# Mathematical Foundations of Machine Learning: Tutorial Sheet 2
# Assignment Submission Deadline: 26/04/2023

## ∗ Problems to be submitted as Assignment

*1. Consider a data generation model

$$x_n = \sum_{k=0}^{N-1} c_k e^{-j\frac{2\pi kn}{N}}, \; n = 0, \ldots, N-1.$$

(a) Write the above equation in matrix-vector form

$$x = Wc.$$

What are the vectors $c$ and $x$, and what is the matrix $W$?

(b) Show that $W$ is orthogonal, i.e.,, $W^H W = I$, where $W^H$ is the conjugate transpose of $W$.

(c) Using (b), derive the least squares regression solution.

*2. Consider a simplified LASSO regression problem:

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \|y - \theta\|^2 + \lambda \|\theta\|_1.$$

Show that the solution is given by

$$\hat{\theta} = \text{sign}(y).\max(|y| - \lambda, 0),$$

where . is the element-wise multiplication.

*3 A one-dimensional signal is corrupted by blur and noise:

$$y_n = \sum_{l=0}^{L} h_l x_{n-l} + e_n.$$

(a) Formulate the least squares regression problem in matrix-vector form $y = Hx + e$. Find $x, y$ and $H$.

(b) Consider a regularization function

$$R(x) = \sum_{n=2}^{N} (x_n - x_{n-1})^2.$$

Show that this regularization is equivalent to $R(x) = \|Dx\|_2$ for some $D$. Find $D$.

(c) Using the regularization in (b), derive the regularized least squares regression result:

$$\underset{x}{\text{minimize}} \;\; \|y - Hx\|^2 + \lambda \|Dx\|^2.$$

*4 Prove that the solution to the Equation

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \; \|X\theta - y\|^2 + \lambda \|\theta\|^2$$

is

$$\hat{\theta} = (X^T X + \lambda I)^{-1} X^T y.$$

*5 Consider a linear model such that

$$y = x^T \theta + e.$$

Show that the average predictor in linear regression is equal to the true predictor.

*6 Prove the following theorems in the reading material for 6th week:
i) Theorem 7.3 ii) Theorem 7.4 iii) Theorem 7.6

*7 The following data yield the amount of hydrogen present (in parts per million) in core drillings of fixed size at the following distances (in feet) from the base of a vacuum-cast ingot.

| Distance | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Amount | 1.28 | 1.50 | 1.12 | 0.94 | 0.82 | 0.75 | 0.60 | 0.72 | 0.95 | 1.20 |

(a) Draw a scatter diagram.

(b) Fit a curve of the form

$$Y = \alpha + \beta x + \gamma x^2 + e$$

to the data minimizing sum square loss.

*8 Derive the maximum likelihood estimates for $\theta$ in the Poisson distribution

$$p(x; \theta) = \frac{1}{x!} \exp(\theta x - e^{\theta}).$$

Note that the Poisson distribution is often specified using $\lambda := \log \theta$ as its rate parameter. In this case we have $p(x; \lambda) = \frac{\lambda^x}{x!} \exp(-\lambda)$.

*9 Let $x_1, x_2, ..., x_N$ be vectors stemmed from a normal distribution with known covariance matrix and unknown mean, that is,

$$p(xk; \mu) = \frac{1}{(2\pi)^{l/2} |\Sigma|^{1/2}} \exp\left(\frac{1}{2}(x_k - \mu)\sigma^{-1}(x_k - \mu)\right).$$

Obtain the ML estimate of the unknown mean vector.

*10 Show that in a binary classification task, the Bayes decision rule minimizes the error probability.

*11 In a two-class, two-dimensional classification task, the feature vectors are generated by two normal distributions sharing the same covariance matrix

$$\Sigma = \begin{bmatrix} 2.1 & 3.3 \\ 1.5 & 3.9 \end{bmatrix}$$

and the mean vectors are $\mu_1 = [0, 0]^T$ , $\mu_2 = [3, 3]^T$, respectively.

(a) Classify the vector $[1.0, 2.2]^T$ according to the Bayesian classifier.

(b) ) Compute the principal axes of the ellipse centered at [0, 0]T that corresponds to a constant Mahalanobis distance $dm = \sqrt{2.952}$ from the center.

*12 In a two-class, two-dimensional classification task, assume that the classes are equiprobable. Suppose the feature vectors are generated by two normal distributions sharing the same covariance matrix that is diagonal with equal elements, and the mean vectors are $\mu_1$ and $\mu_2$, respectively.

(a) Show that the decision surface in the Bayesian classifier is a straight line passing through the point $x_0 = \frac{1}{2}(\mu_1 + \mu_2)$.

(b) Show that the vector $\mu_1 - \mu_2$ is orthogonal to the decision hyperplane.

*13 Derive a Linear programming problem to fit the following data using Robust Linear Regression.

| Speed | Number of Cans Damaged |
|---|---|
| 3 | 54 |
| 3 | 62 |
| 3 | 65 |
| 5 | 94 |
| 5 | 122 |
| 5 | 84 |
| 6 | 142 |
| 7 | 139 |
| 7 | 184 |
| 8 | 25 |

*14 Apply Logistic regression to fit the following two-dimensional binary classification data:

$$X1 = (x1, x2) = \{(4,1), (2,4), (2,3), (3,6), (4,4)\}$$

$$X2 = (x1, x2) = \{(9,10), (6,8), (9,5), (8,7), (10,8)\}$$

(a) Using two iterations of Gradient descent with Learning rate 1.

(b) Using two-iterations of Newton's method.

*15 Show that the negative entropy is 1-strongly convex with respect to the $\| \|_1$ norm on the simplex. Hint: First show that $\phi(t) := (t-1)\log(t) - 2\frac{(t-1)^2}{t+1} \geq 0$ for all $t \geq 0$. Next substitute $t = x_i/y_i$ to show that $\sum_i (x_i - y_i)\log\left(\frac{x_i}{y_i}\right) \geq \|x - y\|_1^2$.

*16 Show that the gradient of $p-$norm

$$f(x) = \frac{1}{2}\|x\|_p^2 = \frac{1}{2}\left(\sum_i x_i^p\right)^{2/p}$$

is

$$\nabla_{x_i} f(x) = \frac{sign(x_i)}{|x_i|^{p-1}\|x\|_p^{p-2}}.$$