2) in MDP we have $\langle X, A, \delta, r \rangle$

X: finite set of state A: finite set of actions

$\delta: X \times A \to X$ is the transition function and $X \times A \to \mathbb{R}$: r → reward function

in POMDP we have $\langle X, A, Z, \delta, r, O \rangle$ where

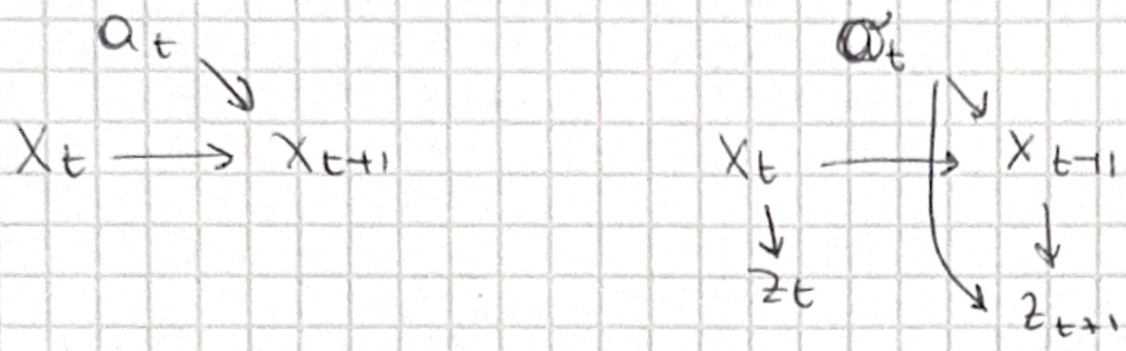
X: finite set of states A: finite set of actions Z: finite set of

$\delta: X \times A \to X$ transition function $r: X \times A \to \mathbb{R}$ is the observation

reward function and O is the probability of observation and

$P(x_0)$ is the probability of the initial state

Ex 3.



Ex. 6

1. $\dim(W_1) = 128 \cdot 10 = 1280$ $\dim(W_2) = 50 \cdot 10 = 500$

2 yes $y(x) = w^T \max(0, w^T x + c) + b$ $h = g(w^T x + c)$

$g(\alpha) = \max(0, \alpha)$ → ReLU

ML EXAMS 18/01/2018

Ex. 1

input image = 1242 × 378 × 3 (RGB) kernel = 5×5 pad = 2 stride = 1

$W_{out} = W_{in} - \frac{W_k}{\text{stride}} + 2P + 1 = 1242$

$d_{out} = d_{in} - \frac{d_{ker}}{\text{stride}} + 2P + 1 = 378$

~~number of trainable parameters: 5 × 5 × 3 × 64~~

$W_{out} = \frac{W_{in} - d_{pol}}{2} + 1 = 621$

$d_{out} = \frac{378 - d_{pol}}{2} + 1 = 189$

} ~~after C1~~

} after pool 1

~~621 × 189 × 64~~

~~310 × 9~~

$W_{out} = 310$ $h_{out} = 84$

$W_{out} = 78$ $h_{out} = 24$

$C_1 \to 1242 \times 378 \times 64$

$P_1 \to 621 \times 189 \times 64$

$C_2 \to 310 \times 84 \times 128$ $P_2 \to 78 \times 24 \times 128$

2) 1 ✕ param = 3×5×5 + 64 = 1664

2 ✕ param = 64 × 3 × 3 × 128 + 128 = ~~73984~~ 73856

EX. 2

1. in a greyscale image of 12×12 pixels so the dimensionality of data space is $2^{12\times12}$. the intrinsic dimensionality is 3 → two coordinates for the center one for rotation

2. we want to maximize the variance of after the projection to some direction u_1 we know the projected points are $u_1^T x_n$ subject to $u_1^T u_1 = 1$

the variance is $\frac{1}{6}\sum_{n=1}^{6}(u_1^T x_n - u_1^T \bar{x_n})^2 = u_1^T S u_1$

with $S = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T$

we want maximize $\max_{u_1} u_1^T S u_1$

if we keep derivative and we impose that is equal to 0

$$Su_1 = \lambda_1 u_1 \Rightarrow u_1^T S u_1 = \lambda_1$$

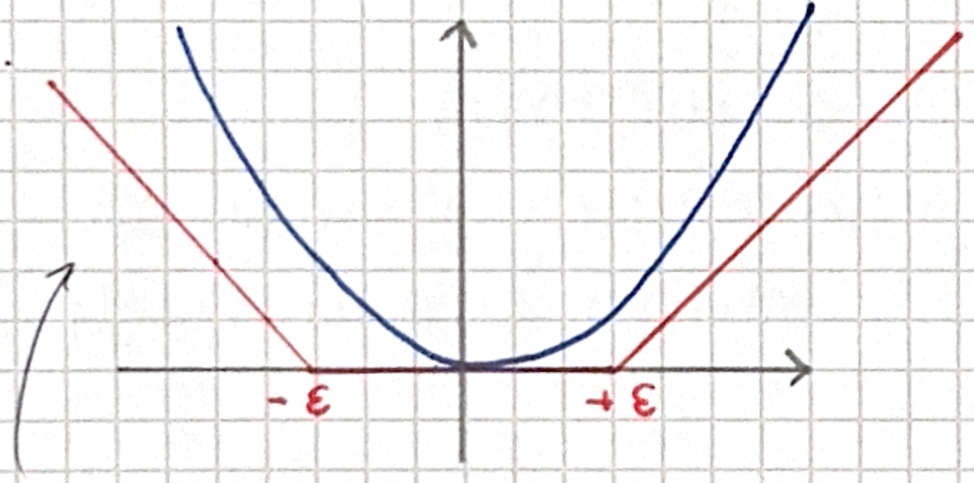
S is the eigenvector associated to the largest eigenvalue λ_1

this is called first principal component

3. No. usually M is greater than the number of intrinsic dimension because the PCA are not latent variable

Ex. 3

1.



doesn't penalize too much outliers

is difficult to solve because is not differentiable

2. $\xi < 0$ is ON or inside the correct margin boundary

$0 < \xi_n \leq 1$ if points inside margine but correct side

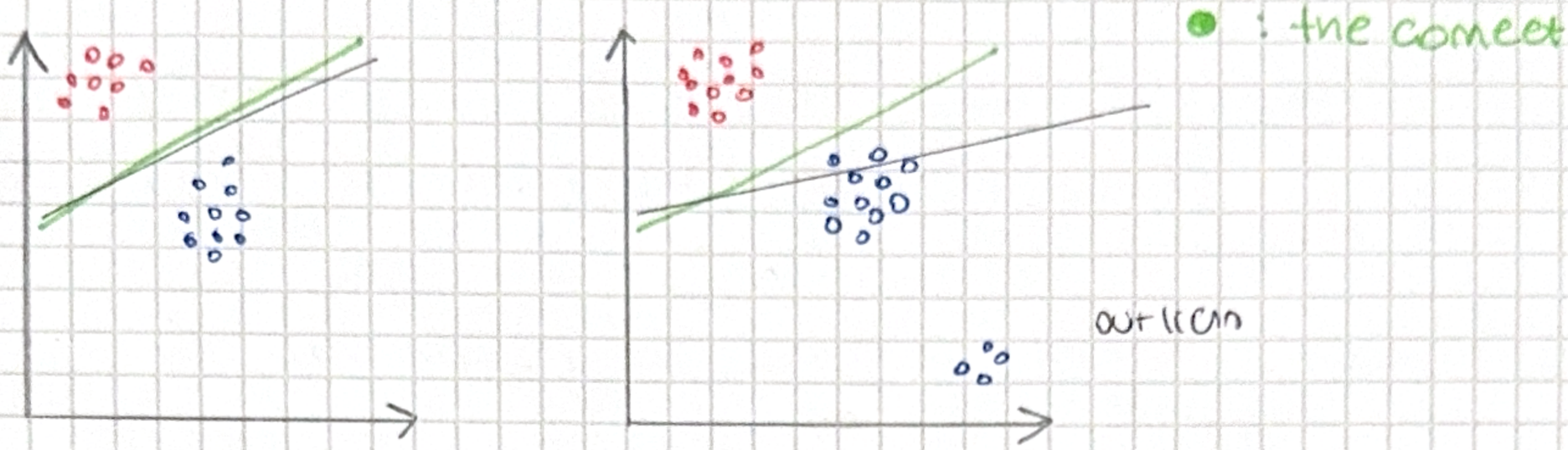
$\xi_n > 1$ if points in the incorrect boundary

optimization with slack variables become:

$$w^*, w_0^* = \text{argmin}\ \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n$$

the slack variables are introduced to allow misclassifications or margin violations in soft margin classifications

EX.4



LEAST SQUARES

given a dataset $D = \{(x_n, t_n)_{n=1}^{N}\}$ find the linear discriminant:

$$y(x) = \tilde{w}^T \tilde{x}$$

the goal is minimize the sum of square error function:

$$E(\tilde{w}) = \frac{1}{2} Tr\{(\tilde{X}\tilde{w} - T)^T (\tilde{X}\tilde{w} - T)\}$$

closed form $\tilde{w} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T T$

learned model $y(x) = \tilde{w}^T \tilde{x} = T^T (\tilde{X}^T)^{\dagger} \tilde{x}$

$$k = \underset{j \in \{1,\dots,K\}}{\operatorname{argmax}} \; y_j(x)$$

is not robust because is simply based on distances

is called in this way because the train operation is simply an sum and the product is between one matrix to other is only the square.

EX.3

1.

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} \; P(h|D) = \underset{h \in H}{\operatorname{argmax}} \; \frac{P(D|h)P(h)}{P(D)} \quad \text{if}$$

$P(h_i) = P(h_j) \quad h_{ML} = \underset{h \in H}{\operatorname{argmax}} \; P(D|h)$

2. BOE is an optimal classifier that returns always the optimal solution given a target function $f: X \to Y$

$$P(v|x, D) = \sum_{h \in H} P(v|x, D, h) P(h|x, D) = \sum_{h_i \in H} P(v|x, h_i) P(h_i|D)$$

$$v_{OE} = \underset{v}{\operatorname{argmax}} \sum_{h_i \in H} P(v|x, h_i) P(h_i|D)$$

3. Boc is an optimal classifier if the hypothesis space is not to large and if have an analitical solutions otherwise is not pratical ~~and~~ because ~~Naive~~ use Naive Bayesian classifier

EX. 6

once the current state is known the evolution of a dynamic system doesn't depend on previous state, actions and observations. the current state contains all information needed to predict the future

future state are conditionally independent of past states and past observations statistically irdipendent

MDP can be describe $\langle X, A, f, r \rangle$ X: finite set of states
A: finite set of actions $f: X \times A \to X$ transition function and
r: reward function

HMM can be describe $\langle X, Z, \Pi_0 \rangle$ where X, Z is set of observations
and Π_0 initial distribution
MDP not the property of fully observability

MDP

$X_t \to X_{t+1}$
$\uparrow$
a_t

HMM

$X_t \to X_{t+1} \to \dots \to X_N$
$\downarrow \quad \downarrow \quad \downarrow$
$z_t \quad z_{t+1} \quad z_N$

ML EXAMS 15/10/2018

EX. 1

1. how many times an element of class C_i in classifies ~~as~~ in the class ej

2.

\	A	B	C
A	40	50	10
B	10	80	10
C	5	5	80

\	A	B	C	← predicted
A	40%	50%	10%	
B	10%	80%	10%	
C	5%	5%	80%	

↳ true values

3. the accuracy in a confusion matrix is the num of elements in the diagonal ~~matrix~~ divide by the all of elements of matrix

accuracy = $(40+80+80)/300 = 0.7$