Exercises for Foundations of Data Science

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

Fall 2020/2021 | Sheet 4 | 23.10.2020

- The solutions will be discussed on Friday 13.11.2020, 14:00-15:45 on Zoom.
- Videos with solutions will be posted on OLAT after the exercise session.

Exercise 4.1 [Optimisation Methods for ℓ_1 Regularisation]

(a) Show that if you use the absolute loss function with ℓ_1 regularisation, the optimisation problem can be solved using linear programming. The objective function is:

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} |\mathbf{w}^\mathsf{T}\mathbf{x}_i - y_i| + \lambda \sum_{i=1}^{D} |w_i|.$$

(b) If we use the squared loss instead of absolute loss, we optimise the Lasso objective:

$$\mathcal{L}_{lasso}(\mathbf{w}) = \sum_{i=1}^{N} (\mathbf{w}^\mathsf{T}\mathbf{x}_i - y_i)^2 + \lambda \sum_{i=1}^{D} |w_i|.$$

In this case, we can no longer use linear programming because of the quadratic term in the objective. Give a quadratic program encoding for this objective function. Show that an optimal solution for your program is minimising the Lasso objective.

(c) We would like to minimise the above Lasso objective function $\mathcal{L}_{lasso}$ using subgradient descent. Write the subgradient descent update rule with step size η, i.e., write how you would obtain $\mathbf{w}_{t+1}$ using $\mathbf{w}_t$ and an (explicitly computed) subgradient of the objective function at $\mathbf{w}_t$ and step-size η.

Exercise 4.2 [Steepest Descent Method]
Let $||\cdot||_p$ be any norm in $\mathbb{R}^D$. We define a *normalised steepest descent direction* (with respect to $||\cdot||_p$) at point $\mathbf{x}$ as

$$\underset{\mathbf{v}}{\operatorname{argmin}}\{\nabla f(\mathbf{x}) \cdot \mathbf{v} \mid ||\mathbf{v}||_p = 1\}$$

where $\nabla f(\mathbf{x}) \cdot \mathbf{v}$ is the directional derivative of f at $\mathbf{x}$ in the direction of $\mathbf{v}$.
Explain how to find a normalised steepest descent direction using the ℓ_∞ norm, and give an interpretation.

Exercise 4.3 [Newton's Method]
Run Newton's method for the following functions using fixed step size 1.

(a) The function $f(x) = \log(e^x + e^{-x})$ has a unique minimiser $x^* = 0$. Run Newton's method starting at $x_0 = 1$ and $x_0 = 1.1$. Show the first few iterates. Plot f and f'.

(b) The function $f(x) = -\log x + x$ has a unique minimiser $x^* = 1$. Run Newton's method starting at $x_0 = 3$. Show the first few iterates. Plot f and f'.

In which of the above cases does the method diverge from the unique minimiser? How can such a behaviour be avoided?

Exercise 4.4 [Naïve Bayes with Mixed Features]
Consider a 3-class naïve Bayes classifier with a binary feature x_1, a continuous feature x_2, and an output label y with classes $c \in \{1, 2, 3\}$. We model (the class prior) $p(y = c \mid \boldsymbol{\pi}) = \pi_c$ using Multinoulli, $p(x_1 \mid y = c, \theta_c)$ using Bernoulli, and $p(x_2 \mid y = c, \mu_c, \sigma_c^2)$ using Gaussian distribution. Let the parameter estimates be as follows:

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3) = (0.5, 0.25, 0.25) \qquad \boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) = (0.5, 0.5, 0.5)$$
$$\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3) = (-1, 0, 1) \qquad \boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 1, 1).$$

(a) Compute $p(y \mid x_1 = 0, x_2 = 0, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. The result should be a vector of 3 numbers that sum up to 1.

(b) Compute $p(y \mid x_1 = 0, \boldsymbol{\pi}, \boldsymbol{\theta})$

(c) Compute $p(y \mid x_2 = 0, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$

(d) Explain any interesting pattern you see in your results.

Exercise 4.5 [Gaussian Decision Boundaries]
Let $p(x \mid y = c, \mu_c, \sigma_c^2) = \mathcal{N}(\mu_c, \sigma_c^2)$ where $c \in \{1, 2\}$, $\mu_1 = 0$, $\sigma_1^2 = 1$, $\mu_2 = 1$, and $\sigma_2^2 = 10^6$. Let the class probability distribution be $p(y = 1 \mid \boldsymbol{\pi}) = p(y = 2 \mid \boldsymbol{\pi}) = 0.5$.

(a) Find the decision region

$$R_1 = \{x \mid p(x \mid y = 1, \mu_1, \sigma_1^2) \geq p(x \mid y = 2, \mu_2, \sigma_2^2)\}.$$

Sketch the result.
Hint: Draw the curves and find where they intersect. Find both solutions of the equation $p(x \mid y = 1, \mu_1, \sigma_1^2) = p(x \mid y = 2, \mu_2, \sigma_2^2)$.

(b) Now suppose $\sigma_2 = 1$ (and all other parameters remain the same). What is R_1 in this case?

Exercise 4.6 [Quadratic Discriminant Analysis]

Consider the following training dataset of heights x (in inches) and gender y (male/female) of some US college students: $\mathbf{x} = (67, 79, 71, 68, 67, 60)$, $\mathbf{y} = (m, m, m, f, f, f)$. You can solve the following tasks by hand or by coding.

(a) Fit a Bayes classifier to the above data by estimating the parameters of the class-conditional probability distributions

$$p(x \mid y = c, \mu_c, \sigma_c^2) = \mathcal{N}(\mu_c, \sigma_c^2)$$

and the class distribution

$$p(y = c, \boldsymbol{\pi}) = \pi_c$$

where $c \in \{m, f\}$. What are your estimates for μ_c, σ_c^2, and π_c for $c \in \{m, f\}$?

(b) Compute $p(y = m \mid x = 72, \mu_m, \sigma_m^2, \pi_m)$.