NPTEL Week-5 Live Session

on Machine Learning and Deep Learning - Fundamentals and Applications (noc24_ee146)

A course offered by: Prof. Manas Kamal Bhuyan, IIT Guwahati

NPTEL Quiz Solution: Week-4

Powered by:

PMRF
Prime Minister’s Research Fellows
Ministry of Education
Government of India

By

Arka Roy
NPTEL PMRF TA

Prime Minister’s Research Fellow
Department of Electrical Engineering, IIT Patna
Web: https://sites.google.com/view/arka-roy/home

2) To avoid the problem of ambiguous region of linear discriminant function for c categories, we can

- Define c linear function $g_i(x)$, one for each class for $i = 1, 2, \ldots, c$ (✓) ✓
- Assign x to w_j if $g_i(x) < g_j(x)$ for all $i \neq j$ (✓)
- Take a linear machine classifier (✓)
- ✓ All the above

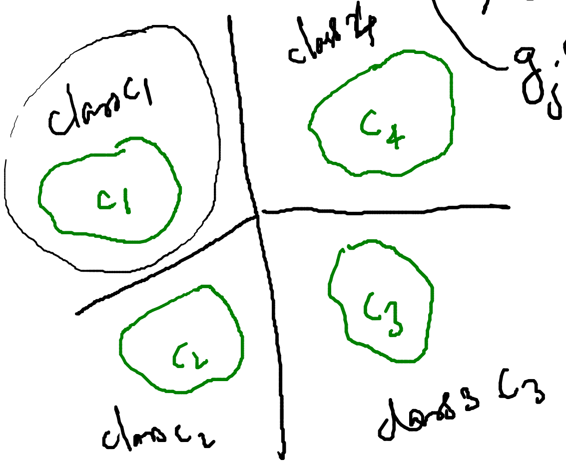
$x \in w_j$; $g_j(x) > g_i(x)$; $i \neq j$

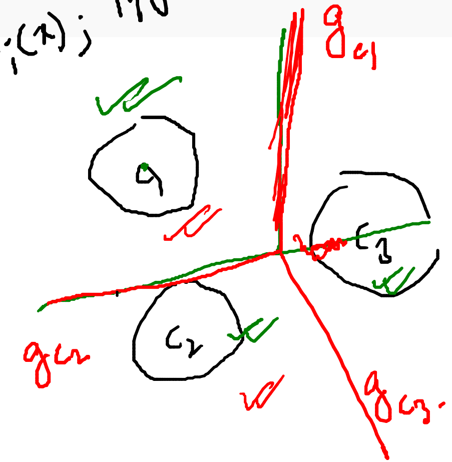
You want to predict

$\underbrace{x \in c_1}$: $g_{c_1}(x)$ ↑↑

$g_{c_2}(x)$

$g_{c_3}(x)$





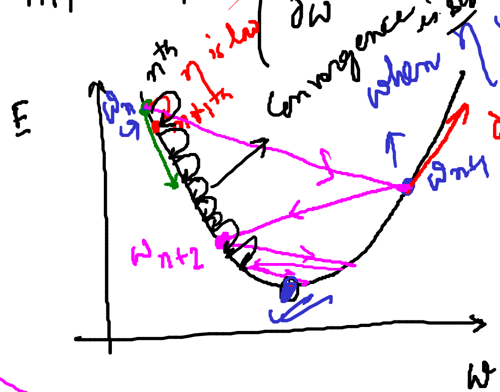
$\underbrace{g_{c_1}(x) > g_{c_2}(x) > g_{c_3}(x)}$

linear function $g_{c_1}(x)$ ↑ ∴ $x \in c_1$

3) Which of the following statements is true about the learning rate in Gradient Descent?

- A very high learning rate may lead to oscillation
- A lower learning rate may lead to faster convergence ⇒ Convergence is slow.
- The learning rate doesn't determine the size of the steps taken towards the minimum
- The learning rate has no effect on the convergence of Gradient Descent

$W_{n+1} = W_n - \eta \frac{\partial E}{\partial W}$ E = error function



η is very low. (η = 0.00001) → The increment of the weight parameter or change in the value of W will be very slow.

→ Convergence of the gradient optimization process is slow.

$W_{n+1} = W_n - \eta \frac{\partial E}{\partial W} \rightarrow (-100)$

$= W_n - (0.00001 \times -100)$

$= W_n + 0.001$ → very small value.

Say $\frac{\partial E}{\partial W} = -100$

η is very high → There will be large no. of oscillations

η ↑↑ η = 100

$W_{n+1} = W_n - (100 \times -100)$

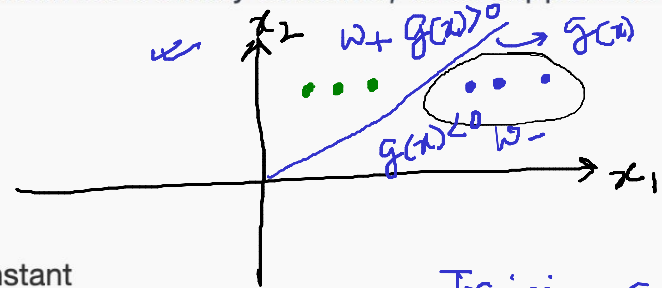
$W_{n+1} = W_n + 10^4$

→ $W_{n+2} = W_{n+1} - (100 \times 10)$

$= W_{n+1} - 10^3$

4) In the Perceptron algorithm for a binary classifier, what happens to the weights when a positive misclassified point is encountered?

- It remains the same
- It is increased
- It is decreased
- It is multiplied by a constant



Rule:- $g(x) > 0$: $x \in \omega_+$
$g(x) < 0$: $x \in \omega_-$

Decision Rule is not unified

To make a uniform decision Rule:-

Training set:-

$x \in \omega_+$: $g(x) > 0$ → What signifies proper classification

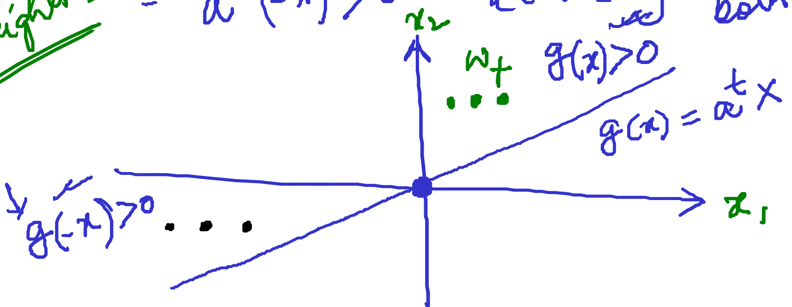
$x \in \omega_-$ ⇒ you will negate the value of x : $g(-x) > 0$ → Then again it signifies that you have done correct classification

Unified decision Rule

Learnable weights

$g(x) = a^t x > 0$: $x \in \omega_+$
$= a^t(-x) > 0$ $x \in \omega_-$

you decision Rule for both the class are now uniform.



For misclassification:-

when $x \in \omega_+$ - $g(x) = a^t x = (-ve)$ Misclassification

$g(x) = a^t x < 0$ → Loss.

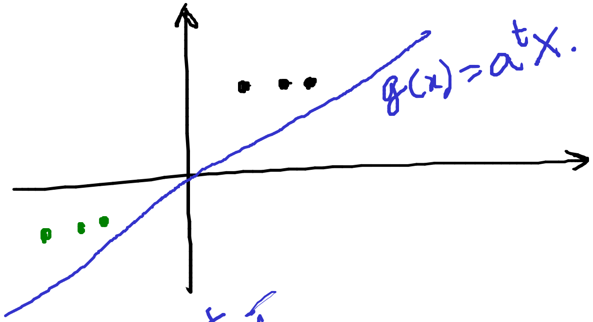
$$\text{Loss} = \sum_{\text{for } \forall x \text{ misclassified}} -a^t x$$

4) In the Perceptron algorithm for a binary classifier, what happens to the weights when a positive misclassified point is encountered?

- ◯ It remains the same
- ✓ It is increased
- ◯ It is decreased
- ◯ It is multiplied by a constant

Loss function $(L(a)) = \sum -a^t x$
$\forall x$ misclassified.

Training decision / Decision Rule $\Rightarrow$ $g(x) = a^t x > 0$; $x \in \omega_1$
$= a^t(-x) > 0$; $x \in \omega_2$



misclassification:-
$x \in \omega_1$: $g(x) = \underbrace{a^t x}_{} < 0$
$= -a^t x$

$L(a) = \sum_{x \in \omega_1} -a^t x$
$= \sum (+ve)\, x\, (-ve)$
$= \sum -a^t x.$
$=$

$a_{n+1} = a_n - \eta \nabla L(a)$

$a_{n+1} = a_n - \eta \frac{\partial L}{\partial a}$

$\frac{\partial L}{\partial a} = \sum -x.$

$= a_n - \eta\left(\sum -x\right)$

$\boxed{a_{n+1} = a_n + \eta \sum x}$

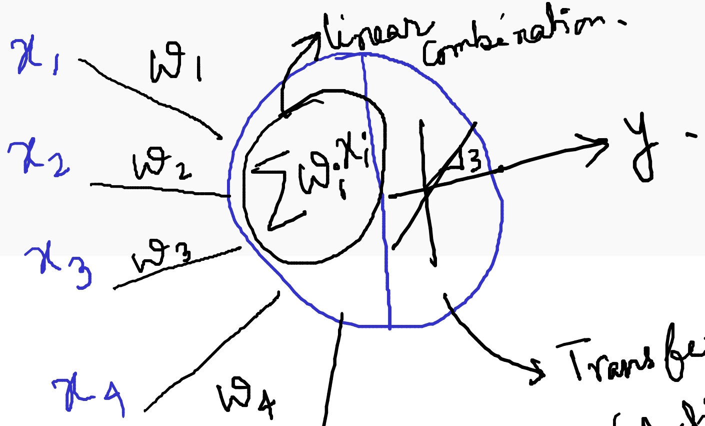
5) Let w_{ij} represents weight between node i at layer k and node j at layer (k-1) of a given multilayer perceptron. The weight updation using gradient descent method is given by: (α and E represent learning rate and Error in the output respectively)

- $W_{ij}(t+1) = W_{ij}(t) + \alpha \frac{\partial E}{\partial W_{ij}}, 0 \le \alpha \le 1$
- $W_{ij}(t+1) = W_{ij}(t) - \alpha \frac{\partial E}{\partial W_{ij}}, 0 \le \alpha \le 1$
- $W_{ij}(t+1) = \alpha \frac{\partial E}{\partial W_{ij}}, 0 \le \alpha \le 1$
- $W_{ij}(t+1) = -\alpha \frac{\partial E}{\partial W_{ij}}, 0 \le \alpha \le 1$

$$\underbrace{W_{ij}(t+1) = W_{ij}(t) - \alpha \frac{\partial E}{\partial W_{ij}}}_{\text{Gradient descent equation}} ; 0 \le \alpha \le 1$$

6) A 4-input neuron has weights 3, 4, 5 and 6. The transfer function is linear with the constant of proportionality being equal to 3. The inputs are 6, 12, 10 and 20 respectively. What will be the output?

- ○ 238
- ○ 76
- ○ 708 ✓
- ○ 123



$$\hat{x} = \sum w_i x_i$$

$$= (w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4)$$

$$= (6\times3 + 4\times12 + 5\times10 + 20\times6)$$

$$\hat{x} = 18 + 48 + 50 + 120 = 236.$$

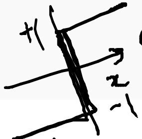
$$\boxed{\hat{x} = 236}$$

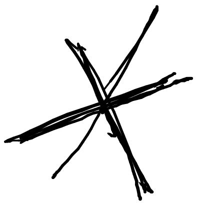
$$\frac{y}{\hat{x}} = 3.$$

$$y = 3\hat{x}$$

$$= 3\times236$$

$$\boxed{y = 708.}$$





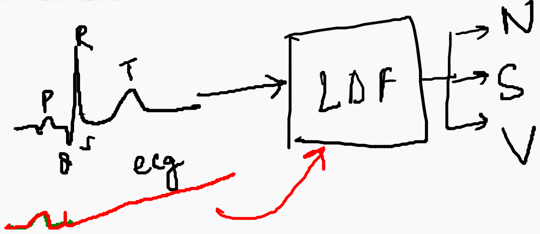
$$\frac{y}{x} = 3.$$

$$y = 3x.$$

$$y = mx$$

7) Which of these is true about discriminant classifiers?

- ~~Assume conditional independence of features~~
- Robust to outliers
- Can perform classification if some missing data points are present
- All the above



$g(x) = P(\omega/x) = P(\omega)\,P(x/\omega)$

→ Distribution.
→ Gaussian Distribution.

$x = [x_1, x_2, \ldots, x_d]$.

$P(x_i/\omega) = \frac{1}{\sqrt{2\pi}\,|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu)\right)$

Covariance matrix

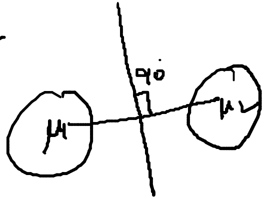
Multivariate gaussian dist.

$\Sigma = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \sigma_3^2 \ldots & \\ 0 & & & \sigma_d^2 \end{bmatrix}$ statistically independent

$\Sigma_i = \Sigma_j \quad \forall i,j$

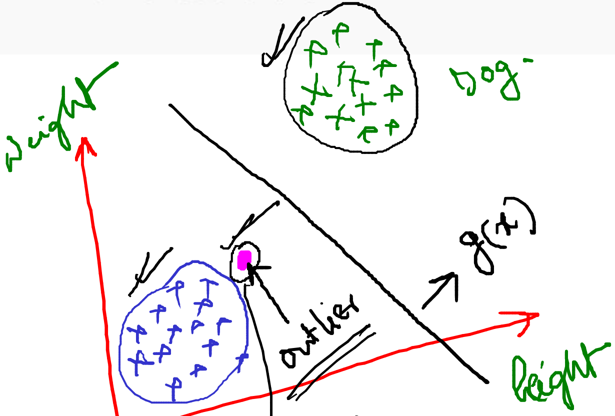
Σ → arbitrary in nature

$(\mu_1 - \mu_2) \cdot w^t = 0$



7) Which of these is true about discriminant classifiers?

- ~~Assume conditional independence of features~~
- ~~Robust to outliers~~ ✗
- ~~Can perform classification if some missing data points are present~~ → Imputation of data
- All the above



Prediction will be = Cat.

↳ Misclassification

Cat → height,
Dog. → weight.

• → Actually Dog.

→ But it's data representation seems to have similar characteristics as the cat class.

8) A set of training samples are given below-

x_1	x_2	y	λ_i
0.38	0.47	+	65.52
0.49	0.61	-	65.52
0.92	0.41	-	0
0.74	0.89	-	0
0.18	0.58	+	0
0.41	0.35	+	0
0.93	0.81	-	0
0.21	0.10	+	0

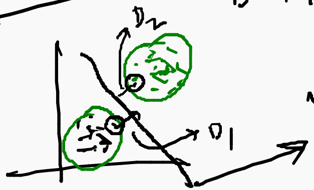
Support vectors from each of the classes.

$w_i = \sum_i \lambda_i y_i x_i$

Class specific equation would be :->

$w^T x + b = 1; \; y \in +$

$D = D_1 + D_2 \uparrow\uparrow$



$w_1 x_1 + w_2 x_2 + b = 0$

Generic equation of line that SVM predicts.

Using Support vector machine algorithm, the Marginal line for the classification can be calculated as-

- $-5.32x_1 - 7.193x_2 + 9.09 = 0$
- $-6.67x_1 + 8.134x_2 - 9.09 = 0$
- $-7.21x_1 - 9.173x_2 + 9.09 = 0$
- $8.21x_1 + 7.12x_2 - 9.09 = 0$

$w_1 = \sum_i \lambda_i y_i x_{1i}$

$w_1 = \lambda_1 y_1 x_{11} + \lambda_2 y_2 x_{12}$

$= 65.52 \times 1 \times (0.38) + 65.52 \times (-1) \times (0.49)$

$w_1 = -7.20$

$w_2 = \sum_i x_{2i} y_i \lambda_i = \lambda_1 y_1 x_{21} + \lambda_2 y_2 x_{22}$

$= 65.52 \times 1 \times 0.47 + 65.52 \times (-1) \times 0.61$

$w_2 = -9.17$

8) A set of training samples are given below-

x_1	x_2	y	λ_i
0.38	0.47	+	65.52
0.49	0.61	-	65.52
0.92	0.41	-	0
0.74	0.89	-	0
0.18	0.58	+	0
0.41	0.35	+	0
0.93	0.81	-	0
0.21	0.10	+	0

$w_1 x_{11} + w_2 x_{21} + b_1 = 1$

$w_1 x_{12} + w_2 x_{22} + b_2 = -1$

$\Rightarrow b_1 = -(-7.2 \times 0.38 - 9.173 \times 0.47) + 1$

$= 8.05$

$b_2 = 10.12.$

$b = \frac{b_1 + b_2}{2}$

$= 9.085$

≈ 9.09

Using Support vector machine algorithm, the Marginal line for the classification can be calculated as-

- $-5.32x_1 - 7.193x_2 + 9.09 = 0$
- $-6.67x_1 + 8.134x_2 - 9.09 = 0$
- $-7.21x_1 - 9.173x_2 + 9.09 = 0$
- $8.21x_1 + 7.12x_2 - 9.09 = 0$

The optimal eq^n of line

$-7.2x_1 - 9.17x_2 + 9.09 = 0$

9) In refer to Q.8, A new test sample (0.5,0.5) is found. The class of the given sample is-

- ✓ Positive
- Negative
- Both class
- Can't say

Q.8. Ans:- $\underbrace{-7.21x_1 - 9.173x_2 + 9.09}_{g(x_1, x_2)} = 0$.

$g(0.5, 0.5) = -7.21 \times 0.5 - 9.173 \times 0.5 + 9.09$

$= 0.8985 = (+ve)$

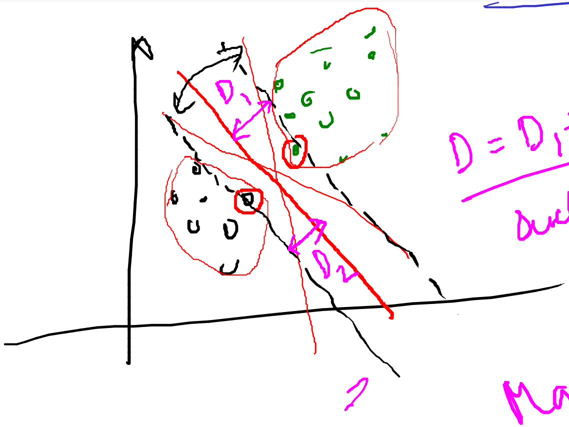
$g(\frac{1}{2}, \frac{1}{2}) > 0$

(+ve) → class identity = + class.

10) What is the main objective of a Support Vector Machine (SVM)?

- ~~To maximize the number of support vectors~~
- ~~To minimize the margin between classes~~
- ~~To maximize the training accuracy~~ effect
- ✓ To find a hyperplane that separates classes with the maximum margin

Cause



$D = D_1 + D_2$ ↑↑

such that $Y_i(w^T x + b) > 0$

Maximize the margin corresponding to a particular decision boundary.