Exam 1

Introduction to Machine Learning, SS 15, U Stuttgart

Prof. Dr. Marc Toussaint

Oct 5, 2015

Name:
Matrikelnummer:

DO NOT OPEN THE EXAM BEFORE THE ANNOUNCEMENT

DIE KLAUSUR NICHT VOR DER ANSAGE ÖFFNEN

- There are 36 points in total.
- You have 180 minutes.
- Write your name on all sheets.
- Put your *Studentenausweis* next to you, so we can check it during the exam.
- **You may only use your pen and scratch paper – no other materials (no textbooks, script, or mobiles) are allowed.**
- Please try to **answer only with equations**, no lengthy text. Of course, we will try to read it if necessary. But usually all answers are well defined in terms of equations.
- Also use the back of the sheets if necessary. Please indicate clearly when you use the back of sheets or extra sheets.

Question 1 — Probabilistic independence (2Pts)

Are binary A and B independent given their joint distribution below? Justify your answer.

	$B = 0$	$B = 1$
$A = 0$	0.02	0.08
$A = 1$	0.18	0.72

Question 2 — Bayesian reasoning (2Pts)

90% of all green party voters and 60% of all Germans refuse nuclear energy. 15% of the Germans vote for green. Assume you know about Anna (a German) that she refuses nuclear energy, what is the probability that she votes for green?

Question 3 — Features & Regularization (4Pts)

a) Given data $D = \{(x_i, y_i)\}_{i=1}^n$ we use a model $f(x) = \phi(x)^\top \beta$. Let $x \in \mathbb{R}$ be 1-dimensional. Please define the following three different kinds of feature vectors $\phi(x)$: (1) linear features, (2) polynomial features of degree 3 (up to cubic), (3) radial basis functions (RBF) at the data inputs. [2]

b) What is the cost function $L^{\text{ridge}}(\beta)$ for ridge regression including a regularization parameter λ? What for Lasso regularization? [2]

Question 4 — Bootstrap & combining learners (3Pts)

a) Given a data set D, describe what bootstrapping means. [1]

b) Given a set of learned models $f_1, .., f_M$ (each is a regression $f_i : \ x \mapsto y$) and the data set D, how can you combine these learned models to a single model $f(x)$ in a way that is better than naive averaging? [2]

Question 5 — Neural Networks (4Pts)

Given data $D = \{(x_i, y_i)\}_{i=1}^N$, where $y_i \in \{1, 2, 3\}$ indicates one of three classes, you want to train a neural network classifier.

a) Define how you can represent a classifier with a 2-layer neural network: Provide exact equations on how an input x is mapped to an output $y \in \{1, 2, 3\}$ depending on the weight matrices (W_0, W_1, W_2) of the 2-layer neural network.[2]

b) Define a cost function for training the neural network on D. (You do not have to derive gradients.)[2]

Question 6 — Clustering (5Pts)

a) Describe the k-means clustering algorithm in terms of the two steps that are alternated.[1]

b) k-means converges to a local minimum of an objective—what is this objective that k-means minimizes?[1]

c) Given data $D = \{x_i\}_{i=1}^n$, describe how you cluster this data using spectral clustering and k-means: First describe how you use D to setup the eigenvalue problem of spectral clustering; then how the output of this eigenvalue problem is used to do the actual clustering.[3]

Question 7 — Logistic regression & log-likelihood gradient (5Pts)

Consider a binary classification problem with data $D = \{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$.

a) What is the probability $P(y=1|x)$ under a logistic regression model with linear discriminative function $f(x) = x^\top \beta$? [1]

b) What is the full data log-likelihood $L(\beta)$ under this model? [1]

c) Derive the partial derivative $\frac{\partial}{\partial \beta} L(\beta)$. Use the fact $\frac{\partial}{\partial z}\sigma(z) = \sigma(z)(1 - \sigma(z))$ for the sigmoid function $\sigma(z)$. [3]

Question 8 — Expectation-Maximization for an outlier regression model (6Pts)

a) Give the general definition of the E-step and M-step in Expectation-Maximization in terms of the expected complete data log-likelihood (or free energy, if you like). [2]

b) Assume we have data $D = \{(x_i, y_i)\}_{i=1}^n$. We assume that for each data point there exists a latent binary variable $c_i \in \{0, 1\}$ that indicates whether the data point is an outlier. More specifically, the likelihood is

$$P(y_i \,|\, x_i, c_i, \beta) = \mathcal{N}(y_i \,|\, x_i^\top \beta, \sigma_{c_i}^2) \tag{1}$$

where β are the parameters of the linear regression, σ_1 is a very large variance (associated to the outliers) and σ_0 describes the variance of inliers. The outlier prior $P(c_i{=}1) = \mu$ is parameterized by $\mu \in [0, 1]$. The model parameters are therefore $(\beta, \sigma_0, \sigma_1, \mu)$.

Derive explicit equations for the E-step and M-step in the Expectation Maximization algorithm to estimate the unknown β. [4]