Lecture 10: Pattern Recognition week 47.

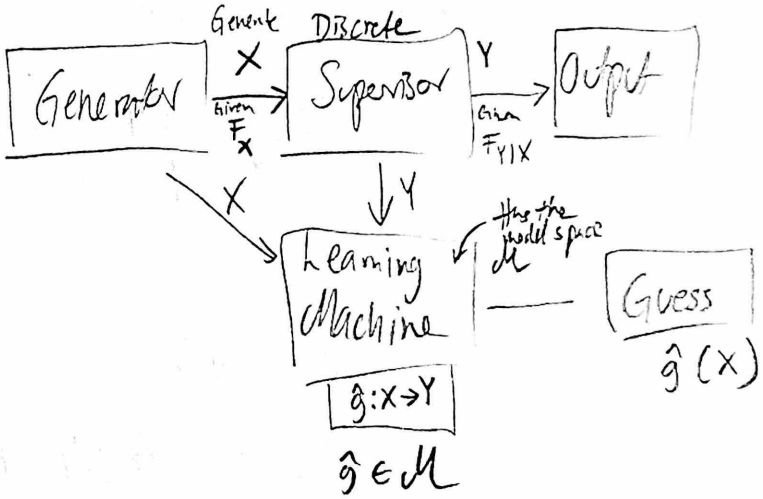
Recap

- Markov chains.
  - Homogenous Markov chains:
    - homogenous in time.
    - dry wet example.
    - Markov can be used on language: what's most likely the next word?
    - Google's Page Rank was based on Markov chains.

state space

The pattern recognition problem

Perceptron later became neural networks



in the pattern recognition problem we are assuming

$F_{Y|X}$ is discrete (a class)

g_λ is a decision function / decision rule / cla

0-1 loss

$$\lambda(z,u) = \begin{cases} 0 & \text{if } y = u \\ 1 & \text{if } y \neq u \end{cases}$$

0-1 loss is mistake cost 1, otherwise 0 co

$z = (x, y)$, u is the guess.

Pattern Recognition

$$R(A) = \int L((x,y), g_A(x))\, dF(x,y)$$

(this is the true risk for $g_A \in \mathcal{A}$.)

The goal is to minimize the true risk $R(A)$

Example

Bayes Classification Rule

Let $r(x) = E[Y|X]$, then the Bayes classification rule h^* is

$$h^*(x) = \begin{cases} 1 & \text{if } r(x) > 1/2 \\ 0 & \text{otherwise} \end{cases}$$

else:

$$Y \in \{0,1\}$$

$$E[Y|X] = P(Y=1|X) = r(x)$$

if $r(x) > 1/2 \Rightarrow h^*(x) = 1$

if $r(x) < 1/2 \Rightarrow h^*(x) = 0$

- This is what scikit-learns LogisticRegression does! (function predict)

Instance & label space

$\mathcal{X}$: instance

$\mathcal{Y}$: label space

d different binary variants → $\mathcal{X}$ can be $\{0,1\}^d$

- $\mathbb{X}$: data space
- $\mathbb{X} = (\mathcal{X}, \mathcal{Y})$ kan skrivas som $\mathbb{X} \times \mathcal{Y}$
- n samples

linear seperator

= a line that seperates. Easy in 1D or 2D. for higher dimensions we need perceptron.

The perceptron algorithm

- a linear classification rule.

- The dimension trick

A prerequisite. Check slides & lecture notes

Perceptron alg

Goal is to find: $(w \cdot x_i)\, y_i > 0$

$$g_w(x) = \begin{cases} 1 & \text{if } w \cdot x > 0 \\ (-1) & \text{if } w \cdot x < 0 \end{cases}$$

$(w_i \cdot x_i)\, y_i > 0$ means that you have guessed the correctly.

if $(w \cdot x_i) y_i > 0 \quad \forall i = 1, \ldots, n$

$$\frac{1}{n} \sum_{i=1}^{n} L\Big((x_i, y_i), g_w(x_i)\Big) = 0.$$

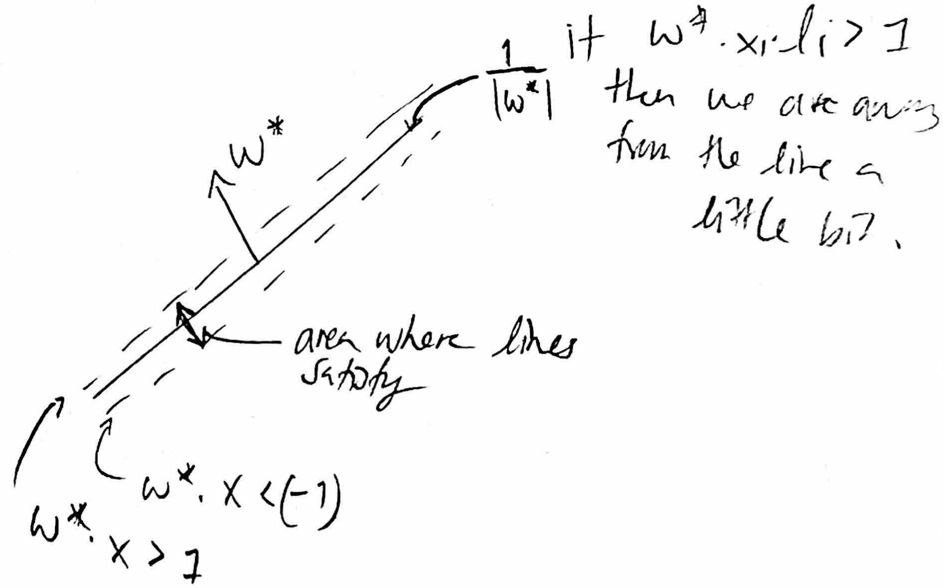
Tentaplugg
Ind. ass.
Problemsessions
Chapter readings
Computer exercises.
Ex. 20 godkänd.
extra poäng = 1/4 p.

the algo

if you misclassify something that should've been 1, then $w := w + x_i y_i$

if you misclassify something that should've been (−1) then $w := w - x_i y_i$.

w^*

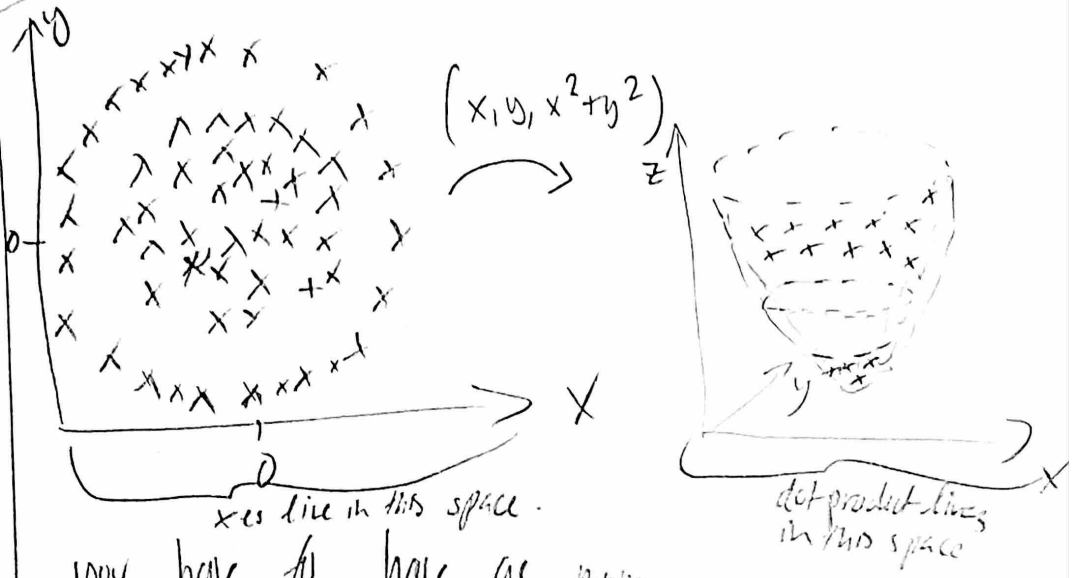


the closer $w^* \cdot x < -1$ or $w^* \cdot x > 1$ are to the line, there are fewer lines that satisfy.

Convergence is guaranteed if the data is separated?

→ check slides.

Kernelization



you have to have as many parameters as you have dimensions.

$l_i = y_i$ i have notes.

A Kernelized perceptron algorithm

$K_{ij} = \phi(x_i) \cdot \phi(x_j) = \phi(x) \cdot \phi(y)$

matrix i have slides

1. $c = 0$
2. While there exists an i such that $(Kc)_i y_i \leq 0$ update $c_i := c_i + y_i$

Kernel functions

we only need to know the kernel to use it, we do not need to know ϕ.

c_i holds the number of times we have added or subtracted x_i.