Foundations of Machine Learning (FOML)
Lecture 1

Kristiaan Pelckmans

September 8, 2015

Overview

Today:

- Overview of the course.
- Support Vector Machines (SVMs) - the separable case.
- Convex Optimization.
- Analysis.
- Kernels.
- SVMs - the inseparable case.

Overview (Ct'd)

Organization:

- 10 Lectures.
- 1 computer lab (mid october). (content)
- $*$ Miniprojects (due end october). (content)
- Participants giving lectures using material.

Overview (Ct'd)

Course:

1. Introduction.
2. Support Vector Machines (SVMs).
3. Probably Approximatively Correct (PAC) analysis.
4. Boosting.
5. Online Learning.
6. Multi-class classification (*).
7. Ranking (*).
8. Regression (*).
9. Stability-based analysis (*).
10. Dimensionality reduction (*).
11. Reinforcement learning (*).
12. Presentations of the results of the mini-projects.

Introduction

Applications and problems:

- Classification.
- Regression.
- Ranking.
- Clustering.
- Dimensionality Reduction or manifold learning.

Introduction (Ct'd)

Definitions and Terminology:

- m Examples.
- Features $x_i \in \mathcal{X}$.
- Labels $y_i \in \mathcal{Y}$.
- Fixed, unknown distribution underlying samples D.
- Training sample $S_m \subset \mathcal{X} \times \mathcal{Y}$.
- Validation sample S'.
- Test sample S''.
- Loss function L.
- Hypothesis set $H = \{h : \mathcal{X} \to \mathcal{Y}\}$.
- Learning algorithm $\mathcal{A}(\theta) : \{S\} \to H : \hat{h}_S$
- where θ are all the free tuning parameters
- (true and average loss) Risk R and R_m.

Introduction (Ct'd)

n-fold Cross-validation

- Let $S_m = \{(x_i, y_i)\}_{i=1}^m$ be the original training set.
- Divide set S_m into n disjunct folds so that every point included once.
- Make n sets with $n-1$ folds, denote them as S_{-i} .
- Let $S_{-i} = \{(x_{ij}, y_{ij})\}_{j=1}^{m_i}$ be the training set of the i-th iteration.
- Hence $\hat{h}_{S_{-i}}$ the outcome of $\mathcal{A}(\theta)$ applied to the i-th training set.

$$\hat{R}_{CV}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{j=1} L\left(\hat{h}_{S_{-i}}(x_{ij}), y_{ij}\right)$$

Introduction (Ct'd)

Learning Scenarios

- Supervised learning.
- Unsupervised learning.
- Semi-supervised learning.
- Transductive inference.
- Online learning.
- Reinforcement learning.
- Active Learning.

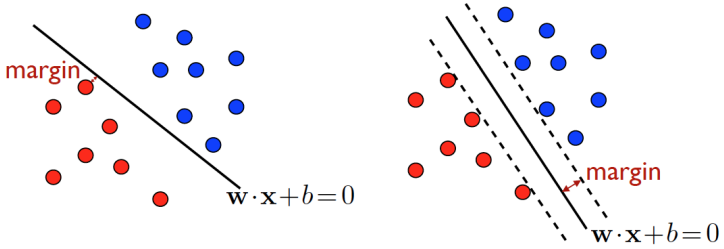
SVM - separable case

Support Vector Machine (SVM)

- Assume that there is a f s.t. $y = f(x)$.
- Find h with minimal

$$R_D = Pr_{x \sim D}\left[h(x) \neq f(x)\right]$$

- Hypothesis set H of all linear separators h characterised by $(\mathbf{w}, b)$:



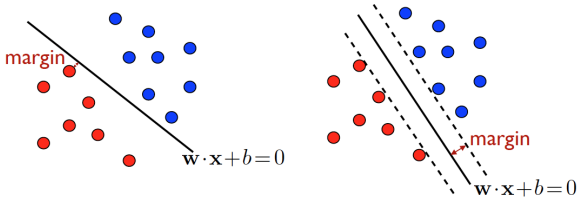
$$\begin{cases} \mathbf{w} \cdot \mathbf{x} + b > 0 & h(\mathbf{x}) = +1 \\ \mathbf{w} \cdot \mathbf{x} + b < 0 & h(\mathbf{x}) = -1 \end{cases}$$

or

$$H = \left\{ \mathbf{x} \to \text{sign}(\mathbf{w} \cdot \mathbf{x} + b) : \ \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R} \right\}$$

- Separable case: $\exists h \in H$ s.t. $f(\mathbf{x})h(\mathbf{x}) = yh(\mathbf{x}) > 0 \ \forall \mathbf{x} \in D$.

SVM - separable case (Ct'd)



Maximal Margin

- Hyperplane $\{\mathbf{x} : \ \mathbf{w}\cdot\mathbf{x} + b = 0\}$.
- Normalise such that $\min_i |\mathbf{w}\cdot\mathbf{x}_i + b| = 1$ (w.l.o.g.).
- Distance point $\mathbf{x}_0$ - margin:

$$\frac{|\mathbf{w}\cdot\mathbf{x}_0 + b|}{\|\mathbf{w}\|}$$

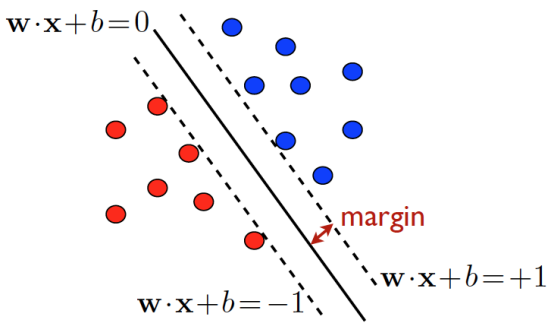
- Thus margin is given as

$$\rho = \frac{\min_i |\mathbf{w}\cdot\mathbf{x}_i + b|}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}.$$

SVM - separable case (Ct'd)

Maximal Margin

- Maximal Hyperplane:



$$\max_{\rho,\mathbf{w},b} \rho \quad \text{s.t.} \quad \begin{cases} y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 0 \ \forall i \\ \rho = \frac{\min_i |\mathbf{w}\cdot\mathbf{x}_i + b|}{\|\mathbf{w}\|} \\ \min_i |\mathbf{w}\cdot\mathbf{x}_i + b| = 1 \end{cases}$$

- Maximal hyperplane:

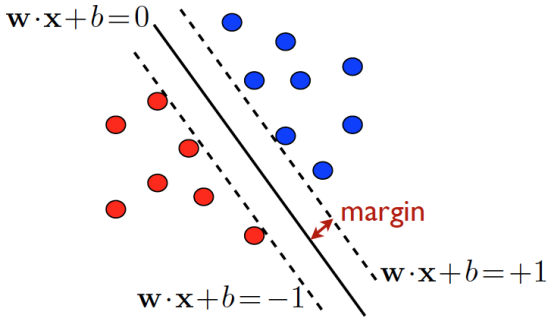
$$\max_{\rho,\mathbf{w},b} \rho \quad \text{s.t.} \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 \ \forall i, \ \rho = \frac{1}{\|\mathbf{w}\|}$$

- Or

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \ \text{s.t.} \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 \ \forall i.$$

- Why? Find the *safest* solution.

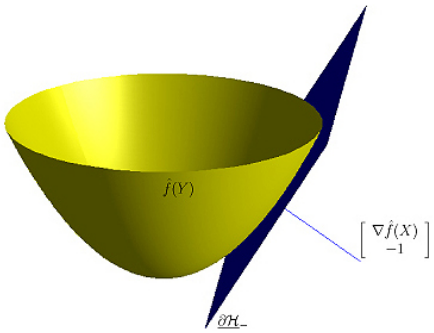
SVM - separable case (Ct'd)



Maximal Margin

- Convex objective.
- Affine inequality constraints.
- Quadratic Programming problem.
- Dual problem: proberties!

Convex Optimization



Convex

- A set $\mathcal{X}$ is convex iff for any two points $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, the segment $\{\alpha\mathbf{x} + (1-\alpha)\mathbf{x}' : 0 \leq \alpha \leq 1\} \in \mathcal{X}$.
- A function $f : \mathcal{X} \to \mathbb{R}$ is convex iff for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and all $0 \leq \alpha \leq 1$ one has that

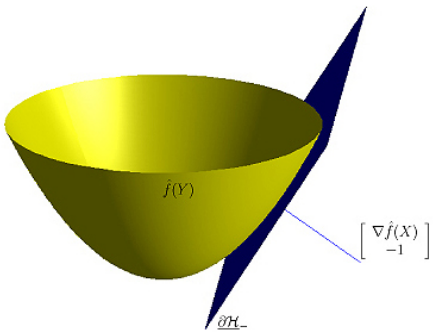
$$f\left(\alpha\mathbf{x} + (1-\alpha)\mathbf{x}'\right) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{x}')$$

- Let f be a differentiable function, then f is convex if and only if $\mathcal{X}$ is convex and

$$\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X} : f(\mathbf{x}') - f(\mathbf{x}) \geq \nabla f(\mathbf{x})(\mathbf{x}' - \mathbf{x})$$

Convex Optimization (Ct'd)

Convex Programming



- Constrained optimisation problem.

$$p^* = \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \text{ s.t. } \quad g_i(\mathbf{x}) \leq 0, \forall i.$$

- Lagrangian:

$$\forall \mathbf{x} \in \mathcal{X}, \alpha \geq 0 : \ \mathcal{L}(\mathbf{x}, \alpha) = f(\mathbf{x}) + \sum_i \alpha_i g_i(\mathbf{x}).$$

- Dual function (concave)

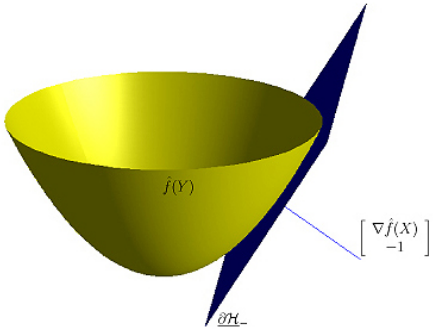
$$\forall \alpha \geq 0 : \ F(\alpha) = \inf_{\mathbf{x} \in \mathcal{X}} \mathcal{L}(\mathbf{x}, \alpha).$$

so that $F(\alpha) \leq p^*$.

- Dual problem:

$$d^* = \max_{\alpha \geq 0} F(\alpha)$$

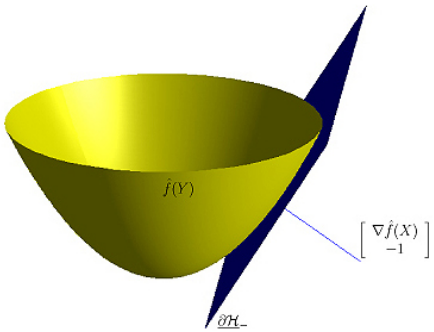
Convex Optimization (Ct'd)



Convex Programming

- Weak duality: $p^* \geq d^*$.
- Strong duality: $p^* = d^*$.
- Duality gap: $p^* - d^*$.
- Strong duality holds when *Constraint qualifications* hold.
  - Strong constraint qualification (Slater): $\exists \mathbf{x} \in \text{int}(\mathcal{C}) : g_i(\mathbf{x}) < 0 \ \forall i$
  - Weak constraint qualification (weak Slater): $\exists \mathbf{x} \in \text{int}(\mathcal{C}) : g_i(\mathbf{x}) < 0$
    or g_i is affine, $g_i(\mathbf{x}) = 0 \ \forall i$

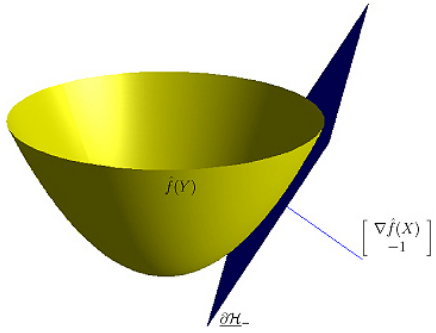
Convex Optimization (Ct'd)



Karush-Kuhn-Tucker (KKT) conditions:
Assume that $f, g_i : \mathcal{X} \to \mathbb{R}$ for all i are convex and differentiable, and that the constraints are qualified, then $\bar{\mathbf{x}}$ is a solution of the constrained program if and only there exists an $\bar{\alpha}$ such that

$$\begin{cases} \nabla_{\mathbf{x}} \mathcal{L}(\bar{\mathbf{x}}, \bar{\alpha}) = 0 \\ \nabla_{\alpha} \mathcal{L}(\bar{\mathbf{x}}, \bar{\alpha}) \leq 0 \\ \alpha_i g_i(\bar{\mathbf{x}}) = 0 \qquad \forall i \end{cases}$$

Analysis of SVMs



- Lagrangian:

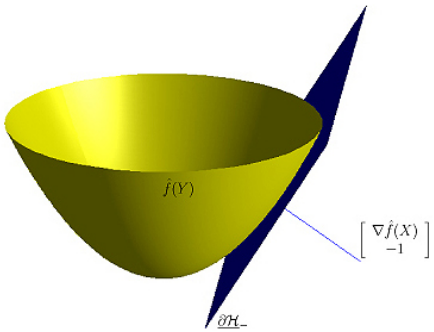
$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i \left(y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1\right)$$

- KKT conditions

$$\begin{cases} \nabla_{\mathbf{w}}\mathcal{L} = 0 \Leftrightarrow \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \\ \nabla_b \mathcal{L} = 0 \Leftrightarrow \sum_{i=1}^{m} \alpha_i y_i = 0 \\ \forall i : \alpha_i \left(y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1\right) = 0 \end{cases}$$

- Support vectors: $N_{SV}(S)$

Analysis of SVMs (Ct'd)



- Dual problem $\max_{\alpha \geq 0} \inf_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha)$
- Eliminate $\mathbf{w}$ and b using KKT conditions:
- Dual problem

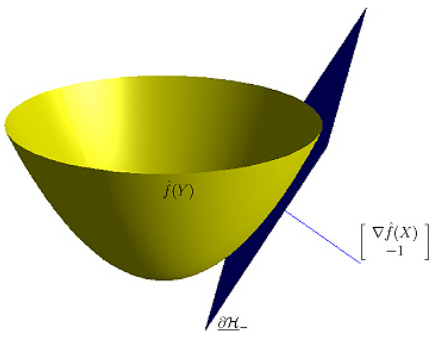
$$\max_{\alpha \geq 0} \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\text{s.t.} \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \quad (1)$$

- And at the optimum $\bar{\mathbf{w}} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$.
- and $b = y_i - \sum_{i=1}^{m} \alpha_j y_j \mathbf{x}_j \cdot \mathbf{x}_i$
- Hence we can predict

$$h(\mathbf{x}) = \text{sign}(\bar{\mathbf{w}} \cdot \mathbf{x} + b)$$

Analysis of SVMs (Ct'd)

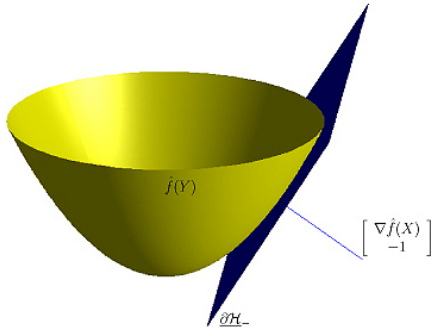


Generalization error

$$R(h_S) = \Pr_{\mathbf{x}\sim D}\left[h_S(\mathbf{x}) \neq f(\mathbf{x})\right]$$

- Leave-one-out analysis.
- In terms of N_{SV}.
- Margin-based analysis.

Analysis of SVMs (Ct'd)



Leave-one-out analysis

$$R_{LOO}(\mathcal{A}(\theta), S) = \frac{1}{m}\sum_{i=1}^{m} 1(h_{S/(\mathbf{x}_i, y_i)}(\mathbf{x}_i) = y_i)$$

- $\mathcal{A}(\theta)(S) = h_S$.
- $1(z) = 1$ iff z is true, $1(z) = 0$.
- In terms of N_{SV}.
- Then

$$E_{S\sim D^m}[R_{LOO}(\mathcal{A}(\theta), S)] = E_{S'\sim D^{m-1}}[R(h_{S'})].$$

Analysis of SVMs (Ct'd)

Proof:

$$
\begin{aligned}
E_{S\sim D^m}\left[R_{LOO}(\mathcal{A}(\theta), S)\right] &= \frac{1}{m}\sum_{i=1}^{m} E_{S\sim D^m}\left[1(h_{S/(\mathbf{x}_i,y_i)}(\mathbf{x}_i) = y_i)\right] \\
&= E_{S\sim D^m}\left[1(h_{S/(\mathbf{x}_1,y_1)}(\mathbf{x}_1) = y_1)\right] \\
&= E_{S\sim D^m}\left[1(h_{S/(\mathbf{x}_1,y_1)}(\mathbf{x}_1) = y_1)\right] \\
&= E_{S'\sim D^{m-1}}\left[E_{\mathbf{x}_1\sim D}\left[1(h_{S'}(\mathbf{x}_1) = y_1)\right]\right] \\
&= E_{S'\sim D^{m-1}}[R(h_{S'})].
\end{aligned}
$$

Analysis of SVMs (Ct'd)

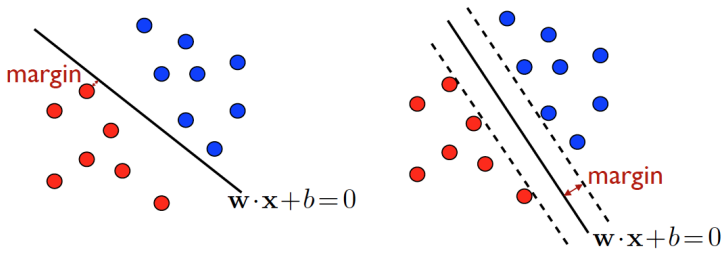
Support Vector Analysis: Let $\mathcal{A}(\theta)(S) = h_S$ be the hypothesis returned by SVMs for a sample S, and let $N_{SV}(S)$ be the number of Support Vectors that define h_S. Then

$$E_{S \sim D^m}[R(h_S)] \leq E_{S' \sim D^{m+1}} \left[\frac{N_{SV}(S')}{m+1} \right]$$

Argument: if $\mathbf{x}_i$ is not a SV, then $h_{S/(\mathbf{x}_i, y_i)} = h_S$ and $h_{S/(\mathbf{x}_i, y_i)}(\mathbf{x}_i) = f(\mathbf{x}_i)$

$$R_{LOO}(\mathcal{A}(\theta), S) \leq \frac{N_{SV}(S)}{m+1}$$

SVM - Margin analysis (Ct'd)



Vapnik-Chervonenkis (VC) dimension:

- Distance point $\mathbf{x}_0$ with label y_0 to a hyperplane $\{\mathbf{x} : \mathbf{w}\cdot\mathbf{x} + b = 0\}$ is

$$\rho(\mathbf{x}) = \frac{y_0(\mathbf{w}\cdot\mathbf{x}_0 + b)}{\|\mathbf{w}\|}$$

- Margin is given as

$$\rho = \min_i \frac{y_i(\mathbf{w}\cdot\mathbf{x}_i + b)}{\|\mathbf{w}\|}$$

- *capacity* of H (Structural Risk Minimisation: see next lecture)
- VC dimension (try) of hyperplane is $N+1$

  ...
- But high-dimensions?

SVM - Margin analysis (Ct'd)

Refined analysis of VC dimension

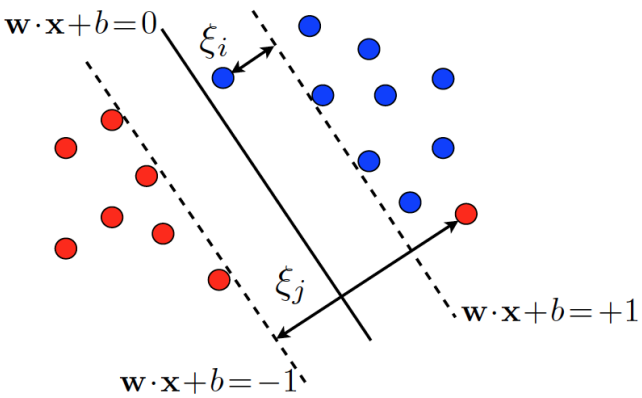
- Margin $\rho = \frac{1}{\|\mathbf{w}\|}$.
- $H = \{h(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b), \|w\| \leq \Lambda\}$.
- How many points can be shattered?

$$d : \exists \{\mathbf{x}_i\}_{i=1}^d, \forall \sigma \in \pm 1^m : \exists h \in H : h(\mathbf{x}_1) = \sigma_1, \ldots, h(\mathbf{x}_d) = \sigma_d,$$

- Measures capacity of H.
- Relates to Rademacher complexity:

$$\hat{\mathcal{R}}_S(H) = \frac{1}{m} E_{\sigma_1, \ldots, \sigma_m} \left[\sup_{h \in H} \sum_{i=1}^m \sigma_i h(\mathbf{x}_i) \right].$$

SVM - non-separable case



Maximal Soft Margin:

- Non-separable case: $\forall(\mathbf{w}, b)$

$$\exists i : y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \ngeq 1$$

- Idea: find *best* $(\mathbf{w}, b)$ with minimal *slack*

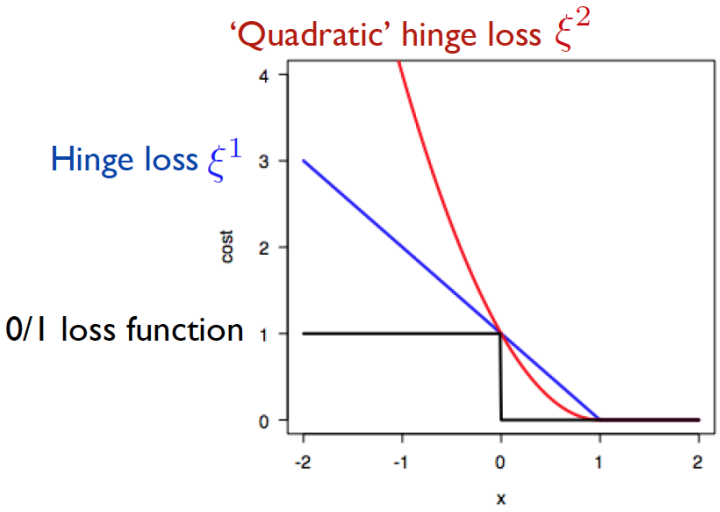
$$y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i$$

- Max Soft Margin:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t.} \quad \begin{cases} y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i & \forall i \\ \xi_i \geq 0 & \forall i. \end{cases} \tag{2}$$

SVM - non-separable case (Ct'd)



Dual problem:

- Dual problem

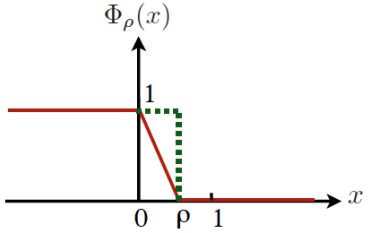
$$\max_{0 \le \alpha \le C} \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\text{s.t.} \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \quad (3)$$

- And at the optimum $\bar{\mathbf{w}} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$.
- and $by_i = 1 - y_i \sum_{i=1}^{m} \alpha_j y_j \mathbf{x}_j \cdot \mathbf{x}_i$ when $\xi_i = 0$.

SVM - Analysis.

Rademacher complexity:

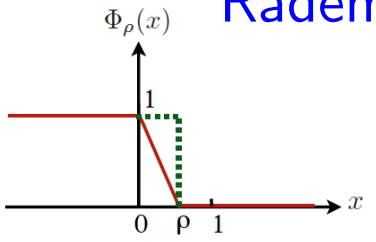


$$R_\rho(h) = \frac{1}{m}\sum_{i=1}^{m} \phi_\rho(h(\mathbf{x}_i) - y_i)$$

- $$\hat{\mathcal{R}}_S(H) = \frac{1}{m} E_{\sigma_1,\ldots,\sigma_m}\left[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i h(\mathbf{x}_i)\right]$$
- $H = \{h(\mathbf{x}) = \text{sign}(\mathbf{w}\cdot\mathbf{x} + b), \|w\| \leq \Lambda, b \in \mathbb{R}\}$.
- Theorem: Let H be a set of real-valued functions, fix $\rho > 0$. For any $\delta > 0$, with probability exceeding $1 - \delta$ one has that

$$\forall h \in H: \; R(h) \leq R_\rho(h) + \frac{2}{\rho}\hat{\mathcal{R}}_S(H) + 3\sqrt{\frac{log\frac{2}{\delta}}{2m}}.$$

SVM - Analysis (Ct'd).



Rademacher complexity:

$$\hat{\mathcal{R}}_S(H) = \frac{1}{m} E_{\sigma_1,\ldots,\sigma_m}\left[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i h(\mathbf{x}_i)\right]$$

- $H = \{h(\mathbf{x}) = \text{sign}(\mathbf{w}\cdot\mathbf{x} + b), \|w\| \leq \Lambda, b \in \mathbb{R}\}$.
- Theorem: Let S be a sample of size m with $\|\mathbf{x}_i\| \leq R$, then

$$\hat{\mathcal{R}}_S(H) \leq \sqrt{\frac{R^2\Lambda^2}{m}}.$$

SVM - Analysis (Ct'd)

Proof:

$$\begin{aligned}
\hat{\mathcal{R}}_S(H) &\triangleq \frac{1}{m} E_\sigma \left[\sup_{\|\mathbf{w}\| \le \Lambda} \sum_{i=1}^{m} \sigma_i (\mathbf{w} \cdot \mathbf{x}_i) \right] \\
&= \frac{1}{m} E_\sigma \left[\sup_{\|\mathbf{w}\| \le \Lambda} \mathbf{w} \cdot \sum_{i=1}^{m} \sigma_i \mathbf{x}_i \right] \\
&\le \frac{\Lambda}{m} E_\sigma \left[\| \sum_{i=1}^{m} \sigma_i \mathbf{x}_i \| \right] \\
&\le \frac{\Lambda}{m} E_\sigma \left[\sum_{i=1}^{m} \sigma_i \|\mathbf{x}_i\|^2 \right]^{1/2} \\
&\le \frac{\Lambda \sqrt{m R^2}}{m} = \sqrt{\frac{\Lambda^2 R^2}{m}}. \quad (4)
\end{aligned}$$

SVM - Analysis (Ct'd)

Talagrand's Contraction Lemma:

$$\hat{\mathcal{R}}_S(\phi \circ H) \leq L\hat{\mathcal{R}}_S(H)$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is L Lipschitz smooth, and H any hypothesis set.

$$\begin{aligned} &\triangleq \frac{1}{m} E_{\sigma_1,\ldots,\sigma_m}\left[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i(\phi\circ h)(\mathbf{x}_i)\right] \\ &= \frac{1}{m} E_{\sigma_1,\ldots,\sigma_{m-1}} E_{\sigma_m}\left[\sup_{h\in H}\sigma_m(\phi\circ h)(\mathbf{x}_m) + u_{m-1}\right] \end{aligned} \quad (5)$$

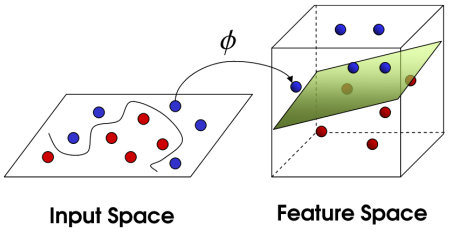
with $u_{m-1}(h) = \sum_{i=1}^{m-1}\sigma_i(\phi\circ h)(\mathbf{x}_i)$.

SVM - Analysis (Ct'd).

$$
\begin{aligned}
&\triangleq \frac{1}{m} E_{\sigma_1,\ldots,\sigma_{m-1}} E_{\sigma_m}\left[\sup_{h\in H} u_{m-1}(h) + \sigma_m(\phi\circ h)(\mathbf{x}_m)\right] \\
&\quad \frac{1}{2}\left[u_{m-1}(h_1) + \phi(h_1(\mathbf{x}_m))\right] + \frac{1}{2}\left[u_{m-1}(h_2) - \phi(h_2(\mathbf{x}_m))\right] \\
&\quad \leq \frac{1}{2}\left[u_{m-1}(h_1) + u_{m-1}(h_2)\right] + sL\frac{1}{2}\left[(h_1(\mathbf{x}_m)) - (h_2(\mathbf{x}_m))\right] \\
&\qquad\qquad \leq E_\sigma\left[\sup_{h\in H} u_{m-1}(h) + \sigma_m L h(\mathbf{x}_m)\right], \quad (6)
\end{aligned}
$$

with $s = \text{sign}(h_1(\mathbf{x}_m)) - (h_2(\mathbf{x}_m))$.

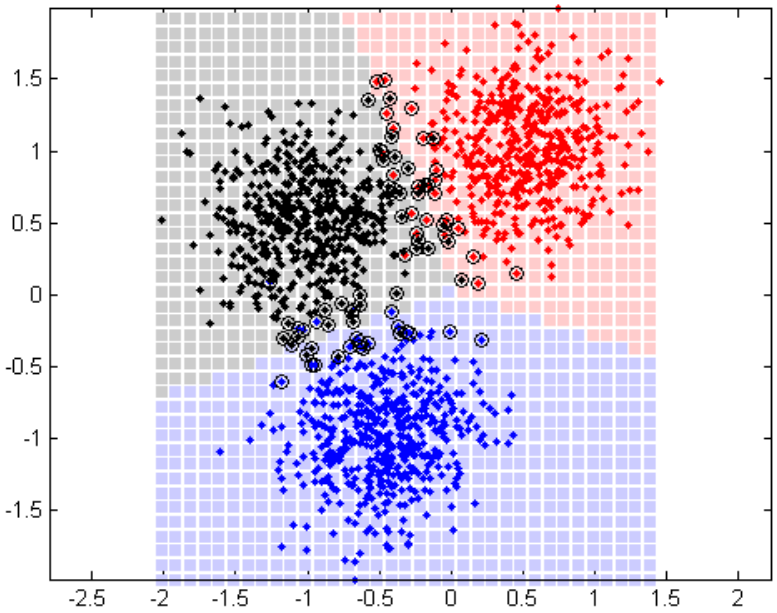
Kernels.



- Note that dual problem and predictor expressed in $(\mathbf{x}_i \cdot \mathbf{x}_j)$.
- Lets generalise it to $(\varphi(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j))$ with $\varphi : \mathbb{R}^N \to \mathbb{R}^*$.
- No explicit mapping, just inner product needed!
- $(\varphi(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j)) = K(\mathbf{x}_i, \mathbf{x}_j)$.
- $\exists \varphi$ iff K PSD!
- Typical choice $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}\right)$
- Rademacher complexity of

$$H = \{\sum_{i=1}^{m} \alpha_i y_i K(\mathbf{x}_i, \cdot),\ \|\alpha\| \leq \Lambda\}.$$

Conclusions



Take home messages:

- SVMs: optimisation, analysis.
- Separable case, non-separable case.
- Linear + kernels.
- Analysis.
- Margin and high-dimensional.