Faculty IV
Neural Information Processing Group

Machine Intelligence 2

Practice exam

Example Solution
16. Juni 2022

Important remarks:

- This set of questions is meant as a preview and means to practice for the actual closed-book written exam.
- It is representative of the style and the difficulty levels of the questions that could come up in the exam.
- It is *not* representative of
  - the number of questions you will get in the actual exam,
  - the duration of the exam. The duration of the actual exam is 90 minutes,
  - the proportion of questions per topic,
  - the proportion of question type (multiple choice vs. free text).
- The following rules apply in the actual exam:
  - This is a closed-book exam. No cheat sheet, no books or any other resources are allowed.
  - You are only allowed to use pens and rulers, *nothing else*!
  - *No* calculators. You will not need one.
  - Do *not* use a pencil (Bleistift) or a pen that writes in red ink. Do *not* use whiteout (Tipp-Ex) or an ink eraser to delete wrong answers. Just cross out what you don't want to be graded. It is therefore best to not bring any of these with you on the day of the exam.
  - Use only the paper included in the exam booklet to answer the questions. Any answers on other sheets of paper are *invalid* and will be confiscated!
  - If you give an answer at any position other than directly following the question, mark clearly where the answer can be found (e.g. "The answer can be found on page XY").
  - Give brief but precise answers.
  - For derivations: show *how* you arrive at the solution.
  - An ambiguous answer is a wrong answer.

Question 1 (8 points) Principal component analysis (PCA).

Mark the correct answers to the following questions. Exactly *one* answer is correct (earning at least 1 point). Wrong answers do not remove any points. Marking more than one answer to the same question invalidates all answers to that question. To remove an already marked answer, fill the entire square and draw an empty square next to it.

Example: The answer is no longer → ☐ ■ Apple

But has been changed to → ☒ Banana

1. (1 point) Let $\underline{\mathbf{C}}$ be the covariance matrix for the data set $\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}, \alpha = 1, \ldots, p$ with $\underline{\mathbf{x}}^{(\alpha)} \in \mathbb{R}^N$. What does $C_{ij} = 0$ imply? The variables x_i and x_j are …

   ☐ independent
   ☒ uncorrelated
   ☐ both
   ☐ neither independent nor uncorrelated

2. (1 point) Which term has to be placed at the dotted space to complete the Lagrangian $\mathcal{L}$ of the PCA problem? $\underline{\mathbf{C}}$ is the covariance matrix of the data, $\{\underline{\mathbf{e}}_a\}$ are the eigenvectors of $\underline{\mathbf{C}}$, λ is the Lagrange multiplier.

$$\mathcal{L} = \underline{\mathbf{e}}_a^\top \underline{\mathbf{C}}\, \underline{\mathbf{e}}_a - \lambda\,(\ldots\ldots\ldots) \overset{!}{=} \max_{\underline{\mathbf{e}}_a}$$

   ☐ $\underline{\mathbf{e}}_a^\top \underline{\mathbf{e}}_a$
   ☒ $\underline{\mathbf{e}}_a^\top \underline{\mathbf{e}}_a - 1$
   ☐ $|\det \underline{\mathbf{C}}\,|$
   ☐ $|\det \underline{\mathbf{C}}\,| - 1$

3. (1 point) Assuming linear PCA has been applied on a data set's covariance matrix, where $\{\lambda_1, \lambda_2, ..., \lambda_N\}$ and $\{\underline{\mathbf{e}}_1, \underline{\mathbf{e}}_2, ..., \underline{\mathbf{e}}_N\}$ are the eigenvalues and principle components (PCs), respectively.
   Which is a *correct* interpretation of the PCs?

   ☐ Each PC represents a cluster in the data.
   ☐ The PCs with lowest λ signify the directions of highest variance in the data.
   ☐ The PCs with highest λ capture outliers in the data.
   ☒ The PCs with highest λ signify the directions of highest variance in the data.

4. (2 points) The stationary and stable state $\underline{\mathbf{w}}^*$ of Oja's rule is the …

   ☐ normalized eigenvector with smallest eigenvalue.
   ☐ unnormalized eigenvector with smallest eigenvalue.
   ☒ normalized eigenvector with largest eigenvalue.
   ☐ unnormalized eigenvector with largest eigenvalue.

5. (2 points) Which statement about kernel PCA is *wrong*?

   ☐ Kernel PCA tries to find nonlinear manifolds by projecting the data implicitly into a high-dimensional feature space.
   ☐ In Kernel PCA the "kernel trick" is applied to circumvent the curse of dimensionality.
   ☒ Standard (linear) PCA is applied after an explicit nonlinear transformation of the data, e.g. $\underline{\mathbf{x}}^{(\alpha)} \mapsto \underline{\phi}\left(\underline{\mathbf{x}}^{(\alpha)}\right)$.
   ☐ The kernel matrix is element of $\mathbb{R}^{p,p}$ where p is the number of data points.

6. (1 point) Given a data set $\left\{\underline{\mathbf{x}}^{(\alpha)} \in \mathbb{R}^N\right\}_{\alpha=1}^{p}$, how many principle components can be obtained by kernel PCA?

   ☐ $p + N$
   ☐ N
   ☐ ∞
   ☒ p

Question 2 (8 points) Independent Component Analysis (ICA).

Let $\widehat{\underline{\mathbf{s}}} = \underline{\mathbf{W}}\,\underline{\mathbf{x}}$ be an estimate of the sources found by ICA for optimally unmixing an observed dataset $\left\{\underline{\mathbf{x}}^{(\alpha)} \in \mathbb{R}^N\right\}, \alpha = 1, \dots, p$. The expression

$$(*) \qquad \int d\widehat{\underline{\mathbf{s}}} P(\widehat{\underline{\mathbf{s}}}) \ln \frac{P(\widehat{\underline{\mathbf{s}}})}{\prod_{i=1}^N P(\widehat{s}_i)}$$

measures the Kullback-Leibler divergence, which is minimized w.r.t. the unmixing matrix $\underline{\mathbf{W}} \in \mathbb{R}^{N,N}$. Here $P(\widehat{\underline{\mathbf{s}}})$ is the joint distribution of the estimated sources and $P(\widehat{s}_i)$ represents the marginal distribution of the i-th estimated source.

1. (1 point)

   What is the value of the expression $(*)$ for a solution that perfectly unmixes the observations?

   Solution:
   zero

2. (5 points)

   Complete the description of the three computational steps that need to be performed in order to directly evaluate the expression $(*)$ for a given matrix $\underline{\mathbf{W}}$. Thereby make clear how the observed data $\{\underline{\mathbf{x}}^{(\alpha)}\}_{\alpha=1}^p$ enters the calculation:

   1. Estimation of ...

      Solution:
      the joint density $P(\widehat{\underline{\mathbf{s}}})$ using the data points $\widehat{\underline{\mathbf{s}}}^{(\alpha)} = \underline{\mathbf{W}}\,\underline{\mathbf{x}}^{(\alpha)}$, $\alpha = 1, \dots, p$.

   2. Marginalization of ...

      Solution:
      the joint density $P(\widehat{\underline{\mathbf{s}}})$ i.e., integrating out all sources except $\widehat{s}_i$ to obtain $P(\widehat{s}_i)$ for each source $\widehat{s}_i$.

   3. Integration of ...

      Solution:
      equation $(*)$ over $\widehat{\underline{\mathbf{s}}} \in \mathbb{R}^N$ using the estimated joint distribution and the individual marginal distributions from above.

3. (2 points)

   Explain why the evaluation of the Infomax cost function is computationally more efficient in comparison to the approach above.

   Solution:
   The Infomax cost function $E^G = \ln|\det \underline{\mathbf{W}}| + \int d\underline{\mathbf{x}} P_{\underline{\mathbf{x}}}(\underline{\mathbf{x}}) \left\{ \sum_{l=1}^{N} \ln \widehat{f}_l' \left(\sum_{k=1}^{N} \mathrm{w}_{lk} \mathrm{x}_k \right) \right\}$ is directly applicable to empirical risk minimization which reduces the practical evaluation of it to empirical averages over the observed data points instead of requiring costly density estimations and marginalizations.

Note that gray parts of the solution are optional, i.e., not required to earn the maximal number of points.

Question 3 (3 points) Stochastic Optimization.

Mark the correct answers to the following questions. Exactly *one* answer is correct (earning at least 1 point). Wrong answers do not remove any points. Marking more than one answer to the same question invalidates all answers to that question. To remove an already marked answer, fill the entire square and draw an empty square next to it.

1. (2 points) Which of the following statements is *wrong* about simulated annealing?

   □ At very high temperature T, all state transitions have similar probabilities regardless of their cost.

   □ At very high T, transtitioning to a state with higher cost occurs with non-zero probability.

   ☒ At very low T, all state transitions have similar probabilities regardless of their cost.

   □ At very low T, transtitioning to a state with higher cost occurs with lower probability.

2. (1 point)
   Which cost function is minimized when deriving the mean-field approximation of simulated annealing?

   □ mean squared distance

   □ negative log-likelihood

   ☒ Kullback-Leibler divergence

   □ None of above

Question 4 (6 points) Clustering.

1. (3 points)

   Specify a continuous schedule for the learning rate ε which enables online algorithms such as online K-means to converge (i.e. it fulfills the Robbins-Munro conditions). Let t denote the number of iterations passed.

   Solution:

   $$\varepsilon(t) = \begin{cases} \varepsilon_0, & t < t_0 \\ \varepsilon_0 + \alpha\left(\frac{1}{t - t_0 + 1} - 1\right), & t \geq t_0, \text{ equally acceptable solution: } \frac{\alpha}{t} \end{cases}$$

   with constants $\varepsilon_0, t_0, \alpha > 0$

2. (3 points)

   Assume a clustering algorithm (e.g. K-means or pairwise clustering) has found a solution which is represented by the binary assignment variables $\{m_q^{(\alpha)}\}$ where $\alpha = 1, \ldots, p$ and $q = 1, \ldots, M$. Each $m_q^{(\alpha)}$ has the value 1 if the object α belongs to cluster q and 0 otherwise.

   Specify a *different* solution $\{\widetilde{m}_q^{(\alpha)}\}$ that shares the same cost as the first solution $\{m_q^{(\alpha)}\}$ for the exact same data set and the same number of clusters M, but that does *not* require running the algorithm again.

   Hint: $\widetilde{m}_q^{(\alpha)}$ can be obtained from $\{m_q^{(\alpha)}\}$

   $\widetilde{m}_q^{(\alpha)} =$

   Solution:

   $$\begin{cases} m_{q+1}^{(\alpha)} \text{ for } q < M \\ m_1^{(\alpha)} \text{ for } q = M \end{cases} \qquad \text{for } \alpha = 1, \ldots, p,\ q = 1, \ldots, M.$$

   (any permutation of clusters is a valid solution)

   or

   $m_{(\mathrm{mod}(q,M)+1)}^{(\alpha)}$ for $\alpha = 1, \ldots, p,\ q = 1, \ldots, M.$ $\mathrm{mod}(a, M) := q \,\%\, M$

   or

   $m_{(M-q+1)}^{(\alpha)}$ for $\alpha = 1, \ldots, p,\ q = 1, \ldots, M.$

Question 5 (4 points) Embedding.

Mark the correct answers to the following questions. Exactly *one* answer is correct (earning at least 1 point). Wrong answers do not remove any points. Marking more than one answer to the same question invalidates all answers to that question. To remove an already marked answer, fill the entire square and draw an empty square next to it.

1. (2 points) Which of the following statements about the learning rate ε and the width σ of a Gaussian neighborhood function is *wrong* for the SOM online algorithm?

   □ A large value of σ implies that an individual data point affects many prototypes in a non-negligible way.

   □ A small value of σ corresponds to significantly changing only those prototypes that are not far in "$\underline{\mathbf{q}}$ space" from the prototype closest to the current data point in "$\underline{\mathbf{x}}$ space".

   □ In the limit $\sigma \to 0$ the standard K-means clustering algorithm is recovered.

   ☒ Annealing ε fast and σ slowly preserves both global and local neighborhood structure.

2. (2 points) Which is *not* a correct statement w.r.t. Locally Linear Embedding?

   □ The data is locally projected onto the tangential space of the data manifold.

   ☒ A nonconvex optimization problem has to be solved involving several locally optimal solutions.

   □ Explicit regularization might be required if the number of neighbors used for reconstruction is larger than the dimensionality of the data.

   □ A balanced search tree can be used to accelerate finding the nearest neighbors.

Question 6 (6 points) Density Estimation.

In parametric density estimation the true unknown probability density $P(\underline{\mathbf{x}})$ for $\underline{\mathbf{x}} \in \mathbb{R}^N$ is approximated by a model $\widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}})$ that is defined through a parameter $\underline{\mathbf{w}} \in \mathbb{R}^M$.
To find the optimal parameter value $\underline{\mathbf{w}}^*$, the Kullback-Leibler divergence between the true distribution and that of the model is minimized:

$$D_{\mathrm{KL}}\left(P \,||\, \widehat{P}\right) = \int d\underline{\mathbf{x}} P(\underline{\mathbf{x}}) \ln \frac{P(\underline{\mathbf{x}})}{\widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}})} \overset{!}{=} \min_{\underline{\mathbf{w}}}$$

1. (3 points)

   Proove that minimizing the KL divergence (above) is equivalent to maximizing the cross entropy (below) w.r.t. the parameter vector $\underline{\mathbf{w}}$:

$$\int d\underline{\mathbf{x}} P(\underline{\mathbf{x}}) \ln \widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}}) \overset{!}{=} \max_{\underline{\mathbf{w}}}$$

Solution:

$$\begin{aligned} D_{\mathrm{KL}}\left(P \,||\, \widehat{P}\right) = \int d\underline{\mathbf{x}} P(\underline{\mathbf{x}}) \ln \frac{P(\underline{\mathbf{x}})}{\widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}})} &= \overbrace{\int d\underline{\mathbf{x}} P(\underline{\mathbf{x}}) \ln P(\underline{\mathbf{x}})}^{\text{constant w.r.t. } \underline{\mathbf{w}}} - \int d\underline{\mathbf{x}} P(\underline{\mathbf{x}}) \ln \widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}}) \overset{!}{=} \min_{\underline{\mathbf{w}}} \\ &\Leftrightarrow - \int d\underline{\mathbf{x}} P(\underline{\mathbf{x}}) \ln \widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}}) \overset{!}{=} \min_{\underline{\mathbf{w}}} \\ &\Leftrightarrow \int d\underline{\mathbf{x}} P(\underline{\mathbf{x}}) \ln \widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}}) \overset{!}{=} \max_{\underline{\mathbf{w}}} \end{aligned}$$

2. (3 points)

   Derive from the cross entropy above for a data set $\left\{\underline{\mathbf{x}}^{(\alpha)} \in \mathbb{R}^N\right\}, \alpha = 1, \ldots, p$ the (empirical) optimization problem that would be solved in practice.
   Hint: The empirical optimization problem is proportional to the maximum log-likelihood formulation.

Solution:
Integral over $P(\underline{\mathbf{x}})$ corresponds to expectation of remaining integrand $\Longrightarrow$ empirical averaging using the data set:

$$E^T = \frac{1}{p} \sum_{\alpha=1}^{p} \ln \widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}}) \overset{!}{=} \max_{\underline{\mathbf{w}}}$$