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).
- Miniprojects (due end October).
- Participants giving lectures using material.
Overview (Ct’d)

Course:
1. Introduction.
2. Support Vector Machines (SVMs).
3. Probably Approximately Correct (PAC) analysis.
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.
Applications and problems:

- Classification.
- Regression.
- Ranking.
- Clustering.
- Dimensionality Reduction or manifold learning.
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 \( \theta \) 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 $A(\theta)$ applied to the $i$-th training set.

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

\[
R_D = \text{Pr}_{x \sim D} [h(x) \neq f(x)]
\]

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

\[
\begin{align*}
    w \cdot x + b > 0 & \quad h(x) = +1 \\
    w \cdot x + b < 0 & \quad h(x) = -1
\end{align*}
\]

or

\[
H = \left\{ x \mapsto \text{sign}(w \cdot x + b) : \ w \in \mathbb{R}^N, \ b \in \mathbb{R} \right\}
\]

- Separable case: \( \exists h \in H \) s.t.

\[
f(x)h(x) = yh(x) > 0 \ \forall x \in D.
\]
SVM - separable case (Ct’d)

Maximal Margin

- Hyperplane $\{x: w \cdot x + b = 0\}$.
- Normalise such that $\min_i |w \cdot x_i + b| = 1$ (w.l.o.g.).
- Distance point $x_0$ - margin:
  $$\frac{|w \cdot x_0 + b|}{\|w\|}$$
- Thus margin is given as
  $$\rho = \frac{\min_i |w \cdot x_i + b|}{\|w\|} = \frac{1}{\|w\|}.$$
SVM - separable case (Ct’d)

Maximal Margin

▶ Maximal Hyperplane:

\[
\begin{align*}
\max_{\rho, w, b} \rho & \quad \text{s.t.} \quad y_i(w \cdot x_i + b) \geq 0 \quad \forall i \\
\rho &= \frac{\min_i |w \cdot x_i + b|}{\|w\|} \\
\min_i |w \cdot x_i + b| &= 1
\end{align*}
\]

▶ Maximal hyperplane:

\[
\begin{align*}
\max_{\rho, w, b} \rho & \quad \text{s.t.} \quad y_i(w \cdot x_i + b) \geq 1 \quad \forall i, \quad \rho = \frac{1}{\|w\|}
\end{align*}
\]

▶ Or

\[
\min_{w, b} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad y_i(w \cdot x_i + b) \geq 1 \quad \forall i.
\]

▶ Why? Find the \textit{safest} solution.
SVM - separable case (Ct’d)

Maximal Margin

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

Convex

- A set $\mathcal{X}$ is convex iff for any two points $x, x' \in \mathcal{X}$, the segment
  \[ \{ \alpha x + (1 - \alpha)x' : 0 \leq \alpha \leq 1 \} \in \mathcal{X}. \]

- A function $f : \mathcal{X} \to \mathbb{R}$ is convex iff for all $x, x' \in \mathcal{X}$ and all $0 \leq \alpha \leq 1$ one has that
  \[ f(\alpha x + (1 - \alpha)x') \leq \alpha f(x) + (1 - \alpha)f(x'). \]

- Let $f$ be a differentiable function, then $f$ is convex if and only if $\mathcal{X}$ is convex and
  \[ \forall x, x' \in \mathcal{X} : f(x') - f(x) \geq \nabla f(x)(x' - x). \]
Convex Optimization (Ct’d)
Convex Programming

▶ Constrained optimisation problem.

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

▶ Lagrangian:

\[ \forall x \in X, \alpha \geq 0 : L(x, \alpha) = f(x) + \sum_{i} \alpha_i g_i(x). \]

▶ Dual function (concave)

\[ \forall \alpha \geq 0 : F(\alpha) = \inf_{x \in X} L(x, \alpha). \]

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): $\exists x \in \text{int}(C) : g_i(x) < 0 \ \forall i$
  - Weak constraint qualification (weak Slater): $\exists x \in \text{int}(C) : g_i(x) < 0$
    or $g_i$ is affine, $g_i(x) = 0 \ \forall i$
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 $\bar{x}$ is a solution of the constrained program if and only there exists an $\bar{\alpha}$ such that

$$
\begin{align*}
\nabla_x \mathcal{L}(\bar{x}, \bar{\alpha}) &= 0 \\
\nabla_\alpha \mathcal{L}(\bar{x}, \bar{\alpha}) &\leq 0 \\
\alpha_i g_i(\bar{x}) &= 0 \quad \forall i
\end{align*}
$$
Analysis of SVMs

- **Lagrangian:**

\[
\mathcal{L}(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{m} \alpha_i (y_i (w \cdot x_i + b) - 1)
\]

- **KKT conditions**

\[
\begin{align*}
\nabla_w \mathcal{L} = 0 & \iff w = \sum_{i=1}^{m} \alpha_i y_i x_i \\
\nabla_b \mathcal{L} = 0 & \iff \sum_{i=1}^{m} \alpha_i y_i = 0 \\
\forall i : \alpha_i (y_i (w \cdot x_i + b) - 1) = 0
\end{align*}
\]

- **Support vectors:** \(N_{SV}(S)\)
Analysis of SVMs (Ct’d)

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

$$\max_{\alpha \geq 0} \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

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

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

$$h(x) = \text{sign}(\bar{w} \cdot x + b)$$
Analysis of SVMs (Ct’d)

Generalization error

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

- Leave-one-out analysis.
- In terms of \( N_{SV} \).
- Margin-based analysis.
Analysis of SVMs (Ct’d)

Leave-one-out analysis

\[ R_{\text{LOO}}(A(\theta), S) = \frac{1}{m} \sum_{i=1}^{m} 1(h_{S/(x_i,y_i)}(x_i) = y_i) \]

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

\[ E_{S \sim D^m} [R_{\text{LOO}}(A(\theta), S)] = E_{S' \sim D^{m-1}} [R(h_{S'})]. \]
Analysis of SVMs (Ct’d)

Proof:

\[
E_{S \sim D^m} [R_{LOO}(A(\theta), S)] = \frac{1}{m} \sum_{i=1}^{m} E_{S \sim D^m} [1(h_{S/(x_i,y_i)}(x_i) = y_i)]
\]

\[
= E_{S \sim D^m} [1(h_{S/(x_1,y_1)}(x_1) = y_1)]
\]

\[
= E_{S \sim D^m} [1(h_{S/(x_1,y_1)}(x_1) = y_1)]
\]

\[
= E_{S' \sim D^{m-1}} [E_{x_1 \sim D} [1(h_{S'}(x_1) = y_1)]]
\]

\[
= E_{S' \sim D^{m-1}} [R(h_{S'})].
\]
Support Vector Analysis: Let $A(\theta)(S) = h_S$ 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 $x_i$ is not a SV, then $h_S/(x_i,y_i) = h_S$ and $h_S/(x_i,y_i)(x_i) = f(x_i)$

$$R_{LOO}(A(\theta), S) \leq \frac{N_{SV}(S)}{m + 1}$$
Vapnik-Chervonenkis (VC) dimension:

- Distance point \( x_0 \) with label \( y_0 \) to a hyperplane \( \{ x : w \cdot x + b = 0 \} \) is

\[
\rho(x) = \frac{y_0(w \cdot x_0 + b)}{\|w\|}
\]

- Margin is given as

\[
\rho = \min_i \frac{y_i(w \cdot x_i + b)}{\|w\|}
\]

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

$$d : \exists \{x_i\}_{i=1}^d, \forall \sigma \in \pm 1^m : \exists h \in H : h(x_1) = \sigma_1, \ldots, h(x_d) = \sigma_d,$$

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

$$\hat{R}_S(H) = \frac{1}{m} E_{\sigma_1, \ldots, \sigma_m} \left[ \sup_{h \in H} \sum_{i=1}^m \sigma_i h(x_i) \right].$$
SVM - non-separable case

Maximal Soft Margin:

- Non-separable case: \( \forall (w, b) \)
  \[
  \exists i : y_i (w \cdot x_i + b) \geq 1
  \]

- Idea: find best \((w, b)\) with minimal slack
  \[
  y_i (w \cdot x_i + b) \geq 1 - \xi_i
  \]

- Max Soft Margin:
  \[
  \min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \xi_i
  \]
  \[
  \text{s.t. } \begin{cases}
  y_i (w \cdot x_i + b) \geq 1 - \xi_i & \forall i \\
  \xi_i \geq 0 & \forall i.
  \end{cases}
  \]
SVM - non-separable case (Ct’d)

Dual problem:

- Dual problem

\[
\begin{align*}
\max_{0 \leq \alpha \leq C} & \quad \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \\
\text{s.t.} & \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \quad (3)
\end{align*}
\]

- And at the optimum \( \bar{w} = \sum_{i=1}^{m} \alpha_i y_i x_i \).
- and \( b y_i = 1 - y_i \sum_{i=1}^{m} \alpha_j y_j x_j \cdot x_i \) when \( \xi_i = 0 \).
SVM - Analysis.

Rademacher complexity:

\[ R_\rho(h) = \frac{1}{m} \sum_{i=1}^{m} \phi_\rho(h(x_i) - y_i) \]

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

\( H = \{ h(x) = \text{sign}(w \cdot x + b), \|w\| \leq \Lambda, b \in \mathbb{R} \}. \)

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

\[ \forall h \in H : \ R(h) \leq R_\rho(h) + \frac{2}{\rho} \hat{R}_S(H) + 3 \sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \]
Rademacher complexity:

\[ \widehat{R}_S(H) = \frac{1}{m} E_{\sigma_1, \ldots, \sigma_m} \left[ \sup_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i) \right] \]

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

\[ \widehat{R}_S(H) \leq \sqrt{\frac{R^2 \Lambda^2}{m}}. \]
SVM - Analysis (Ct’d)

Proof:

\[ \hat{R}_S(H) \triangleq \frac{1}{m} E_\sigma \left[ \sup_{\|w\| \leq \Lambda} \sum_{i=1}^{m} \sigma_i (w \cdot x_i) \right] \]

\[ = \frac{1}{m} E_\sigma \left[ \sup_{\|w\| \leq \Lambda} w \cdot \sum_{i=1}^{m} \sigma_i x_i \right] \]

\[ \leq \frac{\Lambda}{m} E_\sigma \left[ \| \sum_{i=1}^{m} \sigma_i x_i \| \right] \]

\[ \leq \frac{\Lambda}{m} E_\sigma \left[ \sum_{i=1}^{m} \sigma_i \| x_i \|^2 \right]^{1/2} \]

\[ \leq \frac{\Lambda \sqrt{mR^2}}{m} = \sqrt{\frac{\Lambda^2 R^2}{m}}. \tag{4} \]
Talagrand’s Contraction Lemma:

\[ \hat{R}_S(\phi \circ H) \leq L \hat{R}_S(H) \]

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

\[ \triangleq \frac{1}{m} E_{\sigma_1, \ldots, \sigma_m} \left[ \sup_{h \in H} \sum_{i=1}^{m} \sigma_i(\phi \circ h)(x_i) \right] \]

\[ = \frac{1}{m} E_{\sigma_1, \ldots, \sigma_{m-1}} E_{\sigma_m} \left[ \sup_{h \in H} \sigma_m(\phi \circ h)(x_m) + u_{m-1} \right] \quad (5) \]

with \( u_{m-1}(h) = \sum_{i=1}^{m-1} \sigma_i(\phi \circ h)(x_i) \).
\[ \sum \frac{1}{m} E_{\sigma_1, \ldots, \sigma_{m-1}} E_{\sigma_m} \left[ \sup_{h \in H} u_{m-1}(h) + \sigma_m (\phi \circ h)(x_m) \right] \]

\[ = \frac{1}{2} [u_{m-1}(h_1) + \phi(h_1(x_m))] + \frac{1}{2} [u_{m-1}(h_2) - \phi(h_2(x_m))] \]

\[ \leq \frac{1}{2} [u_{m-1}(h_1) + u_{m-1}(h_2)] + sL \frac{1}{2} [(h_1(x_m)) - (h_2(x_m))] \]

\[ \leq E_{\sigma} \left[ \sup_{h \in H} u_{m-1}(h) + \sigma_m L h(x_m) \right], \quad (6) \]

with \( s = \text{sign}(h_1(x_m)) - (h_2(x_m)) \).
Kernels.

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

\[
H = \left\{ \sum_{i=1}^{m} \alpha_i y_i K(x_i, \cdot), \|\alpha\| \leq \Lambda \right\}.
\]
Conclusions

Take home messages:

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