Machine Learning 1 Exam

This is an exam protocol made from memory (Gedächtnisprotokoll). I hope it helps you prepare for the next exams, but keep in mind the solutions are how I solved it and not necessarily 100% accurate.

Each question was worth 20 points.

Good luck :)

Exercise 1 :

Multiple Choice

There is only one correct answer.

— Which of the following is **True** for a Bayes optimal classifier

1. It represents a theoretical framework that gives the lowest possible error rate
2. It refers to a classifier that relies solely on Bayesian probability theory
3. It is synonymous with the Naive Bayes classifier, using independence assumptions
4. It is a specific algorithm that always outperforms other classifiers

Solution: 1

— The expectation minimization algorithm [...]

1. Is usually used for un supervised learning tasks
2. Handles well missing or hidden data
3. Requires labelled data to work
4. ?

Solution: ?

— Which of the following is **True** in the context of bias-variance decomposition

1. Higher bias always leads to lower variance, and vice versa
2. High bias is indicative of underfitting, while high variance suggests overfitting
3. Increasing the model complexity will generally decrease both bias and variance
4. Bias measures the algorithm's flexibility, while variance measures accuracy

Solution: 1

— Why does PCA maximize eigenvalues ?

1. To classify the data into different clusters

2. To directly compute the mean and standard deviation of each variable
3. To normalize (or whiten) the data
4. To identify the direction that maximizes the variance in the dataset

Solution: 4

— Which of the following is **True** : In the soft-margin SVM, the parameter C controls ?

1. Number of training points that are allowed to be misclassified
2. Number of test points that are allowed to be misclassified
3. By what amount the training points can lie not on the correct side of the margin
4. How nonlinear the margin is allowed to be

Solution: 3

Exercise 2 :

Parameter Estimation

The average time to get a letter at the post office follows the following distribution : $p(x|\theta) = \theta(1-\theta)^{(x-1)}$. The variable X is a positive integer (Z^+, and θ is a real number.

(a) Define the likelihood function $p(D|\theta)$
(b) Calculate the likelihood of $D = \{1, 1, 2, 1\}$
(c) Now consider a Bayesian approach, with the following probability distribution :

$p(\theta) = 1$, for $\theta \in [0, 1]$
$p(\theta) = 0$, elsewhere.

Prove that the posterior can be defined as $30 * \theta^4(1-\theta)$

(d) Evaluate the probability of $P(x > 1)$ with $\int p(x|\theta)p(\theta|D)$

Solution:

(a) $\prod p(x_k|\theta) = \prod_i \theta(1-\theta)^{(x_i-1)}$
(b) $\theta_{hat} = \frac{4}{5}$(log likelihood)
(c) For $\theta \in [0, 1]$: $\frac{p(D|\theta)p(\theta)}{\int_0^1 p(D|\theta)p(\theta)} = \frac{\theta^4(1-\theta)}{\frac{1}{5}-\frac{1}{6}} = 30 * \theta^4(1-\theta)$
(d) $P(x > 1) = 1 - P(X <= 1) = 1 - P(X = 1)$, given that x is a positive integer (0 not included). $\theta = 1 - \frac{5}{7} = \frac{2}{7}$

Exercise 3 :

Kernel

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined on a set $\mathcal{X}$ is called a *Positive Semi-Definite (PSD) Kernel* if, for any finite set of points $\{x_1, x_2, \ldots, x_n\} \subseteq \mathcal{X}$ and any corresponding set of coefficients $\{c_1, c_2, \ldots, c_n\} \subseteq \mathbb{R}$, the following condition holds :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$$

for all $n \in \mathbb{N}$ and for all choices of $\{x_1, x_2, \ldots, x_n\}$ and $\{c_1, c_2, \ldots, c_n\}$.

(a) Given the following kernel : $k_f = f(x)k(x, x')f(x')$. Prove it is a psd kernel.

(b) Show that the Gaussian kernel is also a psd kernel, with $k_f = exp(\gamma \cdot \frac{1}{2}||x - x'||^2)$. Also define function f(x) for this case. *Hint : you can use the following kernel definition :* $k(x, x') = exp(\gamma \cdot x \cdot x')$*, and use your answers from a).*

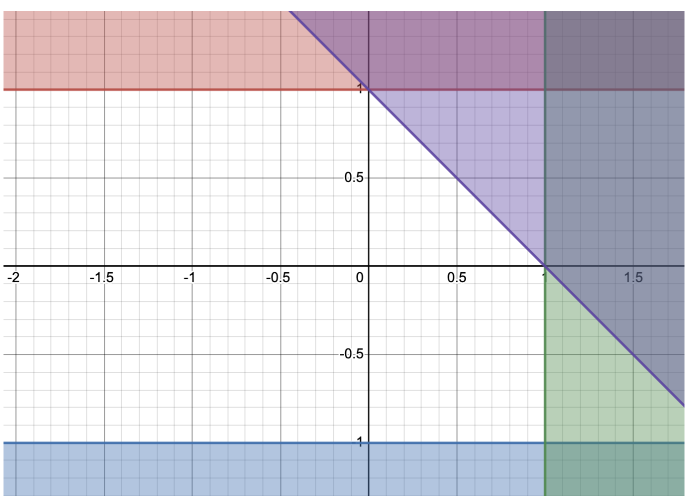
Solution:
(a)
(b)

Exercise 4 :

Neural Networks

Consider the following Neural Network with activation function :

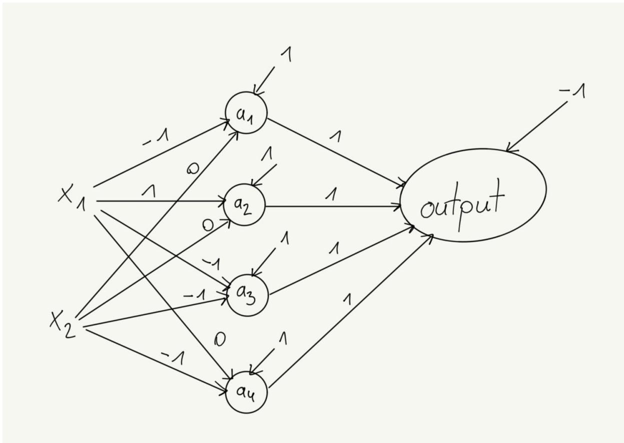
$$\text{step}(x) = \begin{cases} 1 & \text{if } a_i > 0 \\ 0 & \text{if } a_i \leq 0 \end{cases}$$



(a) Give all weights and biases.

(b) Describe values for all activated neurons for $x = (1, 1)$.

Solution:

(a) 

(b) a1 = 1, a2 = 0, a3 = 1, a4 = 1, output = 1

Exercise 5 :

Ridge Regression. Using a Quadratic solver

Given a labeled dataset $((x_1, y_1), ..., (x_N, y_N))$ we consider the regularized regression problem :

$$\min_{\mathbf{w}} ||\mathbf{y} - \mathbf{w}^T\mathbf{X}||^2$$

subject to $\quad 0 \leq w_i \leq C \quad$ and $\forall i : \quad \sum_i w_i \leq D$,

with $C, D \in \Re, w \in \Re^d$ and $X \in \Re^{Nxd}$.

(a) Show that this problem is equivalent to a problem of this type :
$\max_v \mathbf{v}^T(\mathbf{X}^T\mathbf{X})\mathbf{v} - 2\mathbf{y}^T\mathbf{X}\mathbf{v}$, subject to the same constraints.

(b) Implement a code in Python to calculate w. You can use the cvxopt.qp solver, that already implements the optimization problem in the following format :

$\max_v \mathbf{v}^T\mathbf{Q}\mathbf{v} - \mathbf{l}^T\mathbf{v}$ s.t. $\mathbf{A}\mathbf{v} \leq \mathbf{b}$

Solution:

(a) min $y^Ty - 2y^TXw^TX + w^TwX^TX \equiv$ min $v^TX^TXv - 2y^Xv^T$ for v = w, and considering that it is an optimization problem, so all terms independent of v are irrelevant for the solution.

(b) Code given :

def Regression (X, y, C, D) :

```
return QP(Q, l, A, b)
```

Exercise 2 :

Parameter Estimation

The average time to get a letter at the post office follows the following distribution : $p(x|\theta) = \theta(1-\theta)^{(x-1)}$. The variable X is a positive integer (Z^+, and θ is a real number.

(a) Define the likelihood function $p(D|\theta)$

(b) Calculate the likelihood of $D = \{1, 1, 2, 1\}$

a) $P(D|\theta) = \prod_i P(x_i|\theta) = \prod_i \theta(1-\theta)^{x_i - 1}$

b) $P(D|\theta) = \theta^4(1-\theta)$

$\log P(D|\theta) = 4\log\theta + \log(1-\theta)$

$\frac{\partial}{\partial\theta}\log P(D|\theta) = \frac{4}{\theta} - \frac{1}{1-\theta} = \frac{4-4\theta-\theta}{\theta(1-\theta)} = 0$

$\hat{\theta} = \frac{4}{5}$

(c) Now consider a Bayesian approach, with the following probability distribution :

$p(\theta) = 1$, for $\theta \in [0, 1]$

$p(\theta) = 0$, elsewhere.

Prove that the posterior can be defined as $30 * \theta^4(1-\theta)$

(d) Evaluate the probability of $P(x > 1)$ with $\int p(x|\theta)p(\theta|D)$

(c) $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta}$

$= \frac{\theta^4(1-\theta)}{\int_0^1 \theta^4(1-\theta)\,d\theta}$

$= \frac{\theta^4(1-\theta)}{\frac{1}{5}\theta^5 - \frac{1}{6}\theta^6 \Big|_0^1} = 30\theta^4(1-\theta)$

(d) $P(x>1|D) = 1 - P(x=1|D) = 1 - \int P(x=1|\theta)P(\theta|D)$

$= 1 - \int \theta(1-\theta)^0 \cdot 30\theta^4(1-\theta)\,d\theta$

$= 1 - \int 30\theta^5(1-\theta)\,d\theta$

$= 1 - 30\left(\frac{1}{6}\theta^6 - \frac{1}{7}\theta^7\right)\Big|_0^1$

$= 1 - \cancel{30}^{5} \cdot \frac{1}{\cancel{6}\cdot 7} = \frac{2}{7}$

Exercise 3 :

Kernel

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined on a set $\mathcal{X}$ is called a *Positive Semi-Definite (PSD) Kernel* if, for any finite set of points $\{x_1, x_2, \ldots, x_n\} \subseteq \mathcal{X}$ and any corresponding set of coefficients $\{c_1, c_2, \ldots, c_n\} \subseteq \mathbb{R}$, the following condition holds :

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$$

for all $n \in \mathbb{N}$ and for all choices of $\{x_1, x_2, \ldots, x_n\}$ and $\{c_1, c_2, \ldots, c_n\}$.

(a) Given the following kernel : $k_f = f(x)k(x, x')f(x')$. Prove it is a psd kernel.

(b) Show that the Gaussian kernel is also a psd kernel, with $k_f = exp(\gamma \cdot ||x - x'||^2)$. Also define function f(x) for this case. *Hint : you can use the following kernel definition :* $k(x, x') = exp(\gamma \cdot x \cdot x')$*, and use your answers from a).*

(a)

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j f(x_i) k(x_i, x_j) f(x_j) = \sum_i \sum_j c_i c_j f(x_i) \sum_k^d \Phi_k(x_i) \Phi_k(x_j) f(x_j)$$

$$= \sum_k^d \sum_i c_i f(x_i) \Phi_k(x_i) \cdot \sum_j c_j f(x_j) \Phi_k(x_j)$$

$$= \sum_k^d \left(\sum_i c_i f(x_i) \Phi_k(x_i) \right)^2 \geq 0$$

(b)

$$k_f = exp(\gamma \tfrac{1}{2} (x-x')^T (x-x')) = exp(\gamma \; [x^T x - 2x^T x' + x'^T x'])$$

$$= \underbrace{exp(\gamma ||x||^2)}_{f(x)} \cdot \underbrace{exp(\gamma x^T x')}_{k(x,x')} \underbrace{exp(\gamma ||x'||^2)}_{f'(x)}$$

$$\sum_i \sum_j c_i c_j \, exp(\gamma x_i^T x_j) = \sum_i \sum_j c_i c_j \, exp(\gamma \sum_k^d x_{ik} \cdot x_{jk})$$

$$= \sum_i \sum_j c_i c_j \prod_k^d exp(\gamma x_{ik} x_{jk})$$

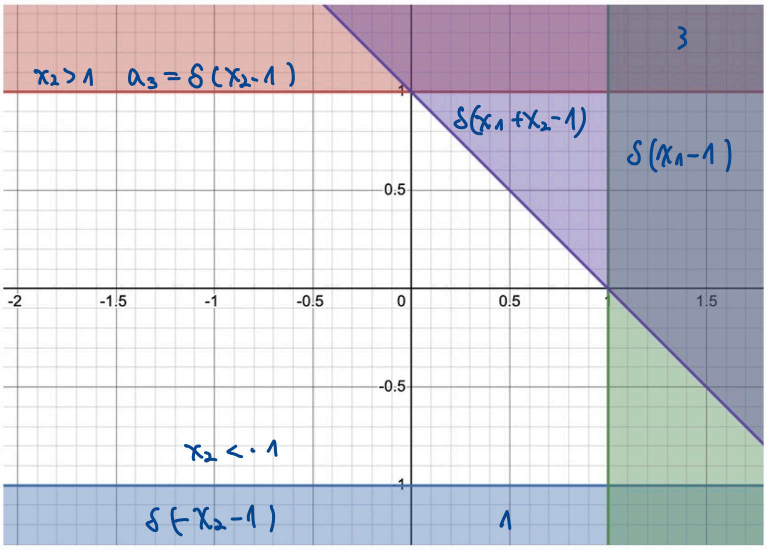
$$= \prod_k^d \sum_i \sum_j exp(\log c_i + \log c_j + \gamma x_{ik} x_{jk}) \geq 0$$

Exercise 4 :

Neural Networks

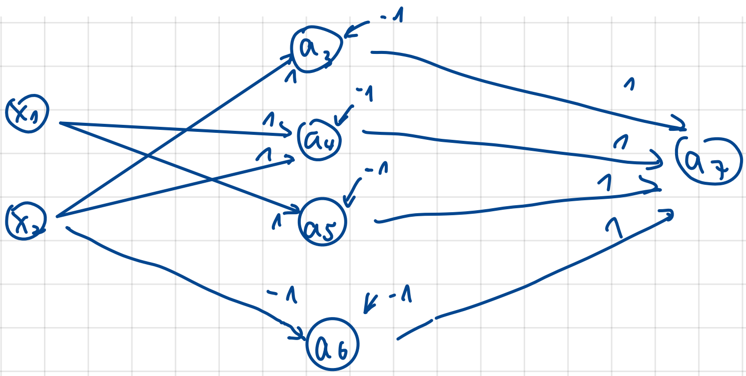
Consider the following Neural Network with activation function :

$$\text{step}(x) = \begin{cases} 1 & \text{if } a_i > 0 \\ 0 & \text{if } a_i \leq 0 \end{cases}$$



(a) Give all weights and biases.

(b) Describe values for all activated neurons for $x = (1, 1)$.



$a_3 + a_4 + a_5 + a_6 = 3$

1

Exercise 5 :

Ridge Regression. Using a Quadratic solver

Given a labeled dataset $((x_1, y_1), ..., (x_N, y_N))$ we consider the regularized regression problem :

$$\text{max} \cancel{\min}_{\mathbf{w}} ||\mathbf{y} - \mathbf{w}^T\mathbf{X}||^2$$

subject to $0 \leq w_i \leq C$ and $\forall i :$ $\sum_i w_i \leq D,$

with $C, D \in \Re, w \in \Re^d$ and $X \in \Re^{Nxd}$.

(a) Show that this problem is equivalent to a problem of this type : $\max_v \mathbf{v}^T(\mathbf{X}^T\mathbf{X})\mathbf{v} - 2\mathbf{y}^T\mathbf{X}\mathbf{v}$, subject to the same constraints.

(b) Implement a code in Python to calculate w. You can use the cvxopt.qp solver, that already implements the optimization problem in the following format :

$\max_v \mathbf{v}^T\mathbf{Q}\mathbf{v} - \mathbf{1}^T\mathbf{v}$ s.t. $\mathbf{A}\mathbf{v} \leq \mathbf{b}$

(a) $$\min_w ||y - w^T X||^2 = \min_w (y - Xw)^T (y - Xw)$$

$$= \min_w y^T y - 2y^T X w + w^T X^T X w$$

$$= \min_w w^T \underbrace{X^T X}_{dxd} w - 2y^T X w$$