ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Computational Intelligence Lab

2015

Prof. T. Hofmann

Final Exam

17th August 2015

First and Last name: _____

Student ID (Legi) Nr: _____

Signature: _____

General Remarks

- Please check that you have all 16 pages of this exam.
- There are 120 points, and the exam is 120 minutes. **Don't spend too much time on a single question!** The maximum of points is not required for the best grade!
- Remove all material from your desk which is not permitted by the examination regulations.
- Write your answers directly on the exam sheets. If you need more space, make sure you put your **student-ID**-number on top of each supplementary sheet.
- Immediately inform an assistant in case you are not able to take the exam under regular conditions. Later complaints are not accepted.
- Attempts to cheat/defraud lead to immediate exclusion from the exam and can have judicial consequences.
- Please use a black or blue pen to answer the questions.
- Provide only one solution to each exercise. Cancel invalid solutions clearly.

	Topic	Max. Points	Points Achieved	Visum
1	Dimensionality Reduction	30		
2	Clustering, Mixture Models, NMF	30		
3	Sparse Coding and Dict. Learning	30		
4	Optimization / Robust-PCA	30		
Total		120		

Grade: .....

1 Dimensionality reduction (30 pts)

1.1 SVD and PCA

Which of the following claims are true/false? *(1 point per correct answer, -1 point per incorrect answer, non-negative total points in any case)*

7 pts

a) PCA is a dimensionality reduction which tries to minimize the variance of the data.
[ ] True [ ] False

b) The first principal direction is the eigenvector of the data matrix $\mathbf{X}$ with largest associated eigenvalue.
[ ] True [ ] False

c) When using PCA, one typically discards the small eigenvalues.
[ ] True [ ] False

d) *(counts for 2/-2pts).* We first perform SVD on a data matrix $\mathbf{X}$. We then rotate the data matrix $\mathbf{X}$. The singular vectors of the rotated matrix will be the same.
[ ] True [ ] False

e) *(counts for 2/-2pts).* Consider SVD for collaborative filtering for a matrix $\mathbf{A}$ of rank r, where we assume all unobserved entries are marked by zero. When using $K = r$ latent concepts, the system will predict zero for all unobserved entries.
[ ] True [ ] False

1.2 SVD

Recall the definition of the matrix operator norm: $\|\mathbf{A}\|_2^2 := \max_{\mathbf{x}} \|\mathbf{Ax}\|_2^2$ where $\|\mathbf{x}\|_2^2 = 1$.

a) Using the SVD decomposition of $\mathbf{A} = \mathbf{USV}^\top$, prove that $\|\mathbf{A}\|_2^2 = \|\mathbf{S}\|_2^2$

4 pts

. . . . . .

. . . . . .

. . . . . .

b) Using SVD, prove that

$$rank(\mathbf{A}) = rank(\mathbf{AA}^\top) = rank(\mathbf{A}^\top\mathbf{A})$$

4 pts

. . . . . .

. . . . . .

. . . . . .

1.3 PCA

Part 1

Consider the following data matrix $\mathbf{X} \in \mathbb{R}^{D \times N}$, where $N = 3$ is the number of datapoints, and $D = 2$ is the dimension of each datapoint.

$$\mathbf{X} = \begin{bmatrix} 2 & -1 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

Compute $\mathbf{X}\mathbf{X}^\top$ and then compute the first two principal axes of the data matrix $\mathbf{X}$.

6 pts

. . . . . .

. . . . . .

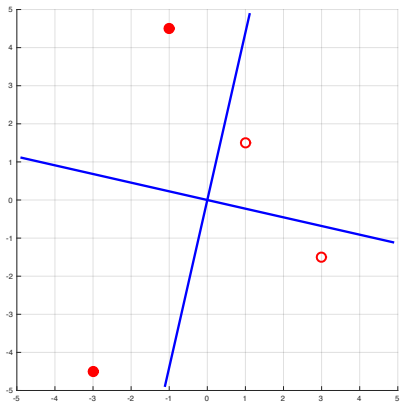
. . . . . .

Part 2

Now consider the following data matrix $\mathbf{X} \in \mathbb{R}^{D \times N}$ where $D = 2$ and $N = 4$.

$$\mathbf{X} = \begin{bmatrix} 1 & -3 & -1 & 3 \\ 1 & -5 & 4 & -2 \end{bmatrix}$$

The centered data $\hat{\mathbf{X}}$ (i.e. $\mathbf{X}$ from which we subtracted the mean) is represented in the figure below. The two lines represent the principal axes of $\hat{\mathbf{X}}$.



The covariance of $\hat{\mathbf{X}}$ is given by

$$\mathbf{S} = \begin{bmatrix} 20 & 6 \\ 6 & 45 \end{bmatrix}$$

What dimension has the most variance?

2 pts ☐

. . . . . .

Project the two points represented as filled-in circles on the two principal axes (draw directly on the figure). Can you conclude which principal axis is a better choice to get a 1-D projection of the data $\mathbf{X}$?

3 pts ☐

. . . . . .

We now consider a new data matrix $\mathbf{Y} = \begin{bmatrix} \mathbf{X} & 2\mathbf{X} \end{bmatrix} \in \mathbb{R}^{D \times 2N}$. Write down the covariance of the centered data $\hat{\mathbf{Y}}$ and draw the principal axes on the same figure as $\mathbf{X}$.

4 pts ☐

. . . . . .

. . . . . .

2 Clustering, Mixture Models, NMF (30 pts)

2.1 K-means clustering

Consider a data matrix $\mathbf{X} \in \mathbb{R}^{D \times N}$ of N data points in D dimensions. We want to perform the K-means algorithm with the Euclidean distance on $\mathbf{X}$ with the addition that we also want to minimize the ℓ_2-norm of each cluster center $\mathbf{u}_k$.

$$\min_{\mathbf{U},\mathbf{Z}} \left[J(\mathbf{U},\mathbf{Z}) := \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn}\|\mathbf{x}_n - \mathbf{u}_k\|_2^2 + \frac{1}{2}\sum_{k=1}^{K}\|\mathbf{u}_k\|^2 \right]$$

Derive the update rule for the cluster centers $\mathbf{u}_k$.

7 pts

2.2 Mixture model

Suppose we have a set of N observations $(x_1, \ldots, x_N)$ and we want to model this data using a mixture of k Poisson distributions defined as

$$p_\lambda(x) = \sum_{i=1}^{k} \pi_i g(x; \lambda_i),$$

where $\lambda = (\lambda_1, \ldots, \lambda_k)$.

Recall that a random variable X has a Poisson distribution with parameter λ if its probability mass function is given by:

$$g(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

Write down the expression for the log-likelihood $\mathcal{L}$.

2 pts ☐

. . . . . .

Write down the update equation for the naive attempt at maximizing $\mathcal{L}$ with respect to a specific mixture parameter λ_i.

6 pts ☐

. . . . . .

. . . . . .

. . . . . .

Explain the problem with the aforementioned approach.

3 pts ☐

. . . . . .

. . . . . .

"The Expectation maximization algorithm maximizes a lower bound on the log-likelihood". Explain what this lower bound is.

2 pts ☐

. . . . . .

. . . . . .

2.3 Nonnegative Matrix Factorizations

Which of the following claims are true/false? *(2 points per correct answer, -2 points per incorrect answer, non-negative total points in any case)*

10 pts

a) Consider NMF for clustering a set of datapoints $\mathbf{X}$. In the solution to the NMF problem, every datapoint is assigned to at most one cluster.
[ ] True [ ] False

b) For non-negative, non-zero matrices $\mathbf{X} \in \mathbb{R}^{D\times N}$, $\mathbf{U} \in \mathbb{R}^{D\times K}$ and $\mathbf{Z} \in \mathbb{R}^{K\times N}$ with $N, D > 1$, furthermore $rank(\mathbf{X}) = \min\{D, N\}$ and $K < \min\{D, N\}$, one can always find $\mathbf{U}$ and $\mathbf{Z}$ such that $\mathbf{X} = \mathbf{UZ}$ exactly.
[ ] True [ ] False

c) In the NMF Algorithm, the update for $\mathbf{U}$ is given as

$$u_{dk} \leftarrow u_{dk} \frac{\left(\mathbf{XZ}^\mathsf{T}\right)_{dk}}{\left(\mathbf{UZZ}^\mathsf{T}\right)_{dk}}$$

[ ] True [ ] False

d) Consider a non-negative $\mathbf{X}$ and a quadratic cost function as in k-means:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{Z}} \quad & J(\mathbf{U},\mathbf{Z}) = \frac{1}{2}\|\mathbf{X} - \mathbf{UZ}\|_F^2. \\ \text{s.t.} \quad & u_{dk} \in [0,\infty) \ \forall d,k \qquad (1) \\ & z_{kn} \in [0,\infty) \ \forall k,n. \qquad (2) \end{aligned}$$

Semi-NMF is obtained by relaxing the positivity assumption (1) on $\mathbf{U}$
[ ] True [ ] False

e) Semi-NMF — when additionally requiring all columns of $\mathbf{Z}$ to sum up to one, and all rows of $\mathbf{Z}$ are orthogonal — is equivalent to k-means clustering.
[ ] True [ ] False

3 Sparse Coding and Dictionary Learning (30 pts)

3.1 Sparse coding

Part 1

Which of the following claims are true/false? *(1 point per correct answer, -1 point per incorrect answer, non-negative total points in any case)*

4 pts ☐

Consider the following signal reconstruction problem

$$\mathbf{z}^* \in \underset{\mathbf{z}}{\operatorname{argmin}} \|\mathbf{z}\|_0, \quad \text{s.t.} \quad \mathbf{y} = \Theta\mathbf{z}$$

a) The above minimization of $\|\mathbf{z}\|_0$ is convex but NP-hard [ ] True [ ] False

b) One can always replace $\|\mathbf{z}\|_0$ by $\|\mathbf{z}\|_1$ and obtain the same solution [ ] True [ ] False

c) Using an overcomplete dictionary yields sparser solutions [ ] True [ ] False

d) Recall: the coherence of $\mathbf{U}$ is defined as

$$m(\mathbf{U}) = \max_{i,j:i\neq j} \left|\mathbf{u}_i^\top \mathbf{u}_j\right|.$$

For an orthogonal basis $\mathbf{B}$, we have $m(\mathbf{B}) = rank(\mathbf{B})$ [ ] True [ ] False

Part 2

Consider a signal $\mathbf{x} \in \mathbb{R}^D$ which we want to represent as a linear combination of basis vectors $\mathbf{U} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N\}$. In other words, $\mathbf{x} = \mathbf{U}\mathbf{z}$ for $\mathbf{z} \in \mathbb{R}^N$. We also assume that $D = N$ and the matrix $\mathbf{U}$ is orthogonal (i.e. $\mathbf{U}\mathbf{U}^\top = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix).

Prove that given $\mathbf{x}$, the transformed representation in the new basis is $\mathbf{z} = \mathbf{U}^T\mathbf{x}$.

4 pts ☐

. . . . . .

. . . . . .

Consider two signals $\mathbf{x}$ and $\mathbf{x}'$ with their corresponding transforms $\mathbf{z}$ and $\mathbf{z}'$ (more precisely $\mathbf{x} = \mathbf{U}\mathbf{z}$ and $\mathbf{x}' = \mathbf{U}\mathbf{z}'$). For an orthogonal matrix $\mathbf{U}$, prove that the change of basis preserves pairwise distances i.e. $\|\mathbf{x} - \mathbf{x}'\|_2 = \|\mathbf{z} - \mathbf{z}'\|_2$.
(*Hint:* Use the fact that $\|\mathbf{x}\|_2 = \|\mathbf{z}\|_2$ when $\mathbf{U}$ is orthogonal.)

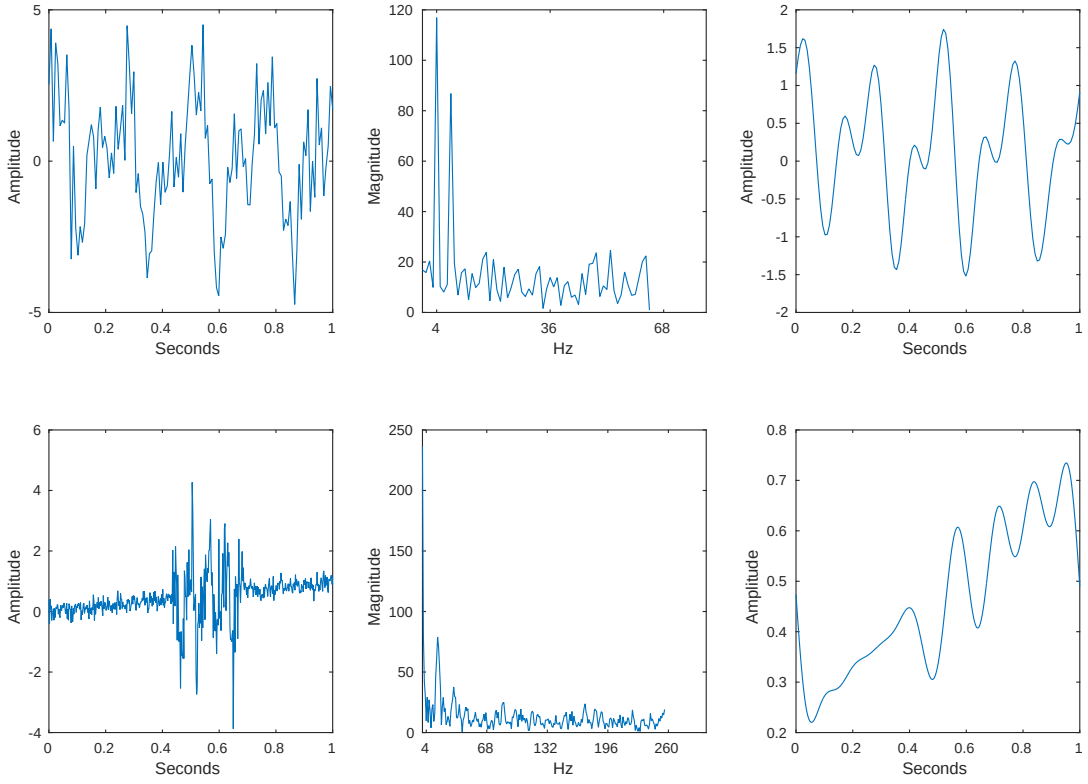
7 pts

. . . . . .

. . . . . .

. . . . . .

Part 3



The figure above shows two different 1-D signals (left column) with their corresponding spectrum obtained using the FFT (middle column). In the right column, we show a signal obtained by discarding part of the frequencies in the spectrum.

(1) Write down the formula to obtain the spectrum in the middle column of the previous figure, in terms of linear transformation or change of basis (assuming a given basis $\mathbf{U}$) applied to the original signal $\mathbf{x}$.
(2) Write down the inverse formula to obtain the reconstructed signal in the right column in terms of linear transformation (change of basis) applied to the filtered spectrum $\hat{\mathbf{z}}$.

3 pts ☐

. . . . . .

. . . . . .

What part of the signal would you discard to obtain the reconstructed signal? Draw a rectangle on each spectrum in the middle column where everything inside the rectangle is kept for the reconstruction.

4 pts ☐

. . . . . .

The Wavelet transform is a better choice than Fourier for the first signal (top row).
[ ] True [ ] False

2 pts ☐

The Wavelet transform is a better choice than Fourier for the second signal (bottom row).
[ ] True [ ] False

2 pts ☐

Looking at the middle figure in the top row, what do the first peaks in the spectrum correspond to?

4 pts ☐

. . . . . .

4 Optimization and Robust PCA (30 pts)

4.1 Lagrange Duality

Consider the optimization problem given as $\min_{\mathbf{x}} \|\mathbf{x}\|^2$, $\mathbf{x} \in \mathbb{R}^N$ under the constraints $A\mathbf{x} \leq \mathbf{b}$, for $A \in \mathbb{R}^{D \times N}$, $\mathbf{b} \in \mathbb{R}^D$.

a) Write down the Lagrangian $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ corresponding to this optimization problem.

2 pts ☐

. . . . . .

b) Derive the dual function $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$ as a function of $\mathbf{x}$.

2 pts ☐

. . . . . .

c) Find an explicit expression for $\mathbf{x}$ *and* $d(\boldsymbol{\lambda}, \boldsymbol{\nu})$.

5 pts ☐

. . . . . .

. . . . . .

. . . . . .

4.2 Convex Optimization

Which of the following claims are true/false? *(1 point per correct answer, -1 point per incorrect answer, non-negative total points in any case)*

4 pts ☐

a) The union of two convex sets is convex.
[ ] True [ ] False

b) The *epigraph* of a function $f : \mathbb{R}^D \to \mathbb{R}$ is a set in $\mathbb{R}^D$.
[ ] True [ ] False

c) *(counts for 2/-2pts).* The function $f(\mathbf{v}) := g(\mathbf{v}\mathbf{v}^T)$ is convex over the vectors $\mathbf{v} \in \mathbb{R}^2$, when $g : \mathbb{R}^{2\times 2} \to \mathbb{R}$ is defined as $g(\mathbf{X}) = X_{12} + X_{21}$.
[ ] True [ ] False

4.3 Gradient Descent for Ridge Regression

Consider the optimization problem

$$\min_{\mathbf{x}} \left[f(\mathbf{x}) := \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2 + \frac{\lambda}{2}\|\mathbf{x}\|^2 \right] ,$$

for given $A \in \mathbb{R}^{D\times N}$, $\mathbf{b} \in \mathbb{R}^D, \lambda \in \mathbb{R}$.

a) Write down the update for one step of *gradient descent*, starting at $\mathbf{x}^{(k)}$, with stepsize γ.

2 pts

. . . . . .

b) Write down the update for one step of *stochastic gradient descent*, starting at $\mathbf{x}^{(k)}$, with stepsize γ. *Hint:* Write f as a sum, and assume the i-th term of the sum is randomly selected.

4 pts

. . . . . .

. . . . . .

4.4 ADMM

a) Consider the same ridge regression problem as in the previous exercise, i.e.

$$\min_{\mathbf{x}} \left[f(\mathbf{x}) := \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2 + \frac{\lambda}{2}\|\mathbf{x}\|^2 \right] .$$

Write it in separable form suitable for ADMM.

2 pts

. . . . . .

b) Provide the augmented Lagrangian L_ρ for the separable form of this optimization problem.

2 pts

. . . . . .

c) Write down the updates of the two variables $\mathbf{x}_1$ and $\mathbf{x}_2$ in the ADMM algorithm, for given A and $\mathbf{b}$. Here for simplicity you are *not* required to find an explicit formula for each minimum.

3 pts ☐

$\mathbf{x}_1 :=$ .....

$\mathbf{x}_2 :=$ .....

4.5 RPCA for Collaborative Filtering

Consider the robust completion problem of a matrix $\mathbf{X}$ where X_{ij} are observed matrix entries for $(i, j) \in \Omega_{obs}$. Write down the robust PCA problem for this task, as well as the convex relaxation of this problem.

2 pts ☐

.....

State in one sentence why for this task, the RPCA approach is sometimes preferred to standard low-rank matrix completion.

2 pts ☐

.....