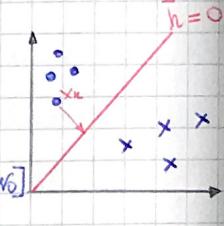


EXAM 12.02.19 B

Question 1

1). SVM aims at maximum margin with the better accuracy. Let x_n be the closest point to the separation surface $\bar{h} = \bar{w}_0 + \bar{w}^T x = 0$. We define

the margin as $\frac{|y(x_n)|}{\|\bar{w}\|}$. To find the margin we have to find $\min_{n=1 \dots N} \frac{1}{\|\bar{w}\|} t_n [\bar{w}^T x_n + \bar{w}_0]$



$w^*, w_0^* = \operatorname{argmax}_{w_0, w} \frac{1}{\|w\|} \min_{n=1 \dots N} t_n [\bar{w}^T x_n + \bar{w}_0]$. If we scale we will not affect the solution.

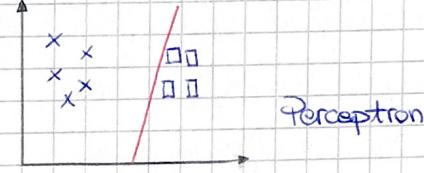
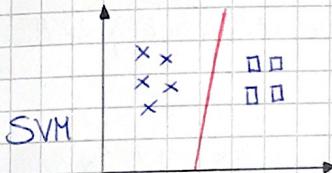
$t_n y(x_n) = 1$ and $t_n y(x_n) \geq 1 \forall n = 1 \dots N \Rightarrow$ The closest point has margin 1.

$w^*, w_0^* = \operatorname{argmin}_{w_0, w} \frac{1}{2} \|w\|^2$. Now we move to a Lagrangian domain; we know that, for the

KKT hypothesis, if $t_n y(x_n) \geq 1 \Rightarrow \alpha_n^* = 0$. Follows that $w^* = \sum_{n=1}^N \alpha_n^* t_n x_n$ and:

$$y(x) = w_0^* + \sum_{j \in SV} \alpha_j^* t_j x_j^T x = 0 \quad \text{with } SV \equiv \{x_n \in D : t_n y(x_n) = 1\}$$

2).



Perceptron

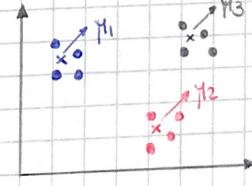
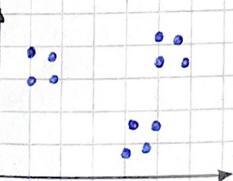
3). The solution with maximum margin is preferred because in this mode the probability that we will misclassify a sample is very low because we have a "safety mask". In perceptron the probability of misclassification is very high because usually we have a separation surface that is very close to the samples of one class. In this mode the result produced by SVM is more stable than the result of perceptron.

Question 2

1). The Gaussian Mixture model (GMM) is $P(x) = \sum_{k=1}^K \pi_k N(\mu_k, \Sigma_k)$

With π_k prior probability of the class k , μ_k mean of the gaussian and Σ_k covariance matrix.

2).



3). 2 unknown parameters for the priors, the third is automatically determined.

3 matrices μ_1, μ_2, μ_3 (2×1 vectors)

3 matrices Σ . 2×2 Symmetric matrices \Rightarrow unknown parameters each matrix

$$2 + (3 \cdot 2) + (3 \cdot 3) = 14$$

Question 3

True. Accuracy is a good performance metric if the dataset is balanced; otherwise no. We do an example:

Binary classification task with $f: X \rightarrow \{+, -\}$ and $D = \{(x_i, y_i)\}_{i=1}^N$.

The dataset is unbalanced, 90% of positive samples and 10% of negative.

$h_1(x) = +$ $h_2(x) =$ computed with a learning algorithm (85% of accuracy)

It's possible to see that the accuracy of h_1 is 90%. But this is not a true accuracy, is given by the fact that the dataset is unbalanced. The accuracy of h_2 is true.

Question 4

- Convolutional stage: in this stage we apply the convolution operation between the input and the kernel: $I * k(i, j) = \sum_{m \in \text{res}_1} \sum_{n \in \text{res}_2} I(m, n) k(i-m, j-n)$

We adopt two principles in a convolutional stage:

- Sparse connectivity: outputs depend only on a few inputs. We are not connecting all the units to all the successive units (like in FC layers).
- Parameter sharing: we force some parameters to share the same value. In this mode we reduce computational complexity and we improve invariance to translation.

We can have also padding to convolve without skipping any pixel.

- Detection stage: non linear activation function (ReLU, sigmoid).

- Pooling stage: max pooling returns the maximum of a rectangular region, average pooling averages

Question 5

1). In exploration the agent explores the environment doing random actions, while in exploitation the agent uses the knowledge obtained from exploration to maximize the reward.

In fact the goal of the agent is to find an optimal policy function, that is a function that maximizes the reward. It's very important to find a trade-off between exploration and exploitation because if we do too much exploration we can not converge to the optimal policy, if we do too much exploitation it's possible that our policy seems optimal because maximizes the reward but is not optimal.

2). We can use the ϵ -greedy policy. We do exploration with probability ϵ and exploitation with probability $1-\epsilon$. ϵ becomes a hyperparameter of our algorithm.