UNIVERSITY OF AMSTERDAM

Faculty of Science

Exam

Machine Learning 1

Final Exam
Date: October 24, 2016
Time: 18:00-21:00

Number of pages: 11 (including front page)
Number of questions: 5
Maximum number of points to earn: 44
At each question is indicated how many points it is worth.

BEFORE YOU START

- Please **wait** until you are instructed to open the booklet.
- Check if your version of the exam is complete.
- Write down **your name, student ID number**, and if applicable the **version number** on **each sheet** that you hand in. Also **number the pages**.
- Your **mobile phone** has to be switched off and in the coat or bag. Your **coat and bag** must be under your table.
- **Tools allowed**: 1 handwritten double-sided A4-size cheat sheet, pen.

PRACTICAL MATTERS

- The first 30 minutes and the last 15 minutes you are not allowed to leave the room, not even to visit the toilet.
- You are obliged to identify yourself at the request of the examiner (or his representative) with a proof of your enrollment or a valid ID.
- During the examination it is not permitted to visit the toilet, unless the proctor gives permission to do so.
- 15 minutes before the end, you will be warned that the time to hand in is approaching.
- If applicable, please fill out the evaluation form at the end of the exam.

Good luck!

1 Multiple Choice Questions

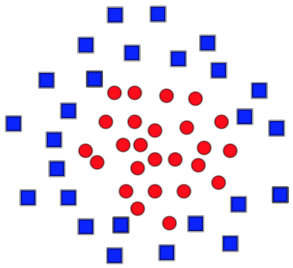
/20

For the evaluation of each question note: Several answers might be correct and at least one is correct. You are granted one point if every correct answer is 'marked' **and** every incorrect answer is 'not marked'. In all other cases zero points are granted. A box counts as 'marked' if a clearly visible symbol is written in there or if the box is blackened out. In the case you want to change an already marked box write 'not marked' next to the box.

1. Consider a neural network where we replace every activation function in every layer by a linear function. Which of the following classes of functions can we expect to fit well using such a network with one hidden layer? **/1**

   - ☐ Polynomials of degree 1.
   - ☐ Polynomials of degree 2.
   - ☐ A constant function.
   - ☐ Piecewise Constant functions.

2. Which of the following classifiers can learn the decision boundary shown in Figure 1?



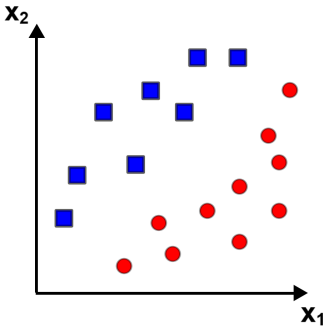
Figuur 1: Two-class data set with class denoted by color/shape.

/1

   - ☐ Neural networks.
   - ☐ Standard logistic regression (with linear features).
   - ☐ Naive Bayes.
   - ☐ Support Vector Machine (with linear kernel).

3. Which of the following classifiers can learn the decision boundary shown in Figure 2? **/1**

   - ☐ Neural networks.
   - ☐ Standard logistic regression (with linear features).
   - ☐ Naive Bayes.
   - ☐ Support Vector Machine (with linear kernel).



Figuur 2: Two-class data set with class denoted by color/shape.

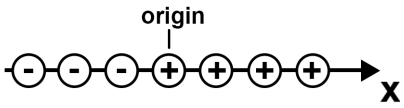
4. Which of the following properties of an activation function are desirable for a neural network to learn complex functions with gradient based methods? **/1**

   - ☐ Differentiability (almost everywhere).
   - ☐ Non-linearity.
   - ☐ Evaluation of the function as well as of its gradient is cheap.
   - ☐ Periodicity.

5. Consider the following two-class SVM optimization problem with data set $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$ with $x_i \in \mathbb{R}$ and $y_i \in \{-1, 1\}$:

$$\begin{aligned} \text{minimize}_{\{w,\xi_i,b\}} \quad & \frac{1}{2}w^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to} \quad & y_i(w \cdot x_i - b) - 1 + \xi_i \geq 0 \ \forall i \\ & \xi_i \geq 0 \ \forall i \end{aligned}$$

   The data consists of 4 positive data points $\{0, 1, 2, 3\}$ (i.e. with label $+1$) and 3 negative data points $\{-3, -2, -1\}$ (i.e. with label -1). The data set is depicted in Figure 3. If we



Figuur 3: Two-class data set in 1-d with class denoted by +/-.

   now choose $C = 0$, how many support vectors do we have? **/1**

   - ☐ 7
   - ☐ 4
   - ☐ 3
   - ☐ 2

6. Consider the exact same setup as in the previous question. If $C \to \infty$, how many support vectors do we have? **/1**

- [ ] 7
- [ ] 4
- [ ] 3
- [ ] 2

7. Which of the following statements is/are true? **/1**

- [ ] Logistic regression is a convex optimization problem.
- [ ] Classification using a deep neural network is a non-convex optimization problem.
- [ ] Regression using a deep neural network is a convex optimization problem.
- [ ] Finding the support vectors for SVM classification is a convex optimization problem.

8. Suppose we have training data that we want to classify using a maximum margin classifier. Which of the following statements is/are true? **/1**

- [ ] Slack variables can be used when the training data is not linearly separable in feature space.
- [ ] Slack variables add a penalty for misclassification based on the distance from the boundary.
- [ ] In this case, the classification constraint is given by $t_n y(x_n) \leq \xi_i$
- [ ] When $C \to \infty$, the optimization problem $\frac{1}{2}||w||^2 + C\sum_{n=1}^{N} \xi_n$ reduces to the support vector machine for linearly separable data.

9. What is the advantage of using kernels in support vector machines? **/1**

- [ ] They can simulate an infinite dimensional features space.
- [ ] They will always reduce the number of support vectors.
- [ ] They reduce the risk of getting stuck in local minima.
- [ ] They make it possible to do non-linear separations.

10. Select the conditions that a valid kernel $\mathbf{K}$ must satisfy: **/1**

- [ ] Symmetry: $\mathbf{K} = \mathbf{K}^T$.
- [ ] Orthogonality: $\mathbf{K}^{-1} = \mathbf{K}^T$.
- [ ] Positive semi-definiteness: $\mathbf{v}^T\mathbf{K}\mathbf{v} \geq 0 \ \forall \mathbf{v}$.
- [ ] Negative semi-definiteness: $\mathbf{v}^T\mathbf{K}\mathbf{v} \leq 0 \ \forall \mathbf{v}$.

11. Which of the following feature mappings $\phi(x)$ correspond to the kernel $k(y, z) = (y^T z)^2$? $(x, y, z \in \mathbb{R}^2)$: **/1**

- ☐ $\phi(x) = (x_1^3, x_1 x_2^2, x_2 x_1^2, x_2^2)$.
- ☐ $\phi(x) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2)$.
- ☐ $\phi(x) = (x_1, x_2)$.
- ☐ $\phi(x) = (x_1^3, x_1^2 x_2, \sqrt{2} x_1^2 x_2, \sqrt{3} x_1 x_2^2, x_2^3)$.

12. Which of the following constructions might not lead to a valid kernel. Assume that k_a, k_b are valid kernel functions. **/1**

- ☐ $k(\mathbf{x}, \mathbf{y}) = k_a(\mathbf{x}, \mathbf{y}) + k_b(\mathbf{x}, \mathbf{y})$.
- ☐ $k(\mathbf{x}, \mathbf{y}) = k_a(\mathbf{x}, \mathbf{y}) \cdot k_b(\mathbf{x}, \mathbf{y})$.
- ☐ $k(\mathbf{x}, \mathbf{y}) = k_a(\mathbf{x}, \mathbf{y}) - k_b(\mathbf{x}, \mathbf{y})$.
- ☐ $k(\mathbf{x}, \mathbf{y}) = \exp(k_a(\mathbf{x}, \mathbf{y}))$.

13. Which of the following statements are true? **/1**

- ☐ A Gaussian Process defines a distribution over infinitely many function values.
- ☐ Gaussian Processes maximize a margin.
- ☐ Ridge regression is a special case of a Gaussian Process.
- ☐ Inference in Bayesian Linear Regression is more expensive than in Gaussian processes.

14. Which of the following statements are true? **/1**

- ☐ The Gaussian Process model is a non-parametric method.
- ☐ The Gaussian Process model is a kernel-based method.
- ☐ The Gaussian Process prior is defined by a mean function and a covariance function.
- ☐ The Gaussian Process model optimizes model parameters.

15. Which of the following statements is/are correct? **/1**

- ☐ PCA is scale-invariant (i.e. result is insensitive to feature scaling)
- ☐ K-Means is scale-invariant (i.e. result is insensitive to feature scaling)
- ☐ K-Means always converges in a finite number of iterations
- ☐ For PCA, the implicit assumption that variance is uninformative has to be made.

16. The covariance matrix $\mathbf{S}$ of a data set $\mathcal{D} = \{\mathbf{X}\}$ with $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}]^T$ (i.e. an N by D matrix, where D is the number of features) is given by: **/1**

- ☐ $\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x_n} - \frac{1}{N}\sum_{n=1}^{N}\mathbf{x_n})(\mathbf{x_n} - \frac{1}{N}\sum_{n=1}^{N}\mathbf{x_n})^T$.
- ☐ $\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x_n} - \frac{1}{N}\sum_{n=1}^{N}\mathbf{x_n})^T(\mathbf{x_n} - \frac{1}{N}\sum_{n=1}^{N}\mathbf{x_n})$.
- ☐ $\frac{1}{N}\mathbf{X}\mathbf{X}^T$, if the data is zero-centered (i.e. mean subtracted).
- ☐ $\frac{1}{N}\mathbf{X}^T\mathbf{X}$, if the data is zero-centered (i.e. mean subtracted).

17. Consider a general Poisson mixture model for a dataset $\{x_1, \ldots, x_n\}$. Which of the following statements is/are true? **/1**

- ☐ The responsibilities are given by $r_{nk} = \frac{\pi_k P(x_n|\lambda_k)}{\sum_{l=1}^{K}\pi_l P(x_n|\lambda_l)}$.
- ☐ In the E-step of the EM-algorithm we calculate the MLE or MAP for π_k and λ_k.
- ☐ In the M-step of the EM-algorithm we calculate the MLE or MAP for π_k and λ_k.
- ☐ The maximization of the MLE is an unconstrained optimization problem.

18. Which of the following are true about boosting? **/1**

- ☐ Boosting is an ensemble method.
- ☐ Boosting averages results from multiple models.
- ☐ Boosting uses errors (or their derivatives) from previous rounds.
- ☐ It is possible to apply boosting to any classifier.

19. Which of the following are true about bagging? **/1**

- ☐ Bagging is an ensemble method.
- ☐ Bagging averages results from multiple models.
- ☐ Bagging uses errors (or their derivatives) from previous rounds.
- ☐ It is possible to apply bagging with any classifier.

20. Which of the following are true about decision trees? **/1**

- ☐ Decision trees cannot handle continuous outcomes.
- ☐ A decision tree will always perfectly fit the training set given arbitrary breadth and depth.
- ☐ A decision tree can be used in both bagging and boosting.
- ☐ A decision tree cannot handle mixed type features (e.g. categorical, ordinal and continuous).

2 Principal Component Analysis (PCA) /4

Consider the following data set of three data points in 2-d space:

$$\mathcal{D} = \left\{ \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \end{bmatrix} \right\}$$

Answer the following questions by providing a numerical result (not just a general expression):

1. What is the (normalized) first principal component $\mathbf{u}_1$? You do not need to solve the eigenvalue problem to answer this question. **/1**
2. If we want to project the original data points into 1-d space by using the first principle component (or any other vector of your choice if you couldn't answer the previous question), what is the variance of the projected data? **/2**
3. For the projected data in 1-d space (using the first principal component), now if we represent them in the original 2-d space, what is the reconstruction error? **/1**

3 K-Means /6

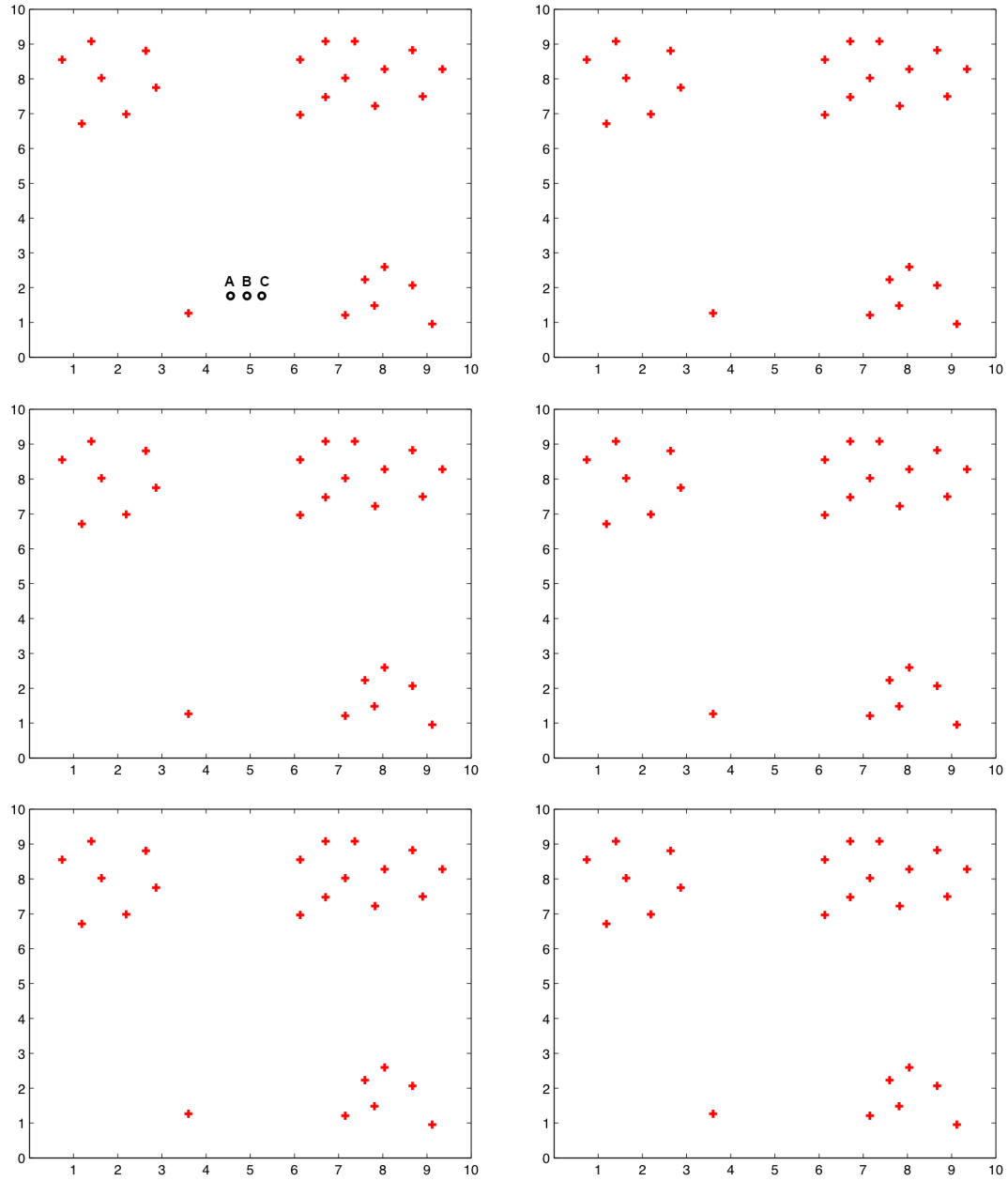
Consider the K-Means algorithm with the following cost function:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\mathbf{x}_n - \boldsymbol{\mu}_k||^2$$

with data points $\mathbf{x}_n$, assignment indicator variables $r_{nk} \in \{0, 1\}$ and cluster centers $\boldsymbol{\mu}_k$.

Now consider the data set in Figure 4. The '+' symbols indicate data points and the (centers of the) circles A, B, and C indicate the starting cluster centers.

1. K-Means can be understood as a sequence of E-M iterations. Explain the updates performed by the K-Means algorithm in both the E- and the M-step in a few words. **/1**
2. Show the results of running the K-means algorithm (until convergence) on this data set. Indicate your solution directly on the figures provided (see Figure 4). For each iteration: indicate which data points will be associated with each of the clusters and show (*approximately*) the locations of the updated cluster centers. A cluster center will not move during the M-step if it has no points associated with it. Use as many figures as you need until the algorithm converges. **/4**
3. After how many full E-M iterations does the algorithm converge? **/1**



Figuur 4: K-Means data set. Indicate your solution directly on these figures.

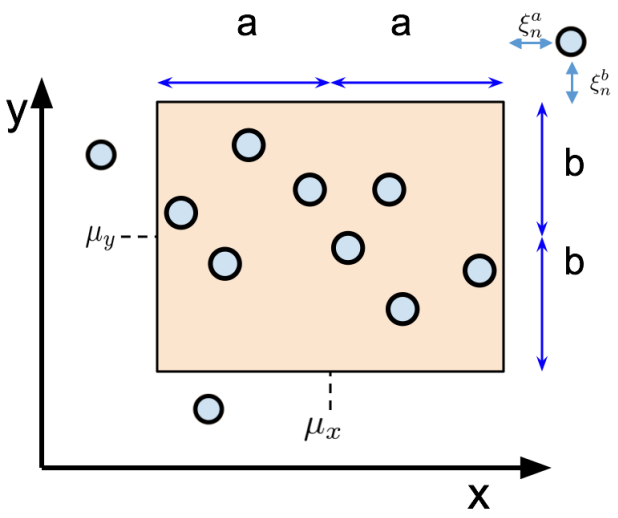
4 Outlier Detection

/7

We are given the following dataset: $\{x_n, y_n\}_{n=1}^N$, where each $x_n \in \mathbb{R}$ and $y_n \in \mathbb{R}$. We want to find the minimum volume enclosing box. We allow outsiders to stay outside of the box, however they will have to pay a penalty proportional to their distance to the box, as well as cost factor. One way of doing this is to minimize the sum of squares of the box widths:

$$\begin{aligned}
\min_{a,b,\mu_x,\mu_y,\xi_n^a,\xi_n^b} \quad & a^2 + b^2 + C_a \sum_n \xi_n^a + C_b \sum_n \xi_n^b \\
\text{subject to:} \quad & a > 0 \quad \text{and} \quad b > 0 \\
& \xi_n^a \geq 0 \qquad \forall n \\
& \xi_n^b \geq 0 \qquad \forall n \\
& (x_n - \mu_x)^2 \leq a^2 + \xi_n^a \qquad \forall n \\
& (y_n - \mu_y)^2 \leq b^2 + \xi_n^b \qquad \forall n
\end{aligned}$$

The parameter μ_x represents center of the box along the x-axis (and similarly for μ_y). A diagram of a *feasible* but non-optimized solution is shown below.



1. Provide an expression for the primal Lagrangian. Use Lagrange multipliers $\{\alpha_n\}$ and $\{\beta_n\}$, as well as A_n and B_n for the slack variables ξ_n^a and ξ_n^b. Include constraints on the Lagrange multipliers. **/2**
2. Write down and solve for all the KKT conditions. **/3**
3. Use these conditions to write down the dual optimization problem including the dual Lagrangian and the dual constraints. **/2**

5 Mixture Models

/7

Consider that you have a data set $X = (x_1, \ldots, x_n)^T$, where each observation x_n contains the time difference (> 0) between two events. You can assume that the observations are independent. You are asked by your employer to partition (cluster) these data points into K clusters so as to better understand the trends of this data set. After a quick search on Wikipedia you find out that an appropriate distribution for this type of data is the Exponential distribution with density:

$$\text{Exp}(x|\lambda) := \lambda \cdot e^{-\lambda \cdot x} \tag{1}$$

where $\lambda > 0$. Armed with this knowledge you try to define a mixture model for the data with a discrete latent variable z with mass function $p(z) = \prod_{k=1}^{K} \pi_k^{[z=k]}$ and $\sum_{k=1}^{K} \pi_k = 1$, $\pi_k \geq 0$ as the prior over the clusters ($[z = k]$ is the indicator function). You assume that each cluster has its own parameter λ_k.

Write down:

1. the conditional probability of a single data point x_n under the parameter of cluster k, and the total probability of data point x_n under this model. **/1**
2. the responsibility of a cluster k generating a data point x_n. **/1**
3. the log-likelihood of the entire data set X (under i.i.d. assumptions). **/1**
4. expressions for the maximum-likelihood estimators λ_k, π_k in terms of the data X and the corresponding responsibilities. **/3**
5. Describe how you can perform maximum-likelihood learning of λ_k, π_k by using an EM-algorithm (Expectation - Maximization) given the data set X and derive the update equations. **/1**