TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

2.2 Linear Support Vector Machines

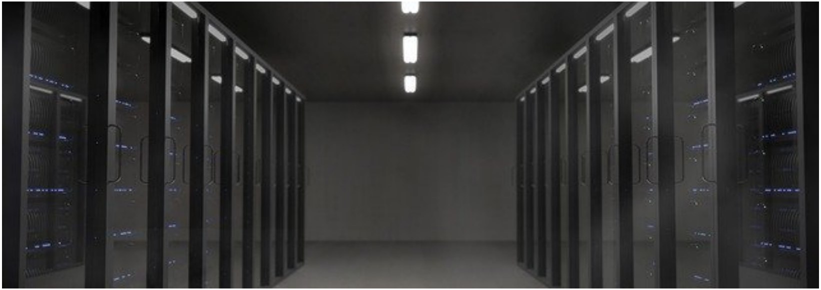
Machine Learning 1: Foundations

Marius Kloft *(TUK)*

28 Apr – 5 May 2020

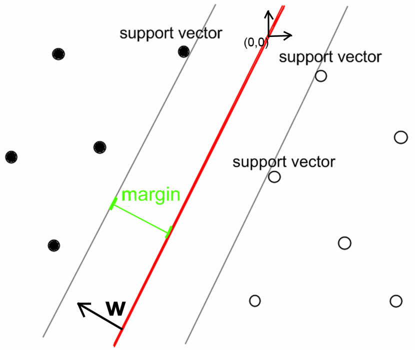
Properties of Linear SVMs: 1. **Fast**

- Can be trained in $O(n \cdot d)$
- Can be trained in a distributed manner (map-reduce)



Properties of Linear SVMs: 2. **Simple**

Geometrical interpretation:

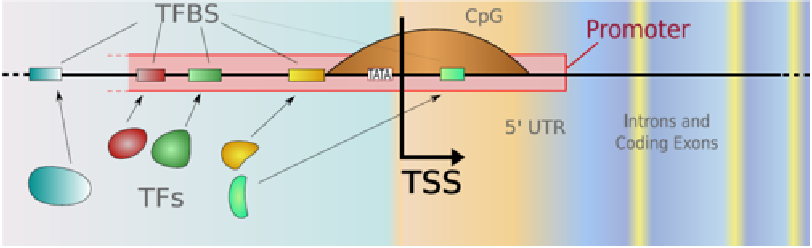


Properties of Linear SVMs: 3. **Accurate** (in ca. 50% of the data out there)

State of the art in many application areas, e.g.:

Gene Finding

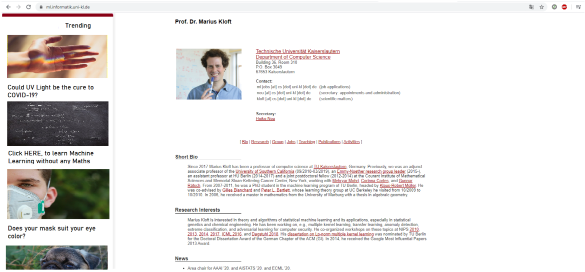
Find in the DNA the positions that impact virtually all important inherited properties of you!
(e.g., intelligence, height, visual appearance, etc.)



SVM State of the Art in...

Ad Click Prediction

Predict the ad that has the highest probability of being clicked by the user

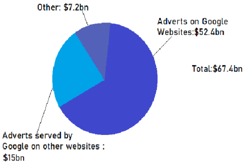


SVM has Generated Zillions and Zillions of Money ...

$75 billion Revenue

$23 billion operating profit

The money comes from:
USA:$34.8bn
UK:$7.1bn
Rest: $33.1bn

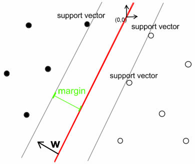


Google in 2015. Source: `http://www.bbc.co.uk/guides/z9x6bk7`

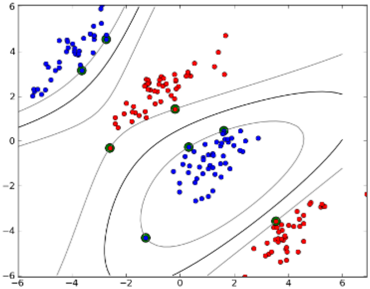
Types of (Linear) SVMs

There are two different sorts of linear SVMs:

1. **Hard-margin** linear SVMs
2. **Soft-margin** linear SVMs



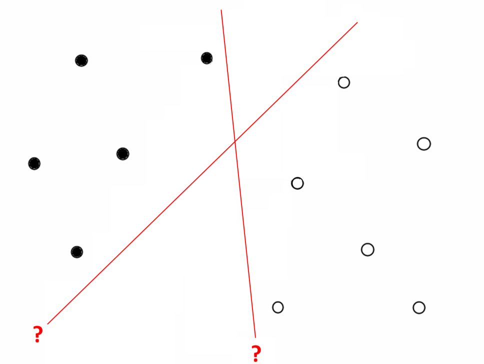
Later in the course, we will also learn about **non-linear** SVMs.



Linear Support Vector Machines

Core idea:

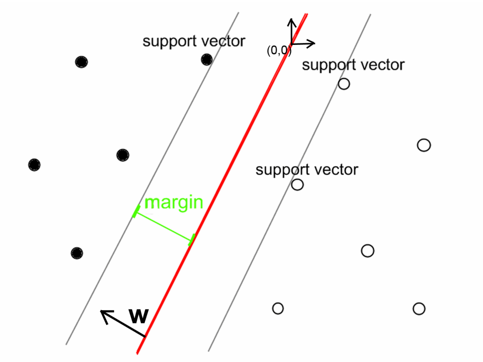
- Which hyperplane to take?



Linear Support Vector Machines

Core idea:

- Which hyperplane to take?
- The one that separates the data with the **largest margin**!



Mathematical Formalization

Maximize margin such that all data points lie outside of the margin — how can we **mathematically** describe this idea?

- Denote margin by γ
- Find hyperplane parameters $\mathbf{w}$ and b that maximize γ
- but make sure that all positive data points lie on one side
  - a point $\mathbf{x}_i$ with $y_i = +1$ lies on correct side of margin if:
    $$d(\mathbf{x}_i, H) \geq +\gamma$$
- and all negative points on the other
  - a point $\mathbf{x}_i$ with $y_i = -1$ lies on correct side of margin if:
    $$d(\mathbf{x}_i, H) \leq -\gamma$$
- Hence, we require for all points $\mathbf{x}_i$,
  $$y_i \cdot d(\mathbf{x}_i, H) \geq \gamma$$

Find hyperplane H with maximal margin γ such that (s.t.) for all training points $\mathbf{x}_i$ it holds: $y_i \cdot d(\mathbf{x}_i, H) \geq \gamma$.

Mathematical Formalization (2)

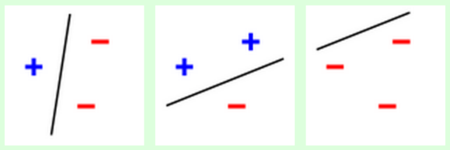
Recall:

- The hyperplane H is parameterized by $\mathbf{w}$ and b through:
$$H = \{\mathbf{x} : \mathbf{w}^\top \mathbf{x} + b = 0\}$$
- Previous proposition: $\quad d(\mathbf{x}, H) = \frac{1}{\|\mathbf{w}\|}\left(\mathbf{w}^\top \mathbf{x} + b\right)$

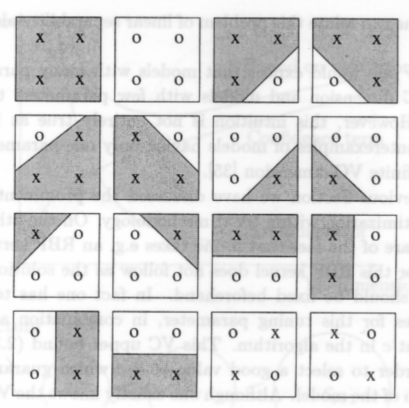
Preliminary formulation of SVM

$$\begin{aligned} \max_{\gamma, b \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d} \quad & \gamma \\ \text{s.t.} \quad & y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq \|\mathbf{w}\| \gamma, \quad \forall i = 1, \ldots, n \end{aligned}$$

Limitations of Hard-Margin SVMs



3 points shattered



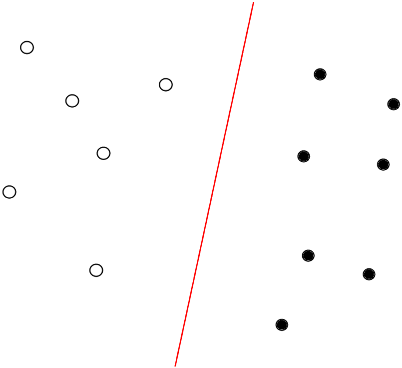
Any three points in the plane $\mathbb{R}^2$ (not lying on a line) can be “shattered ” (separated) by a line (= linear classifier).

But there are configurations of four points which no hyperplane can shatter. More generally:

Any $n+1$ points in $\mathbb{R}^n$ (not lying in a hyperplane) can be “shattered ” by a hyperplane. But there are configurations of $n+2$ points which no hyperplane can shatter.

Limitations Hard-Margin Linear SVMs (continued)

Another Problem is that of outliers potentially corrupting the SVM:



Remedy: **Soft**-Margin Linear SVMs

Core idea:

- Introduce for each input $\mathbf{x}_i$ a **slack variable** $\xi_i \geq 0$ that allows for some (slight violations of the margin separation):

Linear Soft-Margin SVM

$$\max_{\gamma, b \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d, \xi_1, \ldots, \xi_n \geq 0} \quad \gamma - C \sum_{i=1}^{n} \xi_i$$

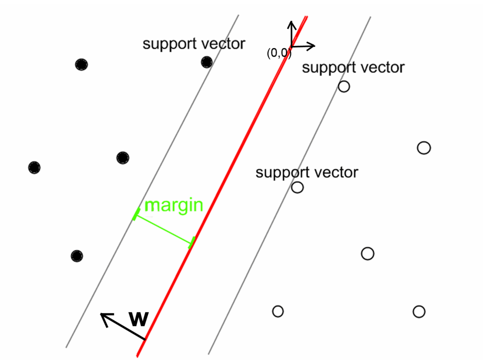
$$\text{s.t.} \quad y_i \left(\mathbf{w}^\top \mathbf{x}_i + b \right) \geq \|\mathbf{w}\| \gamma - \xi_i, \quad \forall i = 1, \ldots n$$

- minimizes $\sum_{i=1}^{n} \xi_i$, to allow for "slight" violations of the margin
- $\sum_{i=1}^{n} \xi_i$ is the total amount of violations of the training points lying inside the margin (measured in distances to the margin)
- $C > 0$ is a trade-off parameter (to be set in advance): the higher C, the more we penalize violations

Why the Name 'Support Vector Machine'?

Denote by γ^* and $\mathbf{w}^*$ the optimal arguments from previous slide

Def.: All vectors $\mathbf{x}_i$ with $y_i \cdot d(\mathbf{x}_i, H(\mathbf{w}^*, b^*)) \leq \gamma^*$ (i.e., lying inside the tube) are called **support vectors**.



The SVM depends only on the support vectors: all other points can be removed from the training data (no impact on classifier)

PanOpto Quiz

Alternatively, one could consider a variation of the SVM, where the penalty term $C\sum_{i=1}^{n}\xi_i$ is replaced by $C\sum_{i=1}^{n}\xi_i^2$. Would this be a reasonable SVM? (i.e., could this work similarly well as the original soft-margin SVM?)

Let us again remove the restriction $\xi_{\geq}0\forall i$ in the SVM's maximization. Let us remove from our training set all data points that lie strictly outside the margin. Will the decision boundary change?

SVM training

How can we train SVMs, that is, how to solve the minimization task?

Next Week: Convex Optimization Problems

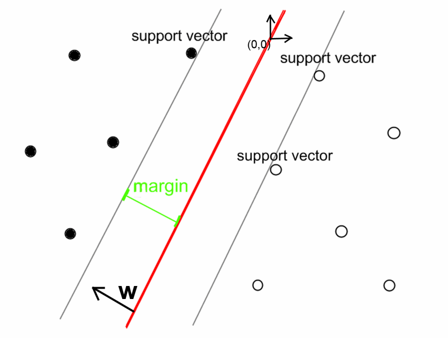
It is known from decades of research in numerical mathematics that so-called **convex optimization problems** (to be introduced in detail next week) can be solved very efficiently.

Will show: we can view the SVM as a convex optimization problem.

Conclusion

Linear Support Vector Machines (SVMs)

- motivated geometrically



Mathematical formalization of this picture:

$$\max_{\gamma, b \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^d} \quad \gamma$$

$$\text{s.t.} \quad y_i \left(\mathbf{w}^\top \mathbf{x}_i + b \right) \geq \|\mathbf{w}\| \gamma, \quad \forall i = 1, \ldots n$$

Suggested Reading

[1]: The Elements of Statistical Learning, Sections 4.5 and Section 12.2

Refs I

T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning, 2nd editiion. Springer series, 2009.