EXAM 19.01.21

Question 1

1). This is a regression problem. The target function is $f: X \to \mathbb{R}$ with $X = \mathbb{R}^3$ and the dataset is $\mathcal{D} = \{(x_n, f_n)_{n=1}^N\}$. We know that our model is $y(x,w) = \sum_{j=1}^{M} w_j \phi_j(x)$. The target value is $f = y(x,w) + \varepsilon$ with ε additive Gaussian noise. Using the Gaussian assumption: $p(\varepsilon|\beta) = N(\varepsilon|0, \beta^{-1}) \longrightarrow p(f|x_1 \dots x_N, w, \beta) = N(f|y(x,w), \beta^{-1})$. Now we consider the iid hypothesis: $p(\{f_1 \dots f_3\}|x_1 \dots x_3, w, \beta) = \prod_{n=1}^{3} (f_n|w^T\phi(x_n), \beta^{-1}) =$

$$= \sum_{n=1}^{3} \ln N(f_n|w^T\phi(x_n), \beta^{-1}) = -\beta \cdot \frac{1}{2} \sum_{n=1}^{3} (f_n - w^T\phi(x_n))^2 - \frac{3}{2} \ln(2\pi\beta^{-1})$$

We want to minimize the error function $E_D(w) = \frac{1}{2} \sum_{n=1}^{3} (f_n - w^T\phi(x_n))^2$

2). We can find $w^* = \underset{w}{\operatorname{argmin}}\, E_D(w)$ using a sequential algorithm:

$\hat{w} \leftarrow \hat{w} + \eta[f_n - w^T\phi(x_n)]\phi(x_n)$ with η very important hyperparameter called learning rat

Question 2

1). Yes, the dataset is linearly separable because exists a separation surface that splits our instance space into two regions such that different classified instances are separated.

2). Least squares, because this method is not robust to outliers and is possible to see in the plot that there are outliers. In fact least squares try to minimize the error function $E(w) = \frac{1}{2} Tr\{(T - \tilde{X}\tilde{w})^T(T - \tilde{X}\tilde{w})\}$. It's possible to see that this function is a ge metric distance so will be affected by samples that are distant from the main probability.

3). SVM, because this method is robust to outliers. In fact SVM aims at maximum margin with the better accuracy, so in this case outliers will not modify the solutio

Question 3

1). Binary classification task $f: X \to \{+, -\}$

Dataset $\mathcal{D} = \{(x_i, y_i)_{i=1}^N\}$ with 90% of positive samples

$h_1(x) = +$ $h_2(x)$ computed with an algorithm (85% of accuracy).

It's possible to see that in this case the accuracy is not a good performance metric because the dataset is unbalanced. $h_1(x) = +$ has the 90% of accuracy, but this is not a true accuracy, is given only by the dataset. A true accuracy is the accuracy

of the second hypothesis.

2). We can introduce other performance metrics:

$$\text{Recall} = \frac{\#\text{True positive}}{\#\text{Real positive}} \qquad \text{Precision} = \frac{\#\text{True positive}}{\#\text{Predicted positive}}$$

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Question 4

1). When the input vector appears only in the form of an inner product $x^T x'$ we can replace the inner product with a kernel function $k(x, x')$.

2). We can consider SVM for classification and we modify this method; for binary classification we know that: $w^* = \sum_{n=1}^{N} a_n^* t_n x_n$ and $w^* = w_0^* + \sum_{x_j \in SV} a_j^* t_j x^T x_j = 0$

with a^* Laplace coefficients, t_n target values and $SV = \{x_n \in D : t_n y(x_n) = 1\}$

Now we try to kernelize this method:

$y(x, w^*) = \text{Sign}(w_0 + \sum_{n=1}^{N} a_n^* x_n^T x)$ is possible to see the inner product

$y(x, w^*) = \text{Sign}(w_0 + \sum_{n=1}^{N} a_n^* k(x_n, x)$ we have applied the kernel trick

$y(x, w^*) = \text{sign}(w_0 + \sum_{n=1}^{N} a_n^* (\beta x_n^T x + \gamma)^d$ with a polynomial kernel function

$w_0 = \frac{1}{|SV|} \sum_{x_i \in SV} (t_i - \sum_{x_j \in SV} a_j^* t_j k(x_i, x_j))$

Question 5

1). The output units are: $y(x, w) = w^T h + b$. We know that $h = g(W^T x + c)$

with $g(a) = \max(0, a)$ ReLU function $\Rightarrow y(x, w) = w^T \max(0, W^T x + c) + b$

A suitable activation function is Softmax and units saturate with a small error.

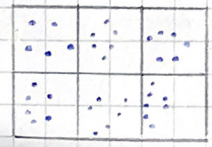
2). The loss function is categorical cross-entropy:

$J(w) = E[-\ln p(t|x, w)]$ if we assume additive gaussian noise:

$p(t|x, w) = N(t|f(x, w), \beta^{-1}) \Rightarrow J(w) = E\left[\frac{1}{2}\|t - f(x, w)\|^2\right]$

Question 6

1). PCA can be used for problems like dimensionality reduction, feature extraction, data compression and visualization.

2). 

We have 12x12 greyscale images. The dimensionality of this dataset is 12x12, but the intrinsic dimensionality is 3, 2 coordinates x and y for the center of the image and one for the rotation.