ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Introduction to Machine Learning
Autumn Semester 2014

Prof. Joachim M. Buhmann

Final Exam

January 20th, 2015

First and Last name: _____

ETH number: _____

Signature: _____

General Remarks

- You have 2 hours for the exam. There are five sections.
- Write your answers directly on the exam sheets. At the end of the exam you will find supplementary sheets, feel free to seperate them from the exam. If you submit the supplementary sheets, put your name and ETH number on top of each.
- Answer the questions in English. Do not use a pencil or red color pen.
- You may provide at most one valid answer per question. Invalid solutions must be canceled out clearly.

	Topic	Max. Points	Points	Signature
1	Kernels & SVMs	24		
2	Unsupervised learning	24		
3	Regression	24		
4	Bayesian inference	24		
5	Ensemble methods	24		
Total		120		

Grade: .....

This page has been intentionally left blank.

Question 1: **Kernels and SVMs (24 pts.)**

Let us consider a dataset with the following positively labeled points: $\{(2,2);(2,-2);(-2,2);(-2,-2)\}$; and negatively labeled points $\{(1,1);(1,-1);(-1,1);(-1,-1)\}$. We shall define a non-linear mapping from the input space to a new feature space according to the following transformation:

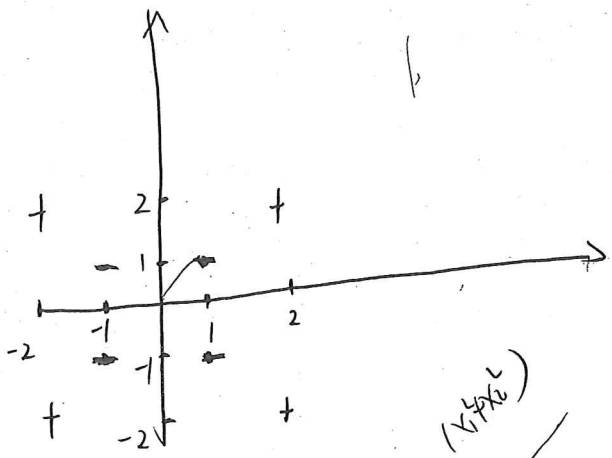
$$\Phi(x_1,x_2)=\begin{cases}(4-x_2+|x_1-x_2|,4-x_1+|x_1-x_2|) & \text{if } \sqrt{x_1^2+x_2^2}>2\\ (x_1,x_2) & \text{otherwise}\end{cases}$$

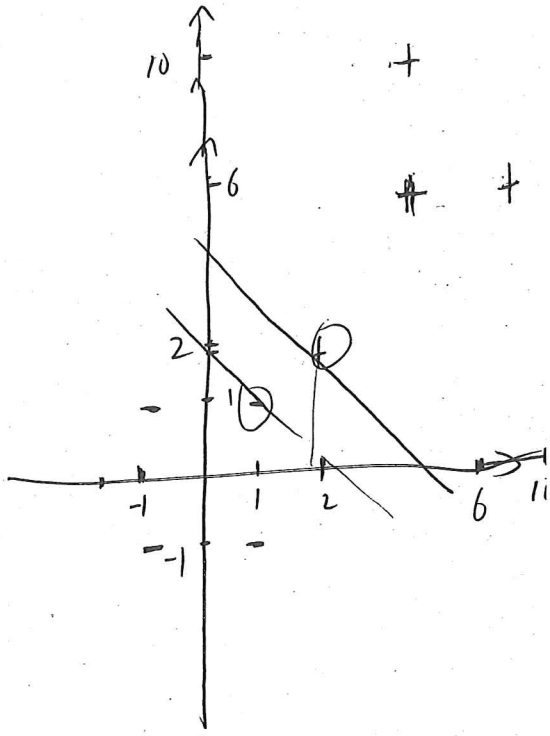
x_1, x_2 refers to the two feature dimensions.

1. Calculate the transformed sample points for each class. In two separate diagrams plot the original data points and the transformed data points. Use separate symbols for different classes (e.g., $+,-$).

+ { (2,2); (10, 6); (6,10) (6,6) }

- { (1,1); (1,-1); (-1,1); (-1,-1) }





$\phi(x_1,x_2) \rightarrow (x_1+x_2^2)$

ϕ^{-1}

3 pts.

$s_2 = (2,2)$ $s_1 = (1,1)$

2. In the above diagram identify the support vectors and denote them as $\mathbf{s}_i$. Define an augmented feature space as $\widetilde{s} = \{\mathbf{s}, 1\}$. Our objective is to find the optimal Lagrangian dual variables α_i for which the following constraints are defined

$\alpha_1 \widetilde{s_1} \cdot \widetilde{s_1} + \alpha_2 \widetilde{s_1} \cdot \widetilde{s_2} = -1$

$\alpha_1 \widetilde{s_1} \cdot \widetilde{s_2} + \alpha_2 \widetilde{s_2} \cdot \widetilde{s_2} = +1$.

Explain how these constraints are derived. From the above constraint equations calculate the values of α_1 and α_2

since for the dual function

we have $w^* = \sum_i \alpha_i z_i \widetilde{s_i}$

for sv we have $z_i w^* s_i = 1$

thus

$z_i \sum \alpha_j z_j \widetilde{s_j} \widetilde{s_i} = 1$

for s_1 $-1 \times (-\alpha_1 \widetilde{s_1} \cdot \widetilde{s_1} + \alpha_2 \widetilde{s_2} \cdot \widetilde{s_1}) = 1$

for s_2 $(-\alpha_1 \widetilde{s_1} \cdot \widetilde{s_2} + \alpha_2 \widetilde{s_2} \widetilde{s_2}) = 1$

$-3\alpha_1 + 5\alpha_2 = -1$

$-5\alpha_1 + 9\alpha_2 = 1$

$\alpha_1 = 7$

$\alpha_2 = 4$

4 pts.

3. Calculate the separating hyperplane $\mathbf{w}$ and plot the decision surface in your diagram from Question 1.1 showing the transformed sample points.

$w = \sum \alpha_i (1,1,1) + \alpha_2 (2,2,1)$

$= -(7,7,7) + (8,8,4) = (1,1,-3)$ **5 pts.**

4. Given a test point $\mathbf{x}$ in the original feature space, write down the classification function after mapping it to the transformed feature space.

$$\begin{cases} \text{sign}[(1,1,3)(x_1, x_2, 1)^T] & \sqrt{x_1^2 + x_2^2} \le 2 \\ \text{sign}[(1,1,3)(4 - x_2 - |x_1 - x_2|, 4 - x_1 + |x_1 - x_2|, 1)^T] & \sqrt{x_1^2 + x_2^2} > 2 \end{cases}$$

3 pts.

(0.92, 1)

5. To which class is $\mathbf{x} = (4, 5)$ assigned? Discuss whether the classification seems intuitively correct. If not, what are the possible reasons for this discrepancy and what changes can improve the generalization of the classifier?

Could not decide

model is not suitable, the

change to use the gaussian kernel

add more data

increase the feature space

kernel function appropriate.

5 pts.

6. Identify, with explanation, whether the following functions are kernels or not

(a) $k(x, y) = \begin{cases} 10 & \text{if } x \neq y \\ 1 & \text{if } x = y \end{cases}$

not considering (1,1) (2,2)

$K = \begin{bmatrix} 1 & 10 \\ 10 & 1 \end{bmatrix}$ delta(K) ≥ 0

not semi-positive

2 pts.

(b) $k(x, y) = \|x - y\|^2 + \|x + y\|^2$

4 1 -1

not considering to sample (1), (2)

$K = \begin{bmatrix} 2 & 5 \\ 5 & 8 \end{bmatrix}$

2 pts.

delta | K| < 0

Question 2: Unsupervised learning (24 pts.)

1. *Histograms and Parzen estimators.*

   (a) Assume you have N data points $x_i \in [0,1], i \in 1,\ldots,N$ (all the points are in the interval from 0 to 1). Give the definition of the histogram with K bins (use equidistant binning strategy). Explain how the resulting histogram can be used for density estimation. Will the "curse of dimensionalty" be a problem in this case and why?

to dive the space [0,1] to k intervels equally $\frac{1}{K}$

which the corresponding histogram of the sample is

$H_j := \#\{x \in s | x \in I_j\}$

after normalization, we could get $A := \frac{1}{N}(h_1, \ldots h_k)$

no, since it is one dimension.

6 pts.

(b) Let k be the unit cube function:

$$k(\mathbf{u}) = \begin{cases} 1 & \textbf{if } |u_i| \leq 1/2 \textbf{ for all } i \in 1, ..., d \\ 0 & \textbf{otherwise} \end{cases}$$

A unit cube function (scaled to size h) around a training data sample $\mathbf{x}_j$ is obtained by $k(\frac{\mathbf{x}-\mathbf{x}_j}{h})$. Using such unit cubes, we can estimate the probability density of a distribution that generated training data $\{\mathbf{x}_j\}_{j=1}^n$ using the formula:

$$p(\mathbf{x}) = \frac{K(\mathbf{x})}{nV}$$

where $K(\mathbf{x})$ is the number of samples falling within a window around $\mathbf{x}$, n is the total number of training samples and V is the area of the window around $\mathbf{x}$. What are the values of $K(\mathbf{x})$ and V for a Parzen window estimate in this setting?

$p(x) = \frac{1}{n} \frac{1}{V_n} \sum_{i=1}^{n} k(\frac{x-x_i}{h})$

$V_n = (h)^d = h$

$K(x) = \sum_{i=1}^{n} k(\frac{x-x_i}{h})$

5 pts.

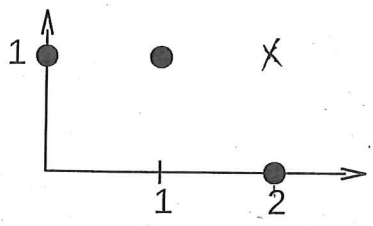


Figure 1: A figure for question (c).

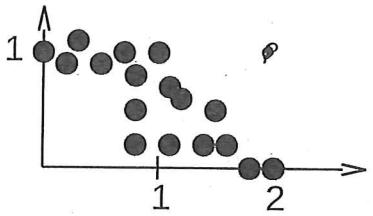


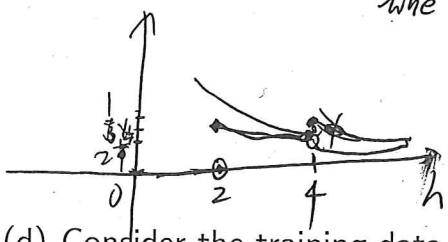
Figure 2: A figure for question (d).

(c) Consider training data with $d = 2$ dimensions and three samples $(0, 1)$, $(1, 1)$ and $(2, 0)$ (see Figure 1):
Using the Parzen window estimate (from the previous point), plot the value of $p(\mathbf{w})$ for $\mathbf{w} = (2, 1)$ with respect to the parameter h.

$p(w) = \frac{1}{3}\frac{1}{h^2}\sum_{i=1}^{3} k\left(\frac{w - x_i}{h}\right)$ when $0 < h \leq 2$, $p(w) = 0$

when $2 \leq h < 4$ $p(w) = \frac{2}{3h^2}$

whe $h \geq 4$ $p(w) = \frac{1}{h^2}$



5 pts.

(d) Consider the training data as shown in Figure 2:
Would you use the same Parzen window estimate for this training data and for the training data from the previous point? Motivate your decision.

no, since the train data became intensive

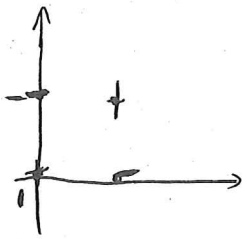
thus the h value could be smaller,

~~to increase~~ thus not to lose detail.

4 pts.

2. *K-Nearest Neighbours.*

Let the dataset for binary classification consist of only four points in $\mathbb{R}^2$: $\mathbf{x}_1 = (0,0), \mathbf{x}_2 = (0,1), \mathbf{x}_3 = (1,1), \mathbf{x}_4 = (1,0)$. The corresponding labels are the following: $y_1 = 1, y_2 = 0, y_3 = 1, y_4 = 0$. Compute leave-one-out cross-validation error of the 3-NN classifier.



$\frac{1}{4} \sum_{i=1}^{4} \mathbb{I}[y_i \neq a_i(x_i)]$

$= \frac{1}{4}(1+1+1+1)$

$= 1$

4 pts.

Question 3: **Regression (24 pts.)**

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n \times p}, \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\epsilon} \in \mathbb{R}^n$.

1. (a) Write down the least squares estimator $\hat{\beta}$ as a function of $\mathbf{y}$ and $\mathbf{X}$.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

1 pts.

(b) Under what statistical assumption is $\hat{\beta}$ unbiased?

$$E(\hat{\beta}) = E((X^T X)^{-1} X^T X\beta + \epsilon)$$

$$= I \cdot \beta + E(\epsilon)$$

when $E(\epsilon) = 0$

2 pts.

(c) Formally show that $\hat{\beta}$ is unbiased given the above assumption.

9 pts.

2. Name at least one reason why a gradient descent method would be preferred over the closed form solution for obtaining $\hat{\beta}$.

it could be time consuming to compute the inverse of (X^TX) matrix

furthermore, when the singular matrix of the X function possem

2 pts.

3. Explain qualitatively how the bias and variance of the estimator are related with under- and over-fitting.

when over-fitting the bias is small while the variance become big

underfitting bias is big while variance small

3 pts.

4. Consider using cross validation for estimating the generalization error.

   (a) Write down the value of the loss function for a new point $\mathbf{x}$ in the test set.

Loss = $(X(X^TX)^{-1}X^TY' - Y)^2$

$(X(X'^TX')^{-1}X'^TY' - Y)^2$

2 pts.

   (b) Formally define the generalization error and suggest a way how to estimate it.

for the test set with n sample

generalization error = $\frac{1}{n}\sum_{i=1}^{n}(x_i^T\hat{\beta} - y_i)^2$

3 pts.

5. What are the advantages of the LASSO over ridge regression?

it LASSO estimates are known to be sparse with few coefficients non-vanishing, thus make it more efficient and be better at interpreting

2 pts.

feature selection

Question 4: **Bayesian inference and maximum likelihood (24 pts.)**

1. Let $X_1, \ldots, X_n$ be n mutually independent and normally distributed random variables, $X_i \sim \mathcal{N}(\mu, \sigma^2)$, with a common unknown mean and common known variance. Let Y be a sum of the random variables $Y = \sum_{i=1}^{n} X_i$.

   (a) How is the likelihood $p(y|\mu, \sigma^2)$ distributed? Derive the density function.

   Hint: The moment generating function of a univariate normal distribution is $M_X(t) = \exp\left(t\mu + \frac{1}{2}\sigma^2 t^2\right)$.

$p(y|\mu,\sigma^2) = \exp(\ldots)$

$= \exp(\ldots)$

Since X_i are iid

Y_i is also a d normally distributed variables

$$E(Y) = \sum_{i=1}^{n} E(X_i) = n\mu$$

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}\, X_i = n\sigma^2$$

$$p(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi n}\,\sigma} e^{-\frac{1}{2n\sigma^2}(y-n\mu)^2}$$

4 pts.

(b) Determine the maximum a posteriori estimate of μ given y. Assume that the prior distribution is a normal distribution $\mathcal{N}(\mu|\mu_0, \sigma_0^2)$.

$$\max p(\mu|y) \propto p(y|\mu, \mu_0, \sigma_0^2, \sigma^2) \cdot p(\mu|\mu_0, \sigma_0^2)$$

$$e^{-\frac{(y-n\mu)^2}{2n\sigma^2}} \cdot e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}$$

$$\max \mathcal{N}(\mu|\mu_n, \sigma_n^2)$$

$$= \mu_n$$

$$\mu_n = \frac{\frac{y}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{y\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}$$

5 pts.

(c) Determine the probability density function of the posterior $p(\mu|y, \mu_0, \sigma_0, \sigma)$.

$$p(\mu|\chi, \mu_0, \sigma_0, \sigma)$$

$$\sim N(\mu|\mu_n, \sigma_n^2)$$

$$\propto e^{-\left(\frac{n}{2\sigma^2}+\frac{1}{2\sigma_0^2}\right)(\mu-\mu_n)^2}$$

$$\frac{n}{2\sigma^2}+\frac{1}{2\sigma_0^2}=\frac{1}{2\sigma_n^2}$$

$$\Rightarrow \sigma_n^2=\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}$$

$$=\frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2}$$

$$\Rightarrow p=\frac{\sqrt{n\sigma_0^2+\sigma^2}}{\sqrt{2\pi}\,\sigma\sigma_0}\cdot e^{-\frac{(n\sigma_0^2+\sigma^2)\left(\mu-\frac{n\sigma_0^2+\mu_0\sigma^2}{n\sigma_0^2+\sigma^2}\right)^2}{2\sigma^2\sigma_0^2}}$$

5 pts.

2. (a) Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a dataset consisting of n samples which are drawn i.i.d. from a normal distribution $\mathcal{N}(x|\mu, \sigma^2)$ with unknown mean and unknown variance. Recall that the maximum likelihood estimate of the mean is given by:

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Determine the maximum likelihood estimate of the variance $\hat{\sigma}^2_{\mathrm{ML}}$. Which estimator has a larger bias, the mean or the variance? Why?

$\log L(X) = \log(\cdots)^n - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum \frac{(x_i - \mu_{ML})^2}{\sigma^2} - \frac{n}{2}\log 2\pi$

$\frac{\partial \log L(X)}{\partial \sigma^2} = -\frac{n}{2}\cdot\frac{2\pi}{2\pi\sigma^2} + \frac{\sum (x_i - \mu_{ML})^2}{2\sigma^4} = 0$

$\Rightarrow \quad \hat{\sigma}^2_{ML} = \frac{\sum (x_i - \mu_{ML})^2}{n}$

$= \frac{1}{n}\sum (x_i - \frac{1}{n}\sum_{j=1}^{n} x_j)^2 = \frac{1}{n}(n-1)\sigma^2$

$E(\hat{\mu}_{ML}) = \frac{1}{n}\cdot n\mu = \mu$

$\mathrm{bias}(\hat{\mu}_{ML}) = 0$

while $\hat{\sigma}^2_{ML}$ is biased, since, μ_{ML} is unknown

3 pts.

(b) Let $\mathcal{X} = \{x_1, \ldots, x_n\}, x_{i:n} \geq 0$ be a dataset consisting of n samples which are drawn i.i.d. from a Weibull distribution:

$$\text{Weibull}(x_i|c,\lambda) = \left(\frac{c}{\lambda}\right)\left(\frac{x_i}{\lambda}\right)^{c-1} \exp\left(-\left(\frac{x_i}{\lambda}\right)^{c}\right).$$

Determine the derivative of the log-likelihood $\frac{\partial}{\partial c}\left(\ln p(\mathcal{X}|c,\lambda)\right)$. Can you find an analytic expression for the maximum likelihood estimate $\hat{c}_{\text{ML}}$? If not, suggest a method to compute $\hat{c}_{\text{ML}}$.

$$\ln p(X|c,\lambda) = \sum_{i=1}^{n} \ln\left(p(x_i|c,\lambda)\right)$$

$$= \sum_{i=1}^{n} \left[\ln c - \ln\lambda + (c-1)\ln x_i - (c-1)\ln\lambda - \left(\frac{x_i}{\lambda}\right)^c\right]$$

$$= n\ln c - nc\ln\lambda + n(c-1)\ln x_i - n\left(\frac{x_i}{\lambda}\right)^c$$

$$f(c) = \frac{\partial \ln p(X|c,\lambda)}{\partial c} = \frac{n}{c} - n\ln\lambda + n\ln x_i - n\left(\frac{x_i}{\lambda}\right)^c \ln\frac{x_i}{\lambda} = 0$$

no analytic expression for $\hat{c}_{ML}$

can use numerical method to approximate it. gradient descent

$f(c_1) > 0 \quad f(c_2) < 0$

the find $f\left(\frac{c_1+c_2}{2}\right) > 0$ or < 0 **4 pts.**

if > 0

let $c_1 \leftarrow \frac{c_1+c_2}{2}$

again $c_{between} = \frac{c_1+c_2}{2}$

see the $f(between)$ value of

if it > 0 ... until $|f(\hat{c}) - 0|$

3. Consider a binomial distribution as likelihood function of x:

$$p(x|\theta, n) = \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x},$$

Determine the posterior distribution $p(\theta|x, n)$ (written in direct closed form) using the following beta distribution as a prior:

$$p(\theta|\alpha, \beta) = B(\alpha, \beta)^{-1}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

The beta function $B(\alpha, \beta)$ is a normalization constant.

$p(\theta|x,n) \propto p(x|\theta,n) \cdot p(\theta|\alpha,\beta)$

$\propto \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

$\propto \theta^{(\alpha+x-1)}(1-\theta)^{(n-x+\beta-1)}$

$p(\theta|x,n) = B(\alpha',\beta')^{-1}\theta^{\alpha'-1}(1-\theta)^{\beta'-1}$

$\alpha' = \alpha + x \quad \beta' = n - x + \beta$

3 pts.

Question 5: Ensemble methods (24 pts.)

bias to large

1. Bagging
   One of the possible theoretical arguments for bagging (bootstrap aggregation) is the fact that the expected error of an averaged model (i.e. committee) is one order less than an average expected error of individual models under certain assumptions.
   However, this error reduction fails often in reality. Please briefly explain why.

   ~~bias~~ $= \left(\frac{1}{B}\sum_{i=1}^{B} E[\hat{f}_i(x)] - E[\hat{f}(x)]\right)^2$ it is assume that each

   $= \text{Var} = \left(\frac{1}{B}\sum_{i=1}^{B}\hat{f}_i(x) - \frac{1}{B}E\sum_{i=1}^{B}\hat{f}_i(x)\right)^2$

   $= \left(\frac{1}{B}\sum_{i=1}^{B}(\hat{f}_i(x) - E\hat{f}_i(x))\right)^2$

   $= \frac{1}{B^2}\sum_{i=1}^{B}\text{Var}(\hat{f}_i(x)) + \frac{1}{B^2}\sum\sum_{i\neq j}\text{cov}(\hat{f}_i(x), \hat{f}_j(x))$

   $\approx \frac{\sigma^2}{B}$ only under the assumption that covariances are small

   in fact, since ~~each~~ bagging base classifiers are trained from the subset of the same ~~big~~ dataset, hence the covariance may not be ideal.

   2 pts.

2. AdaBoost
   (a) Denote the n given training samples as $(\mathbf{x}^{(i)}, y^{(i)})$, $i = 1, \ldots, n$, where $\mathbf{x}^{(i)} \in \mathbb{R}^d$ are points and $y^{(i)} \in \{-1, +1\}$ are labels. Denote the weak learners as $c_b(\mathbf{x}) \in \{-1, +1\}$, $b = 1, \ldots, M$. Sketch the classical AdaBoost algorithm in pseudocode and highlight the training error optimized at each step:

   and the base ~~clas~~ classifier is not ~~clas~~

   initialize $w_i^{(1)} = \frac{1}{n}$

   for $b = 1:B$ do

   train $C_b(x)$ based on training samples with weight $w_i^{(b)}$

   compute the training error $\text{err}_b = \frac{\sum w_i^b \, \mathbb{I}(C_b(x_i) \neq y_i)}{\sum w_i^b}$

   compute $\alpha_b = \log\frac{1-\text{err}_b}{\text{err}_b}$

   update w_i $w_i^{b+1} = \frac{w_i^b \, e^{\alpha_b \mathbb{I}(C_b(x_i) \neq y_i)}}{\sum}$

   end for

   return the ensembled classifier is $H(x) = \text{sign}\left(\sum_{i=1}^{B}\alpha_b C_b(x)\right)$ **4 pts.**

(b) In some formulations of AdaBoost it is required that the training error ϵ_b of each weak binary classifier is < 0.5. How could one guarantee fulfilling such a requirement (i.e. what should be done if $\epsilon_b > 0.5$)?

if $\epsilon_b > 0.5$, in this case $\alpha_b = \log \frac{1-\epsilon_b}{\epsilon_b} < 0$
thus this base classifier will be assigned
a negative weight in the ambined classifier,
the reverse one would have a error rate less than 0.5
base classfer

2 pts.

(c) Explain how AdaBoost could be used for feature selection. Which weak learners would you propose for that?

use decision stumps to train the model

the Vanible importance is measured by
~~how much error it~~ the α_b for each vanible
used to conduct split/branch. we can choose the
feature that possess large α for training

2 pts.

3. Decision Boundaries of AdaBoost

Assume now that the data are 2-dimensional and we use decision stumps as weak learners:

$$c_b(\mathbf{x}) = c_b(\mathbf{x}|k,\theta,\gamma) := \begin{cases} +\gamma & x_k > \theta, \\ -\gamma & \text{otherwise} \end{cases}$$

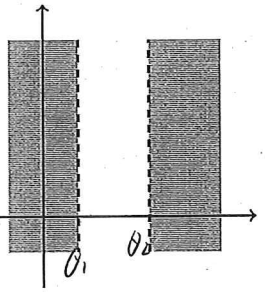
Here, $k = 1, 2$ is the dimension, along which the decision is made, and $\gamma \in \{-1, +1\}$ is the parameter, which provides labelling "direction". A decision stump is a simple binary classifier which slices classes according to a decision line orthogonal to the i-th axis, and orientation of classification is given by γ. The k, θ and γ are *individual* for each step b (i.e. may vary from step to step).

The final classifier is a linear combination of the trained weak classifiers:

$$c_{\text{AdaBoost}}(\mathbf{x}) = \text{sgn}\Big(\sum_{b=1}^{M} \alpha_b c_b(\mathbf{x})\Big)$$

Write down the combination of decision stumps for the following decision boundaries, or explain why it's impossible (Figure convention: decision boundary is a dashed line, and positive class is a shaded region):

(a)



not possible

let's assume

the first θ_1, second axis θ_2

for $C_1(x)$ $k=1$ $\gamma=-1$

$\theta=\theta_1$

$C_2(x)$ $k=1$ $\gamma=1$

$\theta=\theta_2$

in this case

the part that $x>\theta_2$

we have $\sum_{b=1}^{2} \alpha_b C_b(x)$

$\text{sgn}(\alpha_2-\alpha_1)>0$

$\alpha_2>\alpha_1$

for $x<\theta_1$

we have

$\sum_{b=1}^{2} \alpha_b C_b(x)$

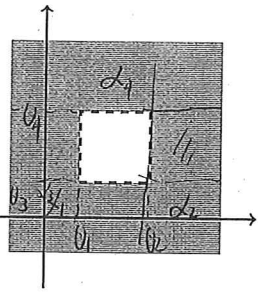
$= 1 \cdot \text{sgn}(\alpha_1-\alpha_2)>$

$\alpha_1>\alpha_2$

which is not possible.

3 pts.

(b)



$\alpha_2 = \alpha_1 \phi - \alpha_4 - \alpha_3 > 0$

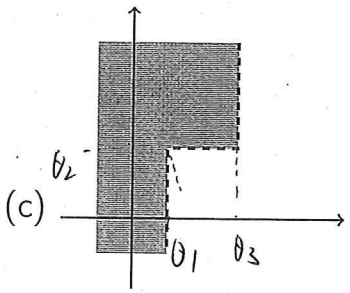
$\alpha_1 - \alpha_2 - \alpha_3 - \alpha_4 > 0$

not likely

not possible

3 pts.

(c)



$C_1(x)$ $k = 1$

$\theta = \theta_1$

$y = -1$

$C_2(x)$ $k = 2$

$\theta = \theta_2$

$y = 1$

$C_3(x)$ $k = 1$

$\theta = \theta_3$

$y = -1$

2 pts.

4. Decision Trees

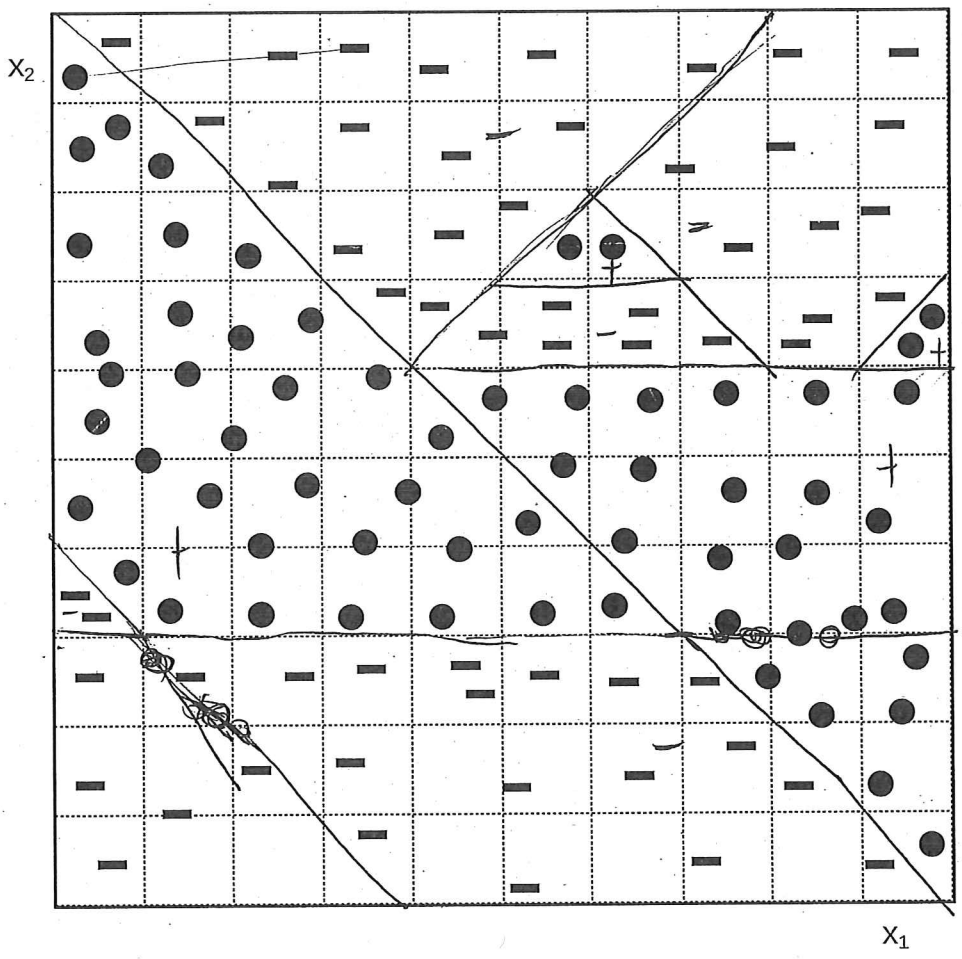
In this exercise, we again consider a binary classification task for a 2-dimensional dataspace. Below you can see a training set for this task.

Consider the set of weak classifiers having the following decision boundaries

$$x_2 = \beta x_1 + \gamma,$$

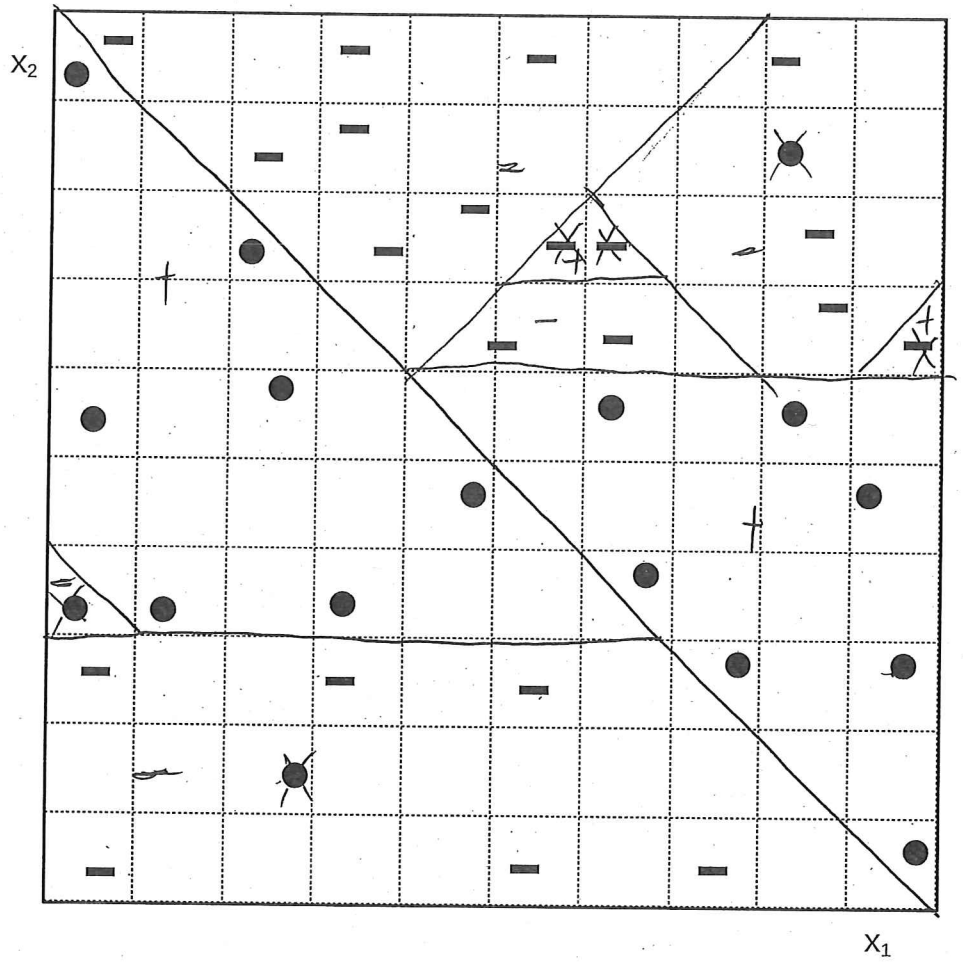
where each weak classifier has *its own* parameters $\beta \in \{-1, 0, 1\}$ and $\gamma \in \mathbb{R}$.

(a) Sketch graphically the flow of a decision tree algorithm with the above-mentioned weak classifiers (i.e. draw the respective decision bounds) on the given training set, until zero training error is achieved.



2 pts.

(b) Transfer the obtained boundary to the test set below and compute the test error.

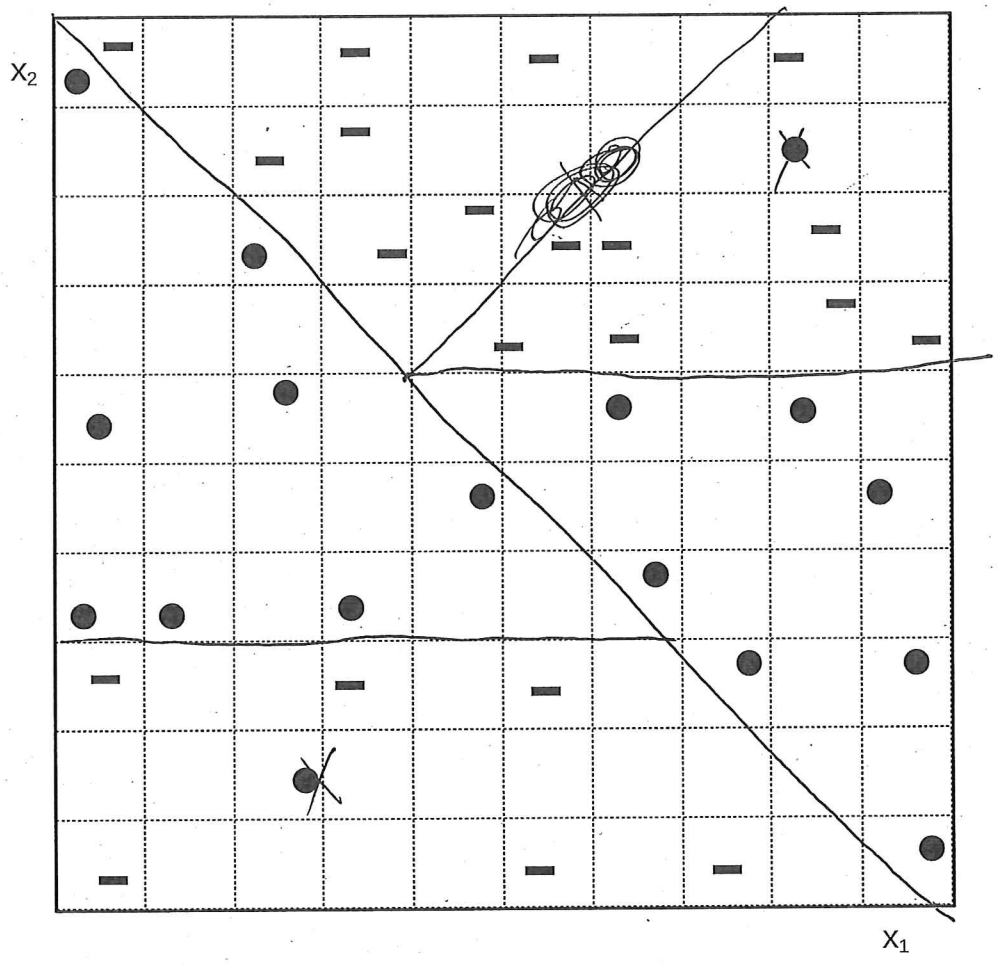


6 misclassification — 38

1 pts.

decrease the complexity

(c) Explain what problems you see with the obtained classifier. How could you modify it in order to make it generalize better? Draw the new one on the test set (it is the same as above) and compute the error.



in this case, only two misclassification

3 pts.

Supplementary Sheet