UNIVERSITY OF AMSTERDAM

Faculty of Science

Exam

Machine Learning 1

Midterm Exam
Date: September 30, 2016
Time: 15:00-17:00

Number of pages: 7 (including front page)
Number of questions: 4
Maximum number of points to earn: 38
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. Which of the following tasks is a supervised learning task? /1

   □ Clustering DNA sequences based on similarities.

   ☒ Predicting tomorrows gold price based on its historical time series and on those of other economical variables.

   □ Clustering spam emails based on the words and language usage and the IP-address of the sender.

   ☒ Examining the win-lose-statistics of soccer teams and predicting which team will win in the next game.

2. Which of the following expressions is equal to 1, given no independence assumption and for (non-trivial) discrete random variables? /1

   □ $\sum_b P(A|B=b)$.

   ☒ $\sum_a P(A=a|B)$.

   □ $\sum_a \sum_b P(A=a|B=b)$.

   □ None of the above.

3. Which of the following equations are correct? /1

   ☒ $\mathbb{E}_{p(x)}[f(x)] = \int f(x)p(x)dx$.

   ☒ $Var_{p(x)}[f(x)] = \int p(x)f(x)^2dx - (\int p(x)f(x)dx)^2$.

   ☒ $Var_{p(x)}[f(x)] = \int p(x)(f(x) - \int p(x)f(x)dx)^2dx$.

   □ $\mathbb{E}_{p(x)}[f(x)] = \int xf(x)p(x)dx$.

4. Which of the following pair of events A, B is independent? /1

   ☒ A: rolling a 6 with a die, B: rolling a 6 with the same die.

   □ A: rolling a 6 with a die, B: the sum of the numbers of the first die and a second die is 8.

   □ A: drawing a black card from a deck of cards, B: drawing a black card from the same deck of cards without replacing the first card.

   ☒ A: drawing a black card from a deck of cards, B: drawing a black card from the same deck of cards after replacing the first card.

5. Let $\sigma(a) = \frac{1}{1+e^{-a}}$ be the sigmoid function. What is its derivative $\frac{d\sigma(a)}{da}$? **/1**

☒ $\sigma(a)(1-\sigma(a))$.

□ $\sigma(a)$.

□ $\frac{\sigma(a)-1}{\sigma(a)}$.

□ $\sigma(a)(\sigma(a)-1)$.

6. You are given a data set $\{\boldsymbol{x}_n, t_n\}$ of N data-points, and basis-functions $\phi_0, \ldots, \phi_M$. You are asked to choose the optimal value for λ in regularized regression. Which of the following approaches is the most suitable. **/1**

□ Train models with a range of different values for λ on the data. Choose the value of λ that gives the smallest error.

☒ Split the data at least into a training and validation set. Train models with a range of different values for λ on the training data. Choose the value of λ that minimizes the prediction error on the validation data.

□ Split the data at least into a training and validation set. Train models with a range of different values for λ on the training data. Choose the value of λ that minimizes the variance on the training set and the bias on the validation set.

□ Train models with a range of different values for λ on the data. Choose the value of λ that minimizes the variance.

7. Consider regularized linear regression with the error function $E(\mathbf{w}, \lambda) = \frac{1}{2}(\mathbf{\Phi w} - \mathbf{t})^T(\mathbf{\Phi w} - \mathbf{t}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$. You want to find the optimal regularization penalty $\lambda \in \{0.01, 0.1, 1\}$ using K-fold cross-validation. You obtain the following cross-validation errors $E(\mathbf{w}, \lambda)$: $E(\mathbf{w}, \lambda_1 = 0.01) = 0.91$, $E(\mathbf{w}, \lambda_2 = 0.1) = 0.23$, $E(\mathbf{w}, \lambda_3 = 1) = 0.74$. Which of the following interpretations is correct? **/1**

□ λ_1: underfitting, λ_2: best fit, λ_3: overfitting.

□ λ_1: overfitting, λ_2: best fit, λ_3: overfitting.

☒ λ_1: overfitting, λ_2: best fit, λ_3: underfitting.

□ λ_1: underfitting, λ_2: best fit, λ_3: underfitting.

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

☒ Overfitting is more likely when the set of training data is small.

☒ If you are given N data points, and use half for training and half for testing, the difference between training error and test error is non-increasing or decreasing as N increases.

□ When the feature space is larger (i.e. more parameters), overfitting is less likely.

□ Linear regression typically has high variance when trying to fit data drawn from a highly non-linear distribution.

9. Consider two polynomial regression models of order M_1 and M_2, with $M_1 > M_2$. Which model is more likely to fit the training data well? Which model is more likely to fit the test data well? /1

   ☒ Training set: Model 1; Test set: impossible to tell.

   □ Training set: Equally likely; Test set: Model 2.

   □ Training set: Model 1; Test set: Model 2.

   □ Training set: Equally likely, Test set: impossible to tell.

10. Which statement is **not** true? /1

    □ The MAP estimator for the mean of a Gaussian distribution is a linear combination of the sample mean and the prior mean.

    □ The maximum likelihood principle states that the most likely explanation of the data $\mathcal{D}$ is given by the index $\boldsymbol{w}$ that maximizes the likelihood.

    □ The maximum a posteriori principle states that the most likely explanation of the data $\mathcal{D}$ is given by the index $\boldsymbol{w}$ that maximizes the a posteriori distribution.

    ☒ The maximum likelihood estimator for the variance of a Gaussian distribution is unbiased.

11. We consider *maximum likelihood* (ML) and *maximum a posteriori* (MAP) inference. Let $\mathcal{D}$ be the data and $\mathbf{w}$ the model parameters. Choose the correct statement: /1

    □ MAP: $\arg\max_{\mathbf{w}} p(\mathcal{D}|\mathbf{w})$, ML: $\arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D})$.

    □ MAP: $\arg\max_{\mathcal{D}} p(\mathbf{w}|\mathcal{D})$, ML: $\arg\max_{\mathcal{D}} p(\mathcal{D}|\mathbf{w})$.

    ☒ MAP: $\arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D})$, ML: $\arg\max_{\mathbf{w}} p(\mathcal{D}|\mathbf{w})$.

    □ MAP: $\arg\max_{\mathcal{D}} p(\mathcal{D}|\mathbf{w})$, ML: $\arg\max_{\mathcal{D}} p(\mathbf{w}|\mathcal{D})$.

12. Given a data set $\mathcal{D} = \{x_n\}_{n=1}^{N}$. The data is normally distributed $\mathcal{N}(x_n|\mu, \sigma^2)$ and we assume a Gaussian prior over $\mu : \mathcal{N}(\mu|0, \sigma_0^2)$. Furthermore, the variance σ^2 is assumed to be known. How does the prior distribution change as (i) $\sigma_0 \to 0$,(ii) $\sigma_0 \to \infty$ (iii) $N \to \infty$? /1

    □ (i) wider (ii) narrower (iii) same.

    □ (i) wider (ii) narrower (iii) narrower.

    □ (i) narrower (ii) wider (iii) narrower.

    ☒ (i) narrower (ii) wider (iii) same.

13. In the same setting as before, how does the posterior distribution change as (i) $\sigma_0 \to 0$,(ii) $\sigma_0 \to \infty$ (iii) $N \to \infty$? /1

    □ (i) wider (ii) narrower (iii) wider.

    □ (i) same (ii) same (iii) narrower.

    ☒ (i) narrower (ii) wider (iii) narrower.

    □ (i) same (ii) same (iii) same.

14. In the same setting as before, how does $|\mu_{\text{MLE}} - \mu_{\text{MAP}}|$ change as (i) $\sigma_0 \to 0$,(ii) $\sigma_0 \to \infty$ (iii) $N \to \infty$? **/1**

   □ (i) decrease (ii) increase (iii) decrease.

   □ (i) decrease (ii) increase (iii) increase.

   ☒ (i) increase (ii) decrease (iii) decrease.

   □ (i) increase (ii) decrease (iii) increase.

15. Let $\mathcal{D}$ be the training dataset and $\mathbf{w}$ the model parameters. Which of the following describes a Bayesian prediction for the target variable of a new datapoint $\mathbf{x}$? **/1**

   □ $t^* = \int\int t \cdot p(t, \mathbf{w}|\mathbf{x}) d\mathbf{w} dt$.

   ☒ $t^* = \int\int t \cdot p(t, \mathbf{w}|\mathcal{D}, \mathbf{x}) d\mathbf{w} dt$.

   □ $t^* = \max_t p(t|\mathbf{x}, \mathbf{w}) p(\mathbf{w})$.

   □ $t^* = \max_t p(t|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathcal{D})$.

16. In classification, there are three approaches, discriminant functions, probabilistic generative models and probabilistic discriminative models. Following are statements about probabilistic generative models and probabilistic discriminative models. Choose the **incorrect** statement. **/1**

   □ Logistic regression is a discriminative model.

   □ In discriminative models, the probability of class C given input x, $p(C|x)$ is modeled directly.

   ☒ In discriminative models, the prior probability of class $p(C)$ is modeled.

   □ Generative models model class conditional probability $p(x|C)$.

17. After fitting a Logistic Regression model with two classes to the training data we calculate the confusion matrix of the model on the test data with N observations: $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$. Which of the following formulas could you use to estimate the misclassification error: **/1**

   □ $\frac{1}{N}(A + D)$.

   ☒ $1 - \frac{1}{N}(A + D)$.

   ☒ $\frac{1}{N}(B + C)$.

   □ $1 - \frac{1}{N}(B + C)$.

18. Consider stochastic gradient descent for Logistic Regression with error function $E(w)$. Which of the following statements is correct? **/1**

   □ If after a few iterations $E(w)$ increases instead of decreases then most likely the learning rate η is too small.

   ☒ If after a few iterations $E(w)$ increases instead of decreases then most likely the learning rate η is too big.

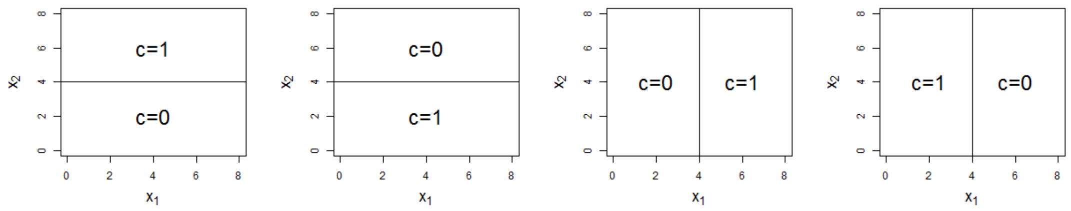
   ☒ Stochastic gradient descent is an online learning method.

   ☒ Stochastic gradient descent is useful for Logistic Regression problems, because no closed-form (analytic) solution for the global minimum exists.

19. Suppose we have fitted a 1-dimensional Linear Regression model (without regularization) to the training data and the computer says that the sum-of-squares error function $E(w)$ is exactly zero on the training set. What does this mean? **/1**

- □ The learned parameter w must be zero.
- □ Our learned prediction function must be the zero function.
- ☒ All of our training data perfectly lies on some straight line.
- □ This is not possible. Our implemented error function must be faulty.

20. After fitting a Logistic Regression model with two classes $\{0, 1\}$ to our 2-dimensional training data we get the function $y(x, w) = \sigma(w_0 + w_1 x_1 + w_2 x_2)$ with $x = (x_1, x_2)$ and learned $w = (w_0, w_1, w_2) = (8, 0, -2)$. How do the decision regions and decision boundary look like in the x_1-x_2-plane?



(a) (b) (c) (d)

Figuur 1: Decision regions.

/1

- □ 1(a).
- ☒ 1(b).
- □ 1(c).
- □ 1(d).

2 Probability Theory and Bayes' Rule

/6

Suppose that 5 men out of 100 and 25 women out of 10 000 are color-blind.

1. Define all random variables (including the values they can take). /1
2. A color-blind person is chosen at random. What is the probability of this person being male? Assume males and females to be in equal numbers. /3
3. How does the probability change if we assume that the number of women is double the number of men? /2

$S \in \{m, f\}$ sex (male, female), $B \in \{b, n\}$ blindness (colorblind, normal).

$$p(m|b) = \frac{p(b|m) \cdot p(m)}{p(b|m) \cdot p(m) + p(b|f) \cdot p(f)}.$$

With numbers we get $p(m|b) = 20/21 = 95.2\%$ and $20/22 = 91\%$.

a) 1 point, b) 1 point for bayes rule, 1 point for numbers, 1 point for result, c) 1 point for numbers, 1 point for result.

3 Maximum Likelihood and A Posteriori Estimates

/7

Consider the experiment of throwing a (possibly biased) coin. The possible outcomes of a coin throw are $\{0, 1\}$. Assume the data you observed so far is $[1, 1, 1, 1, 1]$.

1. Compute the maximum likelihood estimator ρ_{ML} for the unknown probability ρ of throwing a 1. /3
   Hint: The Bernoulli distribution is given by: $\mathrm{Ber}(x|\rho) = \rho^x (1-\rho)^{1-x}$.
2. Now we are adopting a Bayesian view of the previous situation and introduce a prior over ρ. Since it is probably very hard to bias a coin (but you never know), we will formulate our prior to allow for uncertainty on whether the coin is fair or not: $p(\rho)$ =Beta$(\rho|2, 2)$. Compute the maximum a posteriori estimate for ρ. /3
   Hint: The Beta distribution has the form: $\mathrm{Beta}(\rho|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \rho^{a-1}(1-\rho)^{b-1}$.
3. How do you interpret the difference between ρ_{ML} and ρ_{MAP}? /1

$\rho_{ML} = \frac{n}{N} = \frac{5}{5} = 1$, $\rho_{MAP} = \frac{n+1}{N+2} = \frac{6}{7} = 0.8571$, where n is the number of "1"s and N the number of trials. ML solution degenerate for the unlucky outcome of our experiment. The posterior assigns some probability to the ML solution, but including our prior knowledge, we get a more realistic result. We can interpret the posterior as if we have already seen two trials with one "1änd one "0", so in total 6 "1"s and one "0".

a) 1 point for criterion, 1 point for solving, 1 point for solution b) 1 point for criterion, 1 point for solving, 1 point for solution, c) 1 point for making some sense.

4 Weighted Least Squares

/5

Consider the following weighted sum-of-squares error function for a data set where each target value t_n is associated with a weighting factor $r_n > 0$:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} r_n \left\{ t_n - \mathbf{w}^T \boldsymbol{\phi}(x_n) \right\}^2 = \frac{1}{2} \left(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w} \right)^T \mathbf{R} \left(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w} \right), \qquad (1)$$

with a diagonal weighting matrix

$$\mathbf{R} = \begin{bmatrix} r_1 & 0 & 0 & \dots & 0 \\ 0 & r_2 & 0 & \dots & 0 \\ 0 & 0 & r_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & r_N \end{bmatrix} . \tag{2}$$

Find an expression for the solution $\mathbf{w}^*$ that minimizes this error function. In case you need to invert a matrix make the explicit assumption that the matrix is invertible. Furthermore, you may assume that $E(w)$ is a convex function of w. So you don't need to check for second derivatives.

Hint: Recall that $\mathbf{a}^T\mathbf{B}\mathbf{c} = (\mathbf{B}^T\mathbf{a})^T\mathbf{c} = \mathbf{c}^T\mathbf{B}^T\mathbf{a}$ for any two vectors $\mathbf{a}$, $\mathbf{c}$ and an appropriate matrix $\mathbf{B}$. In case you have difficulties deriving the solution in matrix form solve the corresponding scalar version first.

First, we expand the expression:

$$E(\mathbf{w}) = \frac{1}{2}\left(\mathbf{t} - \mathbf{\Phi}\mathbf{w}\right)^T \mathbf{R}\left(\mathbf{t} - \mathbf{\Phi}\mathbf{w}\right) = \tag{3}$$

$$= \frac{1}{2}\left\{\mathbf{t}^T\mathbf{R}\mathbf{t} - 2\mathbf{t}^T\mathbf{R}\mathbf{\Phi}\mathbf{w} + (\mathbf{\Phi}\mathbf{w})^T\mathbf{R}\mathbf{\Phi}\mathbf{w}\right\} = \tag{4}$$

$$= \frac{1}{2}\left\{\mathbf{t}^T\mathbf{R}\mathbf{t} - 2\mathbf{t}^T\mathbf{R}\mathbf{\Phi}\mathbf{w} + (\mathbf{\Phi}^T\mathbf{R}\mathbf{\Phi}\mathbf{w})^T\mathbf{w}\right\} . \tag{5}$$

Next, we take the derivative with respect to $\mathbf{w}$:

$$\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = -\mathbf{\Phi}^T\mathbf{R}\mathbf{t} + \mathbf{\Phi}^T\mathbf{R}\mathbf{\Phi}\mathbf{w} . \tag{6}$$

We want to find a minimum, so we set the derivative to 0:

$$-\mathbf{\Phi}^T\mathbf{R}\mathbf{t} + \mathbf{\Phi}^T\mathbf{R}\mathbf{\Phi}\mathbf{w}^* = 0 \tag{7}$$

$$\mathbf{\Phi}^T\mathbf{R}\mathbf{\Phi}\mathbf{w}^* = \mathbf{\Phi}^T\mathbf{R}\mathbf{t} . \tag{8}$$

And finally, we have:

$$\mathbf{w}^* = \left(\mathbf{\Phi}^T\mathbf{R}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{R}\mathbf{t} . \tag{9}$$

1 point for expanding the expression, 1 point for rearranging the terms (by using the hint), 1 point for the general criterion (derivatives to zero), 1 point for calculating the derivatives, 1 point for solving for w.