- Minimise now

$$C \sum_{n=1}^{N} \xi_n + \frac{1}{2} \|\mathbf{w}\|^2$$

where C controls the trade-off between the slack variable penalty and the margin.

- Any misclassified point contributes $\xi_n >$, therefore $\sum_{n=1}^{N} \xi_n$ is an upper bound on the number of misclassified points.



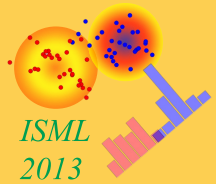
Overlapping Class distributions

- The Lagrangian with the Lagrange multipliers $\mathbf{a} = (a_1, \ldots, a_N)^T$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)^T$ is now

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \mathbf{a}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N} a_n\{t_n\, y(\mathbf{x}_n) - 1 + \xi_n\} - \sum_{n=1}^{N}\mu_n\xi_n.$$

- The KKT conditions for $n = 1, \ldots, N$ are

$$\begin{aligned} a_n &\geq 0 \\ t_n\, y(\mathbf{x}_n) - 1 + \xi_n &\geq 0 \\ a_n(t_n\, y(\mathbf{x}_n) - 1 + \xi_n) &= 0 \\ \mu_n &\geq 0 \\ \xi_n &\geq 0 \\ \mu_n\xi_n &= 0. \end{aligned}$$



Overlapping Class distributions

- The dual Lagrangian after eliminating $\mathbf{w}$, b, $\boldsymbol{\xi}$, and $\boldsymbol{\mu}$ is then

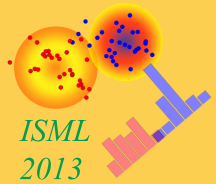
$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n \, a_m \, t_n \, t_m \, k(\mathbf{x}_n, \mathbf{x}_m)$$

$$\sum_{n=1}^{N} a_n \, t_n = 0$$

$$0 \leq a_n \leq C \qquad\qquad n = 1, \ldots, N.$$

- The only change from the separable case, is the box constraint via the parameter C.

Christfried Webers
NICTA
The Australian National University



Overlapping Class distributions

- Equivalent formulation by Schölkopf et. al. (2000)

$$\widetilde{L}(\mathbf{a}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n\, a_m\, t_n\, t_m\, k(\mathbf{x}_n, \mathbf{x}_m)$$

$$\sum_{n=1}^{N} a_n\, t_n = 0$$

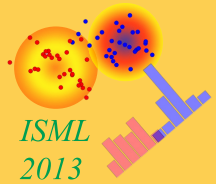
$$0 \leq a_n \leq 1/N$$

$$\sum_{n=1}^{N} a_n \geq \nu \qquad\qquad n = 1, \ldots, N.$$

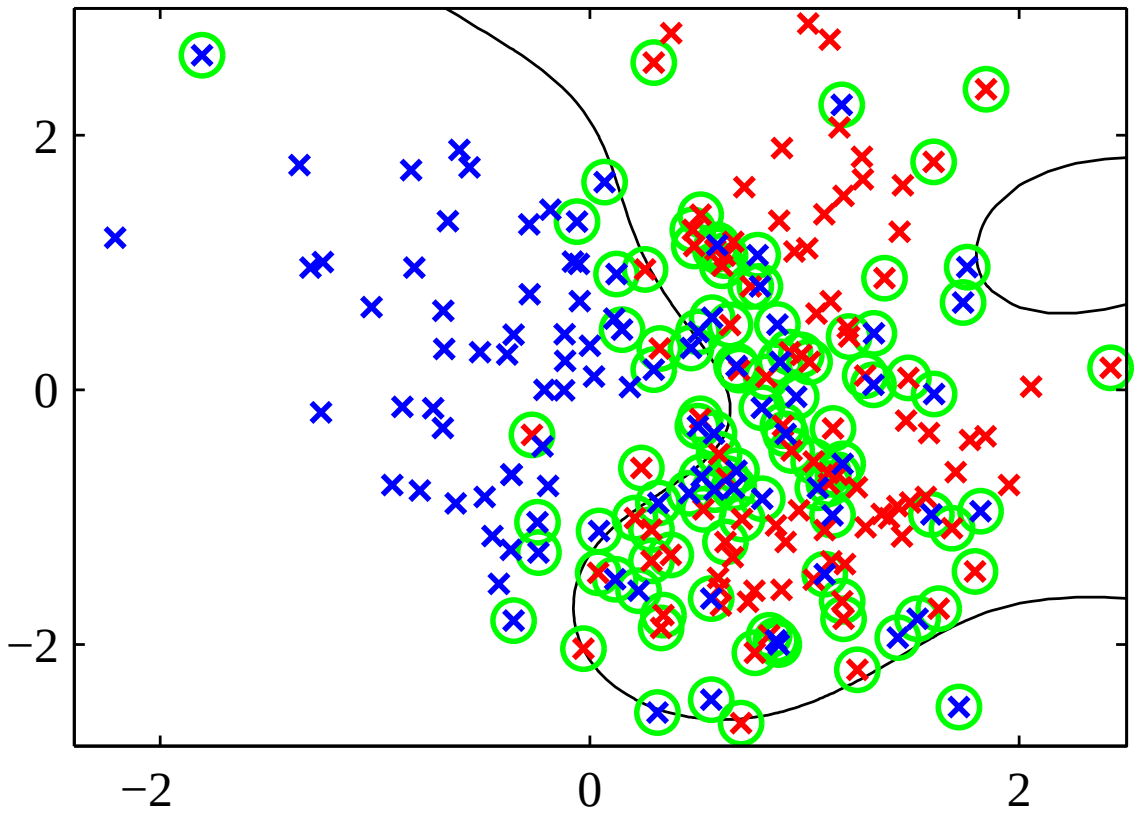
where ν is both

1. an upper bound on the fraction of margin errors, and,
2. a lower bound on the fraction of support vectors.

©2013 Christfried Webers
NICTA
The Australian National University

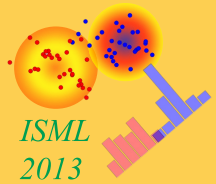


Overlapping Class distributions



The ν-SVM algorithm using Gaussian kernels $\exp(-\gamma\|\mathbf{x}-\mathbf{x}'\|^2)$ with $\gamma = 0.45$ applied to a nonseparable data set in two dimensions. Support vectors are indicated by circles.

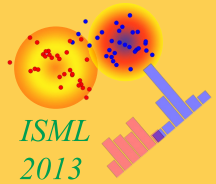
©2013 Christfried Webers
NICTA
The Australian National University



Support Vector Machines - Limitations

- Output are decisions, not posterior probabilities.
- Extension to classification with more than two classes is problematic.
- There is a complexity parameter C (or ν) which must be found (e.g. via cross-validation).
- Predictions are expressed as linear combinations of kernel functions that are centered on the training points.
- Kernel matrix is required to be positive definite.

Christfried Webers
NICTA
The Australian National University



Relevance Vector Machines

- Assume a Bayesian model as in Bayesian Linear Regression with the probability of the target t given by

$$p(t \mid \mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t \mid y(\mathbf{x}), \beta^{-1}),$$

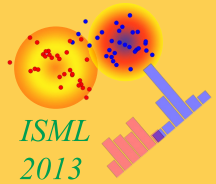
  and the linear model

$$y(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$$

  with fixed nonlinear basis functions $\boldsymbol{\phi}(\mathbf{x})$ (including a constant term to accomodate for the bias).
- But now the prior over $\mathbf{w}$ includes a different precision α_i for each of the components of $\mathbf{w}$

$$p(\mathbf{w} \mid \boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i \mid 0, \alpha_i^{-1}).$$

©2013 Christfried Webers NICTA The Australian National University



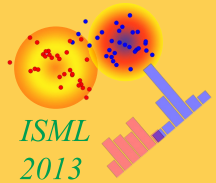
Relevance Vector Machines

- When using this prior for the weights $\mathbf{w}$

$$p(\mathbf{w} \mid \boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i \mid 0, \alpha_i^{-1})$$

  and maximising the evidence with respect to the hyperparameters α_i, a significant portion of the α_i will go to infinity.
- The corresponding w_i will therefore be concentrated at zero, and play no role for the prediction.

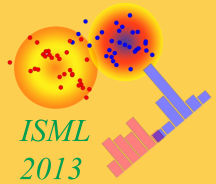
©2013 Christfried Webers
NICTA
The Australian National University



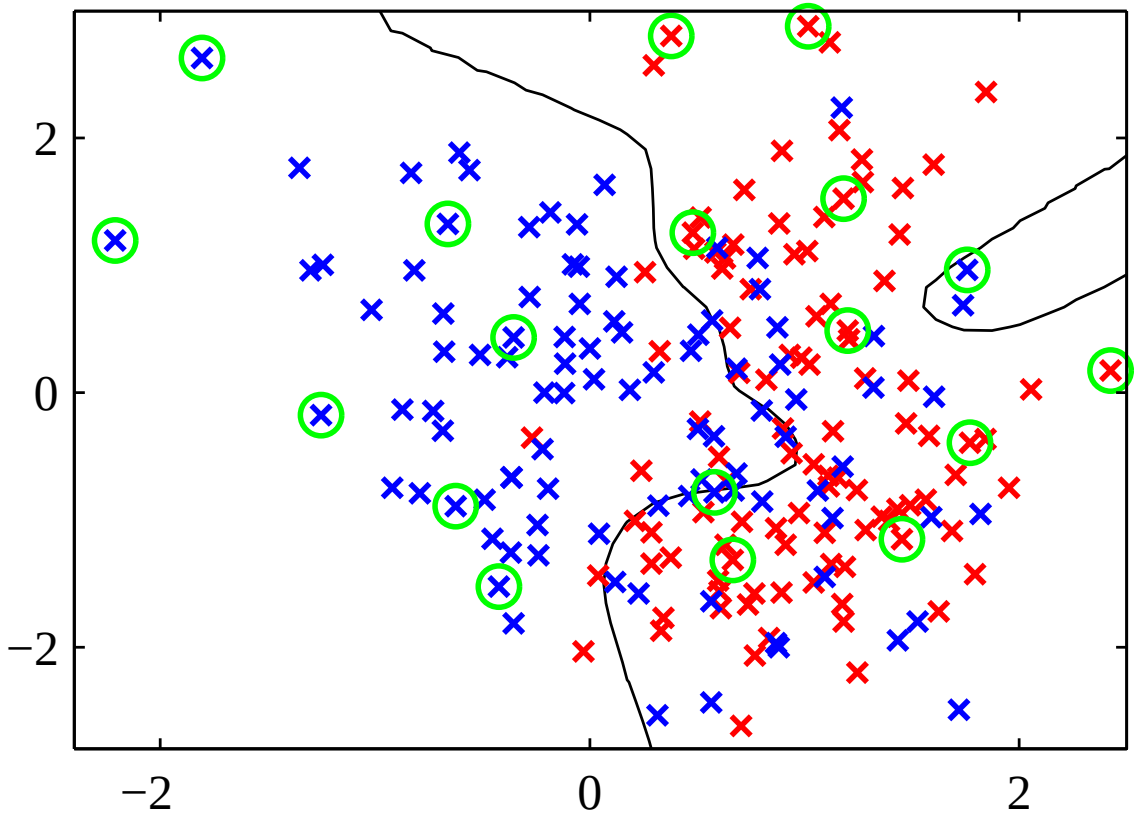
Relevance Vector Machines

- The Relevance Vector Machine iteratively calculates the α_i and β from the training data by optimising a nonconvex function.
- Disadvantages
  1. Nonconvex function means that the algorithm can get stuck in some local minimum.
  2. Training times can be longer than for SVM.
- Advantages
  1. As a Bayesian method, the parameters governing complexity and noise are determined from the input data. (Compare to SVM where we needed cross-validation).
  2. Number of relevance vectors is much smaller than the number of support vectors in a SVM.
  3. Therefore, prediction is much faster than with SVM.
  4. Extends well to multiclass learning.

©2013 Christfried Webers
NICTA
The Australian National University

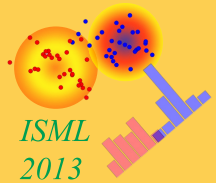


Relevance Vector Machines

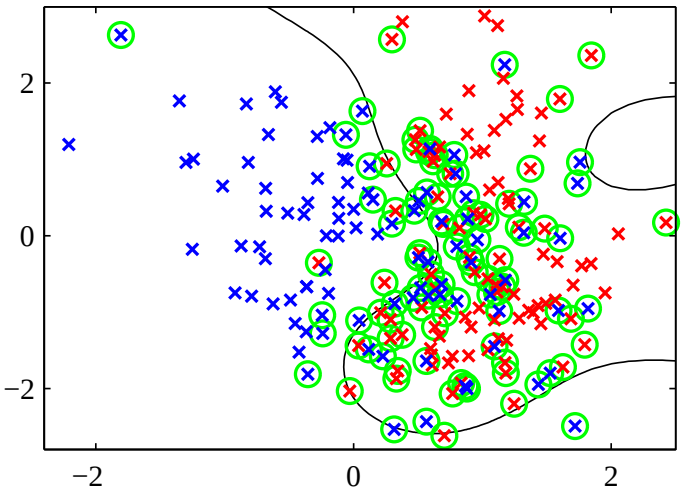


Decision boundaries for a two-class problem found by a Relevance Vector Machine.

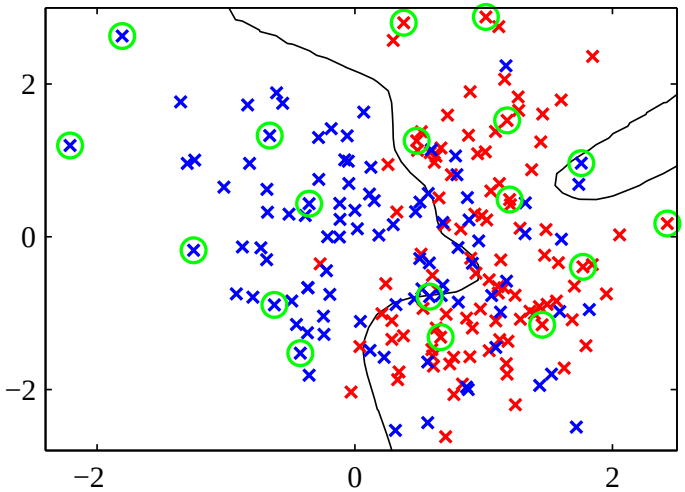
Christfried Webers
NICTA
The Australian National University



Relevance Vector Machines

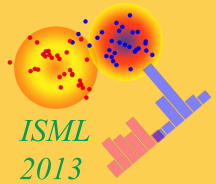


Support Vector Machine

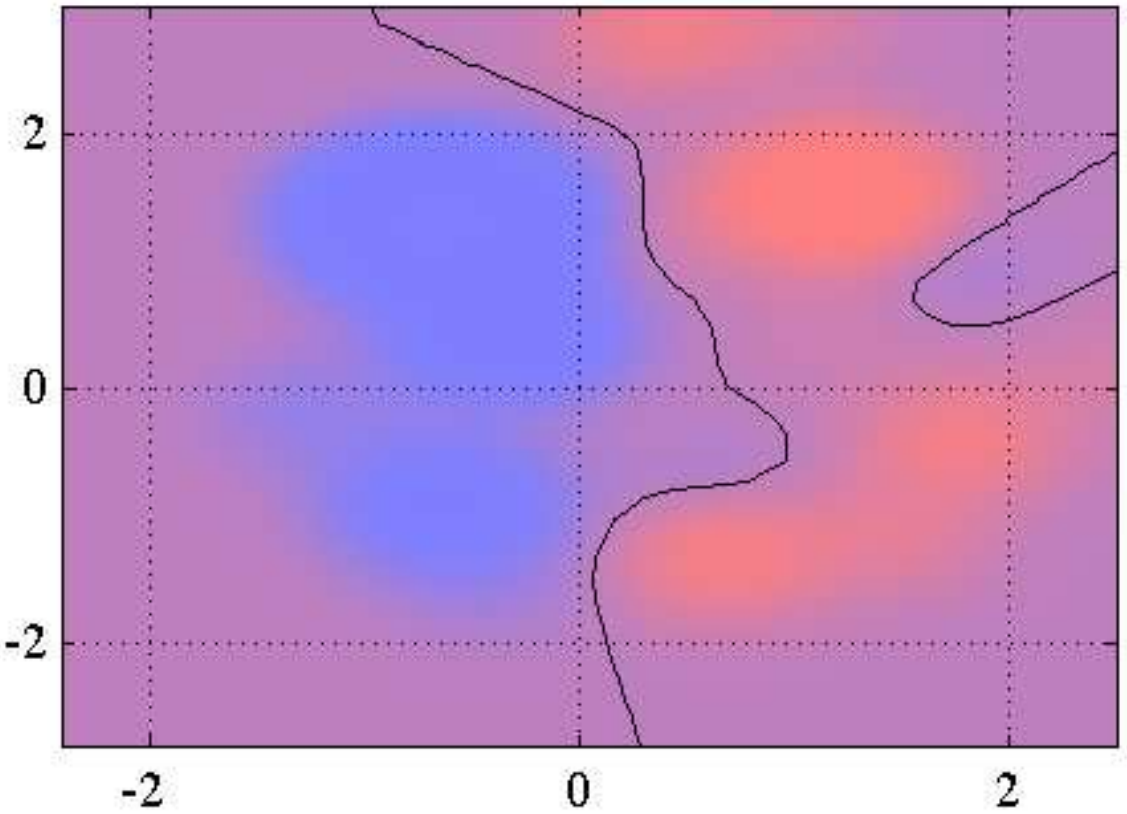


Relevance Vector Machine

Christfried Webers
NICTA
The Australian National University



Relevance Vector Machines



Decision boundaries and posterior probability for a two-class problem found by a Relevance Vector Machine.

Christfried Webers
NICTA
The Australian National University