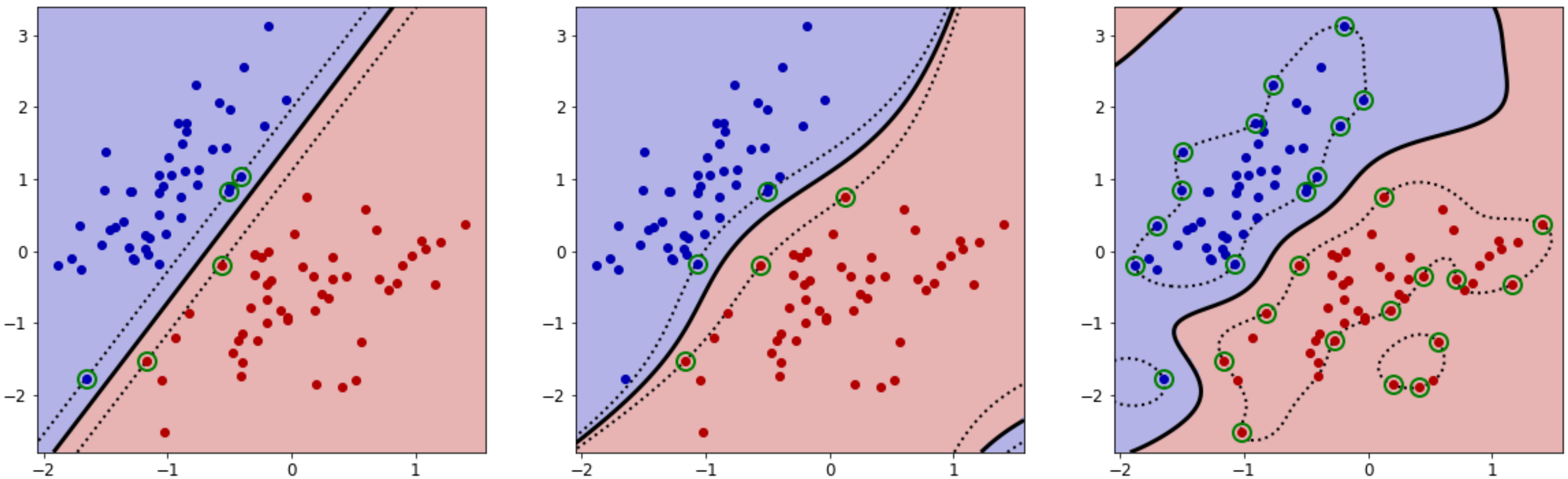
CS178:
Support Vector Machines



Prof. Alexander Ihler

Fall 2023

Support Vector Machines

Large Margin Learning

Lagrangian and Dual Forms

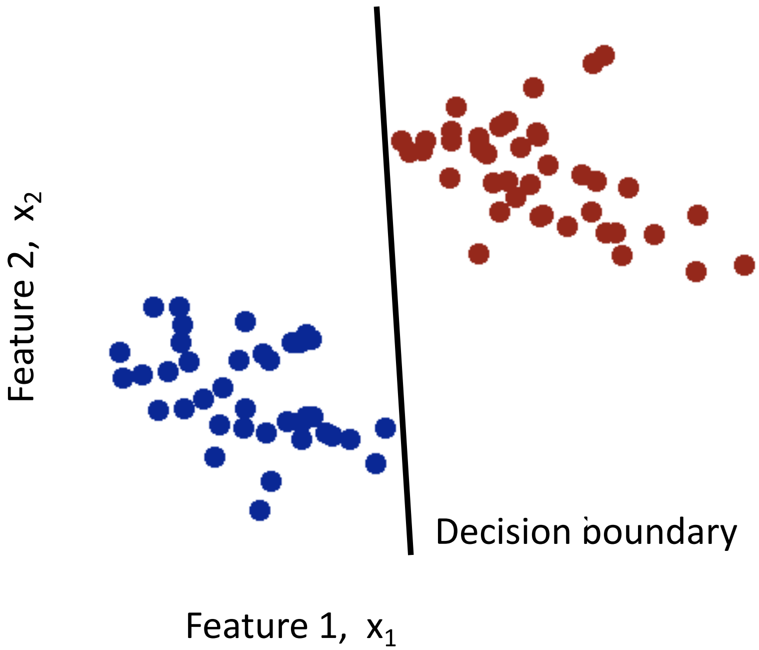
The Kernel Trick

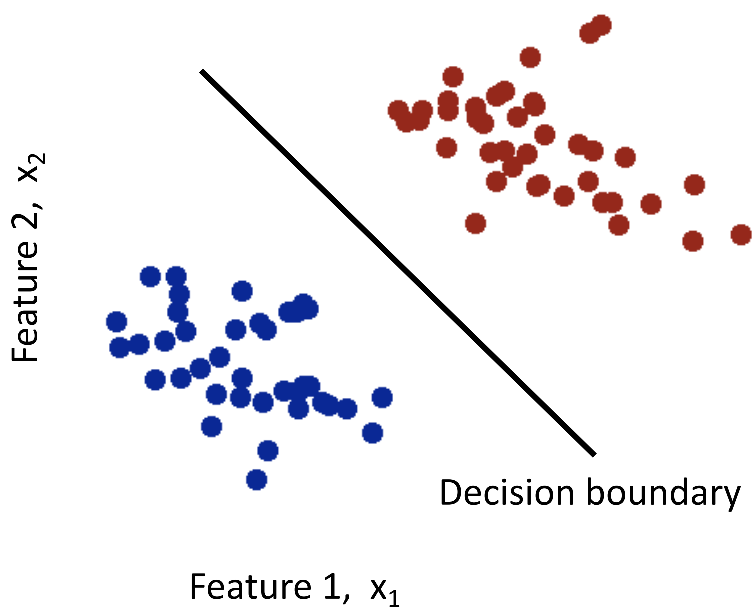
Linear classifiers

- Which decision boundary is "better"?
  - Both have zero training error (perfect training accuracy)
  - But, one of them seems intuitively better…

- How can we quantify "better",

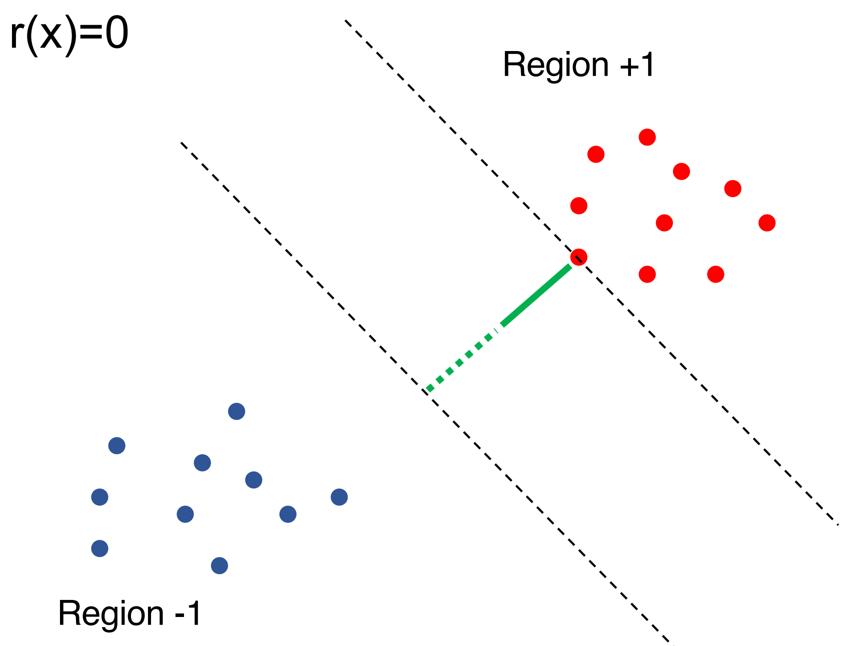
  and learn the "best" parameter settings?





One possible answer…

- Maybe we want to maximize our "margin"
  - How far away is the closest data point? (x 2)
- We could try to optimize this…
  - Move away from the closest point(s)
  - Notice: depends only on the closest data!
  - But – it will be easier to optimize if we do some math first…

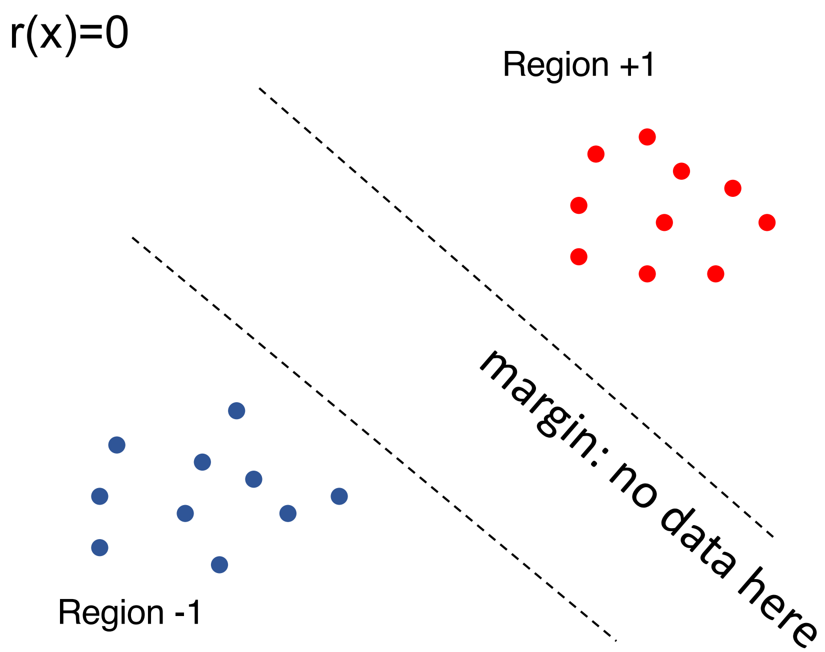


"Support vectors" – data points on margin

One possible answer…

- Maybe we want to maximize our “margin”

- Let’s relate “margin” to the model parameters
  - Define an equivalent optimization problem to the previous one
- Remove the “scale invariance”
  - Define class +1 in some region, class –1 in another
  - Call region in between the “margin”



“Support vectors” – data points on margin

One possible answer…

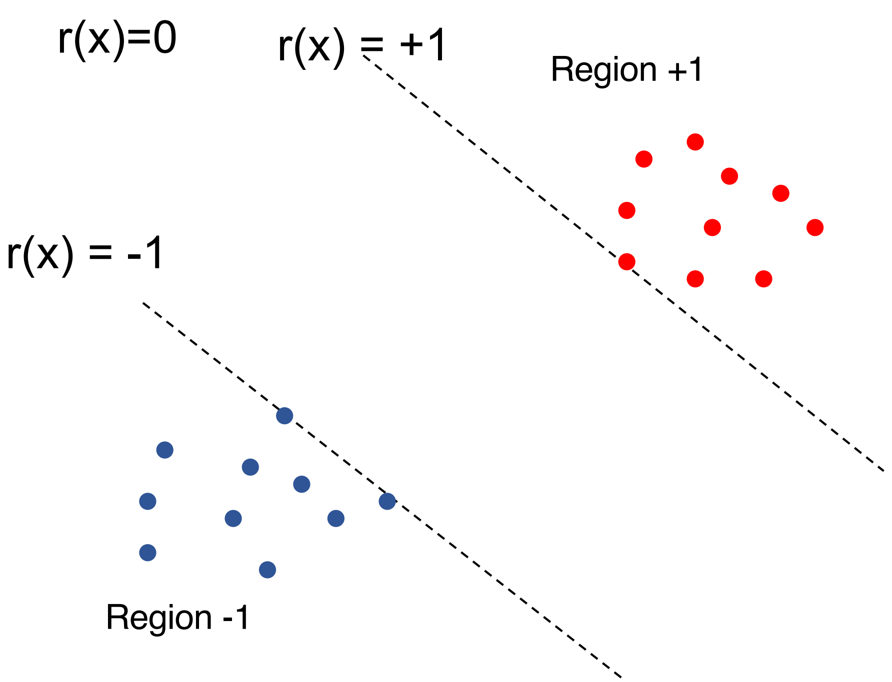
Notation change!

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots$$

$$\Downarrow$$

$$b + w_1 x_1 + w_2 x_2 + \ldots$$

- Maybe we want to maximize our "margin"
- Let's relate "margin" to the model parameters
  - Define an equivalent optimization problem to the previous one
- Remove the "scale invariance"
  - Define class +1 in some region, class –1 in another
  - Call region in between the "margin" – then, make it as **wide as possible**



We could define such a function:

$r(x) = w * x^T + b$

$r(x) > +1$ in region +1
$r(x) < -1$ in region –1

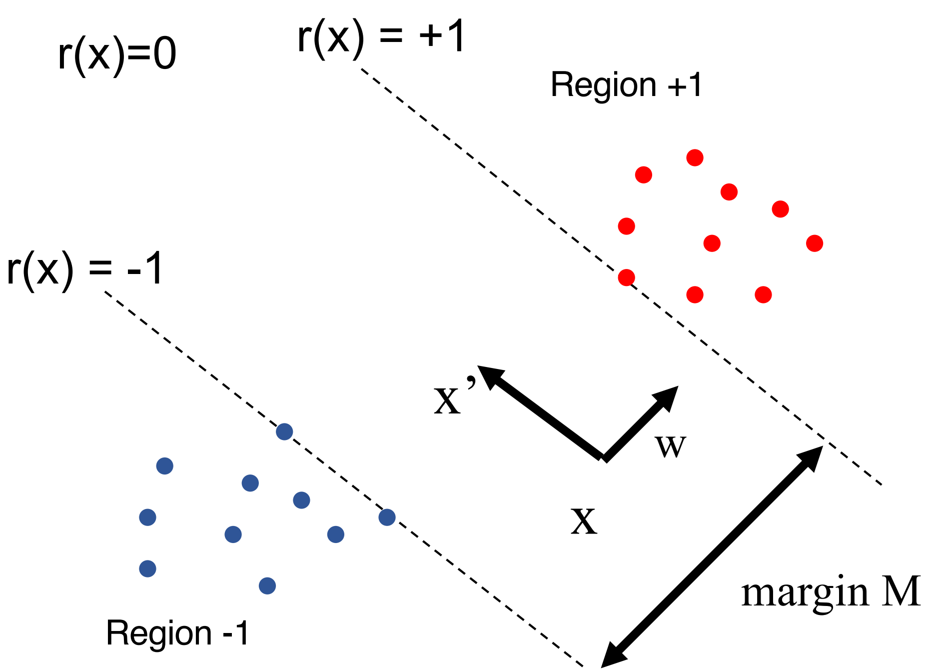
Passes through zero in center…

"Support vectors" – data points on margin

Computing the margin width

- Vector $\underline{w}=[w_1\ w_2\ \ldots]$ is perpendicular to the boundaries (why?)

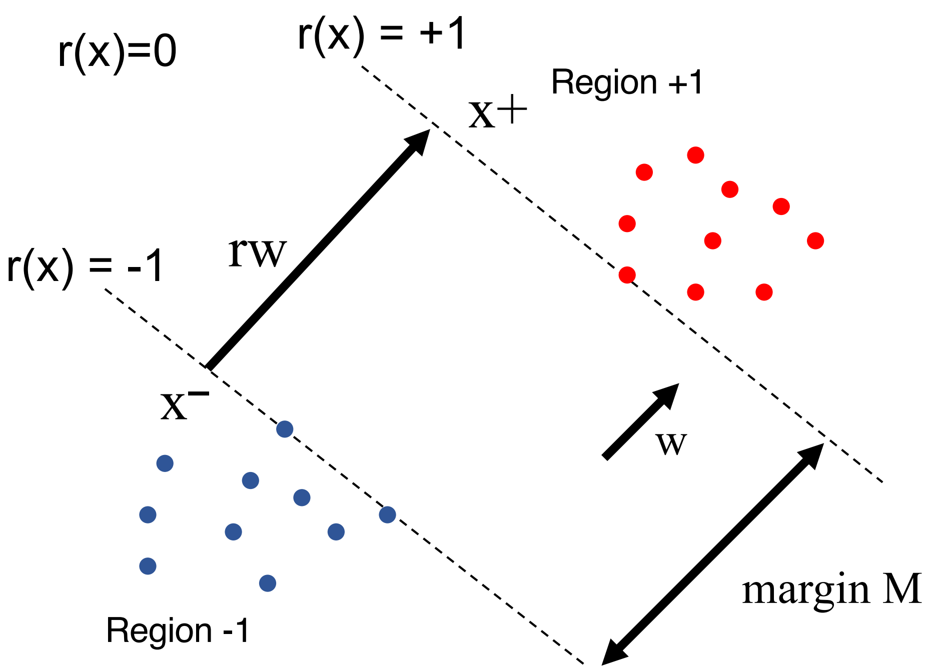
- $w\,x + b = 0$ & $w\,x' + b = 0$ => $w\,(x'-x) = 0$: orthogonal



Computing the margin width

- Vector $\underline{w}=[w_1\ w_2\ \ldots]$ is perpendicular to the boundaries
- Choose $\underline{x}^-$ st $f(\underline{x}^-) = -1$; let $\underline{x}^+$ be the closest point with $f(\underline{x}^+) = +1$
  - $\underline{x}^+ = \underline{x}^- + r * \underline{w}$ (why?)
- Closest two points on the margin also satisfy

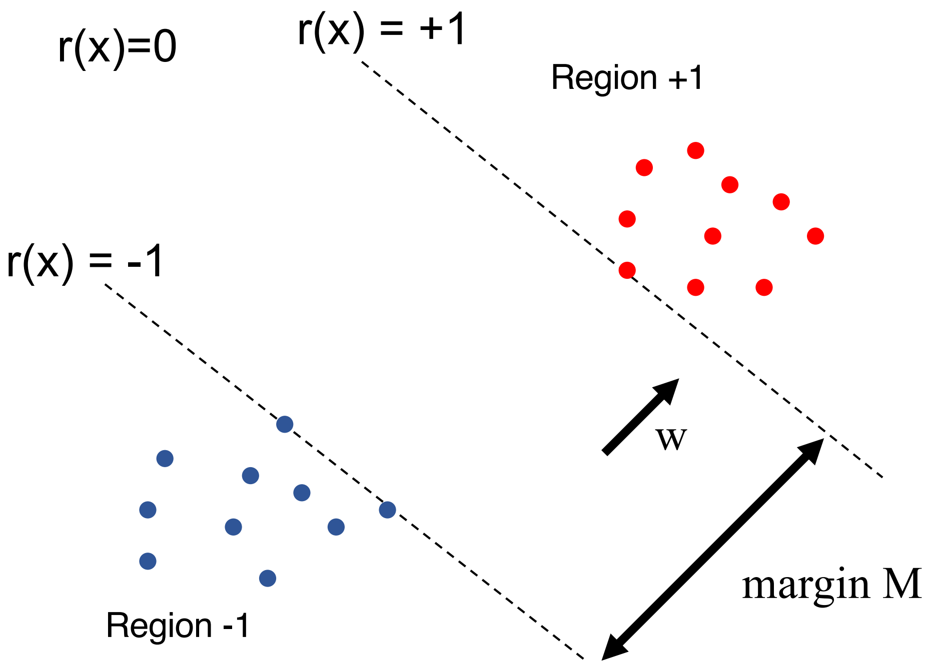
$$w \cdot x^- + b = -1 \qquad\qquad w \cdot x^+ + b = +1$$



Computing the margin width

- Vector $\underline{w}=[w_1\ w_2\ \ldots]$ is perpendicular to the boundaries
- Choose $\underline{x}^-$ st $f(\underline{x}^-) = -1$; let $\underline{x}^+$ be the closest point with $f(\underline{x}^+) = +1$
  - $\underline{x}^+ = \underline{x}^- + r * \underline{w}$
- Closest two points on the margin also satisfy

$$w \cdot x^- + b = -1 \qquad\qquad w \cdot x^+ + b = +1$$



$$w \cdot (x^- + rw) + b = +1$$
$$\Rightarrow r\|w\|^2 + w \cdot x^- + b = +1$$
$$\Rightarrow r\|w\|^2 - 1 = +1$$
$$\Rightarrow r = \frac{2}{\|w\|^2}$$

$$M = \|x^+ - x^-\| = \|rw\|$$
$$= \frac{2}{\|w\|^2}\|w\| = \frac{2}{\sqrt{w^T w}}$$

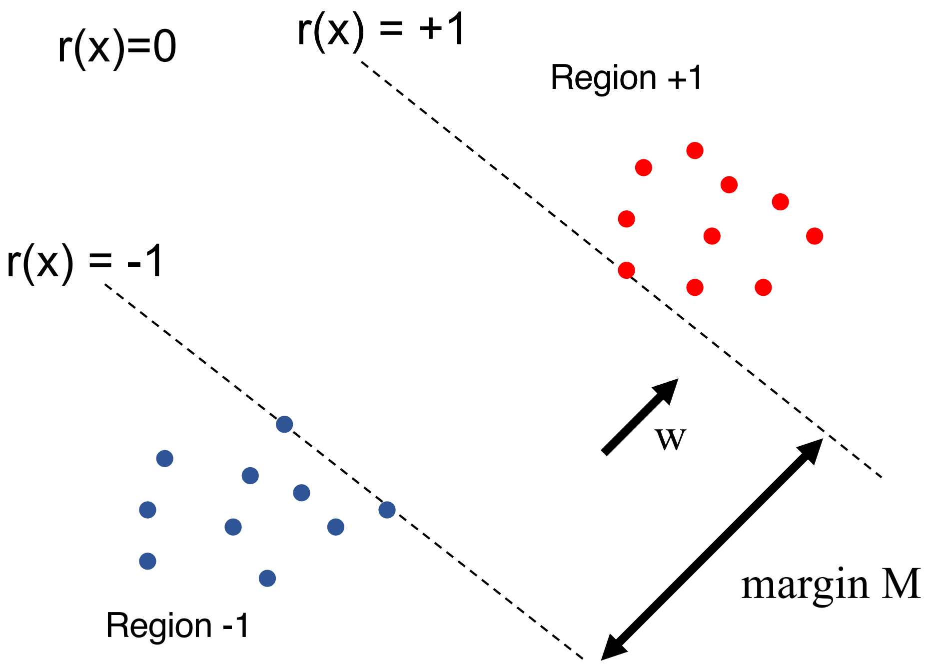
Maximum margin classifier

- Constrained optimization
  - Get all data points correct
  - Maximize the margin

$$w^* = \arg\max_w \frac{2}{\sqrt{w^T w}}$$

such that: "all data on the correct side of the margin"

This is an example of a **quadratic program**: quadratic cost function, linear constraints



Primal problem:

$$w^* = \arg\min_w \sum_j w_j^2$$

$s.t.$

$$y^{(i)} = +1 \Rightarrow \quad w \cdot x^{(i)} + b \geq +1$$
$$y^{(i)} = -1 \Rightarrow \quad w \cdot x^{(i)} + b \leq -1$$

(m constraints)

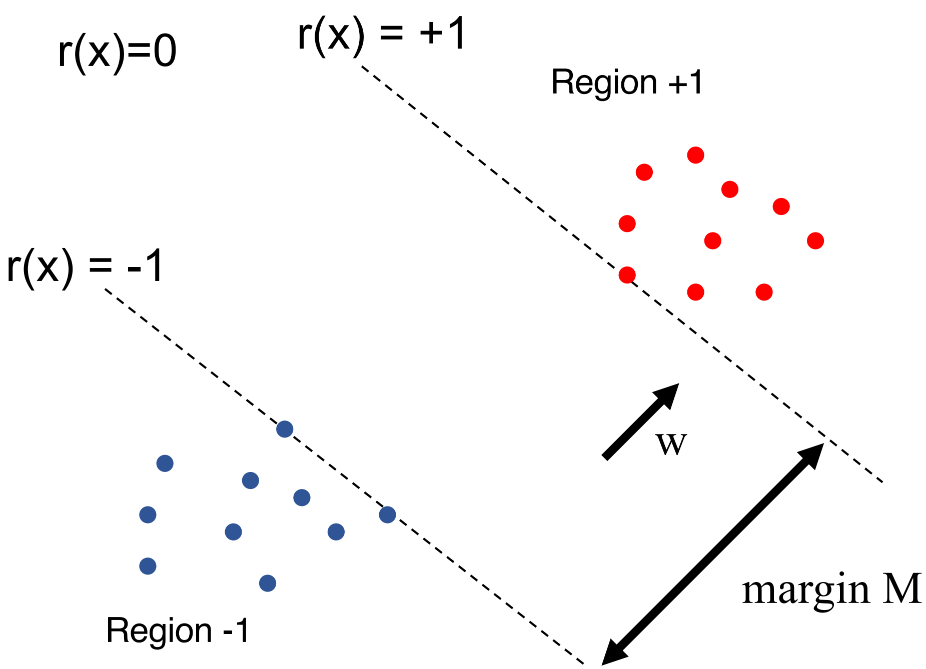
Maximum margin classifier

- Constrained optimization
  - Get all data points correct
  - Maximize the margin

$$w^* = \arg\max_w \frac{2}{\sqrt{w^T w}}$$

such that: "all data on the correct side of the margin"

This is an example of a **quadratic program**: quadratic cost function, linear constraints



Primal problem:

$$w^* = \arg\min_w \sum_j w_j^2$$

$s.t.$

$$y^{(i)}(\, w \cdot x^{(i)} + b\,) \geq +1$$

(m constraints)

A 1D Example

- Suppose we have three data points

  x = -3, y = -1

  x = -1, y = -1

  x = 2, y = 1

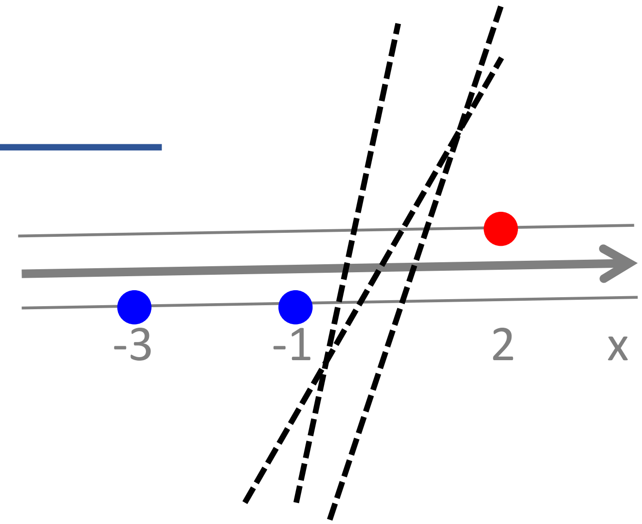
- Many separating perceptrons, T[ax+b]
  - Anything with ax+b = 0 between x=-1 and x=2

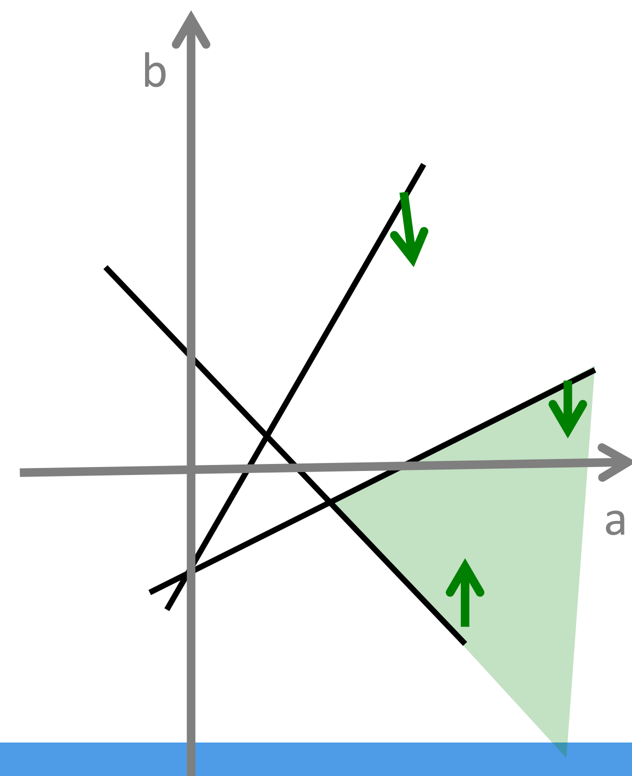
- We can write the margin constraints

  a (-3) + b < -1 => b < 3a - 1

  a (-1) + b < -1 => b < a - 1

  a (2) + b > +1 => b > -2a + 1





A 1D Example

- Suppose we have three data points

  x = -3, y = -1

  x = -1, y = -1

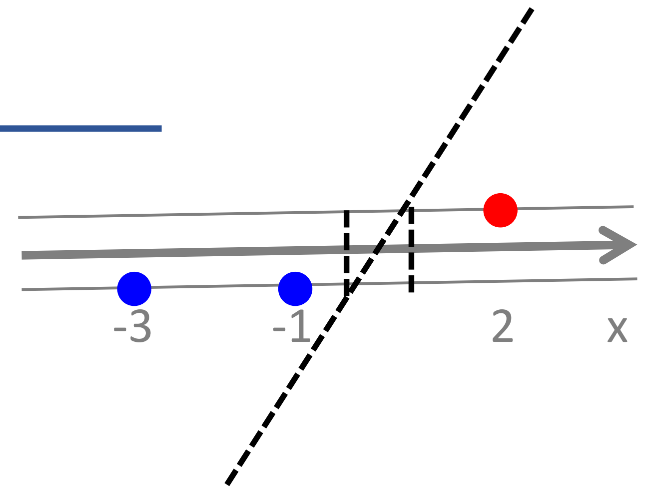
  x = 2, y = 1

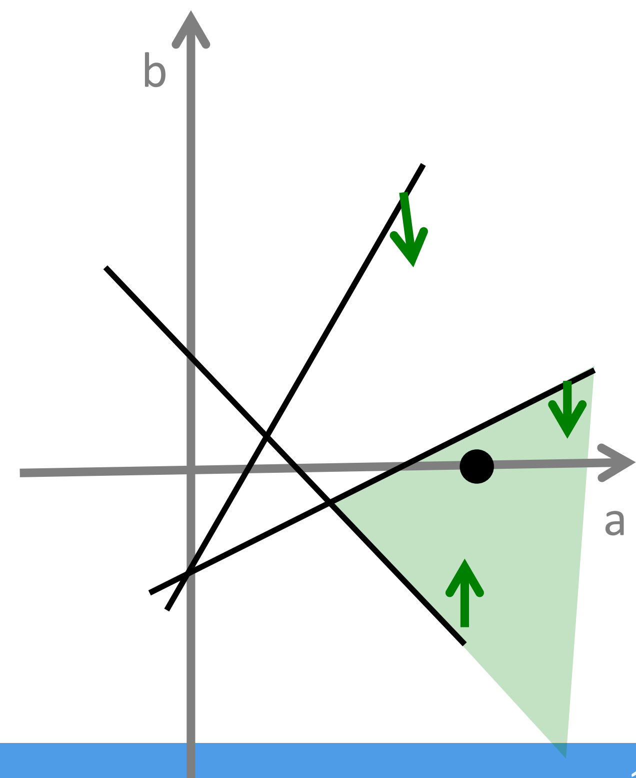
- Many separating perceptrons, T[ax+b]
  - Anything with ax+b = 0 between -1 and 2

- We can write the margin constraints

a (-3) + b < -1	=> b < 3a - 1
a (-1) + b < -1	=> b < a - 1
a (2) + b > +1	=> b > -2a + 1

- Ex: a = 2, b = 0





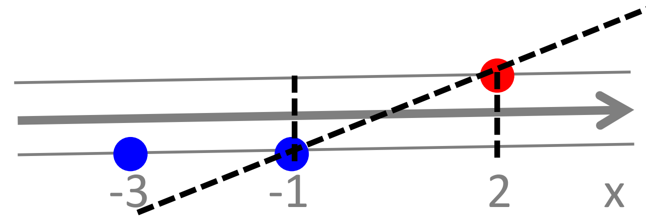
A 1D Example

- Suppose we have three data points

  x = -3, y = -1

  x = -1, y = -1

  x = 2, y = 1



- Many separating perceptrons, T[ax+b]
  - Anything with ax+b = 0 between -1 and 2

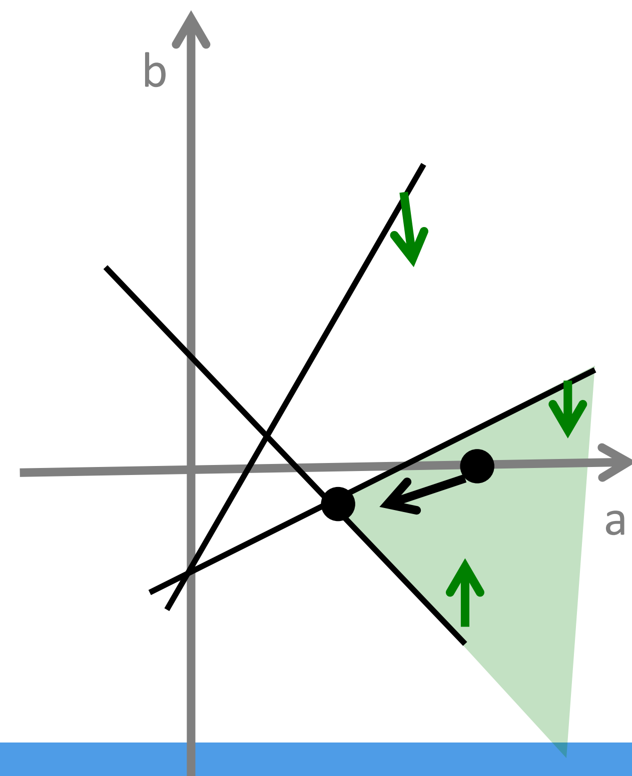
- We can write the margin constraints

  a (-3) + b < -1 => b < 3a - 1

  a (-1) + b < -1 => b < a - 1

  a (2) + b > +1 => b > -2a + 1

- Ex: a = 2, b = 0
- Minimize ||a|| => a = .66, b = -.33
  - Two data on the margin; constraints "tight"



Support Vector Machines

Large Margin Learning

Lagrangian and Dual Forms

The Kernel Trick

Lagrangian optimization

- Want to optimize constrained system:

θ = (w,b)

$$\underbrace{w^* = \arg\min_{w,b} \sum_j w_j^2}_{f(\theta)} \qquad s.t. \qquad \underbrace{1 - y^{(i)}(\, w \cdot x^{(i)} + b\,)}_{g_i(\theta)} \le 0$$

$$g_i(\theta) \le 0$$

- Introduce Lagrange multipliers α (one per constraint)

$$\theta^* = \arg\min_{\theta} \max_{\alpha \ge 0} f(\theta) \;+\; \sum_i \alpha_i \, g_i(\theta)$$

  - Can optimize θ, α jointly over a simpler constraint set (initialization easy)
  - For inner max: $g_i(\theta) \le 0 \;\; : \;\; \alpha_i = 0$

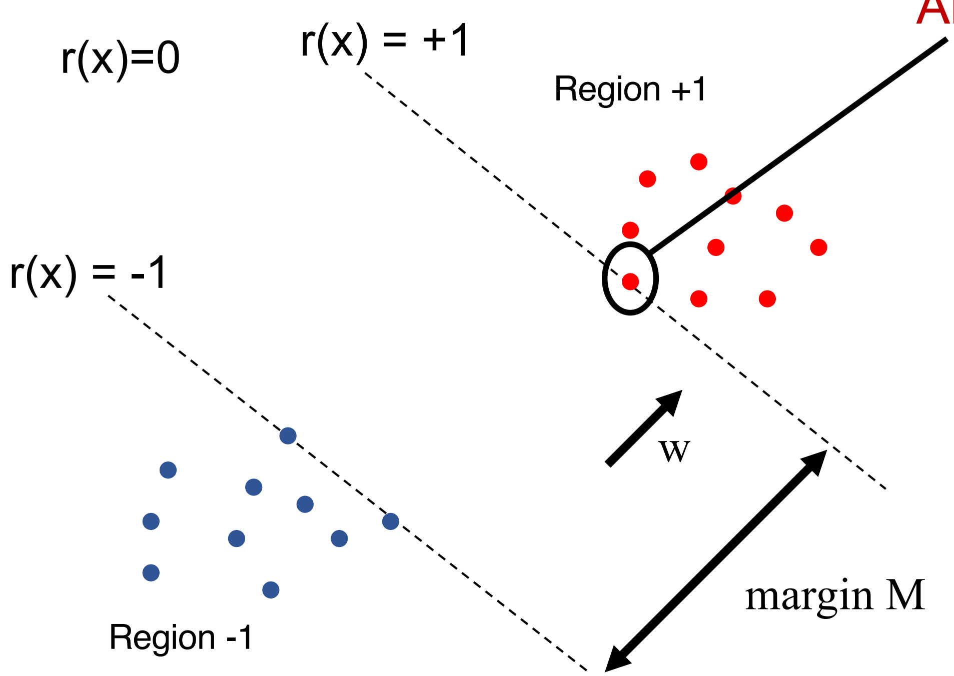
    $g_i(\theta) > 0 \;\; : \;\; \alpha_i \to +\infty$

  - Any optimum of the original problem is a saddle point of the new
  - KKT complementary slackness: $\alpha_i > 0 \;\Rightarrow\; g_i(\theta) = 0$

Optimization

- Use Lagrange multipliers
  - Enforce inequality constraints

$$w^* = \arg\min_w \max_{\alpha \geq 0} \frac{1}{2} \sum_j w_j^2 + \sum_i \alpha_i \left(1 - y^{(i)} (w \cdot x^{(i)} + b) \right)$$



Alphas > 0 only on the margin: "support vectors"

Stationary conditions wrt w:

$$w^* = \sum_i \alpha_i y^{(i)} x^{(i)}$$

and since any support vector has y = wx + b,

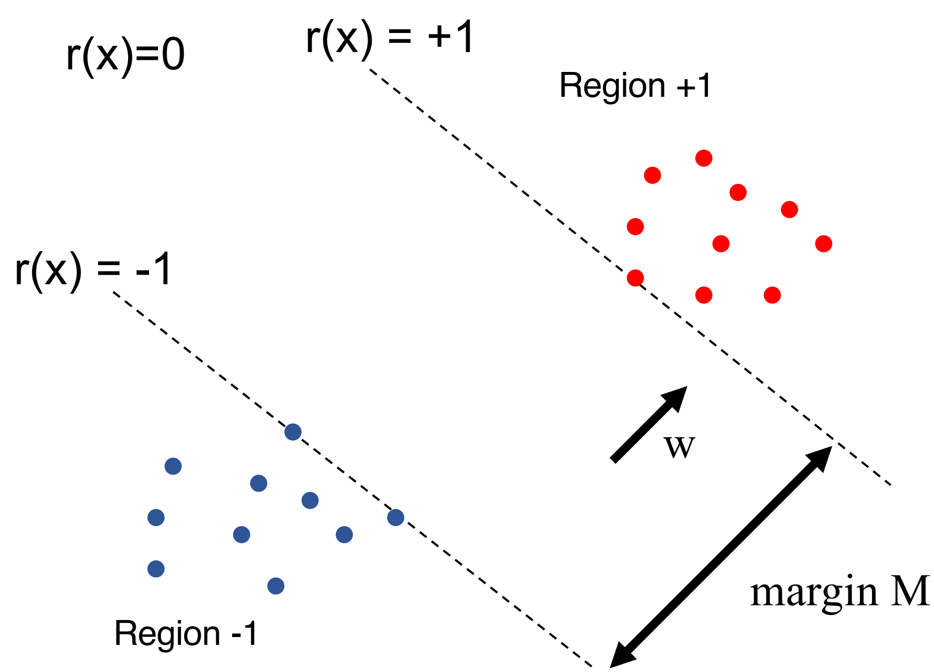
$$b = \frac{1}{Nsv} \sum_{i \in SV} \left(y^{(i)} - w \cdot x^{(i)} \right)$$

Dual form

- Use Lagrange multipliers
  - Enforce inequality constraints
  - Use solution w* to write solely in terms of alphas:

$$\max_{\alpha \geq 0} \sum_i \left[\alpha_i - \frac{1}{2} \sum_j \alpha_i \alpha_j \, y^{(i)} y^{(j)} \left(x^{(i)} \cdot x^{(j)} \right) \right]$$

$$\text{s.t.} \ \sum_i \alpha_i y^{(i)} = 0$$ (since derivative wrt b = 0)



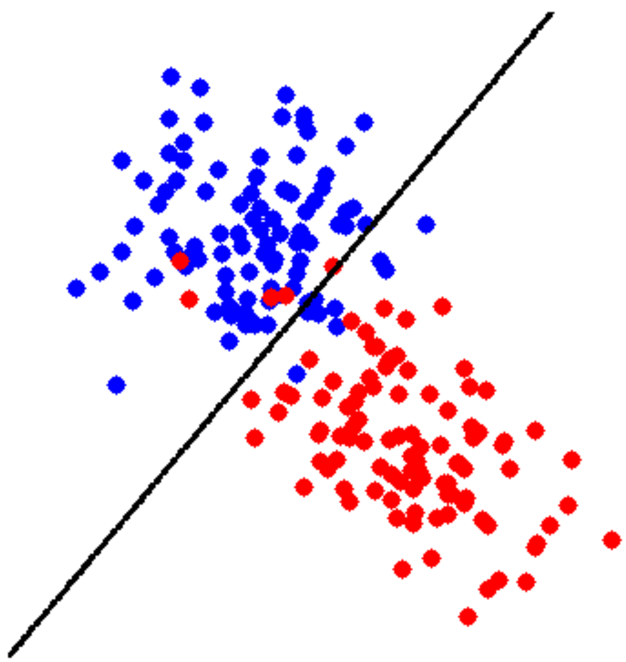
Another quadratic program:
optimize m vars with 1+m (simple) constraints
cost function has m^2 dot products

$$w^* = \sum_i \alpha_i y^{(i)} x^{(i)}$$

$$b = \frac{1}{Nsv} \sum_{i \in SV} \left(y^{(i)} - w \cdot x^{(i)} \right)$$

Maximum margin classifier

- What if the data are not linearly separable?
  - Want a large "margin": $\min_w \sum_j w_j^2$ Want low error: $\min_w \sum_i J(y^{(i)}, \; w \cdot x^{(i)} + b)$
  - "Soft margin" : introduce slack variables for violated constraints



$$w^* = \arg\min_{w,\epsilon} \sum_j w_j^2 + R \sum_i \epsilon^{(i)}$$

$s.t.$

$$y^{(i)}(\, w^T x^{(i)} + b\,) \geq +1 - \epsilon^{(i)}$$ (violate margin by ϵ)

$$\epsilon^{(i)} \geq 0$$

Assigns "cost" R proportional to distance from margin
Another quadratic program!

Soft margin SVM

$$w^* = \arg\min_{w,\epsilon} \sum_j w_j^2 + R \sum_i \epsilon^{(i)}$$

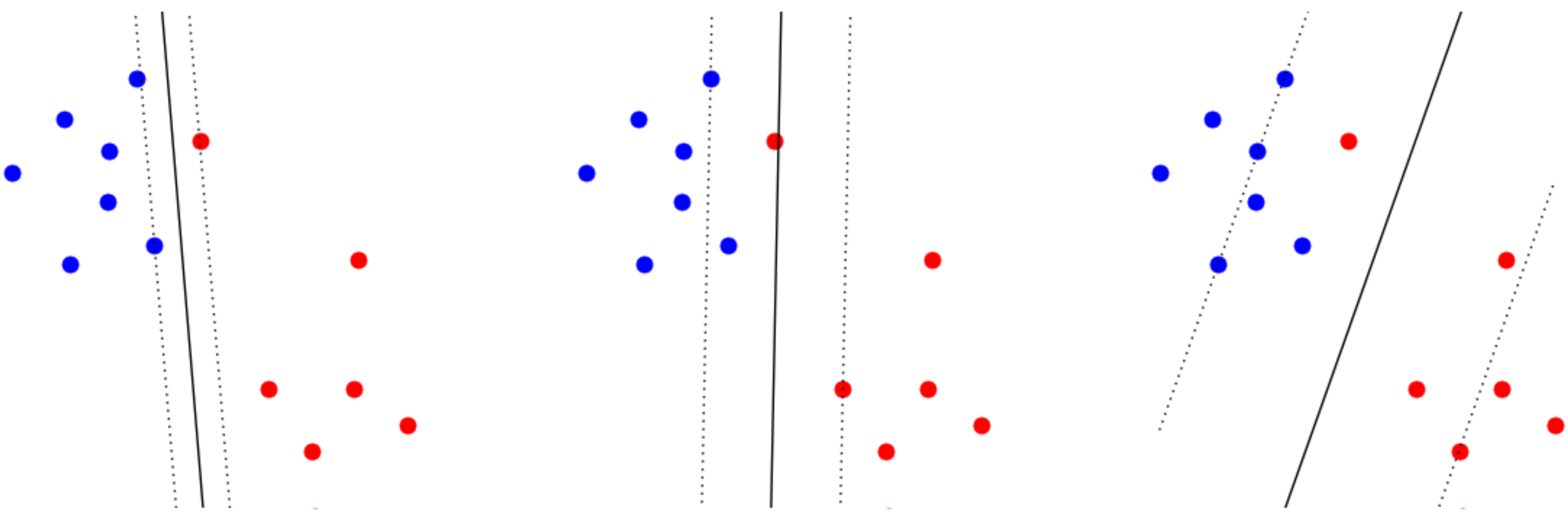
$s.t.$

$$y^{(i)}(w^T x^{(i)} + b) \geq +1 - \epsilon^{(i)}$$

$$\epsilon^{(i)} \geq 0$$

- Large margin vs. Slack variables
- R large = hard margin
- R smaller
  - A few wrong predictions; boundary farther from rest

$R = R_0$ $\quad$ $R = 10^{-2} R_0$ $\quad$ $R = 10^{-4} R_0$



Maximum margin classifier

- Soft margin optimization:

$$w^* = \arg\min_{w,\epsilon} \sum_j w_j^2 + R \sum_i \epsilon^{(i)}$$

  - For *any* weights w, we can choose ϵ to satisfy constraints

$$y^{(i)} (w^T x^{(i)} + b) \geq +1 - \epsilon^{(i)}$$

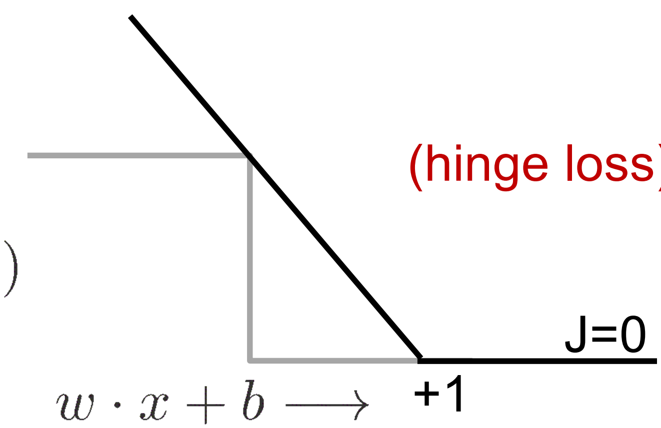
  - Write ϵ* as a function of w (call this J) and optimize directly

- J = distance from the "correct" place

$$J_i = \max[\, 0 \,,\, 1 - y^{(i)} (w \cdot x^{(i)} + b) \,]$$

$$w^* = \arg\min_w \frac{1}{R} \sum_j w_j^2 + \sum_i J_i(y^{(i)} \,,\, w \cdot x^{(i)} + b)$$

(L2 regularization on the weights)



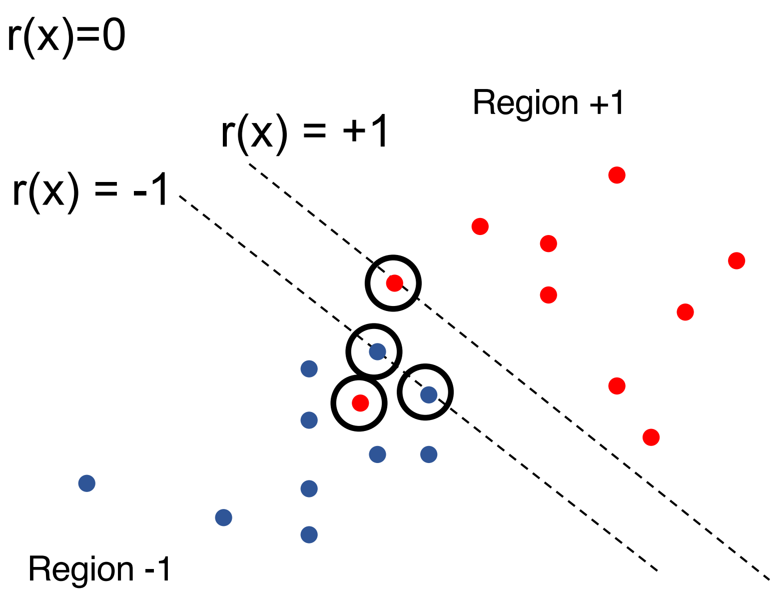
Dual form

- Soft margin dual:

$$\max_{0 \le \alpha \le R} \sum_i \alpha_i - \frac{1}{2} \sum_j \alpha_i \alpha_j \, y^{(i)} y^{(j)} \underbrace{\left(x^{(i)} \cdot x^{(j)} \right)}_{K_{ij}}$$

K_{ij} measures "similarity" of x_i and x_j (their dot product)

$$\text{s.t.} \quad \sum_i \alpha_i y^{(i)} = 0$$



Support vectors now data on or past margin…

Prediction:

$$\hat{y} = w^* \cdot x + b = \sum_i \alpha_i y^{(i)} x^{(i)} \cdot x + b$$

$$w^* = \sum_i \alpha_i y^{(i)} x^{(i)}$$

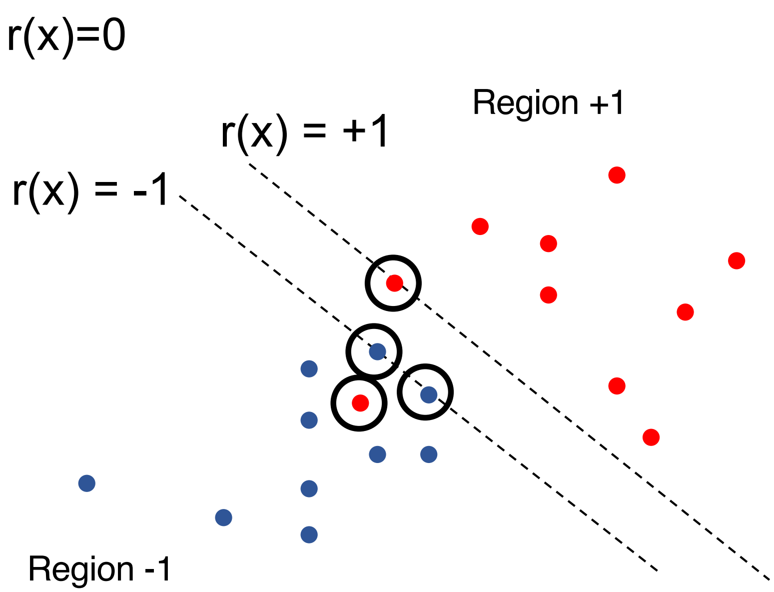
$$b = \ldots$$

More complicated; can solve e.g. using any $\alpha \in$(0,R)

Support Vectors

- The **support vectors** are data points *i* with non-zero weight α_i:
  - Points with minimum margin (on optimized boundary)
  - Points which violate margin constraint, but are still correctly classified
  - Points that are misclassified

  For all other training data, features have **no impact** on learned weights!



Support vectors now data on or past margin…

Prediction:

$$\hat{y} = w^* \cdot x + b = \sum_i \alpha_i y^{(i)} x^{(i)} \cdot x + b$$

$$w^* = \sum_i \alpha_i y^{(i)} x^{(i)}$$

$$b = \dots$$

More complicated; can solve e.g. using any $\alpha \in$ (0,R)

Parametric vs Instance-Based Learners

- Some learners are instance based
  - Save some/all examples, use them to make subs. decisions
  - Ex: KNN

- Other learners are parametric
  - Save some fixed # of parameters
  - Ex: linear regression; perceptrons
  - Contrast with nonparametric: # parameters grows with data size

- SVMs show duality of viewpoint
  - Primal representation: save weight vector w
  - Dual representation: save example data (SVs)

  Either representation gives exactly the same predictions

  In different settings, one or the other may be more efficient!

If a learner can reproduce data points exactly, does it matter if it stores those data points as “x’s” or in “w”?

[Platt, 1998]

Sequential Minimal Optimization (SMO)

- Out-of-the-box QP solvers not very good for SVMs
- Faster: optimize dual QP coordinate-wise over pairs (α_i,α_j)
  - Pick α_i, α_j s.t. α_i violates KKT conditions
  - Solve constrained QP over just (α_i,α_j)

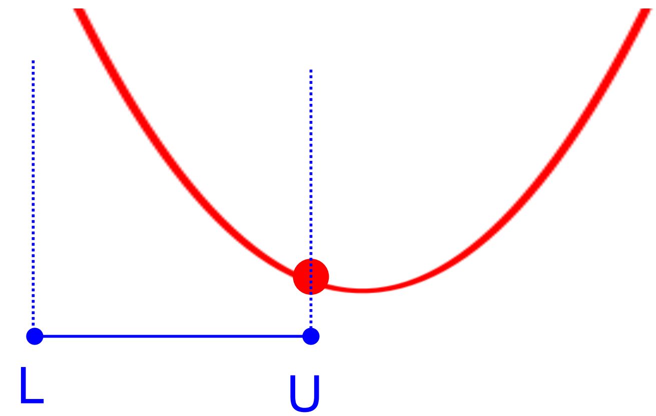
$$\underbrace{\max_{0 \le \alpha \le R}}_{(3)} \sum_i \alpha_i - \frac{1}{2} \sum_j \underbrace{\alpha_i \alpha_j \, y^{(i)} y^{(j)} \left(x^{(i)} \cdot x^{(j)}\right)}_{(2)} \quad \text{s.t.} \quad \overbrace{\sum_i \alpha_i y^{(i)} = 0}^{(1)}$$

(1) $\Rightarrow$ $\alpha_i = c - \alpha_j \, y^{(j)} / y^{(j)}$

(2) Quadratic function of α_j

(1), (3) $\Rightarrow$ $\alpha_j \in [L, U]$

so both α_i, α_j in [0,R]



Multi-class SVMs

- Use standard multi-class linear prediction, 0/1 loss:

$$\hat{y} = f(x;\theta) = \arg\max_y \ \theta \cdot \Phi(x,y)$$

$$\Phi(x,y) = [\ \mathbb{1}[y=0]\,\Phi(x)\ ,\ \mathbb{1}[y=1]\,\Phi(x)\ ,\ \ldots\ ]$$

- Hinge-like loss / slack variable optimization:

$$w^* = \arg\min_{w,b,\epsilon} \sum_j w_j^2 + R \sum_i \epsilon^{(i)}$$

$$w^T\Phi(x^{(i)}, y^{(i)}) - w^T\Phi(x^{(i)}, y) \geq 1 - \epsilon^{(i)} \qquad \forall y \neq y^{(i)}$$

- Can introduce class-specific loss function: $\Delta(y, \hat{y})$

$$w^T\Phi(x^{(i)}, y^{(i)}) - w^T\Phi(x^{(i)}, y) \geq \Delta(y^{(i)}, y) - \epsilon^{(i)} \qquad \forall y \neq y^{(i)}$$

  - Reduces to earlier form for 0/1 loss: $\Delta(y,\hat{y}) = \mathbb{1}[y \neq \hat{y}]$
  - Again, can optimize as QP (e.g., SMO) or hinge-like loss (e.g., SGD)

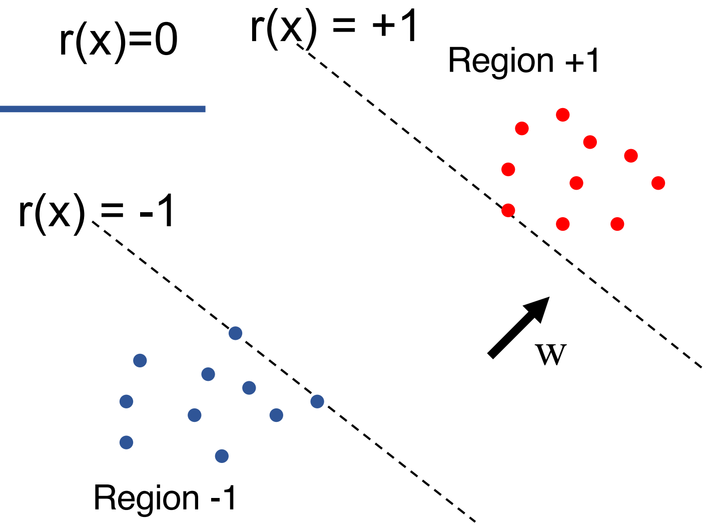
Support Vector Machines

Large Margin Learning

Lagrangian and Dual Forms

The Kernel Trick

Linear SVMs



- So far, looked at linear SVMs:
  - Expressible as linear weights "w"
  - Linear decision boundary
- Dual optimization for a linear SVM:

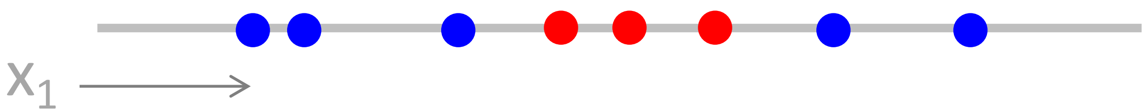
$$\max_{0 \le \alpha \le R} \sum_i \alpha_i - \frac{1}{2} \sum_j \alpha_i \alpha_j \, y^{(i)} y^{(j)} \, (x^{(i)} \cdot x^{(j)}) \qquad \text{s.t. } \sum_i \alpha_i y^{(i)} = 0$$

- Depend on pairwise dot products:
  - Kij measures "similarity", e.g., 0 if orthogonal $K_{ij} = x^{(i)} \cdot x^{(j)}$

Adding features

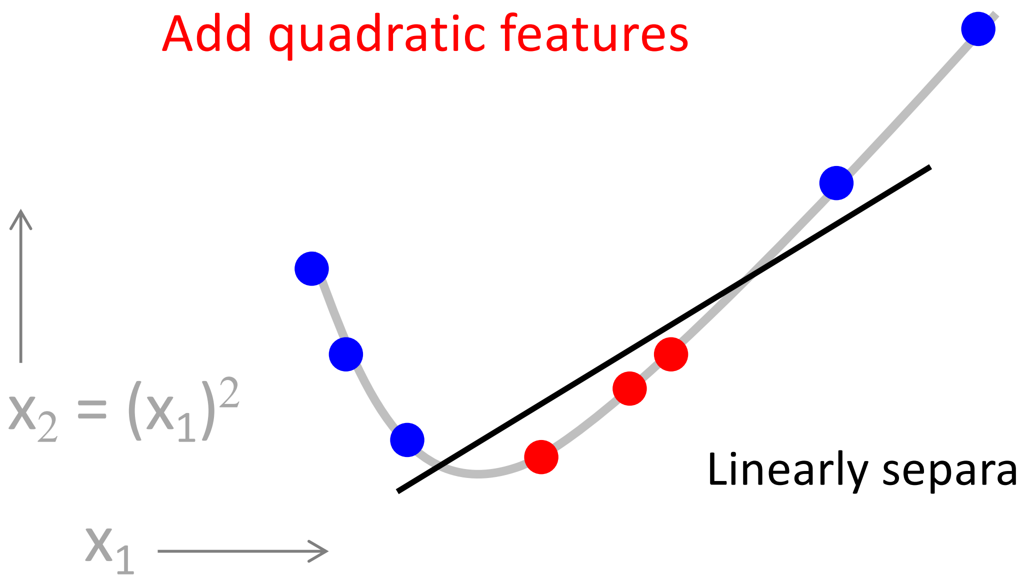
- Linear classifier can't learn some functions

1D example:



Not linearly separable

Add quadratic features



Linearly separable in new features…

Adding features

- Recall: feature function Phi(x)
  - Predict using some transformation of original features

$$\hat{y}(x) = \mathrm{sign}\big[\, w \cdot \Phi(x) + b \,\big]$$

- Dual form of SVM optimization is:

$$\max_{0 \le \alpha \le R} \sum_i \alpha_i - \frac{1}{2} \sum_j \alpha_i \alpha_j \, y^{(i)} y^{(j)} \, \Phi(x^{(i)}) \Phi(x^{(j)})^T \quad \text{s.t.} \; \sum_i \alpha_i y^{(i)} = 0$$

- For example, quadratic (polynomial) features:

$$\Phi(x) = \left(1 \;\; \sqrt{2} x_1 \;\; \sqrt{2} x_2 \;\; \cdots \;\; x_1^2 \;\; x_2^2 \;\; \cdots \;\; \sqrt{2} x_1 x_2 \;\; \sqrt{2} x_1 x_3 \;\; \cdots \right)$$

  - Ignore root-2 scaling for now…
  - Expands "x" to length $O(n^2)$

Implicit features

- Need $\Phi(x^{(i)})\Phi(x^{(j)})^T$

$$\Phi(x) = \begin{pmatrix}1 & \sqrt{2}x_1 & \sqrt{2}x_2 & \cdots & x_1^2 & x_2^2 & \cdots & \sqrt{2}x_1x_2 & \sqrt{2}x_1x_3 & \cdots\end{pmatrix}$$

$$\Phi(a) = \begin{pmatrix}1 & \sqrt{2}a_1 & \sqrt{2}a_2 & \cdots & a_1^2 & a_2^2 & \cdots & \sqrt{2}a_1a_2 & \sqrt{2}a_1a_3 & \cdots\end{pmatrix}$$

$$\Phi(b) = \begin{pmatrix}1 & \sqrt{2}b_1 & \sqrt{2}b_2 & \cdots & b_1^2 & b_2^2 & \cdots & \sqrt{2}b_1b_2 & \sqrt{2}b_1b_3 & \cdots\end{pmatrix}$$

$$\Phi(a)^T\Phi(b) = 1 + \sum_j 2a_jb_j + \sum_j a_j^2b_j^2 + \sum_j\sum_{k>j} 2a_ja_kb_jb_k + \ldots$$

$$= (1 + \sum_j a_jb_j)^2$$

$$= {\color{red}K(a,b)}$$

Can evaluate dot product in only O(n) computations!

Mercer Kernels

- If K(x,x') satisfies Mercer's condition:

$$\int_a \int_b K(a,b)\, g(a)\, g(b)\, da\, db \;\geq 0$$

For all datasets X:

$$g^T \cdot K \cdot g \;\geq 0$$

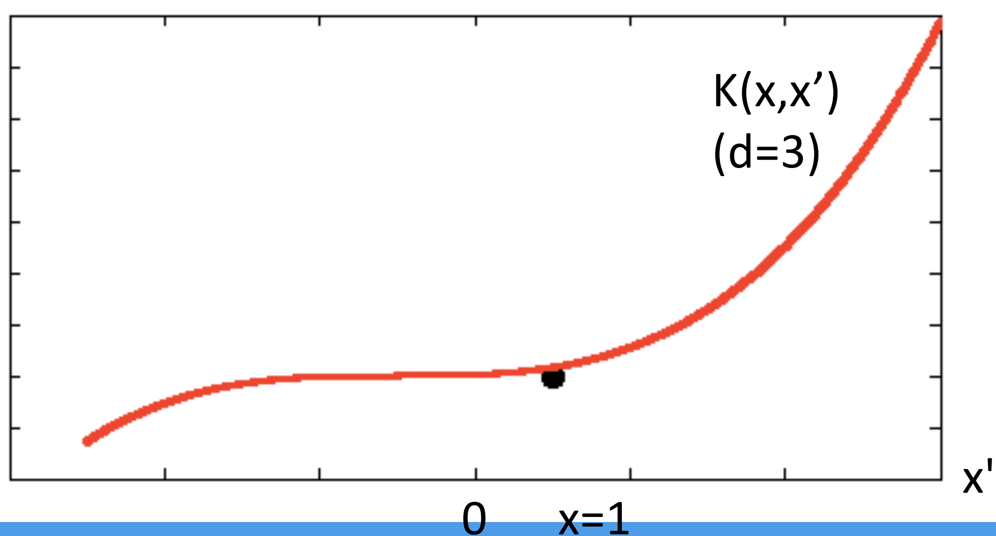
- Then, $K(a,b) = \Phi(a) \cdot \Phi(b)$ for some $\Phi(x)$

- Notably, Phi may be hard to calculate
  - May even be infinite dimensional!
  - Only matters that K(x,x') is easy to compute:
  - Computation always stays $O(m^2)$

Common kernel functions

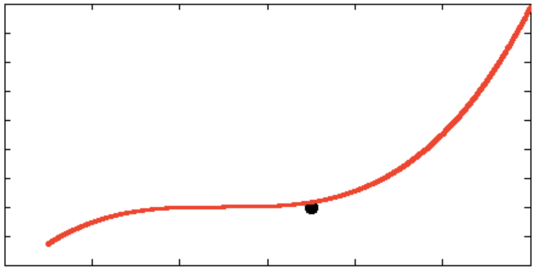
- Some commonly used kernel functions & their shape:

- Polynomial $K(a,b) = (1 + \sum_j a_j b_j)^d$



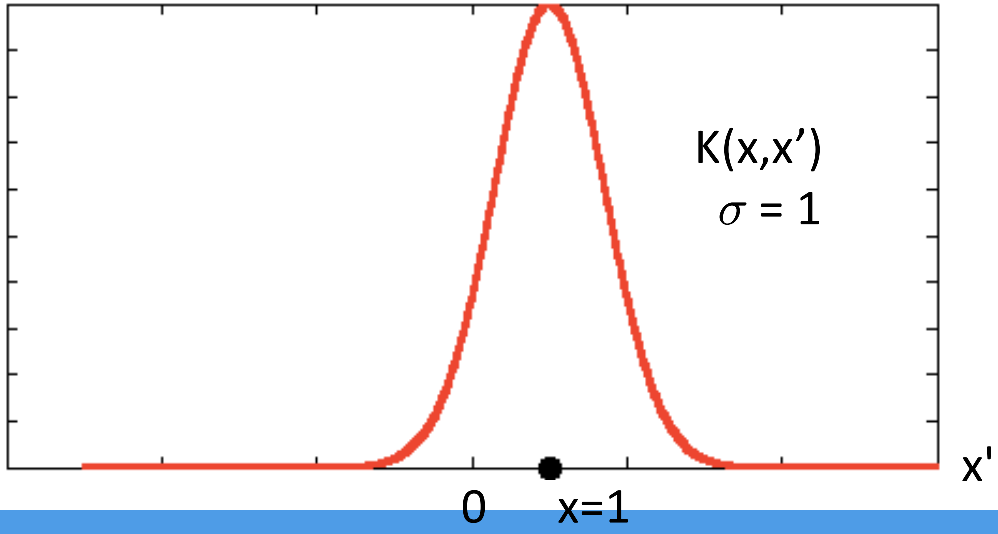
Common kernel functions

- Some commonly used kernel functions & their shape:
- Polynomial $K(a,b) = (1 + \sum_j a_j b_j)^d$



- Radial Basis Functions

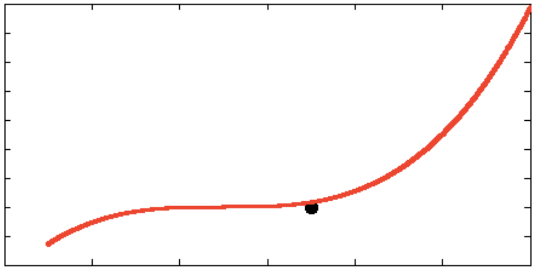
$$K(a,b) = \exp(-(a-b)^2/2\sigma^2)$$



Common kernel functions

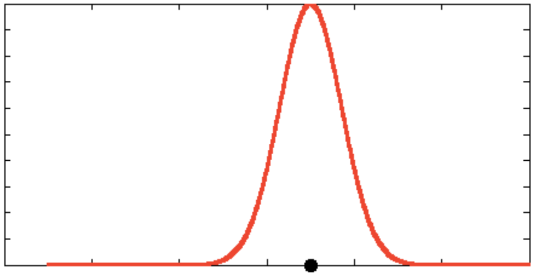
- Some commonly used kernel functions & their shape:

- Polynomial $K(a,b) = (1 + \sum_j a_j b_j)^d$



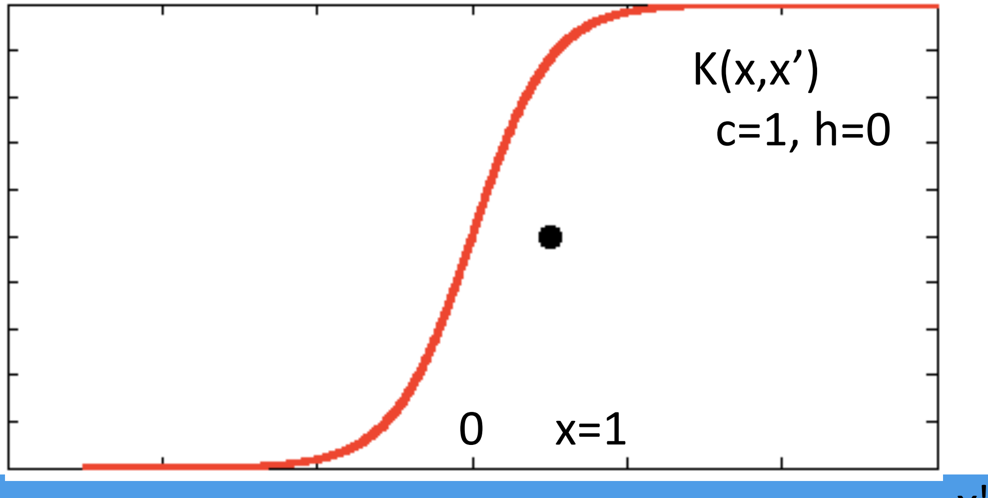
- Radial Basis Functions

$$K(a,b) = \exp(-(a-b)^2/2\sigma^2)$$



- Saturating, sigmoid-like:

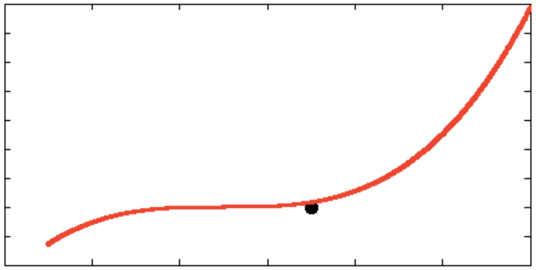
$$K(a,b) = \tanh(ca^T b + h)$$



Common kernel functions

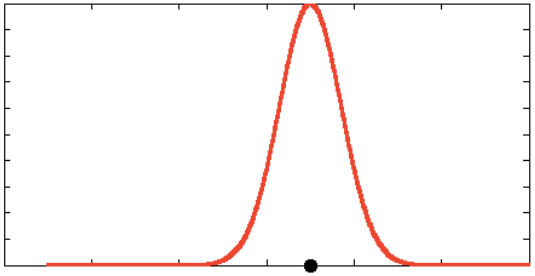
- Some commonly used kernel functions & their shape:

- Polynomial $K(a,b) = (1 + \sum_j a_j b_j)^d$



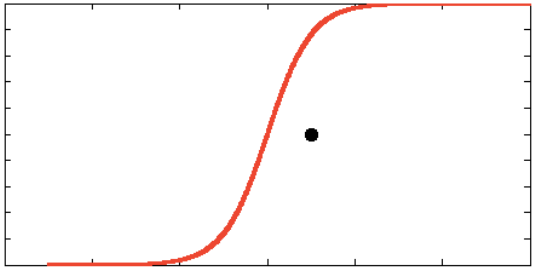
- Radial Basis Functions

$$K(a,b) = \exp(-(a-b)^2/2\sigma^2)$$



- Saturating, sigmoid-like:

$$K(a,b) = \tanh(ca^T b + h)$$

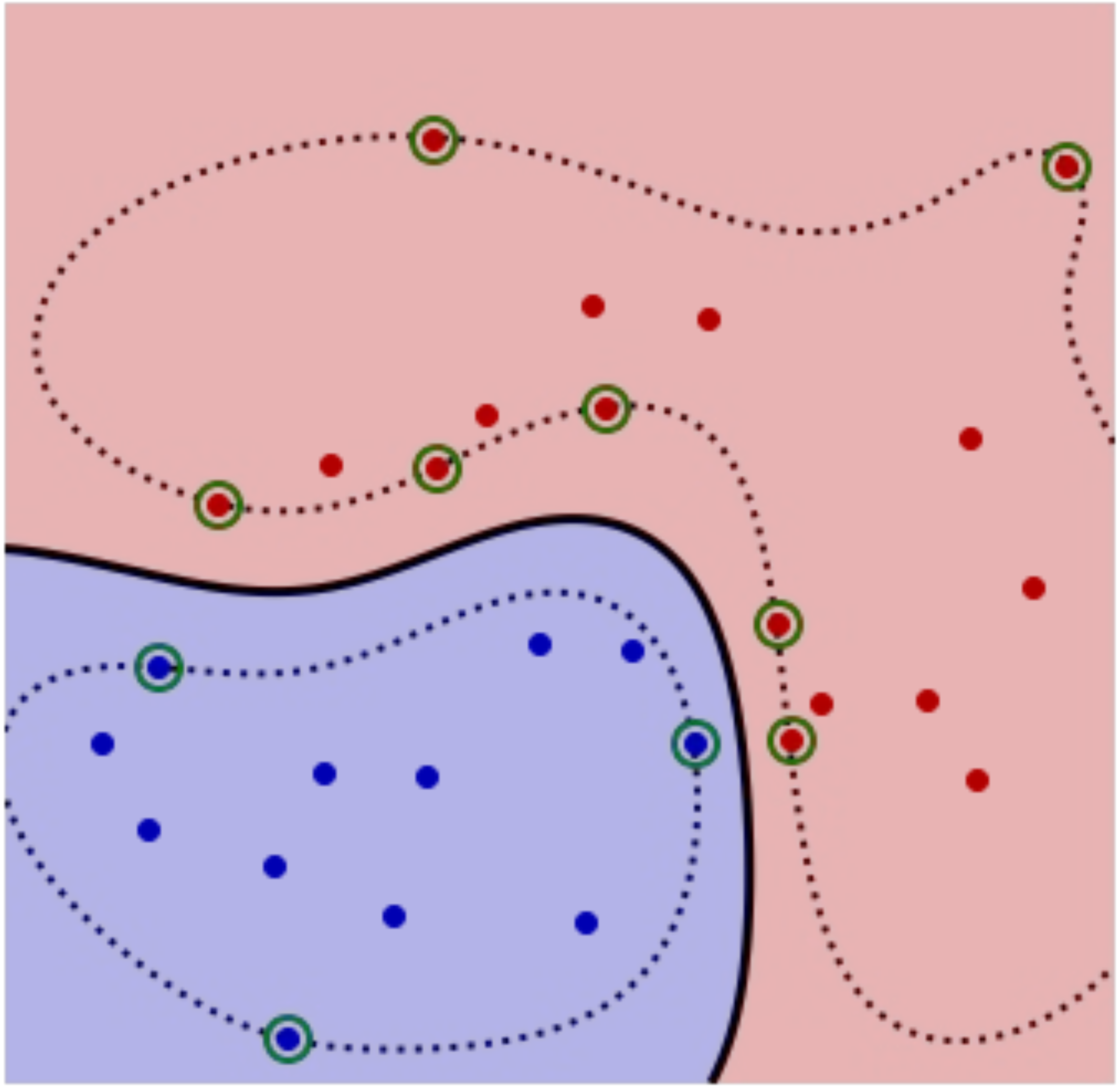


Here, d / σ / {c,h} are **hyperparameters**: influence the meaning of “similarity”

- Many for special data types:
  - String similarity for text, genetics

- In practice, may not even be Mercer kernels…

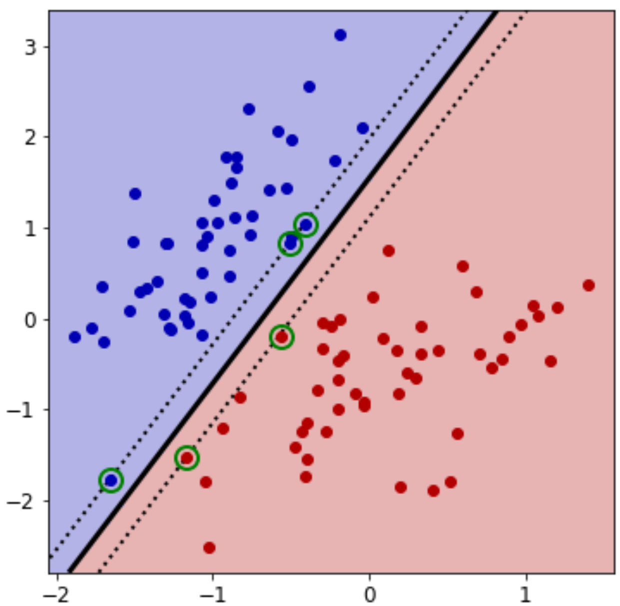
Support Vectors for Kernel SVMs



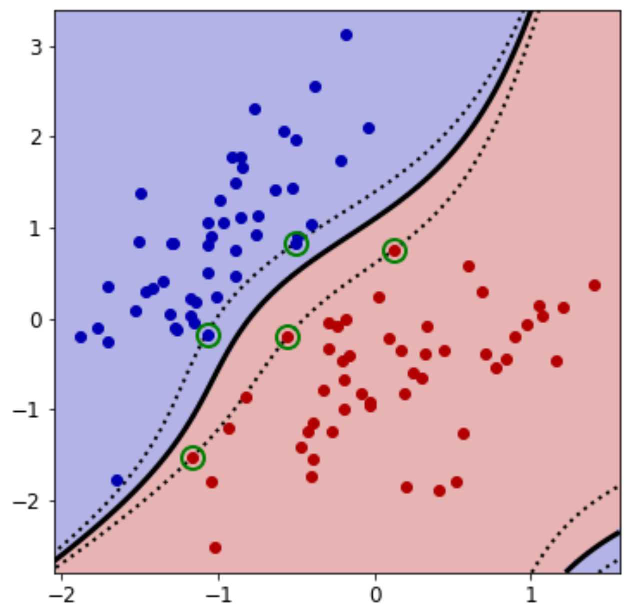
Support vectors (green) for data separable by radial basis function kernels, and non-linear margin boundaries

Support Vectors for Kernel SVMs

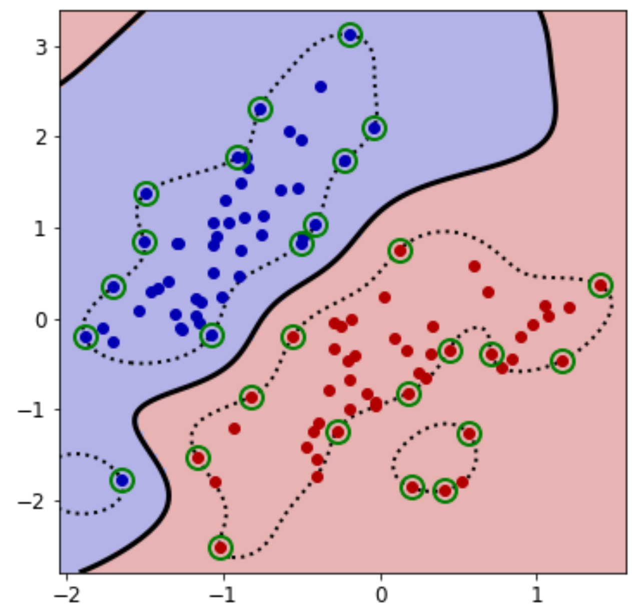
- Separable data



Linear Kernel



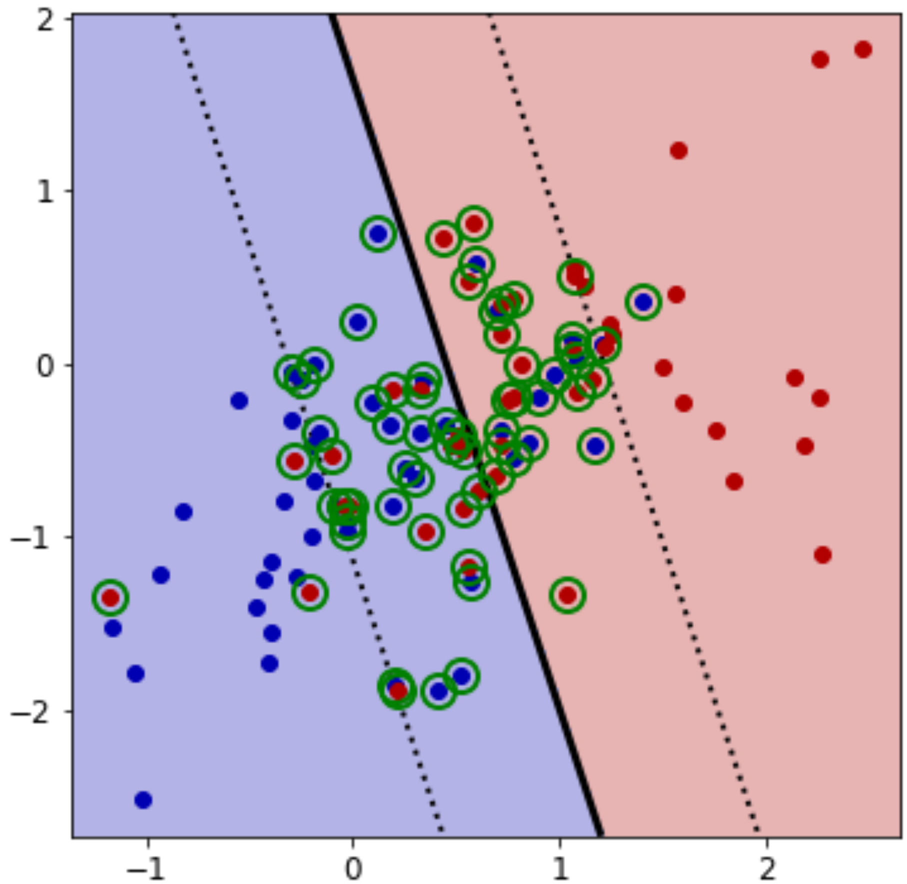
Degree-4 Polynomial



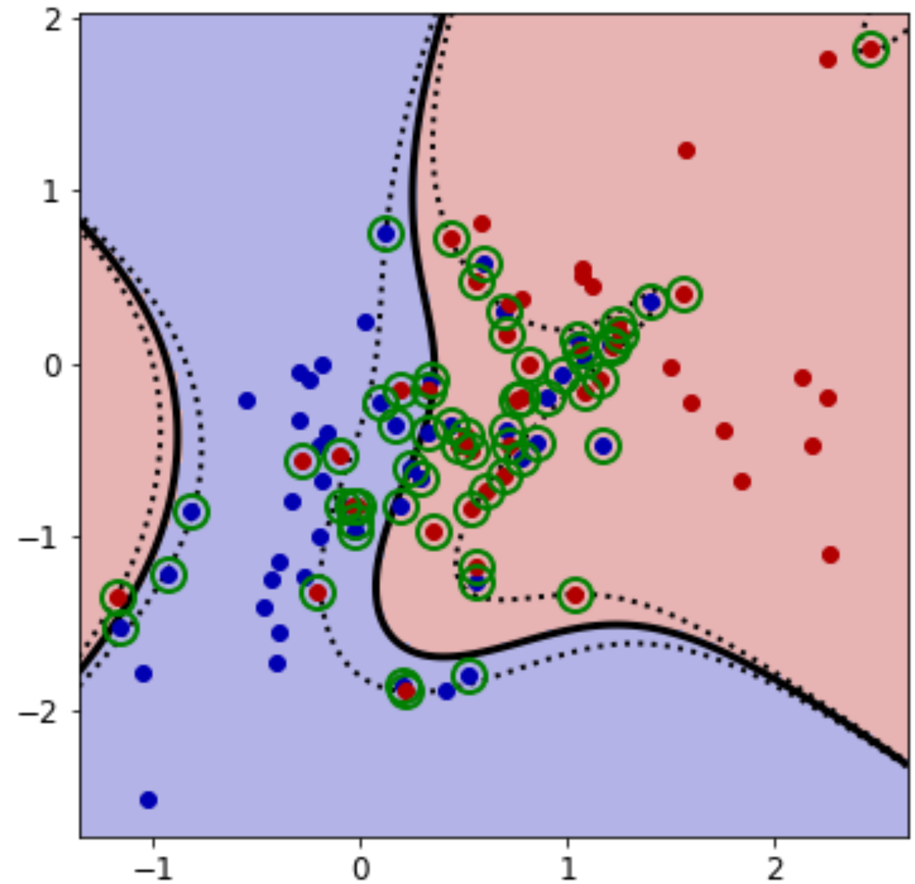
Gaussian Kernel

Support vectors (green) for separable data, along with margin boundaries

How many support vectors?



Linear Kernel



Degree-4 Polynomial

Only need to evaluate the kernel at the support vectors

… but there may be a lot of support vectors

Kernel SVMs

- Linear SVMs
  - Can represent classifier using (w,b) = n+1 parameters
  - Or, represent using support vectors, $x^{(i)}$
- Kernelized?
  - K(x,x') may correspond to high (infinite?) dimensional Phi(x)
  - Typically more efficient to remember the SVs
  - "Instance based" – save data, rather than parameters
- Contrast:
  - Linear SVM: identify *features* with linear relationship to target
  - Kernel SVM: identify *similarity measure* between data

  (Sometimes one may be easier; sometimes the other!)

Kernel Least-squares Linear Regression

- Recall L2-regularized linear regression:

$$\theta = y\,X(X^T X + \alpha I)^{-1}$$

Rearranging,

$$\Rightarrow \ \theta\,(X^T X + \alpha I) = y\,X \longrightarrow \alpha\theta = (y - \theta\,X^T)\,X$$

Define:

$$r = \frac{1}{\alpha}(y - \theta\,X^T) \longrightarrow \underline{\theta} = rX$$

$$\alpha\,r = \underline{y} - \underline{\theta}\,\underline{X}^T = \underline{y} - r\,X X^T$$

Gram matrix: m x m,

$$K_{ij} = \langle x^{(i)}, x^{(j)} \rangle$$

Rearrange & solve for r:

$$r = (X X^T + \alpha I)^{-1} y = (K + \alpha I)^{-1} y$$

Linear prediction:

$$\tilde{y} = \langle \theta, \tilde{x} \rangle = rX\,(\tilde{x})^T = \sum_j r_j \langle x^{(j)}, \tilde{x} \rangle = \sum_j r_j K(x^{(j)}, \tilde{x})$$

Now just replace K(x,x’) with your desired kernel function!

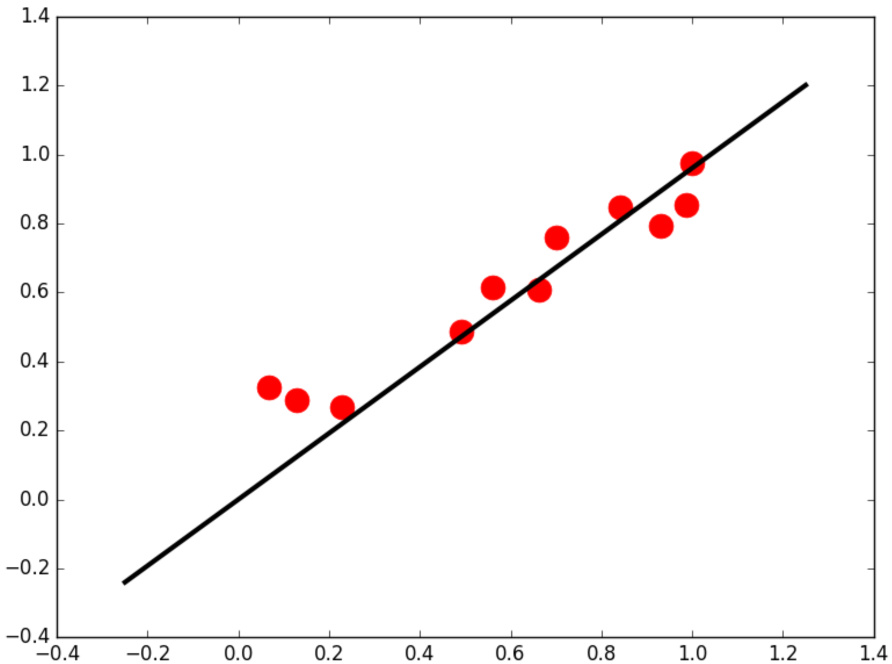
Example: Kernel Linear Regression

- K: MxM

$$r = (K + \alpha I)^{-1} y \qquad \tilde{y} = \sum_j r_j K(x^{(j)}, \tilde{x})$$

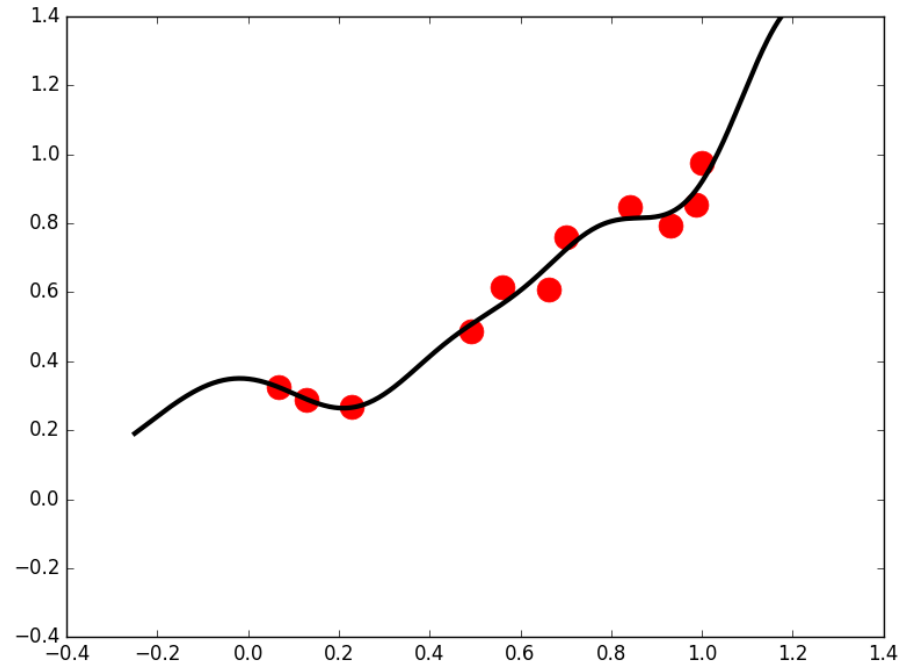
Linear kernel:

$$K(x, x') = x^T \cdot x'$$



Gaussian (RBF) kernel:

$$K(x, x') = \exp(-\gamma(x - x')^2)$$



Summary

- Support vector machines

- "Large margin" for separable data
  - Primal QP: maximize margin subject to linear constraints
  - Lagrangian optimization simplifies constraints
  - Dual QP: m variables; involves m^2 dot product

- "Soft margin" for non-separable data
  - Primal form: regularized hinge loss
  - Dual form: m-dimensional QP

- Kernels
  - Dual form involves only pairwise similarity
  - Mercer kernels: dot products in implicit high-dimensional space