
*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 separate 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., $+$, $-$).

3 pts.

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

$$\alpha_1 \tilde{\mathbf{s}}_1 \cdot \tilde{\mathbf{s}}_1 + \alpha_2 \tilde{\mathbf{s}}_1 \cdot \tilde{\mathbf{s}}_2 = -1$$

$$\alpha_1 \tilde{\mathbf{s}}_1 \cdot \tilde{\mathbf{s}}_2 + \alpha_2 \tilde{\mathbf{s}}_2 \cdot \tilde{\mathbf{s}}_2 = +1 .$$

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

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.

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.

3 pts.

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?

5 pts.

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

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

2 pts.

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

2 pts.

Question 2: Unsupervised learning (24 pts.)

1. *Histograms and Parzen estimators.*

- (a) Assume you have N data points $x_i \in [0, 1], i \in 1, \dots, N$ (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 dimensionality" be a problem in this case and why?

6 pts.

(b) Let k be the unit cube function:

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

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?

5 pts.

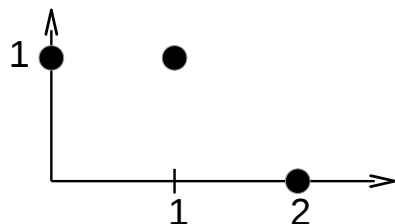


Figure 1: A figure for question (c).

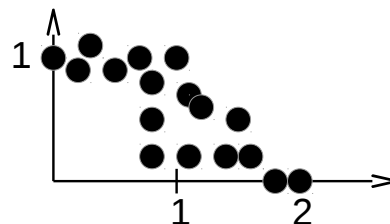


Figure 2: A figure for question (d).

- (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 .

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.

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.

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 $\hat{\boldsymbol{\beta}}$ as a function of \mathbf{y} and \mathbf{X} .

1 pts.

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

2 pts.

- (c) Formally show that $\hat{\boldsymbol{\beta}}$ 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 $\hat{\beta}$.

2 pts.

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

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 x in the test set.

2 pts.

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

3 pts.

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

2 pts.

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

1. Let X_1, \dots, 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 $M_X(t) = \exp\left(t\mu + \frac{1}{2}\sigma^2 t^2\right)$.

4 pts.

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

5 pts.

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

5 pts.

2. (a) Let $\mathcal{X} = \{x_1, \dots, x_n\}$ 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:

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

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

3 pts.

- (b) Let $\mathcal{X} = \{x_1, \dots, x_n\}, x_{i:n} \geq 0$ 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 $\frac{\partial}{\partial c} (\ln p(\mathcal{X}|c, \lambda))$. Can you find an analytic expression for the maximum likelihood estimate \hat{c}_{ML} ? If not, suggest a method to compute \hat{c}_{ML} .

4 pts.

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.

3 pts.

Question 5: Ensemble methods (24 pts.)

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.

2 pts.

2. AdaBoost

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

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 $\epsilon_b > 0.5$)?

2 pts.

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

2 pts.

3. Decision Boundaries of AdaBoost

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

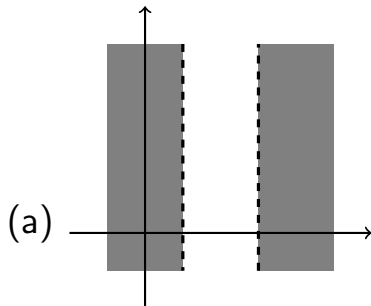
$$c_b(\mathbf{x}) = c_b(\mathbf{x}|k, \theta, \gamma) := \begin{cases} +\gamma & x_k > \theta, \\ -\gamma & \text{otherwise} \end{cases}$$

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).

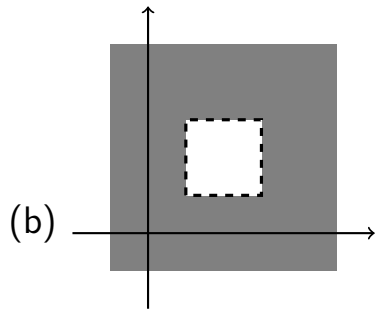
The final classifier is a linear combination of the trained weak classifiers:

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

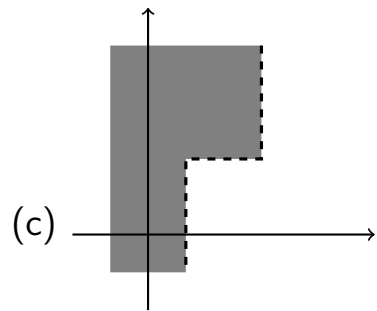
Write down the combination of decision stumps for the following decision boundaries, or explain why it's impossible (Figure convention: decision boundary is a dashed line, and positive class is a shaded region):



3 pts.



3 pts.



2 pts.

4. Decision Trees

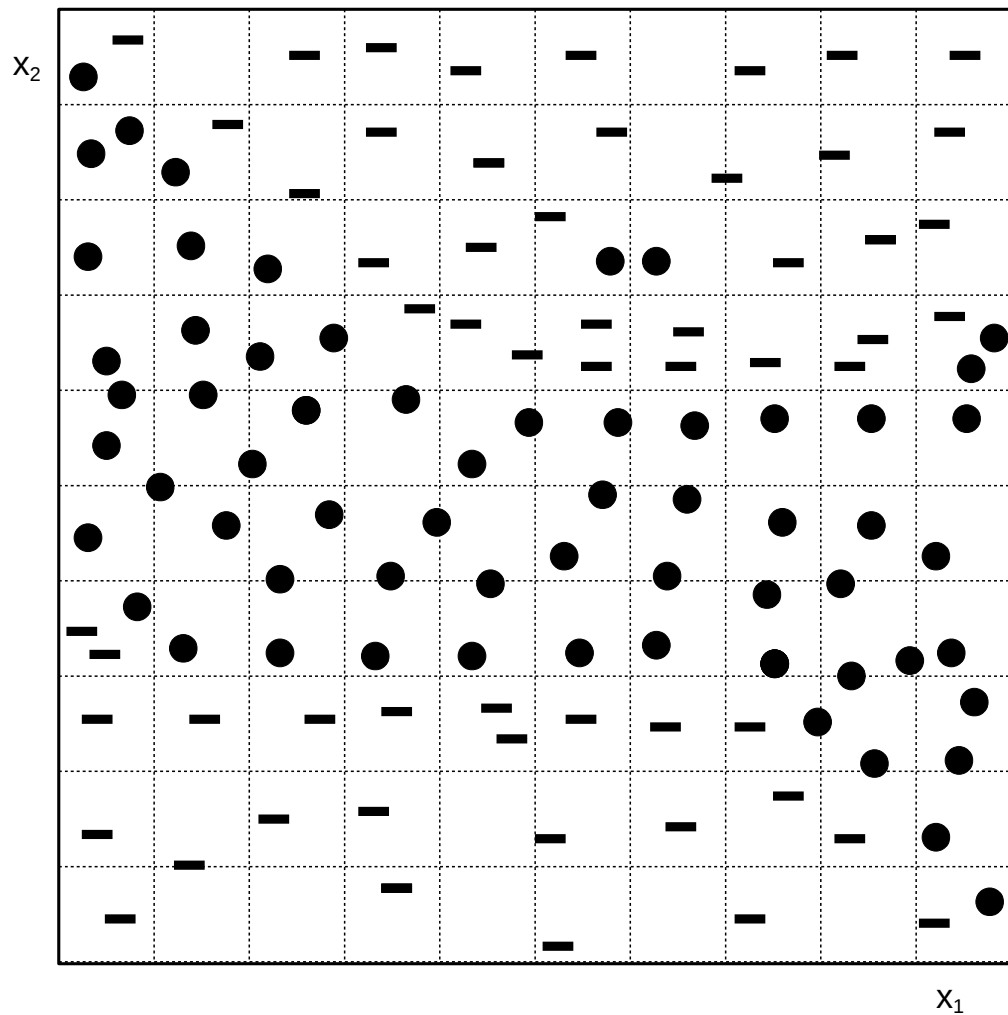
In this exercise, we again consider a binary classification task for a 2-dimensional dataspace. Below you can see a training set for this task.

Consider the set of weak classifiers having the following decision boundaries

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

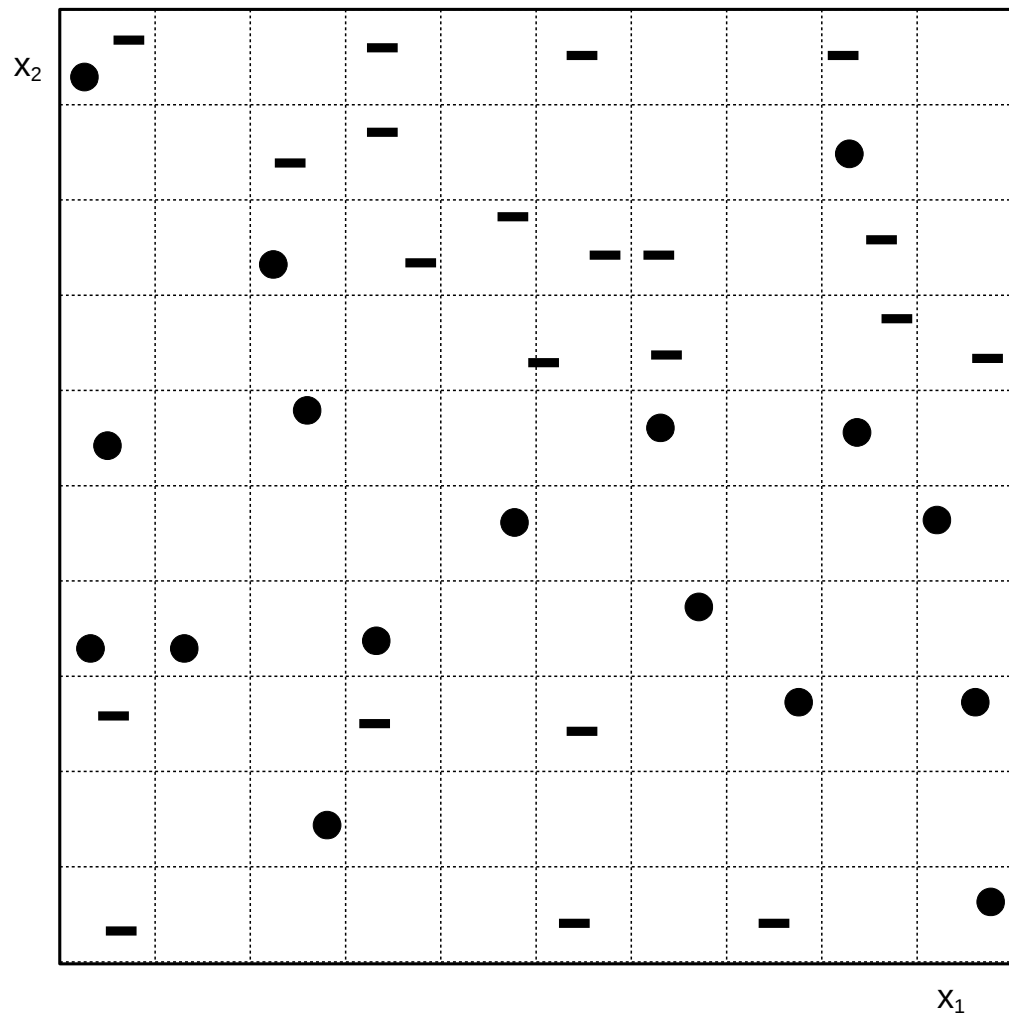
where each weak classifier has *its own* parameters $\beta \in \{-1, 0, 1\}$ and $\gamma \in \mathbb{R}$.

- (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.



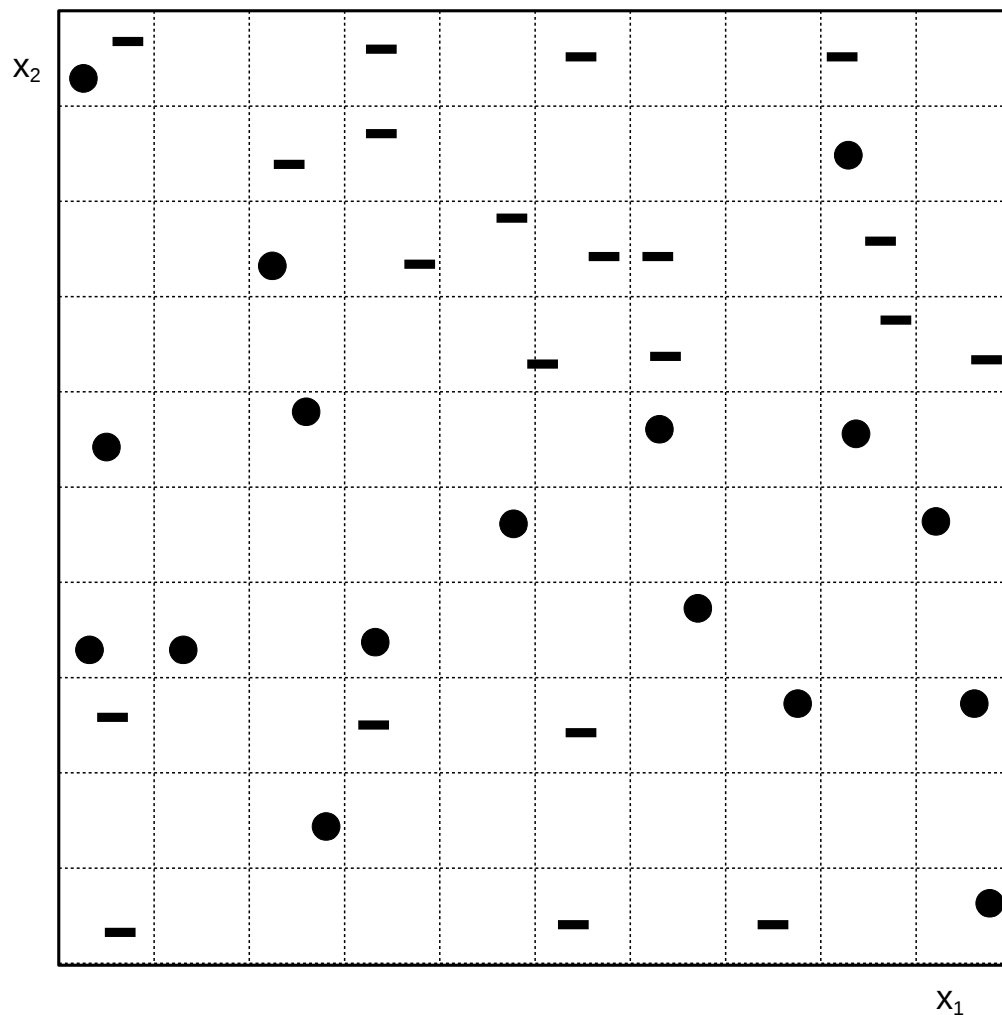
2 pts.

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



1 pts.

- (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.



3 pts.

Supplementary Sheet

Supplementary Sheet

Supplementary Sheet

Supplementary Sheet