(1) Introduction

Study questions

1. Name three topics that fall under the umbrella term "statistics".
2. Name three topics that fall under the umbrella term "machine learning".
3. When was the term "machine learning" coined?
4. What is the difference between statistics and machine learning?
5. Name three topics that fall under the umbrella term "artificial intelligence".
6. What is the difference between machine learning and artificial intelligence?
7. Which machine learning methods were popular in the 1980s, 1990s and 2000s?
8. What does the acronym MVPA stand for?
9. Describe the general logic of MVPA for fMRI data.
10. Which machine learning methods are commonly employed in MVPA for fMRI?

Exercises

1. Study the tables of contents of Bishop (2006), Murphy (2012), Alpaydin (2014), and Goodfellow et al. (2017) and identify a canon of machine learning topics.
2. Study Cox and Savoy (2003).

(2) Optimization

Study questions

1. Write down the definition of a smooth multivariate real-valued function.
2. Write down the definition of the gradient of a multivariate real-valued function.
3. Write down the definition of the Hessian of a multivariate real-valued function.
4. Write down Taylor's theorem in the mean value theorem form.
5. Write down the definition of an unconstrained minimization problem.
6. Why does it suffice to consider minimization problems in optimization?
7. Write down the definition of global, local, and strict local function minimizers.
8. State the first-order necessary condition for a local minimum.
9. State the second-order necessary conditions for a local minimum.
10. State the second-order sufficient conditions for a local minimum.
11. Write down the general form of a gradient descent algorithm and explain its components.
12. Why is the direction of the negative gradient a sensible direction for function minimization?
13. Write down the general form of a line search algorithm and explain its components.
14. Write down the general form of a Newton descent algorithm and explain its components.
15. What motivates the choice of the Newton descent update?
16. Write down the general form of a trust region algorithm and explain its components.

Exercises

1. Implement a gradient descent algorithm for the minmization of the Rosenbrock function

$$f : \mathbb{R}^2 \to \mathbb{R}, (x, y) \mapsto f(x, y) := (1 - x)^2 + 100(y - x^2)^2. \tag{1}$$

   Document and visualize your results.

2. Implement a Newton descent algorithm for the minmization of the Himmelblau function

$$f : \mathbb{R}^2 \to \mathbb{R}, (x, y) \mapsto f(x, y) := (x^2 + y - 11)^2 + (x + y^2 - 7)^2 \tag{2}$$

   Document and visualize your results.

(3) Principal component analysis

Study questions

1. Name a potential use of PCA in data analysis.
2. Write down the definition of the covariance matrix of a random vector.
3. What do the elements of a random vector's covariance matrix encode?
4. Given a data matrix $Y \in \mathbb{R}^{m \times n}$ comprising n samples of an m-dimensional random vector y, write down the unbiased estimator for the random vector's covariance matrix.
5. Write down the definition of an eigenvector and an eigenvalue of a square matrix.
6. Explain, how eigenvalues and eigenvectors of a square matrix can be computed.
7. Write down the definitions of symmetric, diagonal, and orthogonal matrices.
8. What is the inverse of an orthogonal matrix?
9. What can be said about the eigenvalues and eigenvectors of symmetric matrices?
10. State the theorem on the orthonormal decomposition of a symmetric matrix.
11. Write down the definition of a linear combination of vectors.
12. Write down the definition of the linearly independence of a set of vectors.
13. Write down the definition of a basis of a vector space.
14. Write down the definition of a vector basis expansion and of vector coordinates.
15. Write down the definition of an orthonormal basis of $\mathbb{R}^m$.
16. Write down the canonical basis of $\mathbb{R}^m$.
17. For $y \in \mathbb{R}^m$ and a basis B, write down the orthonormal expansion of y in terms of B.
18. State the change of basis and vector coordinate transform theorem.
19. Write down the definition of PCA. What is the relevance of the matrix of eigenvectors of the data random vector's covariance matrix?
20. What can be said about a PCA-coordinate transformed random vector $\tilde{y} = Q^T y$?
21. Write down the definition of the singular value decomposition of a matrix $X \in \mathbb{R}^{n \times m}$.
22. Let $X = USV^T$ be the SVD of a matrix X. What can be said about the entries of U, S and V in relation to the matrix products XX^T and $X^T X$?
23. Explain how to compute a PCA for a data matrix $Y \in \mathbb{R}^{m \times n}$ by means of SVD.

Exercises

1. Specify the covariance matrix parameter of a 5-dimensional Gaussian random vector, sample this vector 30 times, and compute the PCA of the resulting 5×30-dimensional data matrix by means of eigenanalysis. Visualize and document your results.
2. Compute the SVD of the same data matrix and evaluate how well the original data matrix can be reconstructed after setting a subset of the singular values to zero. Visualize and document your results.

(4) Gaussian models

Study questions

1. Write down the definition of a linear Gaussian model (LGM).
2. Write down an LGM in hierarchical form and name its components.
3. Write down the marginal observed data distribution of an LGM.
4. What are the central differences between factor analysis, probabilistic PCA, and PCA from the perspective of LGMs?
5. Which question does "inference" in LGMs address?
6. Which question does "learning" in LGMs address?
7. How can the conditional distribution of the latent random vector x given the observed random vector y be evaluated in LGMs?
8. Write down the formula of the evidence lower bound $\text{ELBO}(q(X), \theta)$ and name its components.
9. Write down the (general) expectation-maximization algorithm.
10. Write down the exact expectation-maximization algorithm.
11. What is the purpose of the application of the exact EM algorithm in LGMs?
12. Describe how the evidence lower bound changes with respect to the log marginal likelihood during the E and M Step of the exact expectation-maximization algorithm.
13. Write down the definition of a factor analysis model and name its components.
14. How does a factor analysis model explain covariance between factors and unique variance of individual factors?
15. Write down the definition of a probabilistic principal component analysis model.
16. How are the B parameter and the orthonormal decomposition $\mathbb{C}(y) = Q\Lambda Q^T$ of the data covariance matrix of a probabilistic principal component analysis model related?
17. Write down the definition of a principal component analysis model.
18. How are the B parameter and the orthonormal decomposition $\mathbb{C}(y) = Q\Lambda Q^T$ of the data covariance matrix of a principal component analysis model related?
19. Write down the definition of an independent component analysis analysis model.
20. Write down an independent component analysis model in hierarchical form and name its components.

Exercises

1. Implement the exact EM algorithm for factor analysis.
2. Implement the exact EM algorithm for probabilistic principal component analysis.
3. Implement the infomax algorithm for ICA parameter estimation.

(5) Support vector machines

Study questions

1. Define the notion of a binary classification training data set for support vector machine training.
2. Write down the definition of a linear discriminant function.
3. Write down the definitions of the decision boundary and decision regions induced by a linear discriminant function.
4. State three geometric relationships between a linear discriminant function's hyperplane and its weight and bias parameters.
5. Write down the definition of a hyperplane margin.
6. Write down the definition of a support vector.
7. Define the concept of equivalent hyperplanes.
8. Write down the definition of the canonical hyperplane.
9. Write down the definition of a linearly separable and a non-linearly separable training set.
10. Write down the nonlinear constrained optimization problem corresponding to support vector machine training for maximum margin classification in the linearly separable case.
11. Write down the nonlinear constrained optimization problem corresponding to support vector machine training for soft margin classification in the not necessarily linearly separable training case.
12. What do the slack variables in soft margin support vector machine training quantify?
13. What does the C parameter in soft margin support vector machine training quantify?

Exercises

1. Rewrite support vector machine training for soft margin classification in its dual Lagrangian form.
2. Create a training data set by sampling from two Gaussian distributions.
3. Train a maximum margin support vector machine using the training data set and cvxopt.solvers.qp.
4. Test the generalization accuracy of the trained support vector machine.

References

Alpaydin, E. (2014). *Introduction to Machine Learning*.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York.

Cox, D. D. and Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) “brain reading”: Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, 19(2):261–270.

Goodfellow, I., Bengio, Y., and Courville, A. (2017). *Deep Learning*. The Mit Press, Cambridge, Massachusetts.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, MA.