DMML, 3 March 2020

Classification – Supervised learning

Sentiment Analysis

Naively – each word has a score (+ve or −ve)

Compute weighted sum of scores of new review

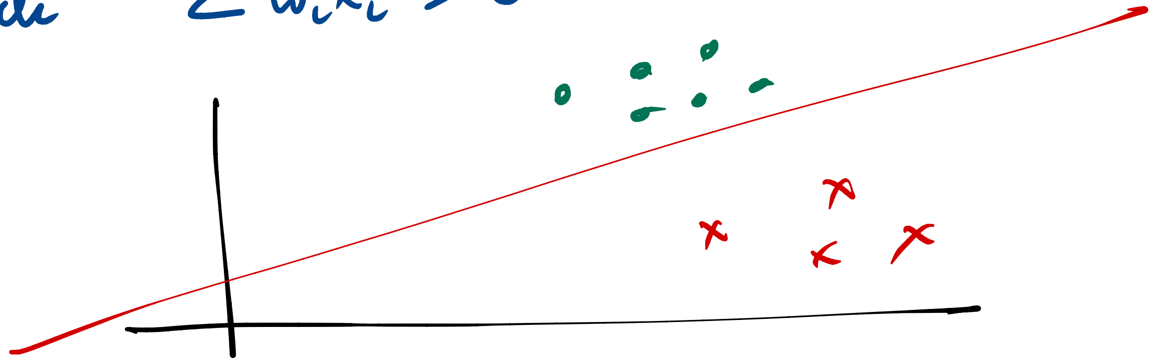
If above threshold – positive

score of word 2

Words are features/attributes : $x_1, x_2, \ldots, x_N$

Compute $\sum w_i x_i$ for words in the document

Check $\sum w_i x_i > t$



Linear separator

Geometric intrepretation of data

Separate categories by a hyperplane

Data items are of the form $\bar{x} = \langle x_1, x_2, \ldots, x_m \rangle$

$x_i \in \mathbb{R}$

Assume classification allows linear separability

$\exists\ \bar{w} = \langle w_1, w_2, \ldots, w_m \rangle$ such that

For each positive $\bar{x}$, $\bar{w} \cdot \bar{x} > t$

negative $\bar{x}$, $\bar{w} \cdot x < t$

How do we find $\bar{w}$?

Instead

$\bar{w}\cdot\bar{x} - t > 0$ Positive $\bar{x}$

$\bar{w}\cdot\bar{x} - t < 0$ Negative $\bar{x}$

Expand $\bar{w}$ to $\langle \bar{w}, -t \rangle \sim \hat{w}$

$\bar{x}$ to $\langle \bar{x}, 1 \rangle \sim \hat{x}$

write as $\bar{w}, \bar{x}$

$$\hat{w}\,\hat{x} > 0$$

$$\hat{w}\,\hat{x} < 0$$

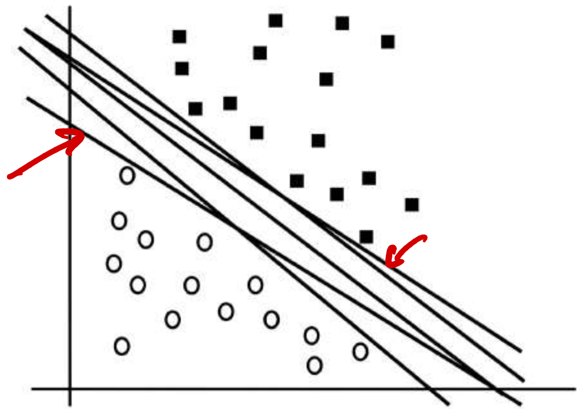
Each input $\overline{x}_i$ has a label — Yes / No

$l_i = +1 \quad -1$

<u>Now</u> $\forall i. \quad \overline{w} \cdot \overline{x}_i \cdot l_i > 0$

Direct way is to use linear programing

But, there is a simpler way



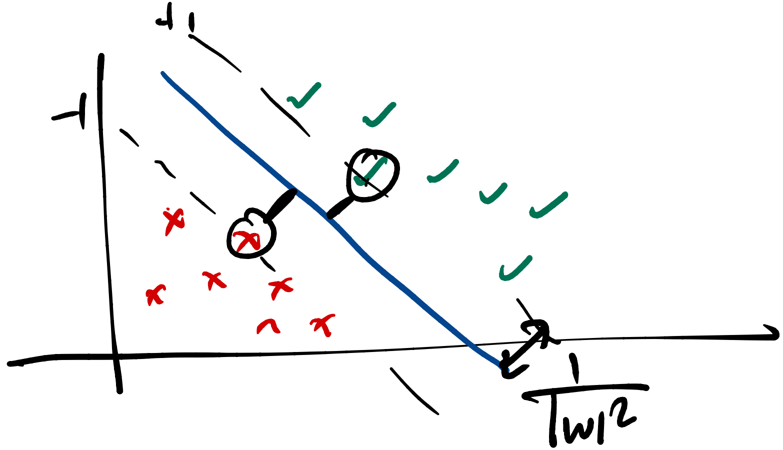
Many possible $\bar{w}$ may work

Right now, we are happy to find any one

Want w^* s.t. $w^* \cdot x_i \cdot l_i > 0 \quad \forall i$

Scale w^* s.t. $w^* \cdot x_i \cdot l_i > 1 \quad \forall i$

Distance of nearest point is $\dfrac{1}{|w_i^*|^2}$



Algorithm [Perceptron]

$w \leftarrow 0$

while there exists x_i s.t. $w \cdot x_i \cdot l_i \le 0$

$$w \leftarrow w + x_i l_i = \begin{array}{l} w + x_i \text{ if } x_i \text{ positive} \\ w - x_i \text{ if } x_i \text{ negative} \end{array}$$

Why does this converge?

Theorem

Suppose $\exists w^*$ s.t. $w^* \cdot x_i \cdot l_i > 1 \;\; \forall i$

Perceptron algorithm finds w s.t. $w \cdot x_i \cdot l_i > 0 \;\forall i$

in at most $r^2 |w^*|^2$ updates, where $r = \max |x_i|$

No guarantee on "quality" of w

Proof Keep track of $w^T w^*$, $|w|^2$

w current estimate

w^* is assumed

1. Each update to w increases $w^T w^*$ by at least 1

$$(w + x_i l_i)^T w^* = w^T w^* + \underbrace{x_i^T l_i w^*}_{\text{at least } +1}$$

Suppose $l_i = +1$ $\quad w^* x_i^T > 1 \quad > 1$

$l_i = -1 \quad w^* x_i^T < -1 \quad > 1$

2. Each update to w increases $|w|^2$ by at most r^2

$$(w + x_i l_i)^T (w + x_i l_i) = |w|^2 + \underbrace{2 x_i^T l_i w}_{\leq 0} + |x_i l_i|^2$$

$$\leq |w|^2 + |x_i|^2$$

Suppose we update w m times

$w^T w^* \geq m$ (grows by at least 1 each update)

$|w|^2 \leq m r^2$ (each update increases by at most r^2)

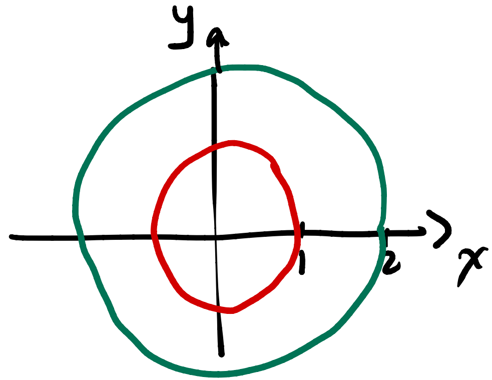
$$m \leq |w||w^*|$$

$$\frac{m}{|w^*|} \leq |w| \leq \sqrt{m}\, r$$

$$\sqrt{m} \leq r \cdot |w^*|$$

$$m \leq r^2 |w^*|^2$$

But what about the assumption of linear separability?



Good points - circle of radius 2

Bad points - circle of radius 1

Not linearly separable.

Geometric transformation.

$$\langle x, y \rangle \longrightarrow \langle x, y, x^2+y^2 \rangle$$

$z = 1.5$ plane separates

Given a transformation $\bar{x} \longmapsto \varphi(\bar{x})$

Perceptron

$$w = 0 \pm x_i \pm x_j \pm x_k \pm \ldots$$

$$w = \langle u_1, u_2, \ldots, u_m \rangle \underbrace{\langle x_1, x_2, \ldots, x_m \rangle}_{x_{train}}$$

Classify a new $\bar{z}$

$$\bar{w} \cdot \bar{z} > 0 \, ? \qquad (\bar{u} \cdot \bar{x}_{train}) \cdot \bar{z} > 0$$

$$(\bar{u} \cdot \overline{\varphi(\bar{x})}) \cdot \varphi(\bar{z}) > 0$$

All we need to know about φ is how to compute dot products $\varphi(\bar{x}_i) \cdot \varphi(\bar{x}_j)$

Suppose we have a function $k(\bar{x}_i, \bar{x}_j)$

s.t. $\forall \bar{x}_i, \bar{x}_j \quad k(\bar{x}_i, \bar{x}_j) = \varphi(\bar{x}_i) \cdot \varphi(\bar{x}_j)$

Such a k is called a <u>kernel</u> <u>function</u>

$\exists \varphi$. s.t. $k(\bar{x}_i, \bar{x}_j) = \varphi(\bar{x}_i) \cdot \varphi(\bar{x}_j) \quad \forall \bar{x}_i, \bar{x}_j$

Strategy Try out various kernels (don't worry about what φ looks like)

When is k a kernel ?

Minimum requirement - symmetric

Non constructive characterization - Mercer's Theorem

Constructive definition in terms of positive semidefinite matrices