Piazza please answer.

Recap :- Logistic Regression.

- $\ell(\bar{w}^*; T)$ ⟶ Log-likelihood of training samples under LR.
- $\nabla_{\bar{w}^*} \ell = \begin{bmatrix} \frac{\partial}{\partial \bar{w}_0^*} \ell(\bar{w}^*; T) \\ \vdots \end{bmatrix}$

$$\frac{\partial \ell(\bar{w}^*; T)}{\partial \bar{w}_j^*} = \sum_{i=1}^{n} \bar{x}_j(i) \cdot \left[f(\bar{x}(i)) - Pr\left(f(\bar{x}(i)) = 1 \mid \bar{x}(i), \bar{w}^* \right) \right]$$

- Compare more change for probability of label 'small'.

(-) Discussion session :- (next)

LR vs Perceptron vs Naive Bayes vs SVM

- what are the positives and negatives.

Support Vector Machines. | * references here

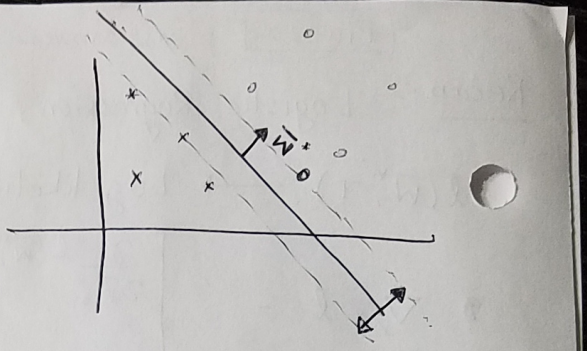
- Vladimir Vapnik - Founder(s) of Statistical learning theory.

1992 :- accuracy comparable to NN on image data sets.

Now :- everywhere...

Agenda :- 1. Maximum margin classifiers.
2. Non-linear classifiers & Kernel methods.

Maximum-margin classifiers.



Suppose the training data is linearly separable.

Margin of a perfect classifier for the data is the "maximum width" that the boundary could be increased before we hit a data point.

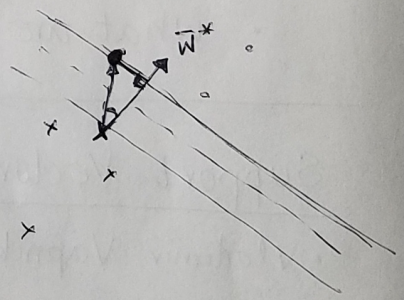
Learn a "~~large~~ maximum margin classifier.

- better generalization

How to find the MMC?

Recall :- decision rule :- $\bar{x}^* \cdot \bar{w}^* > 0 \longrightarrow +1$

$\bar{x}^* \cdot \bar{w}^* < 0 \longrightarrow -1$ }.



Appropriately normalize

For $f(\bar{x}(i)) = 1$, $\bar{x}^*(i) \cdot \bar{w}^* \geq 1$

For $f(\bar{x}(i)) = -1$, $\bar{x}^*(i) \cdot \bar{w}^* \leq -1$.

Suppose, i, i' are such that,

$\bar{x}^*(i) \cdot \bar{w}^* = 1$

$\bar{x}^*(i') \cdot \bar{w}^* = -1$ }

$$\text{distance} = \frac{(\bar{x}^*(i) - \bar{x}^*(i')) \cdot \bar{w}^*}{\|\bar{w}^*\|_2}$$

$$= \frac{(\bar{x}(i) - \bar{x}(i')) \cdot \bar{w}}{\|\bar{w}\|_2} = \frac{2}{\|w\|_2}$$

Formulation :- **Maximize $\frac{2}{\|\bar{w}\|_2}$** subject to

$$f(\bar{x}(i)) \cdot \left(\bar{x}(i) \cdot \bar{w} + w_0 \right) \geq 1.$$

Minimize $\frac{1}{2} \|\bar{w}\|_2^2$, s.t. ↓

Quadratic objective function, and linear constraints.

Lagrangian :-

$$\text{minimize } L_p(\bar{w}, w_0, \alpha_i) = \frac{1}{2} \|\bar{w}\|_2^2 - \sum_{i=1}^{n} \alpha_i \cdot \left(f(\bar{x}(i)) \cdot (\bar{x}(i) \cdot \bar{w} - w_0) - 1 \right)$$

and $\alpha_i \geq 0$.

$$\frac{\partial L_p}{\partial \bar{w}} = \bar{0} \Rightarrow \bar{w} - \sum_{i=1}^{n} \alpha_i \cdot f(\bar{x}(i)) \cdot \bar{x}(i) = 0$$

$$\Rightarrow \bar{w} = \sum_{i=1}^{n} \alpha_i \cdot \left(f(\bar{x}(i)) \cdot \bar{x}(i) \right) \quad -(1)$$

KKT conditions : (Complementary slackness) [Discussion]

At optimum :- $\alpha_i \; ; \; f(\bar{x}(i)) \cdot \left(\bar{x}(i) \cdot \bar{w} - w_0 - 1 \right) = 0$ for all i.

∴ All samples not at the boundary (support) have $\alpha_i = 0$

SV :- support vectors.

$$\bar{w} = \sum_{i \in SV} \alpha_i f(\bar{x}(i)) \cdot \bar{x}(i)$$

LC of things.

what if the data is not linearly classifiable?

Add slack variables for misclassification :-

PRIMAL { minimize $\frac{1}{2}\|\bar{w}\|_2^2 + C \cdot \sum_{i=1}^{n} \xi_i$, such that

$f(\bar{x}(i)) \cdot (\bar{x}^*(i) \cdot \bar{w}^*) \geq 1 - \xi_i$, $\xi_i \geq 0$.

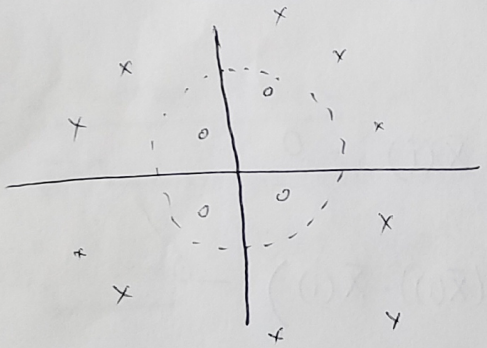
$C \rightarrow$ overfitting parameter

→ GO TO DOTPROD

MNIST data-set :-
Linear SVM :- 8.5% error
Polynomial SVM :- 1% error

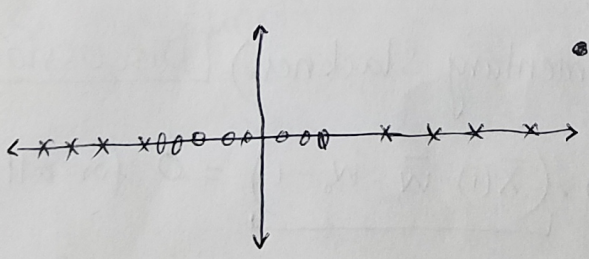
Non-linear SVM :- Kernel trick.

Linear classifiers are not complicated enough sometimes.



- No linear classifier would be great.
- Need non-linear decision regi[illegible].

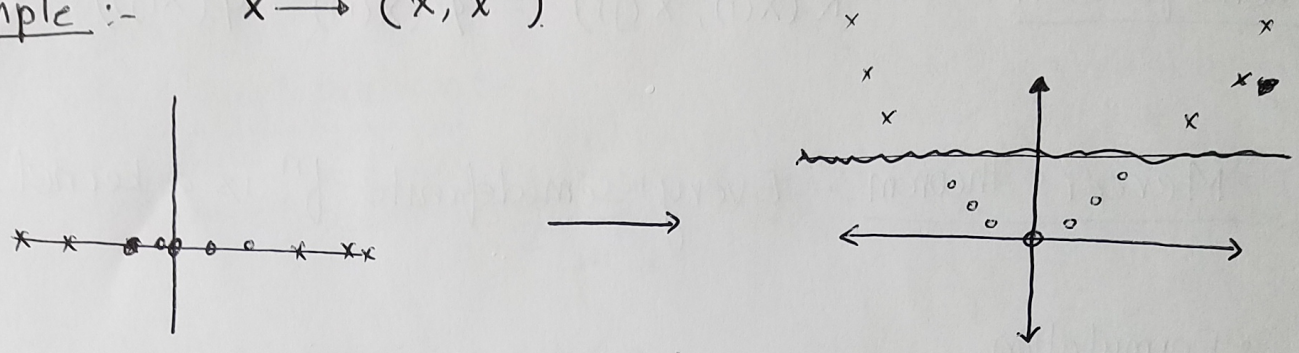
→ ① obtaining this is hard
② ~~po~~ linear SVM's very well understood.

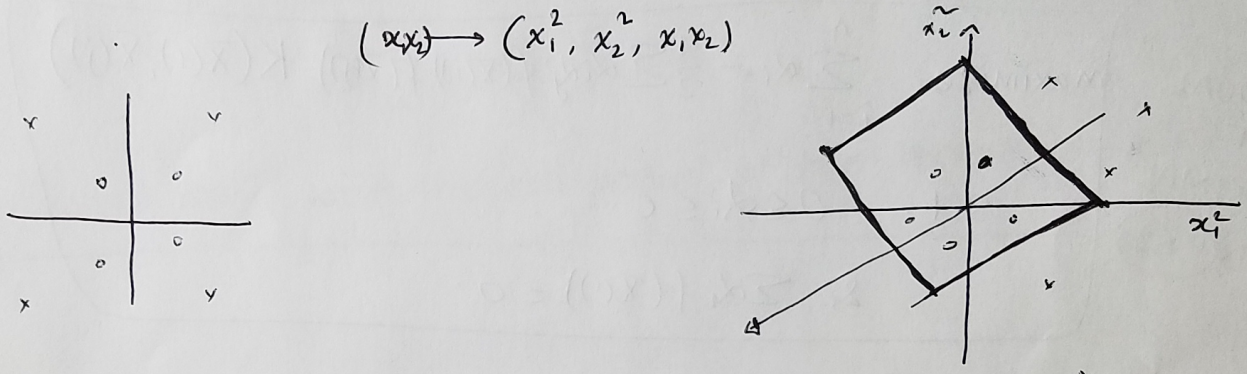


General idea :- ① Map the features to a higher dimensional space.
② Apply a linear classifier in the new space. SVM.

$\phi : R^k \rightarrow R^m$, $(m > k)$ is the function mapping the features to the new space.

Example :- $x \longrightarrow (x, x^2)$ $\quad k=1,\ m=2.$





The kernel trick :-

$\bar{w}^* \longrightarrow$ weight vector after training the SVM in the 'new' space, (i.e, in m dimensions)

$= (\bar{w}_0^*, \ldots, \bar{w}_m)$

So, the function for classification is now :-

$$g(\bar{x}) = \bar{w}^* \cdot \phi(\bar{x}) + \bar{w}_0$$

$$= \left[\sum_{i \in SV} \alpha_i \, \phi(\bar{x}(i)) \right] \cdot \phi(\bar{x}) + \bar{w}_0$$

$$= \sum_{i \in SV} \alpha_i \, \phi(\bar{x}(i))^T \cdot \phi(\bar{x}) + \bar{w}_0$$

We do not need to know the explicit mapping, we only use the dot-product of feature vectors in training & test.

Kernel function :- $K(\bar{x}(i), \bar{x}(j)) = \phi(\bar{x}(i))^T \cdot \phi(\bar{x}(j))$

Mercer's Theorem :- Every ~~te~~ semidefinite f^n is a kernel.

- Formulation

DUAL OF PRIMAL ON (P62)

maximize $\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j f(\bar{x}(i)) f(\bar{x}(j)) K(\bar{x}(i), \bar{x}(j))$

s.t. $0 \le \alpha_i \le C$

& $\sum \alpha_i f(\bar{x}(i)) = 0$

- Linear Kernel :- $k(\bar{x}(i), \bar{x}(j)) = \bar{x}(i)^T \cdot \bar{x}(j)$
- Poly " :- $K(\bar{x}(i), \bar{x}(j)) = (1 + x(i)^T x(j))^p$
- Gaussian /RBF :- $K(\bar{x}(i), \bar{x}(j)) = \exp\left(-\frac{\|\bar{x}(i) - \bar{x}(j)\|_2^2}{2\sigma^2}\right)$

DOT PRODUCT :-

Lagrange Dual problem :-

$$\text{maximize} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j f(\bar{x}(i)) f(\bar{x}(j)) \cdot \boxed{\bar{x}(i)^T \cdot \bar{x}(j)}$$

$$\text{subject to} \quad 0 \le \alpha_i \le C$$

$$\sum_{i=1}^{n} \alpha_i f(\bar{x}(i)) = 0$$

You work only with the dot product of the vectors, nothing else.