Exercises for Foundations of Data Science

University of Zurich[UZH]

Prof. Dan Olteanu,
Dr. Ahmet Kara, Dr. Nils Vortmeier,
Haozhe Zhang

DaST Data•(Systems+Theory)

Fall 2020/2021 | Sheet 7 | 04.11.2020

- This is a revision class sheet.
- The solutions will be discussed on Friday 11.12.2020, 14:00-15:45 on Zoom.
- Videos with solutions will be posted on OLAT after the exercise session.

Exercise 7.1 [Yes/No Questions]
Please decide for each of the following statements whether it is true or false.

(a) Regularisation reduces the training error of a least squares model.

(b) $k(x, y) = k_1(x, y) \cdot k_2(x, y)$ is a Mercer kernel for two Mercer kernels k_1 and k_2.

(c) A neural network can be used to linearly separate a linearly separable dataset.

(d) It is possible to compute the solution to a linear regression problem in closed form.

(e) It is possible that a Support Vector Machine has all data points as support vectors.

Exercise 7.2 [Multiple Choice Questions]
Please decide for each of the following questions which of the given possible answers are true. No, several, or all given answers can be true.

(a) We want to minimise the function

$$f(x) = \begin{cases} x^2 & \text{if } x \leq 0 \\ x & \text{otherwise} \end{cases}$$

(1) Starting from $x = 1$, Newton's method will take us to the minimum in one step.

(2) Starting from $x = -1$, Newton's method will take us to the minimum in one step.

(3) There is a step size such that starting from $x = 1$, gradient descent will take us to the minimum of f in one step.

(4) There exists a step size such that starting from $x = -1$, gradient descent will take us to the minimum of f in one step.

(b) Consider Linear Regression with a data matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ and labels $\mathbf{y} \in \mathbb{R}^N$. If the optimal parameter vector for the least-squares problem is unique, which of the following must be true?

(1) The rank of $\mathbf{X}$ is D.

(2) The rank of $\mathbf{X}^\mathsf{T}$ is D.

(3) $N \leq D$.

(4) $D \leq N$.

(c) Which of the following models are discriminative models?

(1) Logistic Regression

(2) Linear Discriminant Analysis

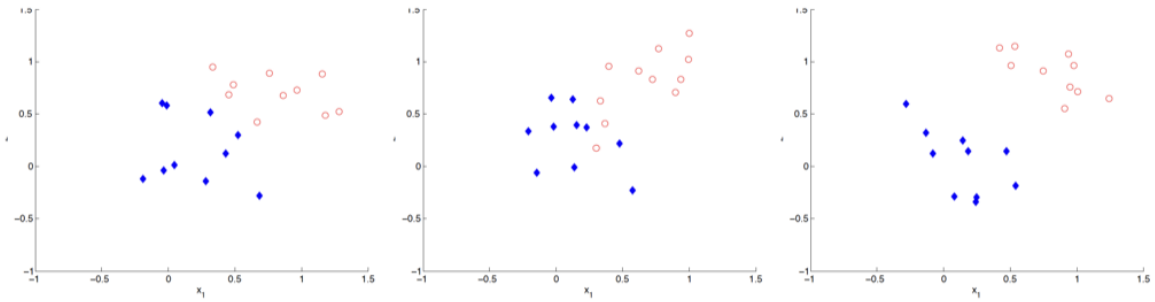
(3) Linear Regression

(4) Naive Bayes

(d) Suppose the perceptron algorithm is run on each of the below datasets in the following manner:

- Make a pass through the data: The points are fed to the algorithm one at a time in random order.
- The above step is repeated until there is an entire pass through the data where the parameter vector does not change.

On which of the datasets will this procedure converge (will not iterate infinitely)?



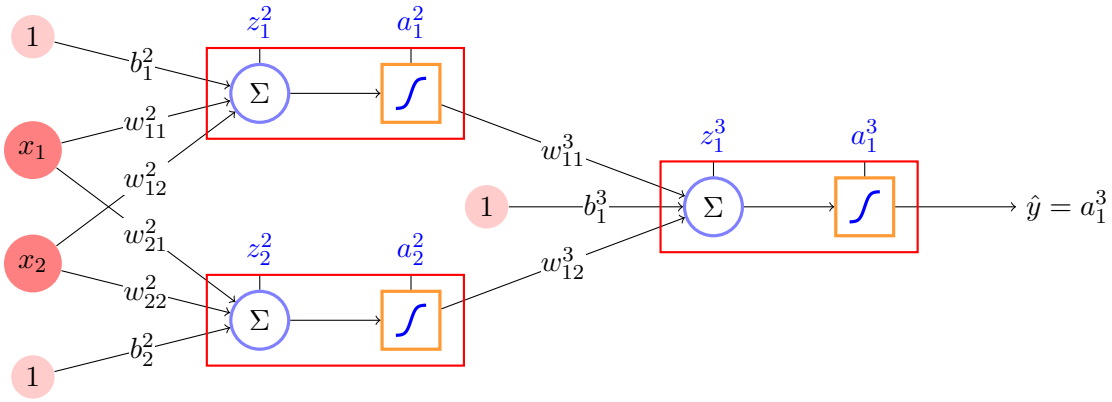
(a) (b) (c)

Exercise 7.3 [Maximum Likelihood Estimation]
Assume that X is a random variable that takes on values from $\mathbb{N}_0$. We say that X is distributed following the Poisson probability distribution if for any $x \in \mathbb{N}_0$, $p(X = x \mid \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$ where $x! = 1 \cdot 2 \cdots x$ is the factorial of x and $\lambda > 0$ is the rate parameter. The Poisson distribution is often used as a model for counts of rare events like radioactive decay and traffic accidents. Assume that X is distributed following the Poisson probability distribution.

(a) Let $\mathcal{D}$ be a dataset that consists of N independent and identically distributed observations $x_1, \ldots, x_N \in \mathbb{N}_0$ for the variable X. Compute the maximum likelihood estimator (MLE) $\widehat{\lambda}$ for λ by maximising the likelihood function $p(\mathcal{D} \mid \lambda)$. Give the intermediate and final steps of your computation.

(b) Assume that $\mathcal{D}$ consists of the following 8 observations: $5, 8, 4, 10, 12, 0, 15, 10$. Give the value $p(\mathcal{D} \mid \widehat{\lambda})$.

Exercise 7.4 [Neural Nets]
Consider the following neural net with three layers: the input layer, one hidden layer with two neurons, and the output layer with one neuron.



All neurons use the *negative sigmoid* $n(z) = -\frac{1}{1+e^{-z}}$ as activation function. It holds $n(z) = -\sigma(z)$, where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the standard sigmoid function with derivative $\sigma'(z) = \sigma(z)(1-\sigma(z))$.

(a) Use the forward equations to express $\mathbf{z}^2$, $\mathbf{a}^2$, $\mathbf{z}^3$ and $\mathbf{a}^3$. Your expression for $\mathbf{a}^2$ may use $\mathbf{z}^2$, your expression for $\mathbf{z}^3$ may use $\mathbf{a}^2$ and $\mathbf{z}^2$, and your expression for $\mathbf{a}^3$ may use all of $\mathbf{z}^2$, $\mathbf{a}^2$ and $\mathbf{z}^3$.

(b) Assume that we use the non-standard loss function $\ell(\hat{y}, y) = \frac{1}{2}(2\hat{y} - 2y)^2$. Use the back-propagation equations to derive expressions for $\frac{\partial \ell}{\partial w_{ij}^k}$ and $\frac{\partial \ell}{\partial b_i^k}$, for all weights w_{ij}^k and biases b_i^k of the network.

(c) Explain how a training step is performed using the expressions you obtained in (a) and (b).

Exercise 7.5 [Support Vector Machines]
Consider the two 1-dimensional data points $x_1 = 0$ and $x_2 = \sqrt{2}$ with labels $y_1 = -1$ and $y_2 = 1$, respectively. Consider the map $\phi(x) = [1, \sqrt{2}x, x^2]^\mathsf{T}$ from the 1-dimensional to the 3-dimensional feature space (this is equivalent to using a degree-2 polynomial kernel). Consider the following primal formulation of the Support Vector Machine optimisation problem:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||\mathbf{w}||_2^2 \\ \text{subject to} \quad & y_1(\phi(x_1) \cdot \mathbf{w} + w_0) \geq 1 \\ & y_2(\phi(x_2) \cdot \mathbf{w} + w_0) \geq 1 \end{aligned}$$

(a) Give a vector that is parallel to the optimal vector $\mathbf{w}$.

(b) What is the value of the margin that is achieved by the above optimisation problem?

(c) Give the values for $\mathbf{w}$ and w_0 that optimise the above optimisation problem.